ACL 2026
SAN DIEGO

JULY 2–7

The 64th Annual Meeting of the Association for Computational Linguistics

Contents

Table of Contents i

1 Conference Overview 2

2 Conference Information 3
Message from the General Chair . . . 3
Message from the Program Co-Chairs . . . 5
ACL Business and Event Manager Message . . . 9
Organizing Committee . . . 11

3 Anti-Harassment Policy 14

4 Code of Ethics 15

5 Social Events 16
Welcome Reception . . . 16
Gala Dinner . . . 17
Meals . . . 18

6 Keynotes 19

7 Panels 23

8 Tutorials 24
Overview . . . 24
T01: Towards Effective and Efficient Multi-Agent Language Model Systems: Foundations, Prospects, and Applications . . . 26
T02: The Data Frontier for Large Language Models: Selection, Synthesis, and Tools . . . 27
T03: Future of Work in the Age of LLMs . . . 28
T04: Current Advances in LLM Reasoning . . . 29

T05: Knowledge Control for Responsible Generative AI: Bridging Academia,Industry, and Society 30
T06: The Interplay between Metaphors and NLP 31

9 Workshops **32**
Overview 32
W01: CustomNLP4U: Second Workshop on Customizable NLP: Progress and Challenges in Customizing NLP for a Domain, Application, Group, or Individual 36
W02: PrivateNLP: Seventh Workshop on Privacy in Natural Language Processing 37
W03: ALVR: 4th Workshop on Advances in Language and Vision Research 38
W04: ArgMining: the 13th Workshop on Argument Mining and Reasoning 39
W05: CODI-CRAC: Joint Workshop on Computational Approaches to Discourse, Context and Document-Level Inferences and Computational Models of Reference, Anaphora and Coreference 40
W06: DravidianLangTech: The Sixth Workshop on Speech and Language Technologies for Dravidian Languages 41
W07: EEUCA: The 9th Workshop on Event Extraction and Understanding: Challenges and Applications 42
W08: LAW XX: The 20th Linguistic Annotation Workshop 43
W09: SELVA: The 1st International Workshop on Sustainable and Efficient Language, Vision, and Action Models 44
W10: SURGeLLM: Structured Understanding, Retrieval, and Generation in the LLM Era 45
W11: KnowFM: 4th Workshop on Towards Knowledgeable Foundation Models 46
W12: CHum: The Second Workshop on Computational Humor 47
W13: SMM4H-HeaRD: Joint 11th Social Media Mining for Health Applications and Health Real-World Data Workshop and Shared Tasks 48
W14: AmericasNLP: The Sixth Workshop on NLP for Indigenous Languages of the Americas . 49
W15: StereACuLT: The 1st Workshop on Stereotypes Across Cultures in Language Technologies 50
W16: SCiL: Society for Computation in Linguistics 51
W17: IWSLT: ACL-IWSLT: 23rd International Conference on Spoken Language Translation . . 52
W18: SemEval: The 20th International Workshop on Semantic Evaluation 53
W19: BioNLP and Shared Tasks 54
W20: StarSEM: *SEM 55
W21: BEA: The 21st Workshop on Innovative Use of NLP for Building Educational Applications 56
W22: CoNLL: The SIGNLL Conference on Computational Natural Language Learning 57
W23: TrustNLP: The 6th Workshop on Trustworthy NLP 58
W24: BigPicture: The Big Picture: Crafting a Research Narrative v2 59
W25: RAG4Reports: The First Workshop on Multilingual Report Generation via Retrieval Augmented Generation 60
W26: LTEDI: Sixth Workshop on Language Technology for Equality, Diversity and Inclusion . . 61
W27: MAGMaR: The Second Workshop on Multimodal Augmented Generation via Multimodal Retrieval 62
W28: GEM: Natural Language Generation, Evaluation, and Metrics 63
W29: NLP4DH: The 6th International Conference on Natural Language Processing for the Digital Humanities 64
W30: ComputEL-9: The Ninth Workshop on the Use of Computational Methods in the Study of Endangered Languages 65
W31: NLP+CSS: Seventh Workshop on NLP and Computational Social Science 66
W32: CLPsych: Eleventh Workshop on Computational Linguistics and Clinical Psychology . . 67
W33: BATCH: Beyond Alignment: Transdisciplinary Conversations on Human-AI Futures . . . 68
W34: CDL: The 1st Workshop on Computational Developmental Linguistics 69
W35: C3NLP: Cross-Cultural Considerations in NLP 70
W36: EvalEval: Second Workshop on Evaluating Evaluations: Examining Best Practices for Utilizing and Developing Generative Model Evaluations 71

W37: MeLLMs: Multilinguality in the Era of Large Language Models . . . 72

10 Birds of a Feather **73**

11 Main Conference Day 1: Sunday, July 5, 2026 **79**
Overview . . . 79
Detailed Program . . . 81
Demo Session A . . . 81
Oral Session A: Computational Social Science, Cultural Analytics, and NLP for Social Good 1 . . . 82
Oral Session A: Ethics, Bias, and Fairness 1 . . . 84
Oral Session A: Industry 1 . . . 87
Oral Session A: Information Extraction and Retrieval 1 . . . 87
Oral Session A: Interpretability and Analysis of Models for NLP 1 . . . 90
Oral Session A: Language Models 1 . . . 92
Oral Session A: Multilinguality and Language Diversity 1 . . . 94
Oral Session A: NLP Applications 1 . . . 96
Oral Session A: Resources and Evaluation 1 . . . 98
Oral Session A: Student Research Workshop 1 . . . 100
Poster Session A (Coronado) . . . 101
Poster Session A (Grand Hall) . . . 115
Poster Session A (Zoom) . . . 155
Demo Session B . . . 156
Oral Session B: AI/LLM Agents 1 . . . 157
Oral Session B: Industry 2 . . . 159
Oral Session B: Information Extraction and Retrieval 2 . . . 161
Oral Session B: Language Models 2 . . . 163
Oral Session B: Multilinguality and Language Diversity 2 . . . 165
Oral Session B: Resources and Evaluation 2 . . . 167
Oral Session B: Safety and Alignment in LLMs 1 (Harbor D) . . . 170
Oral Session B: Safety and Alignment in LLMs 1 (Harbor E-F) . . . 171
Oral Session B: Semantics: Lexical, Sentence-level Semantics, Textual Inference and Other areas . . . 172
Oral Session B: Student Research Workshop 2 . . . 174
Oral Session B: Summarization, Generalizability and Transfer . . . 175
Poster Session B (Coronado) . . . 177
Poster Session B (Grand Hall) . . . 195
Demo Session C . . . 231
Oral Session C: Computational Social Science, Cultural Analytics, and NLP for Social Good 2 . . . 233
Oral Session C: Dialogue and Interactive Systems . . . 235
Oral Session C: Industry 3 . . . 237
Oral Session C: Information Extraction and Retrieval 3 . . . 240
Oral Session C: Linguistic theories, Cognitive Modeling and Psycholinguistics 1 . . . 242
Oral Session C: Machine Learning for NLP 1 . . . 244
Oral Session C: Multilinguality and Language Diversity 3 . . . 246
Oral Session C: NLP Applications 2 . . . 248
Oral Session C: Resources and Evaluation 3 . . . 249
Poster Session C (Coronado) . . . 251
Poster Session C (Grand Hall) . . . 268
Poster Session C (Zoom) . . . 307
Virtual Presentations 1 . . . 308
Virtual Presentations 1 . . . 308

12 Main Conference Day 2: Monday, July 6, 2026 **366**
Overview . . . 366
Detailed Program . . . 368
Virtual Presentations 2 . . . 368
Demo Session D . . . 438
Oral Session D: AI/LLM Agents 2 . . . 440
Oral Session D: Financial Applications and Time Series . . . 442
Oral Session D: Industry 4 . . . 443
Oral Session D: Language Models 3 . . . 444
Oral Session D: Linguistic theories, Cognitive Modeling and Psycholinguistics 2 . . . 446
Oral Session D: Low-resource Methods for NLP 1 . . . 448
Oral Session D: Question Answering 1 . . . 450
Oral Session D: Special Theme: Explainability of NLP Models 1 . . . 452
Oral Session D: Speech Processing and Spoken Language Understanding 1 . . . 454
Poster Session D (Coronado) . . . 457
Poster Session D (Grand Hall) . . . 479
Demo Session E . . . 526
Oral Session E: Multimodality and Language Grounding to Vision, Robotics and Beyond 1 528
Orals Session E: Clinical and Biomedical Applications . . . 528
Orals Session E: Code Models . . . 530
Orals Session E: Human-AI Interaction/Cooperation and Human-Centric NLP 1 . . . 532
Orals Session E: LLM Efficiency 1 . . . 534
Orals Session E: Low-resource Methods for NLP 2 . . . 536
Orals Session E: Machine Learning for NLP 2 . . . 539
Orals Session E: Machine Translation . . . 541
Orals Session E: Multimodality and Language Grounding to Vision, Robotics and Beyond 1 543
Poster Session E (Coronado) . . . 545
Poster Session E (Grand Hall) . . . 565
Poster Session E: Industry (Coronado) . . . 605
Virtual Presentations 3 . . . 618
Virtual Presentations 4 . . . 664

13 Main Conference Day 3: Tuesday, July 7, 2026 **710**
Overview . . . 710
Detailed Program . . . 712
Virtual Presentations 5 . . . 712
Virtual Presentations 5 . . . 712
Demo Session F . . . 763
Oral Session F: Discourse, Pragmatics, and Reasoning . . . 764
Oral Session F: Ethics, Bias, and Fairness 2 . . . 766
Oral Session F: Industry 5 . . . 768
Oral Session F: Interpretability and Analysis of Models for NLP 2 . . . 770
Oral Session F: LLM Efficiency 2 . . . 772
Oral Session F: Multimodality and Language Grounding to Vision, Robotics and Beyond 2 774
Oral Session F: Natural Language Generation 1 . . . 776
Oral Session F: Speech Processing and Spoken Language Understanding 2 . . . 778
Orals Session F: Safety and Alignment in LLMs 2 (Harbor D) . . . 780
Orals Session F: Safety and Alignment in LLMs 2 (Harbor E-F) . . . 781
Poster Session F (Coronado) . . . 783
Poster Session F (Grand Hall) . . . 799
Demo Session G . . . 840
Oral Session G: Hierarchical Structure Prediction, Syntax, and Parsing & Phonology, Morphology and Word Segmentation . . . 841

Oral Session G: Human-AI Interaction/Cooperation and Human-Centric NLP 2 843
Oral Session G: Industry 6 845
Oral Session G: Mathematical, Symbolic, Neurosymbolic, and Logical Reasoning 847
Oral Session G: Natural Language Generation 2 849
Oral Session G: Question Answering 2 851
Oral Session G: Retrieval-Augmented Language Models 853
Oral Session G: Sentiment Analysis, Stylistic Analysis, and Argument Mining 855
Oral Session G: Special Theme: Explainability of NLP Models 2 857
Poster Session G (Coronado) 859
Poster Session G (Grand Hall) 877
Poster Session G (Zoom) 909

14 Awards **910**

15 Local Guide **916**

16 Venue Map **922**

Author Index **929**

Sponsorship **1134**

1 Conference Overview

Date	Time	Event	Location
Tutorials			
Thu, Jul 2	9:00-17:30	Tutorials	Harbor A-I
Workshops			
Fri, Jul 3	9:00-17:30	Workshops Day 1	Harbor, Gaslamp, Balboa & Old Town
Sat, Jul 4	9:00-17:30	Workshops Day 2	Harbor, Gaslamp, Balboa & Old Town
	9:00-17:30	Workshop Posters	Grand Ballroom
	18:00-20:00	Welcome Reception	Sports Court, L4
	9:00 PM	Big Bay Boom Fireworks	San Diego Bay
Main Conference			
Jul 5-7	90-min slots	Birds of a Feather & Affinity	Gaslamp, La Jolla, Old Town
Sun, Jul 5 **Day 1**	9:00-9:30	Welcome	Harbor Ballroom
	9:30-10:30	Keynote: Philip Resnik	Harbor Ballroom
	11:00-17:30	Orals / Demos / Posters	Harbor, Promenade, Regatta
Mon, Jul 6 **Day 2**	1:00-2:00	Keynote: Yue Zhang	Underline (virtual)
	9:00-12:30	Orals / Demos / Posters	Harbor, Promenade, Regatta
	13:30-14:30	Business Meeting	Harbor Ballroom
	14:45-15:30	Panel: Explainability of LLMs	Harbor Ballroom
	15:30-16:15	Presidential Address: Barbara Plank	Harbor Ballroom
	16:45-17:45	Keynote: Tania Lombrozo	Harbor Ballroom
	18:30-22:00	Social Event	USS Midway
Tue, Jul 7 **Day 3**	9:00-12:30	Orals / Demos / Posters	Harbor, Promenade, Regatta
	14:00-15:30	Memorial / ToT / Lifetime Awards	Harbor Ballroom
	16:00-17:30	Award & Closing Session	Harbor Ballroom

Registration & Check-In: Grand Hall Foyer (L1).
Hours: Wed, Jul 1, 17:30-20:00; Thu-Fri, 08:00-16:30; Sat, 08:00-19:30; Sun, 07:30-16:30; Mon-Tue, 08:30-16:30.

2

Conference Information

Message from the General Chair

It is my great pleasure and honor to welcome you to the 64th Annual Meeting of the Association for Computational Linguistics (ACL 2026), held in San Diego, California, USA, July 2–7, 2026. ACL 2026 continues our fields tradition of excellence in scholarship, innovation, and inclusivity, and I am deeply grateful to the many volunteers who have worked tirelessly to bring this event to life.

It is exciting to see the many developments in the field reflected at our conference, such as the advancement of large language models (born out of our community) into many application directions. This is not without challenges, partly due to unsolved problems, and partly due to misuse of the technology. All this is reflected well at ACL 2026. The conference also reflects that the field is not only about advancing technology, but also the science of fundamentally understanding language and how it is used by many different speaker communities.

I want to express my deepest thanks to our Program Chairs — Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens — who have overseen the reviewing and selection process, and shaped a compelling and diverse scientific program. It is impossible to overstate how much work and dedication they put in to make this conference successful. Overall, we received 12,148 submissions and accepted 2,296 at the main conference and 2,163 as findings. This was made possible through close coordination with the ACL Rolling Review (ARR) team and I sincerely thank them. We are also deeply grateful to the many senior area chairs, area chairs, reviewers, and the Best Paper Committee (led by Antonios Anastasopoulos and Eunsol Choi), whose dedication ensured the high quality of our program. Isabelle Augenstein served as Conflict Chair to shepherd papers that were submitted by conference chairs.

I am also grateful to our Technical OpenReview Chairs, Shuoyang Ding and Yi R. (May) Fung, for their behind-the-scenes work. For those joining virtually, our Virtual Infrastructure Chairs — Gael Guibon and Parisa Kordjamshidi — have worked to make the hybrid conference experience inclusive and engaging. I would also like to recognize the efforts of our Ethics Chairs — Margaret Mitchell and Paco Guzman — who have led the work aimed at ensuring that submissions meet the ethical standards of our field.

ACL 2026 hosts a rich set of 37 workshops and 6 tutorials, thanks to the dedicated efforts of our Workshop Chairs — Loic Barrault and Yang Zhao — and our Tutorial Chairs — Kenton Murray and Jacob Andreas. Our Demonstration Track Chairs — Greg Durrett and Ping Jian — have put together an impressive set of system demonstrations of cutting-edge innovations in NLP. I am proud to say that we also have a rich Industry Track, thanks to Yunyao Li, Georg Rehm, and Mei Tu who created a "conference in the conference". Their experience from organizing the industry track last year helped advance the visibility and impact of the event. We also hosted a special linguistic symposium entitled "Linguistics and NLP in the LLM Era" which featured Allyson Ettinger, Richard Futrell, Zoey Liu, and Tom McCoy.

As is customary, TACL and CL papers are presented at the conference. Thanks go to the TACL Editors-in-Chief — Asli Celikyilmaz, Roi Reichart, Dilek Hakkani-Tur, and Yun-Nung (Vivian) Chen — and the CL Editor-in-Chief, Wei Lu, for their coordination efforts.

This conference could only have been made possible thanks to the invaluable work of our ACL Business Manager and Director of Events, Jennifer Rachford, and her team. Our Visa Chairs — David Chiang and Swabha Swayamdipta — have helped numerous participants with their visa needs and any logistical difficulties in attending the conference. I also thank

our Student Volunteer Chairs — Jingbo Shang, Lianhui Qin, Xuezhe Ma, and Muhao Chen — for their behind-the-scenes work, which proved indispensable during the conference.

Additional thanks go to Santosh T.Y.S.S., Juan Diego Rodriguez, and Ona de Gibert, our Student Research Workshop Chairs, who, with the support of faculty advisors Yves Scherrer, Elias Stengel-Eskin, and Snigdha Chaturvedi, devoted their work to supporting and mentoring the next generation of researchers. They also secured additional funding from the Vienna Meeting Fund — congratulations!

Our Publication Chairs — David Stap, Haoran Xu, and Jan Niehues — have ensured timely and high-quality conference proceedings. I also thank the Handbook Chairs, Jia Xu and Ryan Cotterell, for assembling the handbook that will guide attendees through the conference program. Of course our conference could not exist without the support of our sponsors! We are deeply grateful to our Sponsorship Chairs — Baotian Hu and Steve Richardson — and to Joel Tetreault, ACLs Sponsorship Director, for securing the generous support that enables us to keep the conference accessible to as many participants as possible.

Fostering an inclusive environment remains a key goal of ACL. Our Diversity and Inclusion Chairs — Ximena Gutierrez-Vasques, Anietie Andy, Gabrielle Gaudeau, and Felermino Ali — have worked to support diverse participation and organize activities that reflect the global nature of our vibrant community. An inclusive ACL allows researchers from all areas and backgrounds to contribute, collaborate, and thrive — enriching both the scientific conversation and the broader societal impact of our work. We remain committed to building a community where everyone feels welcomed and valued.

Thank you to the Publicity and Social Media Chairs — Raj Dabre, Wei Xu, Flor Miriam Plaza-del-Arco, Debora Nozza, and Yaping Zhang — for their communication work through social media, extending the conferences reach to a wider audience. And this brings me to our Website and App Chairs — Abulhair Saparov and Najoung Kim — who kept our community informed with timely updates about all aspects of the conference. The huge number of emails and communications that we received was smoothly managed throughout the year by our Internal Communication Chairs, Ming Jiang and Margot Mieskes. We are very grateful!

Last but not least, I wish to thank the ACL Executive Committee for their guidance, and my fellow past General Chairs for generously sharing their experience and insights.

ACL 2026 is a community-wide effort. Whether you are presenting a paper, leading a workshop or tutorial, volunteering, mentoring, supporting or attending to learn and connect, your presence and participation make a difference at this conference. Thank you for being part of it!

Welcome to San Diego. Welcome to ACL 2026!

Philipp Koehn
Johns Hopkins University
General Chair, ACL 2026

Message from the Program Co-Chairs

Welcome to the 64th Annual Meeting of the Association for Computational Linguistics! ACL 2026 will be held in a hybrid format, offering attendees the option to join us in person in San Diego, California, or to participate remotely from anywhere in the world.

Review Process. All submissions to ACL 2026 went through a two-stage review process. First, papers were submitted to the ACL Rolling Review (ARR), where they were reviewed by reviewers and received meta-reviews from Area Chairs. Then, authors had the option to commit their reviewed papers to ACL via a separate ACL 2026 commitment site. At this stage, Senior Area Chairs provided recommendations, and the Program Chairs made final acceptance decisions based on SAC recommendations, but also with a view to obtaining a balanced program and prioritizing work at a mature state of development. This process is consistent with previous conferences, including ACL 2025, EMNLP 2025, and EACL 2026. We worked closely with the ARR team, particularly the ARR January 2026 Editors-in-Chief, and served as guest Editors-in-Chief for this round. We helped recruit new reviewers and Area Chairs to ARR, resulting in 8,594 reviewers and 1,434 Area Chairs in the January ARR cycle, which received 10,518 submissions. ACL also recruited 255 Senior Area Chairs to oversee the review and meta-review process. Additionally, we collaborated with the ARR October 2025 cycle, which was shared with EACL 2026, and received 1,630 submissions listing ACL as their preferred venue. We recruited 23 Senior Area Chairs and also served as "super" Senior Area Chairs, overseeing the reviewing process of a large number of submissions.

Coordinating a review process of this scale required the efforts of a large number of volunteers. Delays at multiple stages of the process, including reviewing, meta-reviewing, and recommendation, placed additional burdens on the Program Chairs and ARR Editors. We are particularly grateful to those who absorbed this additional workload, serving as emergency Reviewers, Area Chairs, and Senior Area Chairs. For the ACL commitment phase, Senior Area Chairs made recommendations for 6,564 committed papers based on the reviews, meta-reviews, author responses, and the papers themselves, with final acceptance decisions made by the Program Chairs. The vast majority of papers committed to ACL 2026 came from the ARR January 2026 cycle (88%), followed by ARR October 2025 (10.4%). A small number of papers were committed directly from earlier cycles, including July 2025 (1.2%) and May 2025 (0.4%).

During the ARR review process, reviewers identified many non-existent references, which led to those submissions being desk rejected. During the final consistency checks of the camera-ready versions of papers accepted to ACL 2026, we identified over 100 papers that contained citations to non-existent literature. Since these papers were going to be published and were no longer subject to the anonymity requirements, WisPaper was used to identify potential citation issues, and then all flagged citations were reviewed by human experts (PCs and SACs) to confirm that references did not exist. Whether human-authored or LLM-generated, the inclusion of these non-existent references was a clear violation of the ACL Policy on Publication Ethics. For this reason, we had to reject such papers at this late stage. We recommend that future cycles should run such checks both at the submission stage and at the camera-ready stage to prevent the contamination of the scientific literature with false references and to identify such cases as early as possible.

Acceptance Rate. In total, there are 2,296 papers accepted to the Main Conference and 2,163 papers accepted to Findings. The acceptance rate calculation follows the precedent set by previous conferences that go through ACL Rolling Review (ARR), e.g., ACL 2025. The calculation takes into account the multi-stage process of ARR, in which a paper may be revised in ARR and later committed to the conference. In total, we had 12,148 unique submissions across the October 2025 (1,630) and January 2026 ARR (10,518). The acceptance rate is 18.9% for Main Conference papers and 17.8% for Findings papers.

Special Theme: Explainability of NLP Models. Following the precedent of theme tracks from previous events, ACL 2026 features a special theme designed to reflect on and stimulate discourse regarding the current trajectory of the NLP field. Explainability encompasses the methodologies and techniques intended to render the internal mechanisms of complex NLP models, such as Large Language Models (LLMs), transparent and human-interpretable. It seeks to move beyond the "black box" paradigm, where predictions are accepted without scrutiny, and instead aims to elucidate the underlying reasoning of specific model outputs. Explainability is fundamental to establishing trust, ensuring fairness, and enabling responsible deployment. By identifying a model's potential dependence on spurious correlations or inherent societal biases, explainability empowers developers to diagnose system errors, enhance robustness, and ensure accountability. This is particularly vital in high-stakes domains—including healthcare, finance, and law—where interpreting the rationale behind a decision is as essential as the decision itself.

The theme track invited empirical and theoretical work, as well as surveys and position papers reflecting on the Explainability of NLP Models. Suggested topics of discussion included: (i) How do explainability methods need to be adapted for different model architectures? Can we develop a unified framework to evaluate explanations across these architectures?

(ii) How can we rigorously and quantitatively evaluate the quality of an explanation? What metrics can reliably measure the faithfulness (accuracy of the model's reasoning) and plausibility (human-perceived reasonableness) of an explanation? (iii) Can explanations be used to reliably detect when a model is making a biased prediction based on sensitive attributes? How can input-based explanations help mitigate social biases during model training? (iv) Can we use explanations to systematically find and fix problems in the training data itself, such as spurious correlations or annotation errors? How can explainability facilitate a human-in-the-loop process for iterative data refinement? (v) Can we identify specific directions, mechanisms, patterns, or "knobs" within a model's internal activations that control high-level behaviors like abstaining from unanswerable questions? Can we design models that are inherently more interpretable?
We received 124 submissions to the special theme track during the review phase. Among these, 29 papers were accepted to the main conference and a further 27 to the Findings of ACL 2026. The conference will also feature a panel discussion on this theme, with leading experts from both academia and industry, with one of the keynote talks, by Tania Lombrozo, also focused on explainability.

Best Paper Selection. ACL 2026 implemented the updated ACL award policy that seeks to expand the pool of work recognized as outstanding. In total, 158 papers were nominated by the reviewers, area chairs, and senior area chairs for awards. The best paper committee assessed these papers to select the best papers, outstanding papers, and special awards for social impact and best resource. Separately, the senior area chairs also nominated their favorite papers as SAC Highlights. The selection process was double-blind, using anonymized camera-ready versions of the papers. The final selection was made by the best paper committee, and the winners will be announced during the opening ceremony to provide early recognition of the selected contributions and to maximize their visibility within the conference program. In addition, we have the 2026 ACL Dissertation Award, the CL High Impact Paper Award, and the Test of Time Awards.

Program Composition & Presentation Modes. Based on feedback from the conference support staff and the Underline team after ACL 2025, we decided to hold the virtual presentation sessions during the main conference. This enables us to align the virtual sessions with time slots that suit participants in Europe and Asia. This approach allows virtual attendees to participate concurrently with the physical event, avoiding the need for organizers and attendees to engage with the conference twice and separately. We recognize that many virtual attendees may not be able to attend keynotes during their scheduled times at ACL, so we added a third virtual-only keynote in a timezone aligned with Europe and Asia to allow greater live participation and visibility on this year's theme.
The in-person conference will hold 7 Oral/Poster Sessions running in parallel across 10 rooms. This year, 504 main conference papers were selected for oral presentation by the program chairs. The goal was to create a balanced program that featured a diverse set of topics, rather than selecting papers solely on the basis of their review scores. To accommodate a larger number of oral presentations, presentation time was reduced to 8 minutes, followed by 2 minutes of questions. The in-person conference will have 2,078 poster presentations from Main and Findings papers. Rather than have all posters from a track in the same session, sufficiently-sized tracks were divided into two or more sessions to allow authors presenting in one session to attend the other poster sessions for their track. This year, Findings papers will be presented alongside the Main Conference posters, demos, and the Student Research Workshop. This follows the positive experience and feedback of ACL 2025 of holding Main and Findings posters concurrently. In addition to the main conference papers, the ACL program also includes 4 papers accepted by Computational Linguistics (CL) and 28 papers accepted by Transactions of the ACL (TACL). Among these, 18 journal papers will be presented in person as oral presentations, thematically distributed across appropriate sessions.
The virtual conference will feature over 1,000 presentations, including 752 Main Conference and 176 Findings presentations, as well as Demos, Student Research Workshop, TACL, CL, Industry papers, and a virtual-only keynote talk.

Keynotes. We are happy to have three keynote talks at the conference, two in-person and one virtual.

- Philip Resnik (University of Maryland): A New Balancing Act: Reflections on the Relationship between Computational Linguistics and AI
- Yue Zhang (Westlake University): How Can AI Help Us Manage AI Conferences (where Troubles Are Caused by AI) – virtual
- Tania Lombrozo (Princeton University): Explanation and Understanding: A Perspective from Cognitive Science

Acknowledgments. Organizing ACL 2026 has been a collaborative effort, made possible by the dedication and hard work of thousands of people. We gratefully acknowledge the support and contributions of the following people:

- The General Chair, Philipp Koehn;

- The ARR Editors-in-Chief of the January 2026 cycle (Sarvnaz Karimi, Vincent Ng, Xiaodan Zhu, and Margot Mieskes) and the entire ARR team, especially Holy Lovenia, Freda Shi, and Yuntian Deng from the tech team;
- The technical OpenReview chairs, Shuoyang Ding, May Fung, and the OpenReview support team, in particular, Rachel, Emilia, and Harold, for technical help in setting up ACL 2026 on the OpenReview platform;
- The 255 Senior Area Chairs;
- The 1,434 Area Chairs, the 8,594 reviewers, and 171 ethics reviewers;
- The best paper committee chairs, Antonios Anastasopoulos and Eunsol Choi, and the best paper committee members;
- The ethics chairs, Margaret Mitchell and Paco Guzman;
- The conflicts chairs, Isabelle Augenstein and Joel Tetreault;
- The workshop chairs, Loic Barrault and Yang Zhao;
- The tutorial chairs, Kenton Murray and Jacob Andreas;
- The industry track chairs, Georg Rehm, Yunyao Li, and Mei Tu;
- The demonstration chairs, Greg Durrett and Ping Jian;
- The internal communications chairs, Ming Jiang and Margot Mieskes;
- The website and conference app chairs, Abulhair Saparov and Najoung Kim;
- The publication chairs, David Stap, Haoran Xu, and Jan Niehues;
- The handbook chairs, Jia Xu and Ryan Cotterell;
- The visa chairs, David Chiang and Swabha Swayamdipta;
- The publicity chairs, Raj Dabre, Wei Xu, Flor Miriam Plaza-del-Arco, Debora Nozza, and Yaping Zhang;
- The student research workshop chairs, Santosh T.Y.S.S., Juan Diego Rodriguez, and Ona de Gibert;
- The student research workshop chairs faculty advisors, Yves Scherrer, Elias Stengel-Eskin, and Snigdha Chaturvedi;
- The student volunteer chairs, Jingbo Shang, Lianhui Qin, Xuezhe Ma, and Muhao Chen;
- The diversity and inclusion chairs, Ximena Gutierrez-Vasques, Anietie Andy, Gabrielle Gaudeau, and Felermino Ali;
- The sponsorship chairs, Baotian Hu and Steve Richardson;
- The virtual infrastructure chairs, Gael Guibon, and Parisa Kordjamshidi;
- The ACL Anthology Director Matt Post and his team;
- The TACL editors-in-chief (Asli Celikyilmaz, Roi Reichart, Dilek Hakkani-Tur, and Vivian Chen) and CL Editor-in-Chief Wei Lu for coordinating TACL and CL presentations with us;
- Past program chairs for *CL conferences, in particular Christos Christodoulopoulos (EMNLP 2025), Vera Demberg (EACL 2026), and Ekaterina Shutova (ACL 2025), for information and support;

- Qi Zhang, developer of WisPaper, for checking camera-ready papers to ensure they did not contain hallucinated references;
- Damira Mri and the Underline team for help with the program;
- Jennifer Rachford and the entire conference support staff for their crucial help and patience as we learned our way through running such a complex process;
- All the authors of papers who submitted their papers for review in the ARR 2025 October and ARR 2026 January cycles, and those who committed to the ACL 2026 conference.

We hope you enjoy this year's huge and diverse program!

Maria Liakata (Queen Mary University of London, UK)
Viviane Moreira (UFRGS, Brazil)
Jiajun Zhang (Chinese Academy of Sciences, China)
David Jurgens (University of Michigan, USA)
ACL 2026 Program Co-Chairs

ACL Business and Event Manager Message

On behalf of ACL 2026, I would like to extend my sincere gratitude to the many individuals and partners who have made this conference possible. First and foremost, thank you to the ACL Board, for your continued guidance and support.

ACL 2026 Onsite Team

- Lina Staggs, Event Coordinator
- Hannah Pietsch, Rooming Coordinator
- Paige Bower, On-site Registration
- Sally Stevenson, Brandy Dorsey, and Jason Staggs, On-site Auxiliary Staff
- Lee Hartman & Sons, Support Staff
- Jon Dorsey, AV Director
- Trevor Laffoon, AV Technician

Registration Team
YesEvents—Terah Shaffer and Nichole Ballard

Hybrid Virtual Team
Managing content, streaming, and so much more behind the scenes:

- Damira Mri
- Luka imic
- Borna Bevanda
- Dora Bolanca
- Dora Brodanac
- Rafael Grabovica
- Petra Vizintin
- Leticija Maricic
- Astrid Plio
- Mateo Ruic
- Matija ild
- Petar Vodinac
- Noa Zmazek

Your expertise ensures a seamless experience for both our in-person and virtual attendees.

Venue Partner
Manchester Grand Hyatt San Diego

- Allison Wingard, Sales Manager
- Meghan Muxen, TBHS Senior Event Planning Manager
- Timothy McGraw, Reservations Coordinator

Networking
Single Digits—Marshal Gewarges, Conference Network Manager

AV & Production
Encore—Clarissa Sanchez, Sales Manager

Exhibit Booths & Poster Boards
LVE—Kathy Hahn

Branding & Printing Partners

- Marcelo Viridiano—Graphics & Design
- Arc/Riot—Kelvin Winters, Marketing Leader

Welcome Gifts

- 4Imprint
- Custom Lanyards

Handbooks
Omni Press

Finally, thank you to the entire Organizing and Program Committees for your hard work, dedication, attention to detail, and countless hours of effort. Conferences of this scale require extraordinary coordination and commitment, and I deeply appreciate each and every one of you. ACL 2026 would not be possible without your collective passion and professionalism.

With sincere appreciation,

Jennifer Rachford
ACL Director of Events / Business Manager

Organizing Committee

General Chair

Philipp Koehn, Johns Hopkins University

Program Chairs

Maria Liakata, Queen Mary University of London
Viviane P. Moreira, Institute of Informatics at UFRGS, Brasil
Jiajun Zhang, Chinese Academy of Sciences
David Jurgens, University of Michigan

Workshop Chairs

Loic Barrault, Meta
Yang Zhao, Chinese Academy of Sciences

Tutorial Chairs

Kenton Murray, Johns Hopkins University
Jacob Andreas, Massachusetts Institute of Technology

Demonstration Chairs

Greg Durrett, New York University
Ping Jian, Beijing Institute of Technology

Student Research Workshop Chairs

Santosh T.Y.S.S., Technical University of Munich & Amazon
Juan Diego Rodriguez, University of Texas at Austin
Ona de Gibert, University of Helsinki

Student Research Workshop Chairs: Faculty Advisors

Yves Scherrer, University of Oslo
Elias Stengel-Eskin, University of Texas at Austin
Snigdha Chaturvedi, University of North Carolina

Publication Chairs

David Stap, NXAI
Haoran Xu, Microsoft
Jan Niehues, Karlsruhe Institute of Technology

Handbook Chairs

Jia Xu, Rutgers University
Ryan Cotterell, ETH Zürich

Sponsorship Chairs

Baotian Hu, Harbin Institute of Technology
Steve Richardson, Brigham Young University

Diversity and Inclusion Chairs

Ximena Gutierrez-Vasques, Universidad Nacional Autonoma de Mexico
Anietie Andy, Howard University
Gabrielle Gaudeau, University of Cambridge
Felermino Ali, Microsoft Research Africa

Publicity Chairs

Raj Dabre, Google
Wei Xu, Georgia Tech
Flor Miriam Plaza-del-Arco, University of Leiden
Debora Nozza, Bocconi University
Yaping Zhang, Chinese Academy of Sciences

Website and Conference App Chairs

Abulhair Saparov, Purdue University
Najoung Kim, Boston University

Ethics Chairs

Margaret Mitchell, Huggingface
Paco Guzman, Meta

Student Volunteer Chairs

Jingbo Shang, University of California at San Diego
Lianhui Qin, University of California at San Diego
Xuezhe Ma, University of Southern California
Muhao Chen, University of California at Davis

Technical Open Review Chairs

Shuoyang Ding, NVIDIA
Yi R. (May) Fung, HKUST

Virtual Infrastructure Chairs

Gael Guibon, Université Sorbonne Paris Nord
Parisa Kordjamshidi, Michigan State University

Internal Communication Chairs

Ming Jiang, Indiana University Indianapolis
Margot Mieskes, University of Applied Sciences, Darmstadt

Industry Track Chairs

Georg Rehm, German Research Institute for Artificial Intelligence
Yunyao Li, Adobe
Mei Tu, Samsung

Best Paper Chairs

Antonios Anastasopoulos, George Mason University
Eunsol Choi, New York University

Visa Chairs

David Chiang, University of Notre Dame
Swabha Swayamdipta, University of Southern California

Conflict Chairs

Isabelle Augenstein, University of Copenhagen
Joel Tetreault, Dataminr

ARR Editors-in-Chief

Sarvnaz Karimi, Monash University and CSIRO
Vincent Ng, University of Texas at Dallas
Xiaodan Zhu, Queen's University
Margot Mieskes, University of Applied Sciences, Darmstadt

3

Anti-Harassment Policy

ACL 2026 adheres to the ACL Anti-Harassment Policy. Any participant who experiences harassment or hostile behavior may contact any current member of the ACL Professional Conduct Committee or Jennifer Rachford, who is usually available at the registration desk of the conference. Please be assured that if you approach us, your concerns will be kept in strict confidence, and we will consult with you on any actions taken.

The open exchange of ideas, the freedom of thought and expression, and respectful scientific debate are central to the aims and goals of an ACL conference. These require a community and an environment that recognizes the inherent worth of every person and group, that fosters dignity, understanding, and mutual respect, and that embraces diversity. For these reasons, ACL is dedicated to providing a harassment-free experience for participants at our events and in our programs.

Harassment and hostile behavior are unwelcome at any ACL conference. This includes speech or behavior (including in public presentations and in online discourse) that intimidates, creates discomfort, or interferes with a person's participation or opportunity for participation in the conference. We aim for ACL conferences to be an environment where harassment in any form does not happen, including but not limited to: harassment based on race, gender, religion, age, color, national origin, ancestry, disability, sexual orientation, or gender identity. Harassment includes degrading verbal comments, deliberate intimidation, stalking, harassing photography or recording, inappropriate physical contact, and unwelcome sexual attention.

The ACL board members are listed at:
https://www.aclweb.org/portal/about

The full policy and its implementation are defined at:
https://aclweb.org/adminwiki/index.php/Anti-Harassment_Policy

4 Code of Ethics

ACL adopts the ACM Code of Ethics, in the version adopted June 22, 2018, by the ACM Council: https://www.acm.org/code-of-ethics In its application to ACL, it is to be read in the contextually appropriate interpretation; e.g., "ACM member" is to be read as "ACL member." Sec. 4.2 should be read as follows:

> *4.2 Treat violations of the Code as inconsistent with membership in the ACL. Each ACL member should encourage and support adherence by all members of the CL/NLP community regardless of ACL membership. ACL members who recognize a breach of the Code should consider reporting the violation to the ACL, which may result in remedial action.*

The open exchange of ideas, the freedom of thought and expression, and respectful scientific debate are central to the aims and goals of the ACL. These require a community and an environment that recognizes the inherent worth of every person and group, that fosters dignity, understanding, and mutual respect, and that embraces diversity. For these reasons, ACL is dedicated to providing a harassment-free experience for all members, as well as for participants at our events and in our programs.

Harassment and hostile behavior are unwelcome at any ACL conference, associated event, or ACL-affiliated online discussion. This includes speech or behavior that intimidates, creates discomfort, or interferes with a person's participation or opportunity for participation in a conference or an event. We aim for ACL-related activities to be an environment where harassment in any form does not happen, including but not limited to: harassment based on race, gender, religion, age, color, appearance, national origin, ancestry, disability, sexual orientation, or gender identity. Harassment includes degrading verbal comments, deliberate intimidation, stalking, harassing photography or recording, inappropriate physical contact, and unwelcome sexual attention.

The policy is not intended to inhibit challenging scientific debate, but rather to promote it by ensuring that all are welcome to participate in the shared spirit of scientific inquiry. Vexatious complaints and willful misuse of this procedure will render the complainant subject to the same sanctions as a violation of the anti-harassment policy.

It is the responsibility of the community as a whole to promote an inclusive and positive environment for our scholarly activities. In addition, anyone who experiences harassment or hostile behavior may contact any member of the Professional Conduct Committee. Members of this committee are instructed to keep any such contact in strict confidence, and those who approach the committee will be consulted before any actions are taken.

5 Social Events

Welcome Reception

Dear Authors and Attendees,

We are delighted to invite you to the ACL 2026 Welcome Reception—a wonderful opportunity to kick off the conference in a relaxed and engaging setting.

Venue: Manchester Grand Hyatt (1 Market Pl, San Diego, CA 92101)

Location: Level 4—Follow Hyatt signs to the Sports Court and ACL directional signage

Date & Time: Saturday, July 4, 2026, 18:00-20:00

Join us for an evening of networking, refreshments, and lively conversation with fellow members of the natural language processing community. Whether you're reconnecting with colleagues or meeting new peers, this reception offers the perfect start to an inspiring conference experience.

Enjoy light appetizers and a complimentary drink ticket while connecting with researchers, practitioners, and industry leaders from around the world. First-time attendees and returning participants alike will find this a welcoming and energizing way to begin ACL 2026.

Celebrate July 4th in San Diego

Following the reception, we encourage you to take part in a special local tradition celebrating the 250th Anniversary of the United States—San Diego's Big Bay Boom fireworks show. Beginning at 9:00 PM, the display can be viewed along the bay near the hotel. It is expected to be crowded, but widely regarded as well worth the experience!

Important Information

- Attendance is included for all Main Conference registrants.
- If you are not registered for the Main Conference but wish to attend, you may add this event at the Registration Solutions Desk, located on the lobby level.
- Don't forget: the Opening Session begins at 9:00 the following morning in the Harbor Ballroom, as we officially kick off ACL 2026.

We look forward to welcoming you to San Diego and to an exciting ACL 2026!

Gala Dinner

Dear Authors and Attendees,

We are excited to invite you to the ACL 2026 Social Event, taking place aboard one of San Diego's most iconic landmarks—the historic USS Midway.

Venue: USS Midway Museum (910 N Harbor Dr, San Diego, CA 92101)

Location: A short walk from the Manchester Grand Hyatt along the waterfront

Date & Time: Monday, July 6, 2026, 18:30-22:00

Step aboard this legendary aircraft carrier—once the longest-serving U.S. Navy aircraft carrier of the 20th century—for an unforgettable evening of networking, exploration, and fun. Now a world-renowned floating museum, the USS Midway offers a unique opportunity to experience history firsthand while connecting with colleagues from across the ACL community.

What to Expect On Board

- Explore the expansive flight deck featuring restored aircraft, helicopters, and fighter jets
- Try out interactive flight simulators and hands-on exhibits
- Enjoy breathtaking views of San Diego Bay at sunset and into the evening
- Walk through authentic crew quarters, engine rooms, and command centers
- Engage with docents and exhibits sharing real stories of life at sea

Fun Activities & Experiences

- Themed networking spaces across multiple decks
- Music and entertainment in a truly unique historic setting
- Food and drinks served throughout the evening
- Plenty of opportunities for photos on the flight deck and around the ship

With over 60 exhibits and a massive multi-level layout, the USS Midway provides a one-of-a-kind venue blending history, innovation, and community. We recommend wearing comfortable shoes as you explore the ship's many levels and spaces.

We look forward to welcoming you aboard for an unforgettable evening!

Important Information

- Attendance is included for all Main Conference registrants.
- If you are not registered for the Main Conference but wish to attend, you may add this event at the Registration Solutions Desk, located on the lobby level.
- Don't forget: the Oral/Poster/Demo Sessions begin at 9:00 the following morning.

Meals

To help you plan your time at ACL 2026, please find below an overview of meals, refreshments, and related arrangements during the conference.

Refreshments
July 2-7, 2026. Refreshments will be offered twice daily—mid-morning and mid-afternoon. Refreshments include: coffee, tea, pastries, and fruit.

Refreshment Locations:

- **Tutorial Days:** Harbor Foyer
- **Workshop Days:** Harbor Foyer, Gaslamp Foyer, & Grand Ballroom
- **Main Conference Days:** Harbor Foyer, Coronado Foyer, Grand Ballroom, & Gaslamp Foyer

Please note: Breakfast, lunch, and dinner are not included in conference registration. Kindly refer to individual event announcements for details on included receptions and social gatherings.

For exact timings, please refer to the detailed conference agenda.

Important Reminder
Access to all conference spaces—including refreshments and social events—requires a valid ACL 2026 conference badge and appropriate registration for the specific day, session, or event you are attending.

Additional social events, including the Welcome Reception and Social Event, will be detailed separately.

We look forward to welcoming you to ACL 2026 and hope you enjoy both the program and the vibrant San Diego dining scene!

Warm regards,
The ACL 2026 Organizing Committee

6 Keynotes

Barbara Plank

LMU Munich & IT University of Copenhagen

ACL 2026 Presidential Address

Date & Time: Monday, July 6, 2026, 15:30 **Location:** Harbor Ballroom, Level 2

Bio

Barbara Plank is Professor and co-director of the Center for Information and Language Processing at LMU Munich. She holds the Chair for AI and Computational Linguistics at LMU Munich and is a visiting Professor at the Computer Science department at the IT University of Copenhagen. She directs the MaiNLP research lab (Munich AI and NLP lab) which focuses on Natural Language Processing with an emphasis on human-inspired and data-centric approaches. Her research has been funded by distinguished grants and awards, including an ERC Consolidator Grant (DIALECT) and a DFF Sapere Aude Research Leader grant. She is an ELLIS Fellow and regularly serves on international committees, including ACL and EACL, as well as on scientific advisory boards in academia and industry.

Philip Resnik
University of Maryland

A New Balancing Act: Reflections on the Relationship between Computational Linguistics and AI

Date & Time: Sunday, July 5, 2026, 9:30 **Location:** Harbor Ballroom, Level 2

Abstract
Beginning in the mid-1980s, a "statistical revolution" in our field redefined what it meant to work on language from a computational perspective. Steve Abney, in his essay opening *The Balancing Act: Combining Symbolic and Statistical Approaches to Language* (Klavans and Resnik, eds., MIT Press, 1996), wrote: "In the space of the last ten years, statistical methods have gone from being virtually unknown in computational linguistics to being a fundamental given." We are again caught up in revolutionary times. Except this time around, the revolution is hurtling forward even faster, and it is redefining not only the research we do, but what it means to do research in the first place, the role of research in society, and society itself. What does this mean for ACL—what defines us as a community, what is our role, and where do we go from here?

Bio
Philip is a Professor at Maryland with joint appointments in Linguistics and the Institute for Advanced Computer Studies. He was named an ACL Fellow in 2020 for contributions to symbolic-statistical methods, multilinguality, and interdisciplinary language study. His recent research focuses on computational social science, computational cognitive neuroscience of language, and AI's relationship to human cognition and society. Industry experience includes positions at Bolt Beranek and Newman, Sun Microsystems Laboratories, and IBM; he co-founded CodeRyte (acquired by 3M in 2012) and advises Trustible and FiscalNote. He earned his B.S. from Harvard and Ph.D. from the University of Pennsylvania.

Yue Zhang

Westlake University (Virtual keynote)

How Can AI Help Us Manage AI Conferences (where Troubles Are Caused by AI)

Date & Time: Monday, July 6, 2026, 01:00 **Location: Underline (Virtual)**

Abstract
ACL has become 30+ times larger than two decades ago, and we face issues such as overwhelming participants, outdated papers, and low quality review. The reason can largely be attributed to increasing research in LLM, and AI accelerating research. Can AI help us address challenges in academic publication and evaluation? How can we accommodate AI-driven research? Could the research and publication paradigm be improved? In this talk, I will try to give some thoughts, sharing some of our recent work on AI review automation, paper recommendation, and AI arXiv, before briefly discussing a bit of AI research automation.

Bio
Yue is a Professor at Westlake University and ACL Fellow whose research covers fundamental NLP, machine learning algorithms, and recent work on LLM reasoning and AI scientist development. Major contributions include machine learning for structured prediction (parsing, IE), neural NLP models (lattice and graph LSTM), and generalization research (OOD, logical reasoning). He is coauthor of *Natural Language Processing – a Machine Learning Perspective*, published by Cambridge University Press. He served as PC co-chair for CCL 2020, EMNLP 2022, and LMG 2025; general chair for AACL 2026 and NLPCC 2026; and test-of-time award co-chair for ACL 2024–2025. He is currently editor-in-chief for the LLM journal and action editor for TACL.

Tania Lombrozo

Princeton University

Explanation and Understanding: A Perspective from Cognitive Science

Date & Time: Monday, July 6, 2026, 16:45 **Location:** Harbor Ballroom, Level 2

Abstract

What makes some explanations more satisfying than others? Why are humans so driven to explain? In this talk I'll take a cognitive scientific approach to explanation and understanding, drawing on decades of work in psychology, education, and philosophy to argue for a function-first approach to explanations—one on which we begin by characterizing what it is that explaining achieves. I'll suggest that the process of explaining plays a crucial role in discovery and generalization, and that we can understand the structure and phenomenology of explanations by first appreciating how explaining drives learning and helps construct our understanding of the world. I'll also consider some potential implications for explanation and understanding in AI systems.

Bio

Tania is the Arthur W. Marks '19 Professor of Psychology at Princeton, Associate of the Philosophy Department and University Center for Human Values, director of the Program in Cognitive Science, and co-director of Natural and Artificial Minds within the AI Lab. She earned her Ph.D. in Psychology from Harvard in 2006 following a B.S. in Symbolic Systems and B.A. in Philosophy from Stanford. Her research addresses foundational cognition questions using cognitive psychology and analytic philosophy, focusing on explanation, understanding, conceptual representation, categorization, social cognition, causal reasoning, and folk epistemology. She has received the Stanton Prize, Spence Award, NSF CAREER award, and James S. McDonnell Foundation Scholar Award. Her book *Why We Ask Why: The Science of Explanation and the Human Drive to Understand* releases in October.

7 Panels

Plenary Panel: Explainability of LLMs: Academic vs Industry Perspective

Monday, July 6, 2026, 14:45-15:30, Harbor Ballroom Level 2

Moderator: Julian McAuley (UCSD)

Panelists: Tal Linzen (NYU/Google), Sameer Singh (UCI/Spiffy AI), Kamalika Das (Intuit)

Industry Panel: Careers in NLP

Monday, July 6, 2026, 09:00-10:00, Balboa (L2, Seaport)

Moderator: Yang Liu (Amazon AGI)

Panelists: Roberto Navigli (Sapienza University of Rome and Babelscape), Iz Beltagy (AI2/Envive AI), Heng Ji (UIUC), Trung Bui (Adobe Research)

8 Tutorials

Overview

Thursday, July 2, 2026

9:00-12:30	**T01**: Towards Effective and Efficient Multi-Agent Language Model Systems: Foundations, Prospects, and Applications *Xuan Wang, Shuxiang Cao, Yuchen Zhuang and Wenqi Shi*	Harbor G - I
9:00-12:30	**T02**: The Data Frontier for Large Language Models: Selection, Synthesis, and Tools *Lijun Wu, Wentao Zhang and Conghui He*	Harbor D - F
9:00-12:30	**T03**: Future of Work in the Age of LLMs *Zhiruo Wang, Yijia Shao, David Nguyen and Diyi Yang*	Harbor A - C
12:30–14:00	Lunch	On Own
14:00-17:30	**T04**: Current Advances in LLM Reasoning *Akhil Arora, Vishrav Chaudhary, Julia Kreutzer, Nearchos Potamitis, Nouha Dziri and Niket Tandon*	Harbor A - C
14:00-17:30	**T05**: Knowledge Control for Responsible Generative AI: Bridging Academia,Industry, and Society *Zheyuan Liu, Yixin Wan, Kai-Wei Chang, Meng Jiang, Jieyu Zhao, Nouha Dziri, Yuning Mao, Jiachen Gu, and Jindong Gu*	Harbor D - F
14:00-17:30	**T06**: The Interplay between Metaphors and NLP *Joanne Boisson, Jose Camacho-Collados, Elisa Sanchez-Bayona and Rodrigo Agerri*	Harbor G - I

Message from the Tutorial Co-Chairs

Welcome to the Tutorials Program of ACL 2026. The tutorials session is organized to give conference attendees a comprehensive introduction by expert researchers to topics of importance drawn from our rapidly growing and changing research field.

This year, the call, submission, reviewing, and selection of tutorials were coordinated jointly across multiple conferences. From a pool of submissions, we selected six tutorials for ACL 2026, covering multi-agent language model systems, the future of work in the age of LLMs, data-centric approaches to LLMs, the interplay between metaphors and NLP, knowledge control for responsible generative AI, and current advances in LLM reasoning.

We thank the tutorial authors for their contributions, the reviewers for their thoughtful evaluations, and the conference organizers for their support.

We hope you enjoy the tutorials.

Kenton Murray (Johns Hopkins University)
Jacob Andreas (Massachusetts Institute of Technology)
ACL 2026 Tutorial Co-Chairs

T01: Towards Effective and Efficient Multi-Agent Language Model Systems: Foundations, Prospects, and Applications

Organizers: Xuan Wang, Shuxiang Cao, Yuchen Zhuang and Wenqi Shi
Date & Time: Thursday, July 2, 2026, 9:00-12:30
Room: Harbor G - I

Multi-agent systems powered by large language models (LLMs) offer a promising paradigm for tackling complex reasoning, decision-making, and problem-solving tasks. However, achieving both effectiveness and efficiency in such systems remains a critical challenge. This tutorial introduces recent advances in building effective and efficient multi-agent LLM systems, focusing on three core components. First, we discuss the design of individual LLM agents. We present state-of-the-art techniques for enabling capable agents using efficient and compact LLMs, including model distillation, dynamic routing, and memoryand compute-efficient serving, providing a foundation for scalable and responsive agent design under resource constraints. Second, we cover coordination and communication among agents, crucial for collective performance, highlighting methods for improving multi-agent reasoning and decision-making through prompt and graph optimization, sycophancy mitigation, and structured LLMbased frameworks. Last, we explore real-world applications of LLM agents in areas such as industry, healthcare, quantum computing, and various scientific domains.

T02: The Data Frontier for Large Language Models: Selection, Synthesis, and Tools

Organizers: Lijun Wu, Wentao Zhang and Conghui He
Date & Time: Thursday, July 2, 2026, 9:00-12:30
Room: Harbor D - F

As the development of Large Language Models (LLMs) matures, the focus of the research community is undergoing a critical shift from a purely model-centric to a data-centric paradigm. It is now evident that the quality, diversity, and composition of training datanot merely its scaleare the primary drivers of a models advanced capabilities, from complex reasoning to reliable instruction following. However, acquiring and curating such high-quality data remains a significant bottleneck. This tutorial provides a comprehensive and practical guide to the state-of-the-art in data research directions for LLMs, structured around the two core pillars of modern data strategy: intelligent data selection and advanced data synthesis. In the first part, we delve into methods for curating the most valuable information from vast, noisy datasets, covering techniques like LLM-as-a-judge for automated quality filtering and active learning for maximizing annotation efficiency. The second part explores the synthetic data revolution, detailing paradigms that range from generating complex reasoning traces (e.g., Chain-of-Thought) to deploying sophisticated multi-agent workflows that can autonomously create high-quality, diverse instruction data from raw seeds. Finally, we will conclude with a practical overview of open-source tools and platforms that facilitate these data-centric workflows, empowering researchers and practitioners to build better models through better data. Attendees will leave with a principled framework and actionable insights for designing and implementing the advanced data strategies required to build the next generation of powerful, specialized, and aligned LLMs.

T03: Future of Work in the Age of LLMs

Organizers: Zhiruo Wang, Yijia Shao, David Nguyen and Diyi Yang
Date & Time: Thursday, July 2, 2026, 9:00-12:30
Room: Harbor A - C

The recent development of large language models (LLMs) has revolutionized the landscape of human work. These models possess the ability to follow complex human instructions and operate versatile computer software, enabling them to participate in, augment, or even automate realistic occupational tasks that once thought to be exclusive to humans. As LLMs are increasingly integrated into workplaces, they are already reshaping labor dynamics and raising urgent concerns about job displacement, diminished human agency, and overreliance on automation. As a result, the future of work is undergoing a profound transformation: How will human occupations and task requirements evolve over time? And what roles will LLM-based systems play, as they become increasingly capable collaborators and autonomous workers? And how can we build technological and data infrastructures to support humanAI collaboration? This tutorial will provide an overview of the future of work shaped by the interplay of LLMs and humans, examining the emerging challenges, opportunities, and ethical considerations in this dynamic landscape. We begin by examining the economic landscape of work and how NLP technologies drive automation, followed by methods for developing LLMs that augment human labor and recent advances in LLM-based agents. We then cover evaluation approaches for workplace contexts, including datasets, benchmarks, and metrics, and conclude with open questions on technical, human, and societal implications.

T04: Current Advances in LLM Reasoning

Organizers: Akhil Arora, Vishrav Chaudhary, Julia Kreutzer, Nearchos Potamitis, Nouha Dziri and Niket Tandon

Date & Time: Thursday, July 2, 2026, 14:00-17:30

Room: Harbor A - C

As large language models (LLMs) increasingly tackle reasoning-heavy tasks, from mathematics to commonsense to multilingual understanding, researchers face three pressing questions: How well do models reason? How can we make them reason better? What are the next frontiers in LLM reasoning? This tutorial answers these questions through a unified view of LLM reasoning. It explores comprehensive evaluation strategies to assess the reasoning abilities of models and discusses two types of methods to improve models reasoning: (i) advanced inference-time methods, such as structured and self-improvement inference methods, and (ii) post-training methods, such as RLHF, DPO, and GRPO, that aim to make LLMs think more like humans. The tutorial covers these technical discussions while maintaining a practical outlook through illustrative demos and short guided hands-on exercises, and is designed for both researchers and practitioners seeking practical insights into LLM reasoning.

T05: Knowledge Control for Responsible Generative AI: Bridging Academia,Industry, and Society

Organizers: Zheyuan Liu, Yixin Wan, Kai-Wei Chang, Meng Jiang, Jieyu Zhao, Nouha Dziri, Yuning Mao, Jiachen Gu, and Jindong Gu
Date & Time: Thursday, July 2, 2026, 14:00-17:30
Room: Harbor D - F

This tutorial explores post-training knowledge control as a practical and responsible framework for shaping the behavior of large generative models. It brings together perspectives from academia, industry, and society to address the growing need for scalable, interpretable, and ethically grounded interventions in AI systems. Through discussions of diverse knowledge control techniques like machine unlearning, knowledge editing, and inference-time control, the tutorial illustrates how these approaches can remove or adjust undesirable model behaviors while preserving integrity and usefulness. It emphasizes bridging research, deployment, and societal perspectives, offering attendees a holistic understanding of how advances across academia, industry, and society can jointly promote safer, fairer, and more accountable generative AI.

T06: The Interplay between Metaphors and NLP

Organizers: Joanne Boisson, Jose Camacho-Collados, Elisa Sanchez-Bayona and Rodrigo Agerri
Date & Time: Thursday, July 2, 2026, 14:00-17:30
Room: Harbor G - I

This tutorial will provide an overview of the field of metaphor processing in NLP. We will discuss the influence of various metaphor theories on the creation of annotated resources and on the research projects and algorithm designs. We will particularly focus on recent research made possible by the last generation of transformer-based models and LLMs for metaphor interpretation in multilingual and multimodal settings. Two types of audience may benefit from this tutorial: researchers in the humanities and computational social sciences interested in automatic or semi-automatic metaphor analysis, and NLP researchers interested in understanding metaphor processing and improving metaphor modeling.

9 Workshops

Overview

During the workshop days, **Registration** is held from 08:00. AM and PM coffee breaks are provided; see each workshop for start, lunch, and end times.

Friday, July 3, 2026

Half-day workshop (morning 9:00 - 12:30)

W01: CustomNLP4U: Second Workshop on Customizable NLP: Progress and Challenges in Customizing NLP for a Domain, Application, Group, or Individual	Old Town
W02: PrivateNLP: Seventh Workshop on Privacy in Natural Language Processing	Balboa C

Full-day workshop

W03: ALVR: 4th Workshop on Advances in Language and Vision Research	Harbor H&I
W04: ArgMining: the 13th Workshop on Argument Mining and Reasoning	Harbor E&F
W05: CODI-CRAC: Joint Workshop on Computational Approaches to Discourse, Context and Document-Level Inferences and Computational Models of Reference, Anaphora and Coreference	Regatta
W06: DravidianLangTech: The Sixth Workshop on Speech and Language Technologies for Dravidian Languages	Virtual-only

W07: EEUCA: The 9th Workshop on Event Extraction and Understanding: Challenges and Applications	Pier
W08: LAW XX: The 20th Linguistic Annotation Workshop	La Jolla
W09: SELVA: The 1st International Workshop on Sustainable and Efficient Language, Vision, and Action Models	Virtual-only
W10: SURGeLLM: Structured Understanding, Retrieval, and Generation in the LLM Era	Harbor D
W11: KnowFM: 4th Workshop on Towards Knowledgeable Foundation Models	Virtual-only
W12: CHum: The Second Workshop on Computational Humor	Virtual-only
W13: SMM4H-HeaRD: Joint 11th Social Media Mining for Health Applications and Health Real-World Data Workshop and Shared Tasks	Virtual-only

Half-day workshop (Afternoon 14:00 - 17:30)

W14: AmericasNLP: The Sixth Workshop on NLP for Indigenous Languages of the Americas	Old Town
W15: StereACuLT: The 1st Workshop on Stereotypes Across Cultures in Language Technologies	Balboa C
W16: SCiL: Society for Computation in Linguistics	Promenade
W17: IWSLT: ACL-IWSLT: 23rd International Conference on Spoken Language Translation	Gaslamp C&D
W18: SemEval: The 20th International Workshop on Semantic Evaluation	Harbor B
W19: BioNLP and Shared Tasks	Balboa A&B
W20: StarSEM: *SEM	Harbor A
W21: BEA: The 21st Workshop on Innovative Use of NLP for Building Educational Applications	Virtual-only (Fri) Harbor D (Sat)
W22: CoNLL: The SIGNLL Conference on Computational Natural Language Learning	Harbor C

Saturday, July 4, 2026

Half-day workshop (morning 9:00 - 12:30)

W23: TrustNLP: The 6th Workshop on Trustworthy NLP	Harbor G
W24: BigPicture: The Big Picture: Crafting a Research Narrative v2	Gaslamp A&B
W25: RAG4Reports: The First Workshop on Multilingual Report Generation via Retrieval Augmented Generation	Harbor A

Full-day workshop

W26: LTEDI: Sixth Workshop on Language Technology for Equality, Diversity and Inclusion	Virtual-only
W27: MAGMaR: The Second Workshop on Multimodal Augmented Generation via Multimodal Retrieval	Old Town
W28: GEM: Natural Language Generation, Evaluation, and Metrics	Harbor E&F
W29: NLP4DH: The 6th International Conference on Natural Language Processing for the Digital Humanities	Regatta
W30: ComputEL-9: The Ninth Workshop on the Use of Computational Methods in the Study of Endangered Languages	Balboa C
W31: NLP+CSS: Seventh Workshop on NLP and Computational Social Science	Gaslamp A&B
W32: CLPsych: Eleventh Workshop on Computational Linguistics and Clinical Psychology	Harbor H&I
W33: BATCH: Beyond Alignment: Transdisciplinary Conversations on Human-AI Futures	Pier

Half-day workshop (Afternoon 14:00 - 17:30)

W16: SCiL: Society for Computation in Linguistics	Promenade
W17: IWSLT: ACL-IWSLT: 23rd International Conference on Spoken Language Translation	Gaslamp C&D
W18: SemEval: The 20th International Workshop on Semantic Evaluation	Harbor B
W19: BioNLP and Shared Tasks	Balboa A&B
W20: StarSEM: *SEM	Harbor A
W21: BEA: The 21st Workshop on Innovative Use of NLP for Building Educational Applications	Virtual-only (Fri) Harbor D (Sat)
W22: CoNLL: The SIGNLL Conference on Computational Natural Language Learning	Harbor C

W34: CDL: The 1st Workshop on Computational Developmental Linguistics	Harbor G
W35: C3NLP: Cross-Cultural Considerations in NLP	La Jolla
W36: EvalEval: Second Workshop on Evaluating Evaluations: Examining Best Practices for Utilizing and Developing Generative Model Evaluations	Harbor A
W37: MeLLMs: Multilinguality in the Era of Large Language Models	Gaslamp A&B

W01: CustomNLP4U: Second Workshop on Customizable NLP: Progress and Challenges in Customizing NLP for a Domain, Application, Group, or Individual

Organizers: Sheshera Mysore, Sachin Kumar, Vidhisha Balachandran, Shirley Anugrah Hayati, Hanane Moussa, Alireza Salemi, Faeze Brahman

Venue: Old Town

Date: Friday, July 3, 2026

Half-day workshop (morning 9:00 - 12:30)

Invited speakers: Natasha Jaques (University of Washington), Ramya Korlakai Vinayak (University of Wisconsin-Madison)

https://customnlp4u.github.io

The Second Workshop on Customizable NLP addresses the limitations of generalist language models by focusing on tailoring AI systems to diverse users, domains, and contexts. The workshop encompasses research across multiple dimensions: data collection and privacy considerations, novel modeling approaches for personalization, evaluation methodologies, real-world applications in specialized fields like healthcare and law, and broader ethical implications including security and copyright concerns. Users expectations, values, and workflows can vary significantly across domains, applications, organizations, geography, social groups, cultures, and individuals, requiring development of models that are easily controlled by them and learn over time. The workshop welcomes both short and long papers exploring how customizable language models can deliver improved performance for underserved communities, specialized industries, and individual end-users while maintaining responsible development practices.

W02: PrivateNLP: Seventh Workshop on Privacy in Natural Language Processing

Organizers: Ivan Habernal, Sepideh Ghanavati, Sara Haghighi, Krithika Ramesh, Timour Igamberdiev, Shomir Wilson
Venue: Balboa C
Date: Friday, July 3, 2026
Half-day workshop (morning 9:00 - 12:30)
Invited speakers: Franziska Boenisch (CISPA Helmholtz Center for Information Security)
https://sites.google.com/view/privatenlp2026

The seventh iteration of this workshop convenes researchers and practitioners to address privacy-preserving data analysis in natural language processing, particularly relevant given LLMs reliance on vast datasets. The workshop examines challenges across designing, building, verifying, and testing privacy-preserving NLP systems, with topics ranging from differential privacy and federated learning to machine unlearning and memorization attacks. Submissions may include original research papers, position papers, and fast-track submissions from other venues, all undergoing double-blind review for presentation at the July 2026 event in San Diego.

W03: ALVR: 4th Workshop on Advances in Language and Vision Research

Organizers: Qianqi (Jackie) Yan, Syrielle Montariol, Yue Fan, Jing Gu, Jiayi Pan, Manling Li, Parisa Kordjamshidi, Alane Suhr, Xin Eric Wang
Venue: **Harbor H&I**
Date: Friday, July 3, 2026
Full-day workshop
Invited speakers: Raymond Mooney, Mohit Bansal, Jiajun Wu, Lianhui Qin, Amir Zadeh
https://alvr-workshop.github.io

The ALVR workshop covers topics from self-supervised pre-training, visually-grounded understanding and generation, and multimodal agents, to post-training and alignment, multilingual/culturally-aware LVLMs, ethics and bias, and real-world applications. Features 5 keynote talks (confirmed speakers include Raymond Mooney, Mohit Bansal, Jiajun Wu, Lianhui Qin and Amir Zadeh), spotlight presentations, poster sessions, and a panel discussion. Built on three prior editions at ACL/NAACL (2020, 2021, 2024).

W04: ArgMining: the 13th Workshop on Argument Mining and Reasoning

Organizers: Mohamed Elaraby, Annette Hautli-Janisz, John Lawrence, Elena Musi, Julia Romberg, Federico Ruggeri
Venue: **Harbor E&F**
Date: Friday, July 3, 2026
Full-day workshop
Panel: Argument Mining Meets Reasoning: Understanding and Evaluating Arguments in Both Human and Machine Reasoning
Panelists: Andre Freitas (University of Manchester), Reto Gubelmann (University of Zurich), Priya Pitre (Virginia Tech), Debanjan Ghosh (Analog Devices)
https://argmining-org.github.io/2026

The 13th Workshop on Argument Mining and Reasoning focuses on understanding and evaluating arguments in both human and machine reasoning. The workshop encompasses established argument mining research areas such as automatic identification of argumentative structures (premises, conclusions, and inference schemes), while expanding to include contemporary topics like explainable argumentation, multimodal settings, and the performance of large language models in argumentative tasks. A central goal involves investigating whether LLMs possess genuine understanding of argumentation or merely function as pattern learners. The event features a shared task on reconstructing reasoning in UN resolutions and solicits papers addressing cross-lingual argument mining, LLM capabilities assessment, benchmark development, and real-world applications in legal, scientific, educational, and social domains.

W05: CODI-CRAC: Joint Workshop on Computational Approaches to Discourse, Context and Document-Level Inferences and Computational Models of Reference, Anaphora and Coreference

Organizers: Chloé Braud, Christian Hardmeier, Chuyuan Li, Junyi Jessy Li, Sharid Loáiciga, Vincent Ng, Michal Novák, Maciej Ogrodniczuk, Massimo Poesio, Michael Strube, Amir Zeldes

Venue: Regatta

Date: Friday, July 3, 2026

Full-day workshop

Invited speakers: Ruihong Huang (Texas A&M University), Philippe Laban (Microsoft Research)

https://sites.google.com/view/codi-crac2026

The second joint CODI-CRAC workshop brings together two complementary research communities focused on discourse-level computational linguistics. CODI addresses a broad range of research at the discourse level, while CRAC specializes in computational modeling of reference, anaphora, and coreference. Together, these communities aim to advance discourse research by fostering discussions on dataset development and evaluation methods. The one-day event at ACL 2026 will include invited talks, regular papers, and a shared task, uniting subcommunities working across different levels of linguistic analysis to tackle interconnected challenges in understanding document-level meaning and reference resolution.

W06: DravidianLangTech: The Sixth Workshop on Speech and Language Technologies for Dravidian Languages

Organizers: Bharathi Raja Chakravarthi, Ruba Priyadharshini, Anand Kumar M, Sajeetha Thavareesan, Saranya Rajiakodi, Subalalitha CN, Dhivya Chinnappa, Balasubramanian Palani, Malliga Subramanian, Kogilavani Shanmugavadivel

Venue: Virtual-only

Date: Friday, July 3, 2026

Full-day workshop

https://sites.google.com/view/dravidianlangtech-2026

The sixth iteration of this workshop addresses computational language technologies for Dravidian languages: Tamil, Kannada, Malayalam, Telugu, and others spoken by millions across South India, Sri Lanka, and diaspora communities. Recognizing that regional and under-resourced languages still face significant challenges due to limited computational resources, the workshop seeks to advance speech recognition, multimodal processing, and generative AI applications through collaborative research. Topics span corpus development, hate speech detection, machine translation, speech synthesis, multimodal analysis, and ethical AI considerations. The organizers aim to facilitate collaboration between the global Dravidian language community and international scholars while exploring emerging areas like agentic AI frameworks and assistive technologies for underrepresented communities.

W07: EEUCA: The 9th Workshop on Event Extraction and Understanding: Challenges and Applications

Organizers: Hristo Tanev, Surendrabikram Thapa, Ali Hürriyetolu
Venue: Pier
Date: Friday, July 3, 2026
Full-day workshop
Invited speakers: Usman Naseem (Macquarie University, Australia)

The 9th Workshop on Event Extraction and Understanding brings together researchers working on automatic detection, classification, and reasoning about events in text across domains such as news, social media, and scientific literature. The workshop highlights methodological advances in event detection, schema induction, temporal reasoning, and cross-document event resolution, alongside real-world applications such as crisis monitoring, financial analysis, and biomedical surveillance. Submissions are invited on both established and emerging directions in event-centric NLP, including the use of large language models for event understanding.

W08: LAW XX: The 20th Linguistic Annotation Workshop

Organizers: Yang Janet Liu, Luke Gessler
Venue: La Jolla
Date: Friday, July 3, 2026
Full-day workshop
Invited speakers: Marie-Catherine de Marneffe (UCLouvain), Rachel Rudinger (University of Maryland)
https://sigann.github.io/LAW-XX-2026

The 20th Linguistic Annotation Workshop serves as the annual forum for the ACL and ELRA Special Interest Group on Annotation, bringing together researchers focused on linguistic corpus annotation. The event addresses all aspects of linguistic annotation, including the creation and evaluation of annotation schemes, methods for automatic and manual annotation, software frameworks, data representation, and crowdsourcing methodologies. Participants work collaboratively toward standardization, best practices, and interoperability of annotation information and software, advancing the field that underpins supervised natural language processing research. Taking place in San Diego alongside ACL 2026, LAW XX continues the workshops long-standing tradition of fostering innovation and standardization in annotation practices.

W09: SELVA: The 1st International Workshop on Sustainable and Efficient Language, Vision, and Action Models

Organizers: Jian Cheng, Michele Magno, Zechun Liu, Olivera Kotevska, Yawei Li, Haotong Qin
Venue: Virtual-only
Date: Friday, July 3, 2026
Full-day workshop
Invited speakers: Ruihao Gong (Beihang University), Michele Magno (ETH Zurich), Priyadarshini Panda (University of Southern California), Huanrui Yang (University of Arizona), Yukang Chen (NVIDIA)
https://selva-workshop.github.io

The 1st International Workshop on Sustainable and Efficient Language, Vision, and Action Models addresses the computational costs of scaling Language, Vision, and Action Models by bringing together researchers from NLP, computer vision, robotics, and systems. The workshop covers five core areas: efficient multimodal architectures, model compression techniques, sustainable training and inference, embodied AI efficiency, and measurement standards for energy and environmental impact. SELVA seeks to develop methods that balance high performance with computational feasibility, fostering cross-disciplinary discussion on making advanced AI systems practical for real-time and resource-constrained deployment scenarios.

W10: SURGeLLM: Structured Understanding, Retrieval, and Generation in the LLM Era

Organizers: Vivek Gupta, Kaize Ding, Harsha Kokel, Yue Zhao, Amit Agarwal, Yu Wang, Michael Glass, Yu Zhang, Kavitha Srinivas, Xiusi Chen, Oktie Hassanzadeh, Qi Zhu, Shuaichen Chang, Yuan Luo
Venue: Harbor D
Date: Friday, July 3, 2026
Full-day workshop
Invited speakers: Preslav Nakov (MBZUAI), Dan Roth (University of Pennsylvania), Hamid Palangi (Google), Lisa Amini (IBM Research), Heng Ji (UIUC)
Panel: Structured Data in the LLM Era: From Research Benchmarks to Real-World Industry Systems
Panelists: Kamalika Das (Intuit), Yunyao Li (Adobe), Frederic Sala (Snorkel AI), Lisa Amini (IBM Research), Sujith Ravi (Oracle AI)
https://surgellm.github.io/acl2026

SURGeLLM is an ACL 2026 workshop convening researchers from NLP, information retrieval, data management, and visualization to address how large language models handle structured data. The workshop tackles four primary challenges: enabling robust reasoning over tables and databases, developing structure-aware retrieval mechanisms for various artifact types, generating faithful SQL queries and visualizations, and creating evaluation benchmarks for these systems. Topics span tabular understanding and question answering, natural language interfaces to structured data, and applications and governance for LLMs over structured data, encompassing both methodological advances and real-world deployment considerations. The full-day event features invited talks, panel discussions, oral presentations, and poster sessions, welcoming contributions from students, practitioners, and established researchers, including negative results and lessons learned.

W11: KnowFM: 4th Workshop on Towards Knowledgeable Foundation Models

Organizers: Canyu Chen, Yuji Zhang, Zoey Sha Li, Zihan Wang, Qineng Wang, Priyanka Kargupta, Sara Vera Marjanovi, Jeff Z. Pan, Mohit Bansal, Isabelle Augenstein, Jiawei Han, Heng Ji, Manling Li

Venue: Virtual-only

Date: Friday, July 3, 2026

Full-day workshop

Invited speakers: Luna Dong (Meta), Alexander Spangher (Stanford University), Weijia Shi (Meta), Parisa Kordjamshidi (Michigan State University), Lianhui Qin (UC San Diego), Jiaxin Pei (Stanford University)

https://knowledgeable-lm.github.io

The KnowFM workshop examines the lifecycle of knowledge within foundation models across four stages: emergence through pre-training, injection of external knowledge, updating and editing, and probing/generation. It brings together researchers addressing critical challenges in knowledge-intensive AI systems, including assessing model knowledge, reconciling parametric memory with retrieved information, maintaining up-to-date knowledge, and extending capabilities to multimodal and agentic settings. The workshop welcomes submissions on knowledge representation, retrieval-augmented generation, model editing, hallucination mitigation, and knowledge-intensive agents, with papers selected through double-blind review for presentation at the July 2026 event in San Diego.

W12: CHum: The Second Workshop on Computational Humor

Organizers: Tristan Miller, Ori Amir, Julia Rayz, Tiansi Dong, Christian F. Hempelmann
Venue: Virtual-only
Date: Friday, July 3, 2026
Full-day workshop
Invited speakers: Gil Greengross (Aberystwyth University), Brad Templeton (Electronic Frontier Foundation and rec.humor.funny), Joe Toplyn (Twenty Lane Media, LLC), Laura E. Little (Temple Law School)
Panel: Computational Humor and the Law
Moderator: Tristan Miller (University of Manitoba)
Panelists: Brad Templeton (Electronic Frontier Foundation), Joe Toplyn (Twenty Lane Media), Laura E. Little (Temple Law School)
https://chumweb.org

The Second Workshop on Computational Humor brings together researchers across NLP, linguistics, cognitive science, and the humanities to study the generation, detection, and understanding of humor in text and dialogue. Topics include theoretical models of humor and incongruity, humor generation systems, humor detection benchmarks, cultural and multilingual aspects of humor, and the role of large language models in producing or interpreting jokes, puns, irony, and sarcasm. The workshop aims to foster cross-disciplinary dialogue and to define shared evaluation practices for this notoriously difficult area of language understanding.

W13: SMM4H-HeaRD: Joint 11th Social Media Mining for Health Applications and Health Real-World Data Workshop and Shared Tasks

Organizers: Guillermo Lopez-Garcia, Graciela Gonzalez-Hernandez
Venue: Virtual-only
Date: Friday, July 3, 2026
Full-day workshop
https://healthlanguageprocessing.org/smm4h-2026

The Social Media Mining for Health Applications and Health Real-World Data workshop, now in its 11th edition, convenes researchers to discuss advances in natural language processing, machine learning, and artificial intelligence at the intersection of health and web-based data. Held online and co-located with ACL 2026, this years edition emphasizes privacy-preserving data sharing and real-world evaluation across sources including social media, electronic health records, and biomedical literature. The workshop invites work on interpretable clinical NLP methods, aligning with the ACL theme on explainability, while organizing eight shared tasks spanning adverse drug event detection, clinical documentation, cancer staging, and multilingual clinical entity recognition to address computational challenges in healthcare applications.

W14: AmericasNLP: The Sixth Workshop on NLP for Indigenous Languages of the Americas

Organizers: Manuel Mager, Arturo Oncevay, Abteen Ebrahimi, Minh Duc Bui, Shruti Rijhwani, Luis Chiruzzo, Robert Pugh, Rolando Coto-Solano, John E. Ortega, Katharina von der Wense
Venue: Old Town
Date: Friday, July 3, 2026
Half-day workshop (Afternoon 14:00 - 17:30)
Invited speakers: Jacqueline Brixey (University of Wisconsin-Madison)
https://americasnlp.org/2026_workshop.html

AmericasNLP 2026 aims to encourage and increase the visibility of work on the Indigenous languages of the Americas. The workshop brings together researchers, native speakers, and professionals to advance NLP and computational linguistics for endangered languages, with particular emphasis on approaches suited for low-resource settings. Research topics span dataset creation, machine translation, LLM applications, morphological analysis, and ethical considerations for Indigenous communities. The workshop also features a new shared task on image captioning of culturally relevant images, with results presented during the in-person event co-located with ACL 2026 in San Diego.

W15: StereACuLT: The 1st Workshop on Stereotypes Across Cultures in Language Technologies

Organizers: Weicheng Ma, Soroush Vosoughi, Rolando Coto Solano, Nabeel Gillani
Venue: **Balboa C**
Date: Friday, July 3, 2026
Half-day workshop (Afternoon 14:00 - 17:30)
Invited speakers: Rolando Coto-Solano, Caleb Ziems
https://sites.google.com/view/stereacult-2026

The StereACuLT workshop addresses how stereotypes persist in language technologies, particularly across diverse cultural and linguistic contexts. While existing benchmarks like StereoSet and CrowS-Pairs have advanced bias measurement, they remain predominantly Englishcentric, limiting their validity and applicability in multilingual and multicultural settings. The workshop seeks to develop culturally grounded approaches for defining, evaluating, and mitigating stereotypes, emphasizing methods that incorporate culturally valid operationalizations and account for linguistic and geopolitical variation. A central focus includes examining how alignment techniques like RLHF and policy tuning perform across different cultural contexts, recognizing that culture-agnostic assumptions and heuristics can lead to systematic failures in deployed multilingual models.

W16: SCiL: Society for Computation in Linguistics

Organizers: Rob Voigt, Alex Warstadt, Naomi Feldman, Tal Linzen
Venue: **Promenade**
Date: Friday, July 3, 2026 – Saturday, July 4, 2026
Half-day workshop (Afternoon 14:00 - 17:30)
Invited speakers: Noah Smith (University of Washington), Jennifer Hu (Johns Hopkins University)
Panel: Linguistics and NLP in the LLM Era
Panelists: Allyson Ettinger, Richard Futrell, Zoey Liu, Tom McCoy
https://sites.google.com/view/scil2026

The Society for Computation in Linguistics will hold its ninth annual meeting as a co-located workshop with ACL 2026 in San Diego on July 3-4, 2026. The event seeks high-quality research on computational and mathematical approaches in any area of linguistics, requiring substantial computational or mathematical modeling components and appeal to interdisciplinary audiences. Presentations will be offered both in-person and online across two tracks (papers and abstracts) and two formats (talks and posters). The conference features keynote speakers Jennifer Hu and Noah Smith, organizes crossover sessions with the main ACL conference, and provides NSF-funded registration subsidies for accepted papers to support author participation.

W17: IWSLT: ACL-IWSLT: 23rd International Conference on Spoken Language Translation

Organizers: Marcello Federico, Alex Waibel, Antonios Anastasopoulos, Matteo Negri, Luisa Bentivogli, Marco Gaido, Jan Niehues, Sebastian Stüker, Elizabeth Salesky, Atul Kr. Ojha
Venue: Gaslamp C&D
Date: Friday, July 3, 2026 – Saturday, July 4, 2026
Half-day workshop (Afternoon 14:00 - 17:30)
Invited speakers: Mohit Bansal (University of North Carolina at Chapel Hill), Marine Carpuat (University of Maryland)
Panelists: Marcello Federico (Amazon), Jon May (University of Southern California - ISI), Satoshi Nakamura (The Chinese University of Hong Kong, Shenzhen)
https://iwslt.org/2026

The International Conference on Spoken Language Translation is an annual scientific event where researchers present papers and system descriptions on spoken language translation technologies. The 23rd edition, co-located with ACL 2026 in San Diego on July 3-4, features multiple shared tasks spanning speech-to-text translation, customized speech translation applications, speech generation across diverse languages, instruction following, and speech translation metrics evaluation. The conference aims to advance high-quality, original contributions on theoretical and practical issues of spoken language translation research, technologies, and applications through both in-person and virtual formats, welcoming scientific papers, system descriptions from task participants, and presentations from related ACL-accepted work.

W18: SemEval: The 20th International Workshop on Semantic Evaluation

Organizers: Debanjan Ghosh, Kai North, Shervin Malmasi, Ekaterina Kochmar, Mamoru Komachi, Marcos Zampieri
Venue: **Harbor B**
Date: Friday, July 3, 2026 – Saturday, July 4, 2026
Half-day workshop (Afternoon 14:00 - 17:30)
Invited speakers: Alex Warstadt (UC San Diego)
https://semeval.github.io/SemEval2026

SemEval-2026 is the 20th International Workshop on Semantic Evaluation, a major venue for shared tasks in computational linguistics. The workshop invites task organizers to propose challenges for the broader NLP community, with evaluation running from January through February 2026, followed by a workshop in July co-located with ACL 2026. The event emphasizes the exchange of ideas and rigorous scientific evaluation within the field, sponsored by the SIGLEX Special Interest Group on the Lexicon of the Association for Computational Linguistics.

W19: BioNLP and Shared Tasks

Organizers: Dina Demner-Fushman, Sophia Ananiadou, Kirk Roberts, Jun-ichi Tsujii
Venue: **Balboa A&B**
Date: Friday, July 3, 2026 – Saturday, July 4, 2026
Half-day workshop (Afternoon 14:00 - 17:30)
Invited speakers: Annika Schoene (Northeastern University)
https://aclweb.org/aclwiki/BioNLP_Workshop

The BioNLP 2026 workshop aims to advance natural language processing applications in biomedical and clinical contexts by bringing together researchers focused on health-related language technologies. The event emphasizes evaluation frameworks and metrics that reflect the needs of health-related use cases, prioritizing transparency in generative approaches, factuality verification, and bias detection. Research areas span information extraction, question answering, summarization, synthetic data generation, and reproducibility, with particular attention to underrepresented languages and health disparities. The workshop features four shared tasks addressing medical decision detection, psychological defense mechanism classification, multimodal biomedical question answering with citations, and clinical skill assessment from video, and by co-locating with ACL 2026 exposes mainstream NLP researchers to this growing domain.

W20: StarSEM: *SEM

Organizers: Saif Mohammad, Nedjma Ousidhoum
Venue: **Harbor A**
Date: Friday, July 3, 2026 – Saturday, July 4, 2026
Half-day workshop (Afternoon 14:00 - 17:30)
Invited speakers: Alex Warstadt (UC San Diego), Dallas Card (University of Michigan)
https://starsem2026.github.io

The 15th Joint Conference on Lexical and Computational Semantics (*SEM 2026) is an ACL-affiliated event that brings together researchers focused on understanding and computationally modeling the semantics of natural languages. Organized by SIGLEX and co-located with ACL 2026, the conference addresses the intersection of linguistic meaning and computational approaches, accepting paper submissions across lexical semantics, computational semantics, and their practical applications. *SEM 2026 continues the established tradition of fostering collaboration among researchers in this specialized field.

W21: BEA: The 21st Workshop on Innovative Use of NLP for Building Educational Applications

Organizers: Ekaterina Kochmar, Bashar Alhafni, Stefano Bannò, Marie Bexte, Jill Burstein, Andrea Horbach, Ronja Laarmann-Quante, Anaïs Tack, Victoria Yaneva, Zheng Yuan
Venue: Virtual-only (Fri)
Harbor D (Sat)
Date: Friday, July 3, 2026 – Saturday, July 4, 2026
Half-day workshop (Afternoon 14:00 - 17:30)
Panel: Transitioning from Academia to the EdTech Industry
Panelists: Christine Bagarino (Social AI), Kai North (Cambium Assessment), Keelan Evanini (NBME), Mariano Felice (British Council)
https://sig-edu.org/bea/2026

The 21st BEA Workshop is a leading venue for NLP innovation in the context of educational applications that brings together researchers to advance computational approaches for learning environments. The workshop covers diverse topics including automated evaluation of written and spoken open-ended responses, grammar error correction, intelligent tutoring systems, and the role of generative AI in education. Beyond traditional research presentations, the two-day event features two shared tasks addressing vocabulary difficulty prediction and rubric-based answer scoring, a half-day tutorial on Theory of Mind applications in tutoring, and a panel discussion on transitioning from academia to EdTech careers.

W22: CoNLL: The SIGNLL Conference on Computational Natural Language Learning

Organizers: Claire Bonial, Yevgeni Berzak
Venue: **Harbor C**
Date: Friday, July 3, 2026 – Saturday, July 4, 2026
Half-day workshop (Afternoon 14:00 - 17:30)
Invited speakers: Michael C. Frank (Stanford University), Adele Goldberg (Princeton University)
Panel: Community Discussion
Moderator: Omri Abend, Antske Fokkens, Claire Bonial, Yevgeni Berzak
https://www.conll.org

CoNLL is a premier conference dedicated to theoretically, cognitively, and scientifically motivated approaches to computational linguistics and NLP. The conference welcomes submissions across diverse areas including psycholinguistics, language acquisition, multimodality, speech, semantics, and interpretability of machine learning models for language. CoNLL prioritizes work grounded in cognitive science, linguistic theory, and empirical methodology rather than purely engineering-focused contributions. Taking place July 3-4, 2026 in San Diego alongside ACL, the event accepts both archival papers (up to 8 pages) and non-archival submissions of previously published work, emphasizing rigorous scientific standards.

W23: TrustNLP: The 6th Workshop on Trustworthy NLP

Organizers: Kai-Wei Chang
Venue: **Harbor G**
Date: Saturday, July 4, 2026
Half-day workshop (morning 9:00 - 12:30)
Invited speakers: Michael Johnston (Amazon AGI), Lily Weng (UC San Diego)
https://trustnlpworkshop.github.io

The TrustNLP workshop at ACL 2026 addresses trustworthy natural language processing amid rapid advances in large language models and agentic AI systems. The event invites research on developing models that are explainable, fair, privacy-preserving, causal, and robust, covering topics including fairness evaluation, model interpretability, adversarial robustness, hallucination mitigation, and industry applications. Sponsored by Meta, Capital One, and Amazon, the workshop aims to advance responsible AI development by promoting interdisciplinary approaches that integrate sociopolitical and cultural perspectives alongside technical solutions for safe language generation and deployment.

W24: BigPicture: The Big Picture: Crafting a Research Narrative v2

Organizers: Yanai Elazar, Allyson Ettinger, Nora Kassner, Sebastian Ruder
Venue: Gaslamp A&B
Date: Saturday, July 4, 2026
Half-day workshop (morning 9:00 - 12:30)
Invited speakers: Noah A. Smith (University of Washington and AI2), Tiago Pimentel (ETH Zurich), Aaron Mueller (Boston University)
https://www.bigpictureworkshop.com

The Big Picture Workshop is a dedicated venue for NLP researchers to explore broader research narratives and reflect on how individual contributions connect within the larger research landscape. The workshop aims to enhance the communication and understanding between different lines of work and highlight how works connect and build on each other, recognizing that despite rapid publication rates and concise paper formats, the larger story connecting research has become increasingly difficult to discern. By bringing together diverse research threads, the workshop seeks to generate insights, encourage broader collaboration, and facilitate understanding of field trajectories, particularly benefiting newcomers and outsiders who struggle to grasp connections between isolated papers.

W25: RAG4Reports: The First Workshop on Multilingual Report Generation via Retrieval Augmented Generation

Organizers: Dawn Lawrie, Sean MacAvaney, James Mayfield, Luca Soldaini, Eugene Yang, Andrew Yates

Venue: **Harbor A**

Date: Saturday, July 4, 2026

Half-day workshop (morning 9:00 - 12:30)

Invited speakers: Chris Callison-Burch (University of Pennsylvania)

https://rag4reports.github.io

RAG4Reports is a workshop focused on retrieval-augmented generation systems for generating long-form reports with proper source attribution. The workshop addresses Report Generation as a long-form RAG task with strict attestation requirements, emphasizing three core areas: grounded generation with citation support, multilingual RAG capabilities, and coherent long-form responses. Participants will engage with research papers, shared tasks on automatic report evaluation and multilingual report generation, and a keynote on LLM evaluation frameworks. The workshop aims to develop systems and evaluation methodologies that can advance not only report generation but also broader solutions for other long-form RAG applications.

W26: LTEDI: Sixth Workshop on Language Technology for Equality, Diversity and Inclusion

Organizers: Bharathi Raja Chakravarthi, Bharathi B, Dhairaj Thenmozhi, Salud María Jiménez Zafra, Miguel Ángel García Cumbreras
Venue: Virtual-only
Date: Saturday, July 4, 2026
Full-day workshop
Invited speakers: Laura Hollink (CWI)
https://sites.google.com/view/lt-edi-2026

The Sixth Workshop on Language Technology for Equality, Diversity and Inclusion (LT-EDI2026) addresses bias and fairness in NLP and large language models across multiple dimensions. The workshop brings together researchers, practitioners, and ethicists to advance critical challenges and innovative solutions for creating inclusive language models. Core topics include dataset construction for bias detection, mitigation techniques for gender and racial biases, inclusive language generation, multilingual approaches for underrepresented communities, and frameworks for trustworthy AI systems. Through this interdisciplinary forum, organizers aim to foster Language Technologies for All by designing equitable systems that serve English and under-resourced languages while ensuring fair representation and cultural sensitivity.

W27: MAGMaR: The Second Workshop on Multimodal Augmented Generation via Multimodal Retrieval

Organizers: Reno Kriz; Kenton Murray; Andrew Yates; Desmond Elliott; Francis Ferraro; Xiang Xiang; William Walden; Jeremy Gwinnup; Joel Brogan; Alexander Martin

Venue: Old Town

Date: Saturday, July 4, 2026

Full-day workshop

Invited speakers: Nanyun Peng (UCLA), Chenliang Xu (University of Rochester)

https://nlp.jhu.edu/magmar

The MAGMaR workshop addresses the growing importance of audiovisual media as an information source by focusing on two key areas: retrieving multimodal content spanning text, images, audio, and video; and developing retrieval-augmented generation, with an emphasis on multimodal retrieval and grounded generation. The event features a shared task where systems must identify relevant videos from a multilingual collection and produce coherent written reports grounded in that evidence, with separate tracks for retrieval and generation. Research submissions are welcome on related topics including document retrieval, multimodal question answering, and video understanding. The one-day hybrid workshop will be co-located with ACL 2026 in San Diego on July 4th, 2026, bringing together independent research efforts into a unified community effort.

W28: GEM: Natural Language Generation, Evaluation, and Metrics

Organizers: Simon Mille, Sebastian Gehrmann, Patrícia Schmidtová, Ondej Duek, Marzieh Fadaee, Kyle Lo, Enrico Santus, Gabriel Stanovsky
Venue: **Harbor E&F**
Date: Saturday, July 4, 2026
Full-day workshop
Invited speakers: Vered Shwartz (University of British Columbia), Swabha Swayamdipta (University of Southern California), Chris Callison-Burch (University of Pennsylvania)
https://gem-workshop.com

The fifth edition of the GEM Workshop focuses on evaluation and metrics for natural language generation systems. The workshop addresses how evaluation of language models has grown to be a central theme in NLP research while remaining far from solved, exploring topics including automatic evaluation methods, benchmark creation, human evaluation practices, and emerging challenges like long-context understanding and multimodal evaluation. Beyond traditional archival papers, GEM introduces a special opinion and statement papers track for thought-provoking position pieces that challenge conventional evaluation wisdom, alongside the ReproNLP shared task examining reproducibility of past NLP evaluations. The event emphasizes an interactive, discussion-driven format with themed poster sessions, panel discussions, and keynotes from leading experts, culminating in a Comic-Con Edition celebrating evaluation research in San Diego.

W29: NLP4DH: The 6th International Conference on Natural Language Processing for the Digital Humanities

Organizers: Sil Hamilton, Emily Öhman, Rebecca M. M. Hicke, Yuri Bizzoni, Axel Bax, Jacob A. Matthews, Mika Hämäläinen
Venue: Regatta
Date: Saturday, July 4, 2026
Full-day workshop
Invited speakers: Sophie Hao (Boston University), Luca Soldaini (AI2)
https://www.nlp4dh.com/nlp4dh-2026

The 6th International Conference on Natural Language Processing for the Digital Humanities brings together researchers applying computational text analysis to humanistic inquiry. The event encompasses two tracks: a main track addressing applications of NLP techniques to digital humanities research, including OCR, cultural bias assessment, and heritage collection analysis; and a special track titled Reinterpreting NLP that examines LLM training characteristics and mechanistic interpretability through humanities perspectives. Organized with ACL 2026 in San Diego on July 4, the conference solicits original short and long papers (4–8 pages respectively) that bridge NLP methods with critical humanistic scholarship, with accepted works published in the ACL Anthology and eligible for a special journal issue.

W30: ComputEL-9: The Ninth Workshop on the Use of Computational Methods in the Study of Endangered Languages

Organizers: Sarah Moeller; Godfred Agyapong; Antti Arppe
Venue: **Balboa C**
Date: Saturday, July 4, 2026
Full-day workshop
https://computel-workshop.org

The ComputEL workshop concentrates on the use of computational methods in the study, support, and revitalization of endangered languages. Its central objective is bridging the divide among computational linguists developing methods for endangered language work, field linguists engaged in documentation, and Indigenous communities dedicated to language preservation. The workshop alternates between computational linguistics conferences and language documentation conferences to ensure accessibility across these distinct professional communities. By fostering collaboration among these groups, ComputEL addresses computational challenges specific to language endangerment while grounding technical innovations in real-world documentation and revitalization efforts.

W31: NLP+CSS: Seventh Workshop on NLP and Computational Social Science

Organizers: Dallas Card, Anjalie Field, Katie Keith, Julia Mendelsohn
Venue: Gaslamp A&B
Date: Saturday, July 4, 2026
Full-day workshop

The Seventh Workshop on NLP and Computational Social Science brings together researchers applying language technology to study society, culture, politics, and social behavior at scale. The workshop showcases methodological advances and empirical findings at the intersection of NLP and the social sciences, including work on political discourse analysis, online communities, public health communication, misinformation, social bias, and large-scale text-as-data studies. Both methods-focused and substantive papers are welcome, with an emphasis on rigorous engagement with social science theory.

W32: CLPsych: Eleventh Workshop on Computational Linguistics and Clinical Psychology

Organizers: Aya Zirikly,Kfir Bar, Sean MacAvaney, Molly Ireland, Yaakov Ophir, Dana Atzil-Slonim, Vasudha Varadarajan, Steven Bedrick, and Bart Desmet
Venue: **Harbor H&I**
Date: Saturday, July 4, 2026
Full-day workshop
Invited speakers: Colin Walsh (Vanderbilt University), Dirk Hovy (Bocconi University)
Moderator: Philip Resnik (University of Maryland)
Panelists: Colin Walsh (Vanderbilt University), Justin Tauscher (University of Washington), Anat Brunstein Klomek (Reichman University)
https://clpsych.org

CLPsych is an interdisciplinary workshop series established in 2014 that convenes computational linguists and NLP researchers with mental health practitioners to advance understanding of human emotional and psychological states through language analysis. The workshop aims to use human language as a tool to better understand emotional and mental state, and to reduce emotional suffering and the potential for self-harm. Participants employ computational methods to infer meaning, predict behavioral characteristics, and bridge the gap between algorithmic approaches and clinical practice needs. The 2026 iteration, titled Moving Beyond Labels to Understand Mental Health Dynamics, will convene at ACL in San Diego on July 4th.

W33: BATCH: Beyond Alignment: Transdisciplinary Conversations on Human-AI Futures

Organizers: Matthew Stone, Malihe Alikhani, Camille Gagnier, Lauren M. E. Goodlad, Dan Roth, Mark Sammons
Venue: Pier
Date: Saturday, July 4, 2026
Full-day workshop
Invited speakers: Michele Elam (Stanford University), David Leslie (Alan Turing Institute)
https://criticalai.org/2026/02/27/

BATCH (Beyond Alignment: Transdisciplinary Conversations on Human-AI Futures) is an ACL 2026 workshop designed to unite technologists, humanists, and social scientists around shared challenges in AI development. The event brings together researchers studying LLM alignment efforts with those examining social impacts of AI systems, creating rare opportunities to connect these efforts across disciplines. Participants will explore interdisciplinary evaluation methods, real-world deployment impacts, participatory design approaches, and critical research directions through invited talks, contributed papers, and collaborative sessions aimed at producing a motivating white paper for future interdisciplinary work.

W34: CDL: The 1st Workshop on Computational Developmental Linguistics

Organizers: Martin Ziqiao Ma, Emmy Liu, Jing Liu, Tyler Chang, Abdellah Fourtassi, Alex Warstadt, Michael Hahn, Weiwei Sun, Freda Shi

Venue: **Harbor G**

Date: Saturday, July 4, 2026

Half-day workshop (Afternoon 14:00 - 17:30)

Invited speakers: Mariya Toneva (Max Planck Institute), Brian MacWhinney (Carnegie Mellon University), Kanishka Misra (UT Austin)

https://comp-dev-ling.github.io/

The 1st Workshop on Computational Developmental Linguistics (CDL) aims to bridge machine learning and developmental linguistics by examining how linguistic capabilities emerge in both human learners and artificial systems. The workshop welcomes research on computational modeling of language acquisition and change over time across phonology, lexicon, syntax, and idiolectal shifts. Key topics include pretraining dynamics versus first language acquisition, continual learning versus second language acquisition, knowledge tracing in language models, and applications to language education. By assembling researchers from machine learning, cognitive science, and psychology, the workshop seeks to establish rigorous frameworks for comparing human and machine language development while avoiding anthropomorphizing artificial systems.

W35: C3NLP: Cross-Cultural Considerations in NLP

Organizers: Vinodkumar Prabhakaran, Sunipa Dev, Luciana Benotti, Daniel Hershcovich, Yong Cao, Ife Adebara, Li Zhou
Venue: La Jolla
Date: Saturday, July 4, 2026
Half-day workshop (Afternoon 14:00 - 17:30)
Invited speakers: Helena Gomez Adorno (UNAM), Rachel Rudinger (University of Maryland)
https://c3nlp.github.io/

The 4th Workshop on Cross-Cultural Considerations in NLP brings together researchers and multidisciplinary scholars to address how cultural contexts shape NLP technologies and data. The workshop focuses on three main themes: inclusivity and representation of cultures in NLP, cultural harms of NLP systems, and examining social biases through a culture-sensitive lens. Papers are invited on topics ranging from how culture influences model development and annotation practices to detecting cultural biases, incorporating cultural knowledge into systems, and building bridges between NLP researchers and language communities. The event aims to establish a long-term research agenda that goes beyond multilingual approaches to develop more culturally competent NLP technologies grounded in social science theories.

W36: EvalEval: Second Workshop on Evaluating Evaluations: Examining Best Practices for Utilizing and Developing Generative Model Evaluations

Organizers: Jan Batzner, Usman Gohar, Jennifer Mickel, Leshem Choshen, Michelle Lin, Ichhya Pant, Mubashara Akhtar, Zeerak Talat, Avijit Ghosh
Venue: Harbor A
Date: Saturday, July 4, 2026
Half-day workshop (Afternoon 14:00 - 17:30)
Moderator: Leshem Choshen (MIT / IBM Research)
Panelists: Angelina Wang (Cornell Tech), Sebastian Gehrmann (Bloomberg), Wesley Deng (Microsoft Research)
https://evalevalai.com/events/2026-acl-workshop/

The ACL 2026 Workshop on Evaluating Evaluations addresses the growing gap between evaluation research and real-world AI development practices. Focusing on Examining Best Practices for Utilizing and Developing Generative Model Evaluations, the workshop brings together model developers and evaluation researchers to discuss tensions around speed versus rigor, benchmarks versus deployment needs, and generality versus task-specificity. The scope encompasses three main topic areas: evaluation methodology and measurement theory, infrastructure and stakeholder considerations, and sociotechnical impact assessment across dimensions like bias, privacy, labor, and environmental costs. Through invited panels, oral presentations, poster sessions, and a shared task on standardizing evaluation results, the workshop aims to identify concrete pathways for more aligned and impactful evaluation practices in NLP.

W37: MeLLMs: Multilinguality in the Era of Large Language Models

Organizers: Fengran Mo, Xue Zhang, Pinzhen Chen, Zheyuan Liu, Kaiyu Huang, Zhiqi Huang, Koel Dutta Chowdhury, Barry Haddow, Meng Jiang, Alexandra Birch, Jian-Yun Nie
Venue: **Gaslamp A&B**
Date: Saturday, July 4, 2026
Half-day workshop (Afternoon 14:00 - 17:30)
Invited speakers: Julia Hirschberg (Columbia University), Julia Kreutzer (Cohere Labs)
https://www.mellm.org/

The 1st Workshop on Multilinguality in the Era of Large Language Models seeks to address challenges and opportunities emerging from multilingual LLMs across NLP research. The workshop covers diverse topics including multilingualism in the era of LLMs (reasoning, long-context, instruction following, safety), training approaches, information access, global applications, multimodal systems, and resources. Organizers aim to develop equitable language technologies for the worlds languages to satisfy the needs of users across the globe, with particular focus on constructing better multilingual representations and creating real-world application solutions. The event will feature keynote presentations, oral sessions, and poster presentations from accepted papers across both direct and ACL Rolling Review submissions.

10 Birds of a Feather

Overview

Time	Type	Session	Room
Sunday, July 5			
10:30–12:30	BoF	Social Agents in the Wild: Reasoning, Simulation, and Evaluation	Gaslamp C&D
11:00–12:30	Affinity	Pacific Northwest Researchers	Old Town
12:30–14:00	Affinity	Queer in AI	La Jolla
14:00–15:30	BoF	User Simulation Inherent Group	Old Town
16:00–17:30	BoF	NLP for Education and Workforce Readiness	Gaslamp C&D
16:00–17:30	BoF	AI for Computer Science Education	Old Town
16:00–17:30	BoF	Video Retrieval	La Jolla
Monday, July 6			
9:00–10:30	Affinity	Teaching NLP	Old Town
9:00–10:30	BoF	Who Gets Left Behind? Identity, Culture, and Power in NLP Systems	Gaslamp C&D
9:00–10:30	BoF	Open-Source Foundation Models and Training Infrastructure	La Jolla
11:00–12:30	BoF	AI Agents From Capability to Productivity	Old Town
11:00–12:30	BoF	Shared Interest: From Text to Structure: Interpretable and Auditable Systems with Language Models	La Jolla
11:00–12:30	Mentorship	Mentorship on NLP Research	Gaslamp C&D
Tuesday, July 7			
9:00–10:30	Affinity	Responsible Research Productivity for Early-Career NLP Researchers	Old Town
11:00–12:30	BoF	Christians in NLP	Gaslamp C&D
11:00–12:30	BoF	What Makes an Enterprise Agent Benchmark Useful? A Discussion on Tasks, Tools, and Trust	La Jolla

Sunday, July 5

Social Agents in the Wild: Reasoning, Simulation, and Evaluation

BoF · In-person · Sunday, July 5, 10:30–12:30 · Room: Gaslamp C&D · Organizer: Xiangjue Dong

Social agents — LLM-based systems that interact with humans and each other — are moving rapidly from controlled benchmarks to real-world deployment: customer service, education, deliberation support, and online communities. Yet fundamental questions remain open: How do we build agents that reason coherently about social context? How do we simulate multi-agent interactions that are faithful to human behavior? And how do we evaluate agents not just on task success, but on social alignment, robustness, and trustworthiness?

This Birds of a Feather session brings together researchers working at the intersection of NLP, multi-agent systems, and computational social science to discuss these challenges openly. Topics include (but are not limited to): chain-of-thought and theory-of-mind reasoning in social contexts, persona-consistent and value-aligned agents, LLM-based social simulation at scale, evaluation frameworks beyond accuracy, and the gap between benchmark performance and real-world deployment.

Whether you are building social agents, stress-testing them, or studying the humans they interact with, this is a space to share ongoing work, compare methodologies, and identify open problems together. All career stages and disciplinary backgrounds are welcome. For updates and community discussions, visit https://social-llm-workshop.github.io/ and join our Slack.

Pacific Northwest Researchers

Affinity · In-person · Sunday, July 5, 11:00–12:30 · Room: Old Town · Organizer: Rhitabrat Pokharel

We would like to bring together NLP researchers from the Pacific Northwest (Washington and Oregon) to share their research and build connections across institutions. The session will include informal discussions of current research directions, and opportunities to identify common interests and potential collaborations.

Queer in AI

Affinity · In-person · Sunday, July 5, 12:30–14:00 · Room: La Jolla · Organizer: Shane Storks

Queer in AI affinity group meeting.

User Simulation Inherent Group

BoF · Hybrid · Sunday, July 5, 14:00–15:30 · Room: Old Town · Organizer: Ziwei Gong

This User Simulation Birds-of-a-Feather session aims to bring together researchers interested in simulated users and agent-based simulation. We hope to discuss the current capabilities and limitations of simulated users, including where they work well, where they fail, and how their use cases differ across data generation, system testing, user modeling, and dialogue evaluation.

A central question is how we should understand the gap between simulated users and real human users. Do simulated users need to closely match individual human behavior to be useful, or can many imperfect agents still produce aggregate trends that resemble population-level human behavior?

We would like to focus the discussion on two main themes: first, what simulated users are currently capable of and for which purposes; and second, how we should evaluate them differently depending on those purposes. Rather than asking only whether simulated users are "human-like," we hope to ask what aspects of human behavior they need to capture, when they are useful, and what new evaluation methods are needed.

NLP for Education and Workforce Readiness

BoF · Hybrid · Sunday, July 5, 16:00–17:30 · Room: Gaslamp C&D · Organizer: Vivek Gupta

This Birds-of-a-Feather session will bring together researchers, educators, students, and industry practitioners interested in how NLP and Large Language Models (LLMs) are reshaping education, learning, assessment, and workforce development. The discussion will cover topics such as AI-assisted tutoring, educational agents, automated feedback, academic integrity, personalized learning, multilingual educational support, workforce upskilling, and responsible deployment of generative AI in classrooms and professional environments. The session is intended to be highly interactive and community-driven, providing a space to exchange ideas, discuss emerging challenges and opportunities, and identify future directions for NLP research and deployment in education and workforce ecosystems.

AI for Computer Science Education

BoF · Hybrid · Sunday, July 5, 16:00–17:30 · Room: Old Town · Organizer: Cy (Tianzhu) Xie

This Birds-of-a-Feather session brings together computer science educators, researchers, and students passionate about advancing computer science education. As a community college student, I particularly focus on building stronger bridges between computer science education at high schools and community colleges and resource-rich top-tier universities, to better provide guidance, resources, and academic opportunities for learners in these settings. We aim to discuss innovative teaching strategies, address inclusive curriculum design, explore key challenges, and examine AI-powered adaptive personalized computer science learning—enabled by recent advances in AI and computing—as an alternative to traditional one-size-fits-all, uniform-pace instruction. This includes exploring practical AI-driven implementation models, adaptive syllabus design, and evidence-based strategies to enhance the efficiency and scalability of personalized learning pathways in the era of rapid AI development. The session will feature open discussion, experience sharing, and networking, fostering a supportive community to exchange insights and collaborate on improving the quality and accessibility of computer science education worldwide.

Video Retrieval

BoF · In-person · Sunday, July 5, 16:00–17:30 · Room: La Jolla · Organizer: Sourajit Saha

A casual ACL 2026 meetup for researchers interested in video retrieval, multimodal retrieval, and efficient video understanding. Topics include zero-shot and few-shot retrieval, CoT-guided retrieval, memory/computation-efficient search, compressed representations, test-time adaptation/training, multimodal interaction, and scalable video recognition. We also want to discuss practical challenges: how to retrieve from millions of videos without melting GPUs, how to reduce annotation burden, and how to make retrieval systems both effective and deployable. Expect informal discussions, idea sharing, open problems, and probably too many opinions about embeddings, rerankers, and "just use more compute."

Monday, July 6

Teaching NLP

Affinity · Hybrid · Monday, July 6, 9:00–10:30 · Room: Old Town · Organizer: Matthias ASSenmacher

At this informal meetup, we'll discuss challenges and opportunities in teaching natural language processing in 2026. We hope that this meetup will bring together teachers from different institutions and departments to discuss teaching NLP in varied settings. We invite anyone interested in teaching NLP, from experienced faculty to students who would like to teach NLP in the future.

Who Gets Left Behind? Identity, Culture, and Power in NLP Systems

BoF · In-person · Monday, July 6, 9:00–10:30 · Room: Gaslamp C&D · Organizer: Alberto Testoni

NLP systems increasingly mediate decisions that affect and shape people's lives, but they do not serve everyone equally well. Depending on the user's gender, sexual orientation, geographic origin, religion, or cultural background, the same system can give worse, less appropriate, or less safe responses. These are not merely abstract failure modes. They mirror real-world inequalities and disproportionately affect communities that are already marginalized, at a moment when the rights and protections of many of these groups are being challenged around the world. Yet fairness research tends to study one axis at a time, in isolation, while people live at their intersection. This BoF brings together researchers and practitioners working across these dimensions to compare what they are seeing, confront the shared methodological and ethical challenges, and ask who our systems are actually built to serve. Anyone working on, or simply curious about, fairness across identities and cultures is welcome.

Open-Source Foundation Models and Training Infrastructure

BoF · Hybrid · Monday, July 6, 9:00–10:30 · Room: La Jolla · Organizer: Frank Poon

This session brings together NLP and ML researchers who are building with, contributing to, or advocating for open-source foundation models. We'll discuss the current open-source model landscape, practical challenges in training and fine-tuning (including techniques like DPO, GRPO, and RLHF), self-hosted deployment across heterogeneous hardware, and the role of open infrastructure in keeping cutting-edge NLP research accessible. The session is organized by Transformer Lab, an open-source research initiative, and is open to anyone.

AI Agents From Capability to Productivity

BoF · In-person · Monday, July 6, 11:00–12:30 · Room: Old Town · Organizer: Xuan Wang

This session will explore the gap between the impressive capabilities demonstrated by current AI agents and their limited real-world productivity in high-stakes domains such as healthcare, manufacturing, and enterprise systems. While recent agents can code, write, plan, and use tools effectively in controlled settings, significant challenges remain in reliability, safety, deployment, and evaluation in complex real-world environments. This session aims to bring together researchers and practitioners to discuss the key bottlenecks preventing AI agents from achieving dependable real-world impact, identify important research directions, and foster interdisciplinary collaboration across academia and industry.

Shared Interest: From Text to Structure: Interpretable and Auditable Systems with Language Models

BoF · In-person · Monday, July 6, 11:00–12:30 · Room: La Jolla · Organizer: Tunazzina Islam

This BoF session presents a research agenda centered on developing interpretable and auditable language-based systems for understanding large-scale communication. Modern communication platforms generate massive amounts of unstructured discourse containing implicit themes, arguments, persuasion strategies,

and demographic targeting patterns that are difficult to analyze using traditional methods. Existing approaches often rely on predefined labels, weak semantic representations, or opaque language model outputs, limiting interpretability, validation, and reliable auditing.

To address these challenges, I present a series of works that progressively move from semantic discovery toward structure validation and responsible auditing. First, I introduce machine-in-the-loop frameworks for discovering latent themes and arguments in large-scale discourse using language models. I then present reasoning-based methods for refining and validating unsupervised semantic structure, where language models act as semantic judges to improve coherence, reduce redundancy, and produce interpretable representations. Finally, I show how these semantic structures can support post-hoc auditing of demographic targeting and persuasive framing in both human-generated and AI-mediated communication systems.

Across these directions, the broader goal of my research is to develop language technologies that transform unstructured communication into interpretable, validated, and auditable semantic structure. I conclude by discussing future directions in dynamic discourse analysis, AI-generated communication auditing, and human-AI collaborative systems for reliable and socially responsible language technologies.

Mentorship on NLP Research

Mentorship · In-person · Monday, July 6, 11:00–12:30 · Room: Gaslamp C&D · Organizer: ACL 2026 Diversity \& Inclusion Committee

A mentorship session for early-career NLP researchers, offering guidance and connections with experienced mentors. Organized by the ACL 2026 Diversity \& Inclusion Committee.

Tuesday, July 7

Responsible Research Productivity for Early-Career NLP Researchers

Affinity · Hybrid · Tuesday, July 7, 9:00–10:30 · Room: Old Town · Organizer: Chengguang Gan

This session is intended as an informal discussion space for early-career NLP researchers, especially graduate students and young researchers who have limited access to senior mentorship, large computing resources, or established research networks.

The discussion will focus on how to develop feasible NLP research ideas, organize small-scale but meaningful experiments, write clear papers, improve reproducibility, and navigate the review process responsibly. We also hope to discuss common challenges faced by early-career researchers, such as choosing appropriate research topics, avoiding overclaiming, understanding reviewer feedback, and balancing productivity with research quality.

The goal of this session is not to promote shortcut-based publishing, but to encourage responsible, transparent, and community-supported research practice. Participants will be invited to share experiences, practical advice, and lessons learned from paper writing, submission, rebuttal, and revision.

Christians in NLP

BoF · In-person · Tuesday, July 7, 11:00–12:30 · Room: Gaslamp C&D · Organizer: Pedro Cisneros

It is very likely that as a Christian in NLP, working in industry or academia, you have not found people that shared your faith. Indeed, Christians can be a minority in this field. This affinity group presents you with the opportunity to meet fellow Christians and share our research interests, experiences, and faith. Anyone who wants to meet Christians are welcome to come.

What Makes an Enterprise Agent Benchmark Useful? A Discussion on Tasks, Tools, and Trust

BoF · In-person · Tuesday, July 7, 11:00–12:30 · Room: La Jolla · Organizer: Dhaval Patel

LLM agents are moving from chat into enterprise workflows: industrial asset operations, financial analysis, customer support, and scientific discovery, where evaluation stops being a leaderboard exercise and starts being a question of safety, cost, and trust. Yet most public agent benchmarks still measure protocol fluency or single-tool calls in toy domains, leaving open the harder questions. Does the agent retrieve the right tool? Plan the right sequence? Recover from failure? Hallucinate when it shouldn't? Cost what its owner can afford? This BoF brings together researchers and practitioners building, evaluating, and deploying enterprise-grade agents to share lessons learned and identify shared open problems. Drawing on the experience of AssetOpsBench (NeurIPS 2025; AAAI 2026 Lab), an industrial asset-operations benchmark with 141 scenarios, 1.4k+ GitHub stars, and a CODS 2025 community challenge that drew 300+ submissions across 149 teams, we will discuss what worked, what didn't, and what's still missing.

Discussion topics will include *task design* (what separates an enterprise-realistic task from a toy task; capturing multi-step workflows, ambiguous instructions, and domain knowledge); *evaluation beyond pass@1* (trajectory-level metrics, hallucination taxonomies, execution quality, and cost-aware scoring, and whether leaderboards are even the right format); *tool ecosystems and MCP* (evaluating agents that must discover and orchestrate tools, not just call them, in heterogeneous environments); *privacy-aware online evaluation* (lessons from competitions where private data, hidden test sets, and reproducibility must coexist); *community challenges as benchmark validation* (what 300+ submissions to AssetOpsBench-Live taught us about which design patterns transfer); and *open problems* (multi-agent failure modes, world-model evaluation, domain transfer, and the gap between research benchmarks and production deployment).

The session is structured as a series of talks. Brief framing remarks (about 10 minutes) will be followed by open discussion organized around participant-proposed topics. Researchers, practitioners, and anyone building or breaking enterprise agents are welcome. Bring open questions, war stories, or both.

11

Main Conference Day 1: Sunday, July 5, 2026

Overview

07:30-16:30	**Registration & Conference Check In (Full Conference & Main Conference Attendees)**	Grand Hall Foyer (L1)
09:00-09:30	**Session 1: Welcome**	Harbor Ballroom (L2)
09:30-10:30	**Session 1: Keynote: Philip Resnik (University of Maryland)**	Harbor Ballroom (L2)
10:30-11:00	Coffee	Grand Hall
11:00-12:30	**Session 2: Orals / Demos / Poster Session A**	
	Demo, Main & Finding Posters	
	See Tracks on Whova or Underline for room assignment	Grand Hall (Ground), Coronado (L4)
	Multilinguality and Language Diversity 1	Harbor A (L2)
	Resources and Evaluation 1	Harbor B&C (L2)
	Interpretability and Analysis of Models for NLP 1	Harbor D (L2)
	NLP Applications 1	Harbor E&F (L2)
	Information Extraction and Retrieval 1	Harbor G (L2)
	Language Models 1	Harbor H&I (L2)
	Ethics, Bias, and Fairness 1	Promenade (L3)
	Computational Social Science, Cultural Analytics, and NLP for Social Good 1	Regatta (L4)
	Industry 1: Opening - Keynote - Orals	Balboa (L2, Seaport)
	SRW 1: Opening - Orals	Gaslamp A&B (L2, Seaport)
	BoF/Affinity: Please check the online schedule	Gaslamp C&D, Old Town A&B, La Jolla A&B
12:30-14:00	Lunch	On Own
14:00-15:30	**Session 3: Orals / Demos / Poster Session B**	

	Demo, Main & Finding Posters	
	See Tracks on Whova or Underline for room assignment	Grand Hall (Ground), Coronado (L4)
	Multilinguality and Language Diversity 2	Harbor A (L2)
	Resources and Evaluation 2	Harbor B&C (L2)
	Safety and Alignment in LLMs 1	Harbor D (L2)
	AI/LLM Agents 1	Harbor E&F (L2)
	Information Extraction and Retrieval 2	Harbor G (L2)
	Language Models 2	Harbor H&I (L2)
	Summarization, Generalizability and Transfer	Promenade (L3)
	Semantics: Lexical, Sentence-level Semantics, Textual Inference and Other areas	Regatta (L4)
	Industry 2: Keynote - Orals	Balboa (L2, Seaport)
	SRW 2: Orals	Gaslamp A&B (L2, Seaport)
	Linguistics Symposium - Linguistics and NLP in the LLM Era	
	Allyson Ettinger, Richard Futrell, Zoey Liu and Tom McCoy	Gaslamp C&D (L2, Seaport)
	BoF/Affinity: Please check the online schedule	Old Town A&B, La Jolla A&B
15:30-16:00	Coffee	Grand Hall
16:00-17:30	**Session 4: Orals / Demos / Poster Session C**	
	Demo, Main & Finding Posters	
	See Tracks on Whova or Underline for room assignment	Grand Hall (Ground), Coronado (L4)
	Multilinguality and Language Diversity 3	Harbor A (L2)
	Resources and Evaluation 3	Harbor B&C (L2)
	Information Extraction and Retrieval 3	Harbor D (L2)
	NLP Applications 2	Harbor E&F (L2)
	Computational Social Science, Cultural Analytics, and NLP for Social Good 2	Harbor G (L2)
	Linguistic theories, Cognitive Modeling and Psycholinguistics 1	Harbor H&I (L2)
	Dialogue and Interactive Systems	Promenade (L3)
	Machine Learning for NLP 1	Regatta (L4)
	Industry 3: Orals	Balboa (L2, Seaport)
	SRW 3: Panel Discussion: The next big questions in NLP: What should students work on? Chair: Juan Diego Rodriguez. Panelists: Jordan Boyd-Graber (University of Maryland), Yulia Tsvetkov (University of Washington), Iryna Gurevych (TU Darmstadt), Preslav Nakov (MBZUAI)	Gaslamp A&B (L2, Seaport)
	Closing Remarks	Gaslamp A&B (L2, Seaport)
	BoF/Affinity: Please check the online schedule	Gaslamp C&D, Old Town A&B, La Jolla A&B
18:00-19:30	**Session 5: Virtual Presentations 1**	Underline

Each paper's title links to its full paper, and the ▶ ***Underline Video*** link opens its talk. Recordings, live schedule, and program are fully available on Underline. powered by underline

Detailed Program

Session 2: Oral/Posters/Demos A - 11:00-12:30

Demo Session A

11:00-12:30 (Grand Hall)

[Demo] AutoForest: Automatically Generating Forest Plots from Biomedical Studies with End-to-End Evidence Extraction and Synthesis ▶ *Underline Video*

Massimiliano Pronesti, Angelo Miculescu, Mohsin Kapdi, Paul Flanagan, Oisín Redmond, Joao H Bettencourt-Silva, Gurdeep Singh Mannu, Spiros Denaxas, Rui Providencia, Anya Belz and Yufang Hou;

Systematic reviews rely on forest plots to synthesise quantitative evidence across biomedical studies, but generating them remains a fragmented and labour-intensive process. Researchers must interpret complex clinical texts, manually extract outcome data from trials, define appropriate interventions and comparators, harmonise inconsistent study designs, and carry out meta-analytic computations, typically using specialised software that demands structured inputs and domain expertise. While recent work has demonstrated that large language models can extract study-level data from unstructured text, no existing system automates the complete pipeline from raw documents to synthesised forest plots. To address this gap, we introduce AutoForest, the first end-to-end system that generates publication-ready forest plots directly from biomedical papers. Given one or more study papers, AutoForest automatically suggests ICO (Intervention, Comparator, Outcome) elements, extracts outcome data, performs statistical synthesis, and renders the final forest plot. We describe the system architecture, user interface and demonstrate its effectiveness on real-world examples through a user study involving clinicians, showing how AutoForest accelerates evidence synthesis and substantially lowers the barrier to conducting meta-analyses.

[Demo] DialogGuard: Multi-Agent Psychosocial Safety Evaluation Interface of Sensitive LLM Responses ▶ *Underline Video*

Han Luo and Guy Laban

LLM-based agents are increasingly deployed for mental-health support and crisis coun- selling, yet recent evaluations reveal that commercial therapy chatbots respond appropriately only about half the time in clinical scenarios. Clinicians and safety engineers are called upon to audit these systems, but existing tools do not surface the subtler psychosocial harms (manip- ulation, discrimination, psychological distress) nor produce the explainable rationales that prac- titioners need. We present DialogGuard, an open-source system that lets practitioners inspect, stress-test, and create audit trails for prompted LLM agents across five psychosocial safety dimensions. The system wraps around arbitrary generative models through four LLM- as-a-judge pipelines (single-agent scoring, dual- agent correction, multi-agent debate, and majority voting), each grounded in shared three- level rubrics. Through its web interface, practitioners evaluate agents in two modes (Live Chat and Manual Input) and review per- dimension risk scores with natural-language rationales. Experiments on PKU-SafeRLHF show that dual-agent correction provides the best accuracy-robustness trade-off, and a for- mative study with 12 practitioners confirms that the system supports prompt auditing, safety inspection, and supervisory decision-making. Code and demo: https://anonymous.4open.science/r/dialogguard-web-CE7E.

[Demo] dLLM: Simple Diffusion Language Modeling ▶ *Underline Video*

Zhanhui Zhou, Lingjie Chen, Hanghang Tong and Dawn Song;

Although diffusion language models (DLMs) are evolving quickly, many recent models converge on a set of shared components. These components, however, are distributed across ad-hoc research codebases or lack transparent implementations, making them difficult to reproduce or extend. As the field accelerates, there is a clear need for a unified framework that standardizes these common components while remaining flexible enough to support new methods and architectures. To address this gap, we introduce dLLM, an open-source framework that unifies the core components of diffusion language modeling, training, inference, and evaluation, and makes them easy to customize for new designs. With dLLM, users can reproduce, finetune, deploy, and evaluate open-source large DLMs such as LLaDA and Dream through a standardized pipeline. The framework also provides minimal, reproducible recipes for building small DLMs from scratch with accessible compute, including converting any BERT-style encoder or autoregressive LM into a DLM. We also release the checkpoints of these small DLMs to make DLMs more accessible and accelerate future research.

[Demo] Graph of Trace: Visualizing Execution Traces of Scientific Agents ▶ *Underline Video*

Haoxuan Li

Scientific AI agents can autonomously carry out complex research workflows, yet these unfolded workflows often remains difficult for humans to inspect and review, limiting interpretable, controllable and effective human-AI collaboration. To address this challenge, we present a monitoring and visualization framework that records fine-grained execution events and organizes them into a directed graph that make agent workflows explicit as they proceed. The system records intermediate steps (e.g. tool calls and code executions), and renders them

as real-time updated visual traces that expose workflow structure. This allows domain experts to examine how results are produced, identify where failures emerge, and better understand agent behavior across different stages of the research process. We conduct an evaluation on complex research tasks with domain experts of interdisciplinary background in AI, neuroscience and biology. Experts report that structured traces visualization improves understanding of agent workflows, perceived interpretability, and usability for analysis and further interaction.

[Demo] IDP Accelerator: Agentic Document Intelligence from Extraction to Compliance Validation ▶ *Underline Video*

Md Mofijul Islam, Md Sirajus Salekin, Joe King, Priyashree Roy, Vamsi Thilak Gudi, Spencer Romo, AKHIL NOONEY, Bob Strahan, Boyi Xie and Diego A. Socolinsky

Understanding and extracting structured insights from unstructured documents remains a foundational challenge in industrial NLP. While Large Language Models (LLMs) enable zero-shot extraction, traditional pipelines often fail to handle multi-document packets, complex reasoning, and strict compliance requirements. We present IDP (Intelligent Document Processing) Accelerator, a framework enabling agentic AI for end-to-end document intelligence with four key components: (1) DocSplit, a novel benchmark dataset and multimodal classifier using BIO tagging to segment complex document packets; (2) configurable Extraction Module leveraging multimodal LLMs to transform unstructured content into structured data; (3) Agentic Analytics Module, compliant with the Model Context Protocol (MCP) providing data access through secure, sandboxed code execution; and (4) Rule Validation Module replacing deterministic engines with LLM-driven logic for complex compliance checks. The interactive demonstration enables users to upload document packets, visualize classification results, and explore extracted data through an intuitive web interface. We demonstrate effectiveness across industries, highlighting a production deployment at a leading healthcare provider achieving 98% classification accuracy, 80% reduced processing latency, and 77% lower operational costs over legacy baselines. IDP Accelerator is open-sourced with a live demonstration available to the community.

[Demo] JobMatchAI - An Intelligent Job Matching Platform Using Knowledge Graphs, Semantic Search and Explainable AI ▶ *Underline Video*

Mayank Vyas, Abhijit Chakraborty and Vivek Gupta

Recruiters and job seekers rely on search systems to navigate labor markets, making candidate matching engines critical for hiring outcomes. Most systems act as keyword filters, failing to handle skill synonyms and nonlinear careers, resulting in missed candidates and opaque match scores. We introduce JobMatchAI, a production-ready system integrating Transformer embeddings, skill knowledge graphs, and interpretable reranking. Our system optimizes utility across skill fit, experience, location, salary, and company preferences, providing factor-wise explanations through resume-driven search workflows. We release JobSearch-XS benchmark and a hybrid retrieval stack combining BM25, knowledge graph and semantic components to evaluate skill generalization. We assess system performance on JobSearch-XS across retrieval tasks, provide a demo video, a hosted website and installable package.

[Demo] RiskLab: A Controlled Toolkit for Probing Emergent Risks in LLM-Based Multi-Agent Systems ▶ *Underline Video*

Yue Huang, Yu Jiang, Wenjie Wang, Yanbo Wang, Zhenhong Zhou, Xiuying Chen, Yang Liu, Pin-Yu Chen, Wei Wang and Xiangliang Zhang

Large language model (LLM) agents increasingly operate in multi-agent settings where failures emerge from interaction dynamics rather than isolated model errors. We introduce RiskLab, an open-source toolkit for instantiating, probing, and measuring emergent risks in LLM-based multi-agent systems under controlled conditions. Each experiment is defined as a structured topology-environment-protocol-agent-task quintuple, enabling reproducible studies of how communication structure, coordination mechanisms, and incentives shape system-level risks. RiskLab provides flexible communication topologies, swappable interaction protocols, trajectory-grounded evaluation, and extensible registries for risk detectors and agent backends. We demonstrate the toolkit across representative risks, including collusion, resource overreach, semantic drift, and strategic misreporting, and support one-file reproducibility via configuration.

[Demo] ScheMatiQ: From Research Question to Structured Data through Interactive Schema Discovery ▶ *Underline Video*

Shahar Levy, Eliya Habba, Reshef Mintz, Barak Raveh, Renana Keydar and Gabriel Stanovsky

Many disciplines pose natural-language research questions over large document collections whose answers typically requires structured evidence, traditionally obtained by manually designing an annotation schema and exhaustively labeling the corpus, a slow and error-prone process. We introduce ScheMatiQ, which leverages calls to a backbone LLM to take a question and a corpus to produce a schema and a grounded database, with a web interface that lets steer and revise the extraction. In collaboration with domain experts, we show that ScheMatiQ yields outputs that support real-world analysis in law and computational biology. We release ScheMatiQ as open source with a public web interface, and invite experts across disciplines to use it with their own data. All resources, including the website, source code, and demonstration video, are available at: www.ScheMatiQ-ai.com.

Oral Session A: Computational Social Science, Cultural Analytics, and NLP for Social Good 1

11:00-12:30 (Regatta)

[Long] A Theoretically Grounded Approach to Summarizing Conversation Dynamics for Forecasting the Derailment of Online Conversations ▶ *Underline Video*

Yingxue Fu and Anaïs Ollagnier

Conversation derailment prediction represents a new paradigm of toxicity detection, where a system predicts from the start of a conversation whether it will derail into toxic exchanges, allowing moderators and users to act preemptively before harm is done. This approach requires a deep understanding of conversation dynamics. Previous work relies on linguistic features rooted in linguistic and social theories. While these features provide signals of conversation dynamics, they are exploratory in nature and potentially reflect a fraction of the overall pragmatic devices shaping the conversation trajectory. To capture the pragmatic dimension of conversations systematically, we start with a framework for annotating pragmatic information of conversations systematically and design summary generation methods to capture conversation trajectory dynamically. We achieve about 10% performance increase over a simple baseline, and 6.47% increase over a strong baseline on a dataset, and a slight performance increase on a benchmark dataset for the task of summarizing conversation dynamics.

[Long] CAPC-CG: A Large-Scale, Expert-Directed LLM-Annotated Corpus of Adaptive Policy Communication in China ► *Underline Video*

Bolun Sun, Charles Chang, Yuen Yuen Ang, Ruotong Mu, Yuchen Xu and Zhengxin Zhang

We introduce CAPC-CG, the Chinese Adaptive Policy Communication (Central Government) Corpus, the first open dataset of Chinese policy directives annotated with a five-color typology of policy signals, capturing clarity and ambiguity, grounded in the theory of adaptive policy communication. Spanning 1949-2023, this corpus includes laws, regulations, and rules issued by Chinese central authorities, segmented into 3.3 million paragraph units. We further propose and validate an expert-directed LLM annotation method that integrates codebook design, structured training, a two-step workflow, and LLM-based scaling. Alongside the corpus, we release metadata and a gold-standard labeled set developed by trained coders. Inter-annotator agreement achieves a Fleiss kappa of kappa = 0.86 on directive labels, indicating high reliability. We provide baseline classification results with several large language models (LLMs), together with our codebook, and describe patterns from the data. This release enables downstream tasks and multilingual NLP research in communication strategies under complexity and uncertainty.

[Long] Psychological Steering in LLMs: An Evaluation of Effectiveness and Trustworthiness ► *Underline Video*

Amin Banayeeanzade, Ala N. Tak, Fatemeh Bahrani, Anahita Bolourani, Leonardo Blas, Emilio Ferrara, Jonathan Gratch and Sai Praneeth Karimireddy

The ability to control LLMs' emulated emotional states and personality traits is an essential step in enabling rich, human-centered interactions in socially interactive settings. We introduce PsySET, a Psychologically-informed benchmark to evaluate LLM Steering Effectiveness and Trustworthiness across the emotion and personality domains. Our study spans four models from different LLM families paired with various steering strategies, including prompting, fine-tuning, and representation engineering. Our results indicate that prompting is consistently effective but limited in intensity control, whereas vector injections achieve finer controllability while slightly reducing output quality. Moreover, we explore the trustworthiness of steered LLMs by assessing safety, truthfulness, fairness, and ethics, highlighting potential side effects and behavioral shifts. Notably, we observe idiosyncratic effects; for instance, even a positive emotion like joy can degrade robustness to adversarial factuality, lower privacy awareness, and increase preferential bias. Meanwhile, anger predictably elevates toxicity yet strengthens leakage resistance. Our framework establishes the first holistic evaluation of emotion and personality steering, offering insights into its interpretability and reliability for socially interactive applications.

[TACL] Retain or Reframe? A Computational Framework for the Analysis of Framing in News Articles and Reader Comments ► *Underline Video*

Matteo Guida, Yulia Otmakhova, Eduard Hovy and Lea Frermann

When a news article describes immigration as an "economic burden" or a "humanitarian crisis," it selectively emphasizes certain aspects of the issue. Although this framing shapes how the public interprets such issues, audiences do not absorb frames passively but actively reorganize the presented information. While this relationship between source content and audience response is well-documented in the social sciences, NLP approaches often ignore it, detecting frames in articles and responses in isolation. We present the first computational framework for large-scale analysis of framing across source content (news articles) and audience responses (reader comments). Methodologically, we refine frame labels and develop a framework that reconstructs dominant frames in articles and comments from sentence-level predictions, and aligns articles with topically relevant comments. Applying our framework across eleven topics and two news outlets, we find that frame reuse in comments correlates highly across outlets, and that readers often selectively engage with frames of the articles. We release a frame classifier that performs well on both articles and comments, a dataset of article and comment sentences manually labeled for frames, and a large-scale dataset of articles and comments with predicted frame labels.

[Long] Splits! Flexible Sociocultural Linguistic Investigation at Scale ► *Underline Video*

Eylon Caplan, Tania Chakraborty and Dan Goldwasser

Variation in language use, shaped by speakers' sociocultural background and specific context of use, offers a rich lens into cultural perspectives, values, and opinions. For example, Chinese students discuss healthy eating with words like timing, regularity, and digestion, whereas Americans use vocabulary like balancing food groups and avoiding fat and sugar, reflecting distinct cultural models of nutrition (Banna et al., 2016). The computational study of these Sociocultural Linguistic Phenomena (SLP) has traditionally been done in NLP via tailored analyses of specific groups or topics, requiring specialized data collection and experimental operationalization, a process not well-suited to quick hypothesis exploration and prototyping. To address this, we propose constructing a "sandbox" designed for systematic and flexible sociolinguistic research. Using our method, we construct a demographically/topically split Reddit dataset, Splits!, validated

by self-identification and by replicating several known SLPs from existing literature. We showcase the sandbox's utility with a scalable, two-stage process that filters large collections of potential SLPs (PSLPs) to surface the most promising candidates for deeper, qualitative investigation.

[Long] Tears or Cheers? Benchmarking LLMs via Culturally Elicited Distinct Affective Responses ▶ *Underline Video*

Chongyuan Dai, Yaling Shen, Jinpeng Hu, Zihan Gao, Jia Li, Yishun Jiang, Yaxiong Wang, Liu Liu and Zongyuan Ge

Culture serves as a fundamental determinant of human affective processing and profoundly shapes how individuals perceive and interpret emotional stimuli. Despite this intrinsic link extant evaluations regarding cultural alignment within Large Language Models primarily prioritize declarative knowledge such as geographical facts or established societal customs. These benchmarks remain insufficient to capture the subjective interpretative variance inherent to diverse sociocultural lenses. To address this limitation, we introduce CEDAR, a multimodal benchmark constructed entirely from scenarios capturing Culturally Elicited Distinct Affective Responses. To construct CEDAR, we implement a novel pipeline that leverages LLM-generated provisional labels to isolate instances yielding cross-cultural emotional distinctions, and subsequently derives reliable ground-truth annotations through rigorous human evaluation. The resulting benchmark comprises 10,962 instances across seven languages and 14 fine-grained emotion categories, with each language including 400 multimodal and 1,166 text-only samples. Comprehensive evaluations of 17 representative multilingual models reveal a dissociation between language consistency and cultural alignment, demonstrating that culturally grounded affective understanding remains a significant challenge for current models.

[Long] The Proxy Presumption: From Semantic Embeddings to Valid Social Measures ▶ *Underline Video*

Baishi Li, Ta Yu, Kelvin J.L. Koa and Ke-Wei Huang

Natural Language Processing is rapidly evolving into a primary instrument for Computational Social Science, with researchers increasingly using embeddings to measure latent constructs such as novelty, creativity, and bias. However, this transition faces a fundamental validity challenge: the "Proxy Presumption," or the reliance on geometric properties (e.g., cosine distance) as direct measures of social concepts. We argue that without explicit validation, unsupervised representations remain entangled mixtures of the target construct (C) and confounding attributes (Z) like topic, style, and authorship. To bridge the gap between semantic embeddings and valid social measures, we introduce the Construct Validity Protocol (CVP). Drawing on causal representation learning and psychometrics, the CVP offers a rigorous pipeline from conceptualization to quantitative verification. We further propose Counterfactual Neutralization, a novel method using LLMs to reduce confounding in embedding space. By providing a standardized Validity Suite—including tests for discriminant, incremental, and predictive validity—this work offers the community a toolkit to transform heuristic proxies into robust, scientifically defensible instruments.

[Long] Uncovering Temporal Framing in the News ▶ *Underline Video*

Tarek Mahmoud, Veronika Solopova, Premtim Sahitaj, Ariana Sahitaj, Max Upravitelev, Mervat Abassy, Hana Fatima Shaikh, Neda Foroutan, Vera Schmitt and Preslav Nakov

Temporal language does more than place events on a timeline. In news discourse, references to the past, present, and future can function as rhetorical devices that shape interpretation and persuasion. Here, we study temporal framing, defined as the persuasive use of time-related language to structure meaning rather than to report chronology. We propose a taxonomy of eight temporal frames grounded in prior work on temporality and framing, and we realize it through expert annotation of a multilingual news corpus. The resulting dataset includes 458 English and German news articles, with over 2K temporally framed sentences and approximately 3K temporal framing annotations identified from a corpus of more than 20K sentences. We analyze frame prevalence, co-occurrence patterns, and lexical cues, and evaluate temporal framing detection using supervised fine-tuning and zero-shot classification. Our experiments show that temporal framing is learnable at the sentence level, with supervised models substantially outperforming zero-shot approaches. We publicly release the corpus to support future research on temporal framing: https://mbzuai-nlp.github.io/temporal-framing/.

[Long] What is a protest anyway? Codebook conceptualization is still a first-order concern in LLM-era classification ▶ *Underline Video*

Andrew Halterman and Katherine Keith

Generative large language models (LLMs) are now used extensively for text classification in computational social science (CSS). In this work, we focus on the steps before and after LLM prompting: conceptualization of the categories to classify and using LLM predictions in downstream statistical inference. We argue these steps have been overlooked in much of LLM-era CSS and LLMs can tempt analysts to skip conceptualization altogether. For example, a political scientist classifying "protest" with LLMs may never be forced to craft a definition: unlike human annotators who would ask clarifying questions, an LLM can silently accept an underspecified concept to classify and return plausible-looking labels. Using simulations, we show that conceptualization failures induce downstream inferential bias that cannot be corrected solely by a more accurate LLM or post-hoc bias correction methods. We conclude by reminding CSS analysts that conceptualization is still a first-order concern in the LLM-era and provide concrete advice for pursuing low-cost, unbiased, low-variance downstream estimates.

Oral Session A: Ethics, Bias, and Fairness 1

11:00-12:30 (Promenade)

[Long] A Game-Theoretica Negotiation Framework for Cross-Cultural Consensus ▶ *Underline Video*

Guoxi Zhang, Jiawei Chen, Tianzhuo Yang, Jiaming Ji, Yaodong Yang and Juntao Dai

Large language models (LLMs) are shaping global values, yet they frequently exhibit a pronounced WEIRD (Western, Educated, Industrialized, Rich, Democratic) cultural bias, marginalizing diverse viewpoints and posing challenges for reconciling diverse populations with varying cultural backgrounds and value systems. In this work, we move beyond simple alignment methods to propose a new paradigm for cross-cultural fairness. We introduce a Nash Consensus Negotiation framework under the formulation of cross-cultural consensus as a Nash Equilibrium. Each LLM iteratively proposes and refines natural-language guidelines, guided by a utility function balancing self-consistency with mutual acceptance, while penalizing redundancy. The process expands the proposal space and converges to a consensus, yielding fair and interpretable consensus outcomes. We evaluate our framework against baselines using quantitative metrics, qualitative analysis, and large-scale human studies. Experiments demonstrate that our framework generates higher-quality and more balanced consensus, effectively mitigating assimilation toward WEIRD values. Furthermore, we finetune diverse LLM architectures with negotiation data via preference optimization and supervised reasoning, reducing cultural distances by up to 95.53%. Overall, our work offers a systematic path to mitigate cultural bias in LLMs by guiding them toward self-consistency, mutually-acceptable equilibria.

[Long] AwarenessBench: Assessing Cognitive Capabilities of Language Models ▶ *Underline Video*

Xiaojian Li, Rongwu Xu, Tianyun Zhang, Yue Wang, Shuo Chen, Qiner Lyu, Briana Zhang, Peiran Yang, Kyle Xue Chen, Haoyuan Shi, Yu Wang and Wei Xu

As language models (LMs) exhibit increasingly consciousness-like behaviors, evaluating their cognitive abilities becomes essential. We introduce AwarenessBench, the first comprehensive benchmark for assessing the cognitive abilities of LMs in four dimensions: metacognition, self-awareness, social awareness, and situational awareness, covering 15 cognitive functions and 14,381 samples. Evaluating 18 state-of-the-art LMs, we find that all consistently surpass random baselines, with more advanced models performing better. We further compare LMs with human performance across three demographic groups, where the best-performing model surpasses human averages overall, but most still fall markedly short in metacognition and self-awareness. Finally, we show that awareness is a distinct capability: progress in language modeling or reasoning does not necessarily translate into improved cognition.

[Long] CiPO: Counterfactual Unlearning for Large Reasoning Models through Iterative Preference Optimization ▶ *Underline Video*

Junyi Li, Yongqiang Chen and Ningning Ding

Machine unlearning has gained increasing attention in recent years, as a promising technique to selectively remove unwanted privacy or copyrighted information from Large Language Models that are trained on a massive scale of human data. However, the emergence of Large Reasoning Models (LRMs), which emphasize long chain-of-thought (CoT) reasoning to address complex questions, presents a dilemma to unlearning: existing methods either struggle to completely eliminate undesired knowledge from the CoT traces or degrade the reasoning performances due to the interference with the reasoning process. To this end, we introduce Counterfactual Unlearning through iterative Preference Optimization (CiPO), a novel framework that redefines unlearning as the targeted intervention of the CoT reasoning in LRMs. More specifically, given a desired unlearning target answer, CiPO instructs LRMs to generate a logically valid counterfactual reasoning trace for preference tuning. As the LRM adjusts to the counterfactual trace, CiPO iteratively updates the preference learning data to increase the discrepancy from the original model. This iterative loop ensures both desirable unlearning and smooth optimization, effectively mitigating the dilemma. Experiments on challenging benchmarks demonstrate that CiPO excels at unlearning, completely removing knowledge from both the intermediate CoT steps and the final answer, while preserving the reasoning abilities of LRMs.

[Long] DIA-HARM: Dialectal Disparities in Harmful Content Detection Across 50 English Dialects ▶ *Underline Video*

Jason S Lucas, Matt Murtagh White, Ali Al-Lawati, Uchendu Uchendu, Adaku Uchendu and Dongwon Lee

Harmful content detectors—particularly disinformation classifiers—are predominantly developed and evaluated on Standard American English (), leaving their robustness to dialectal variation unexplored. We present, the first benchmark for evaluating disinformation detection robustness across 50 English dialects spanning U.S., British, African, Caribbean, and Asia-Pacific varieties. Using Multi-VALUE's linguistically-grounded transformations, we introduce blueD-CUBE (Dialectal Disinformation Detection Corpus), a core corpus component of comprising 195K samples derived from established disinformation benchmarks. Our evaluation of 16 detection models reveals systematic vulnerabilities: human-written dialectal content degrades detection by 1.4–3.6% F1, while AI-generated content remains stable. Fine-tuned transformers substantially outperform zero-shot LLMs (96.6% vs. 78.3% best-case F1), with some models exhibiting catastrophic failures exceeding 33% degradation on mixed content. Cross-dialectal transfer analysis across 2,450 dialect pairs shows that multilingual models (mDeBERTa: 97.2% average F1) generalize effectively, while monolingual models like RoBERTa and XLM-RoBERTa fail on dialectal inputs. These findings demonstrate that current disinformation detectors may systematically disadvantage hundreds of millions of non-speakers worldwide. We release the bluebenchmark, including the https://anonymous.4open.science/r/dia-harm-61D9/, and evaluation tools.

[Long] Identifying Bias in Machine-generated Text Detection ▶ *Underline Video*

Kevin Stowe, Svetlana Afanaseva, Rodolfo Raimundo, Yitao Sun and Kailash Patil

The meteoric rise in text generation capability has been accompanied by parallel growth in interest in machine-generated text detection:

the capability to identify whether a given text was generated using a model or written by a person. While detection models show strong performance, they have the capacity to cause significant negative impacts. We explore potential biases in English machine-generated text detection systems. We curate a dataset of student essays and assess 16 different detection systems for bias across four attributes: gender, race/ethnicity, English-language learner (ELL) status, and economic status. We evaluate these attributes using regression-based models to determine the significance and power of the effects, as well as performing subgroup analysis. We find that while biases are generally inconsistent across systems, there are several key issues: several models tend to classify disadvantaged groups as machine-generated, ELL essays are more likely to be classified as machine-generated, economically disadvantaged students' essays are less likely to be classified as machine-generated, and non-White ELL essays are disproportionately classified as machine-generated relative to their White counterparts. Finally, we perform human annotation and find that while humans perform generally poorly at the detection task, they show no significant biases on the studied attributes.

[Long] Identity-Robust Language Model Generation via Content Integrity Preservation ▶ *Underline Video*

Miao Zhang, Kelly Chen, Mehrab Tanjim and Rumi Chunara

Large Language Model (LLM) outputs often vary across user sociodemographic attributes, leading to disparities in factual accuracy, utility, and safety, even for objective questions where demographic information is irrelevant. Unlike prior work on stereotypical or representational bias, this paper studies identity-dependent degradation of core response quality. We show empirically that such degradation arises from biased generation behavior, despite factual knowledge being robustly encoded across identities. Motivated by this mismatch, we propose a lightweight, training-free framework for identity-robust generation that selectively neutralizes non-critical identity information while preserving semantically essential attributes, thus maintaining output content integrity. Experiments across four benchmarks and 18 sociodemographic identities demonstrate an average 66.3% reduction in identity-dependent bias compared to vanilla prompting and outperforms existing prompt-based defenses. Our work addresses a critical gap in mitigating the impact of user identity cues in prompts on core generation quality.

[Long] SPAGBias: Uncovering and Tracing Structured Spatial Gender Bias in Large Language Models ▶ *Underline Video*

Binxian Su, Haoye Lou, Shucheng Zhu, Weikang Wang, Ying Liu, Dong Yu and Pengyuan Liu

Large language models (LLMs) are being increasingly used in urban planning, but since gendered space theory highlights how gender hierarchies are embedded in spatial organization, there is concern that LLMs may reproduce or amplify such biases. We introduce SPAGBias , the first systematic framework to evaluate spatial gender bias in LLMs. It combines a taxonomy of 62 urban micro-spaces, a prompt library, and three diagnostic layers: explicit (forced-choice resampling), probabilistic (token-level asymmetry), and constructional (semantic and narrative role analysis). Testing six representative models, we identify structured gender-space associations that go beyond the public-private divide, forming nuanced micro-level mappings. Story generation reveals how emotion, wording, and social roles jointly shape spatial gender narratives. We also examine how prompt design, temperature, and model scale influence bias expression. Tracing experiments indicate that these patterns are embedded and reinforced across the model pipeline (pre-training, instruction tuning, and reward modeling), with model associations found to substantially exceed real-world distributions. Downstream experiments further reveal that such biases produce concrete failures in both normative and descriptive application settings. This work connects sociological theory with computational analysis, extending bias research into the spatial domain and uncovering how LLMs encode social gender cognition through language.

[Long] Thinking beyond the anthropomorphic paradigm benefits LLM research ▶ *Underline Video*

Lujain Ibrahim and Myra Cheng

Anthropomorphism, or the attribution of human traits to technology, is an automatic and unconscious response that occurs even in those with advanced technical expertise. In this position paper, we analyze hundreds of thousands of research articles to present empirical evidence of the prevalence and growth of anthropomorphic terminology in research on large language models (LLMs). We argue for challenging the deeper assumptions reflected in this terminology , which, though often useful, may inadvertently constrain LLM development , and broadening beyond them to open new pathways for understanding and improving LLMs. Specifically, we identify and examine five anthropomorphic assumptions that shape research across the LLM development lifecycle. For each assumption (e.g., that LLMs must use natural language for reasoning, or that they should be evaluated on benchmarks originally meant for humans), we demonstrate empirical, non-anthropomorphic alternatives that remain under-explored yet offer promising directions for LLM research and development.

[Long] VIGNETTE: Socially Grounded Bias Evaluation for Vision-Language Models ▶ *Underline Video*

Chahat Raj, Bowen Wei, Aylin Caliskan, Antonios Anastasopoulos and Ziwei Zhu

While bias in large language models (LLMs) is well-studied, similar concerns in vision-language models (VLMs) have received comparatively less attention. Existing VLM bias studies often focus on portrait-style images and gender-occupation associations, overlooking broader and more complex social stereotypes and their implied harm. This work introduces Vignette, a large-scale VQA benchmark with 30M+ images for evaluating bias in VLMs through a question-answering framework spanning four directions: factuality, perception, stereotyping, and decision making. Beyond narrowly-centered studies, we assess how VLMs interpret identities in contextualized settings, revealing how models make trait and capability assumptions and exhibit patterns of discrimination. Drawing from social psychology, we examine how VLMs connect visual identity cues to trait and role-based inferences, encoding social hierarchies, through biased selections. Our findings uncover subtle, multifaceted, and surprising stereotypical patterns, offering insights into how VLMs construct social meaning from inputs.

Oral Session A: Industry 1

11:00-12:30 (Balboa)

[Industry] AgenticQwen: Training Small Agentic Language Models with Dual Data Flywheels for Industrial-Scale Tool Use ► *Underline Video*

Yuanjie Lyu, Chengyu Wang, Haonan Zheng, Yuanhao Yue, Junbing Yan, Ming Wang and Jun Huang

Modern industrial applications increasingly demand language models that act as agents, capable of multi-step reasoning and tool use in real-world settings. These tasks are typically performed under strict cost and latency constraints, making small agentic models highly desirable. In this paper, we introduce the AgenticQwen family of models, trained via multi-round reinforcement learning (RL) on synthetic data and a limited amount of open-source data. Our training framework combines reasoning RL and agentic RL with dual data flywheels that automatically generate increasingly challenging tasks. The reasoning flywheel increases task difficulty by learning from errors, while the agentic flywheel expands linear workflows into multi-branch behavior trees that better reflect the decision complexity of real-world applications. We validate AgenticQwen on public benchmarks and in an industrial agent system. The models achieve strong performance on multiple agentic benchmarks, and in our industrial agent system, close the gap with much larger models on search and data analysis tasks.

[Industry] SimRPD: Optimizing Recruitment Proactive Dialogue Agents through Simulator-Based Data Evaluation and Selection ► *Underline Video*

Zhiyong Cao, Dunqiang Liu, Qi Dai, Haojun Xu, Huai Yuen Khor, Hao Wang, Huan He, Yafei Liu, Ke Ma, Ruqian Shi, Sicheng Zhou and Sijia Yao

Task-oriented proactive dialogue agents play a pivotal role in recruitment, particularly for steering conversations towards specific business outcomes, such as acquiring social-media contacts for private-channel conversion. Although supervised fine-tuning and reinforcement learning have proven effective for training such agents, their performance is heavily constrained by the scarcity of high-quality, goal-oriented domain-specific training data. To address this challenge, we propose SimRPD, a three-stage framework for training recruitment proactive dialogue agents. First, we develop a high-fidelity user simulator to synthesize large-scale conversational data through multi-turn online dialogue. Then we introduce a multi-dimensional evaluation framework based on Chain-of-Intention (CoI) to comprehensively assess the simulator and effectively select high-quality data, incorporating both global-level and instance-level metrics. Finally, we train the recruitment proactive dialogue agent on the selected dataset. Experiments in a real-world recruitment scenario demonstrate that SimRPD outperforms existing simulator-based data selection strategies, highlighting its practical value for industrial deployment and its potential applicability to other business-oriented dialogue scenarios.

[Industry] Thinking with Reasoning Skills: Fewer Tokens, More Accuracy ► *Underline Video*

Guangxiang Zhao, Qilong Shi, Xiangzheng Zhang, Xusen Xiao, Tong Yang and Lin Sun

Reasoning LLMs often spend substantial tokens on long intermediate reasoning traces (e.g., chain-of-thought) when solving new problems. We propose to summarize and store reusable reasoning skills distilled from extensive deliberation and trial-and-error exploration, and to retrieve these skills at inference time to guide future reasoning. Unlike the prevailing reasoning from scratch paradigm, our approach first recalls relevant skills for each query, helping the model avoid redundant detours and focus on effective solution paths. We evaluate our method on coding and mathematical reasoning tasks, and find that it significantly reduces reasoning tokens while improving overall performance. The resulting lower per-request cost indicates strong practical and economic potential for real-world deployment.

[Industry] Towards Faithful Industrial RAG: A Reinforced Co-adaptation Framework for Advertising QA ► *Underline Video*

wenwei li, Ming Xu, Tianle Xia, Lingxiang Hu, Yiding Sun, Linfang Shang, Liqun Liu, Peng Shu, Huan Yu and Jie Jiang

Industrial advertising question answering (QA) is a high-stakes task in which hallucinated content, particularly fabricated URLs, can lead to financial loss, compliance violations, and legal risk. Although Retrieval-Augmented Generation (RAG) is widely adopted, deploying it in production remains challenging because industrial knowledge is inherently relational, frequently updated, and insufficiently aligned with generation objectives. We propose a reinforced co-adaptation framework that jointly optimizes retrieval and generation through two components: (1) Graph-aware Retrieval (GraphRAG), which models entity-relation structure over a high-citation knowledge subgraph for multi-hop, domain-specific evidence selection; and (2) evidence-constrained reinforcement learning via Group Relative Policy Optimization (GRPO) with multi-dimensional rewards covering faithfulness, style compliance, safety, and URL validity. Experiments on an internal advertising QA dataset show consistent gains across expert-judged dimensions including accuracy, completeness, and safety, while reducing the hallucination rate by 72%. A two-week online A/B test demonstrates a 28.6% increase in like rate, a 46.2% decrease in dislike rate, and a 92.7% reduction in URL hallucination. The system has been running in production for over half a year and has served millions of QA interactions.

Oral Session A: Information Extraction and Retrieval 1

11:00-12:30 (Harbor G)

[Long] ARK: Answer-Centric Retriever Tuning via KG-augmented Curriculum Learning ▶ *Underline Video*

Jiawei Zhou, Hang Ding and Haiyun Jiang

Retrieval-Augmented Generation (RAG) has emerged as a powerful framework for knowledge-intensive tasks, yet its effectiveness in long-context scenarios is often bottlenecked by the retriever's inability to distinguish sparse yet crucial evidence. Standard retrievers, optimized for query-document similarity, frequently fail to align with the downstream goal of generating a precise answer. To bridge this gap, we propose a novel fine-tuning framework that optimizes the retriever for Answer Alignment. Specifically, we first identify high-quality positive chunks by evaluating their sufficiency to generate the correct answer. We then employ a curriculum-based contrastive learning scheme to fine-tune the retriever. This curriculum leverages LLM-constructed Knowledge Graphs (KGs) to generate augmented queries, which in turn mine progressively challenging hard negatives. This process trains the retriever to distinguish the answer-sufficient positive chunks from these nuanced distractors, enhancing its generalization. Extensive experiments on 10 datasets from the Ultradomain and LongBench benchmarks demonstrate that our fine-tuned retriever achieves state-of-the-art performance, improving 14.5% over the base model without substantial architectural modifications and maintaining strong efficiency for long-context RAG. Our work presents a robust and effective methodology for building truly answer-centric retrievers.

[Long] CoQuIR: A Comprehensive Benchmark for Code Quality-Aware Information Retrieval ▶ *Underline Video*

Jiahui Geng, Fengyu Cai, Shaobo Cui, Qing Li, Liangwei Chen, Chenyang Lyu, Haonan Li, Derui Zhu, Alexander Pretschner, Heinz Koeppl and Fakhri Karray

Code retrieval is vital to modern software engineering as it boosts reuse and speeds up debugging. However, current benchmarks primarily emphasize functional relevance while neglecting code quality. To address this gap, we introduce CoQuIR, the first large-scale, multilingual benchmark specifically designed to evaluate quality-aware code retrieval across four critical dimensions: correctness, efficiency, security, and maintainability. CoQuIR includes fine-grained quality annotations over 42,725 queries and 134,907 code snippets in 11 programming languages. Evaluating 23 retrievers (both open-source and proprietary) shows that even state-of-the-art models often fail to separate buggy or insecure code from robust counterparts. We further investigate methods for explicitly training retrievers to recognize code quality, demonstrating that quality-aware metrics can be improved without loss of semantic relevance; downstream code generation benefits from these gains. CoQuIR underscores the importance of embedding quality signals into retrieval systems as a crucial component for more trustworthy developer tools.

[Long] CRAFT: Training-Free Cascaded Retrieval for Tabular QA ▶ *Underline Video*

Adarsh Singh, Kushal Raj Bhandari, Jianxi Gao, Soham Dan and Vivek Gupta

Open-Domain Table Question Answering (TQA) involves retrieving relevant tables from a large corpus to answer natural language queries. Traditional dense retrieval models, such as DTR and DPR, not only incur high computational costs for large-scale retrieval tasks but also require retraining or fine-tuning on new datasets, limiting their adaptability to evolving domains and knowledge. In this work, we propose CRAFT, a zero-shot, cascaded retrieval approach that first uses a sparse retrieval model to filter a subset of candidate tables before applying more computationally expensive dense models as re-rankers. To improve retrieval quality, we enrich table representations with descriptive titles and summaries generated by Gemini Flash 1.5, enabling richer semantic matching between queries and tabular structures. Our method outperforms state-of-the-art (SOTA) sparse, dense, and hybrid retrievers on the NQ-Tables dataset. It also demonstrates strong zero-shot performance on the more challenging OTT-QA benchmark, achieving competitive results at higher recall thresholds, where the task requires multi-hop reasoning across both textual passages and relational tables. This work establishes a scalable and adaptable paradigm for table retrieval, bridging the gap between fine-tuned architectures and lightweight, plug-and-play retrieval systems. Code and data are available at: https://coral-lab-asu.github.io/CRAFT/

[Long] Don't Be Misled by Style: A Style-Adaptive Reranker for Capturing Effective Knowledge in Retrieval-Augmented Generation ▶ *Underline Video*

Ruwen Zhang, Bo Liu, Shengxiang Zhang, Yida Chen, Hantao Zhao, Ding Ding, Jiahui Jin and Jiuxin Cao

Rerankers are critical in Retrieval-Augmented Generation (RAG) for filtering evidence that enhances the accurate generation of LLMs. With the extension to open-domain scenarios, rerankers are inevitably deployed on mixed-style corpora, whereas most existing rerankers are mainly trained on well-edited texts. A rarely explored issue lies in enabling rerankers to maximally capture the effective knowledge for downstream LLMs without being misled by stylistic features. To address this issue, we propose SARK (Style-Adaptive Reranker with Knowledge Prioritization), a style-augmented multi-task framework that prioritizes effective knowledge over stylistic perturbations. SARK performs multi-granular knowledge mining by using an LLM to derive passage-level supervision on whether a passage helps or harms answer correctness, and list-level relative ranking preferences over candidate passages. It then jointly optimizes the reranker model with passage-level classification and list-level ranking objectives via style-augmented multi-task learning, encouraging the model to focus on the information needed for answering under mixed-style scenarios. Extensive experiments demonstrate that SARK improves generation performance across multiple LLMs under mixed-style conditions.

[Long] EXCEEDS: Extracting Complex Events via Nugget-based Grid Modeling in Scientific Domain ▶ *Underline Video*

Yi-Fan Lu, Xian-Ling Mao, Bo Wang, Xiao Liu and Heyan Huang

It is crucial to understand a specific domain by events. Extensive event extraction research has been conducted in many domains such as news, finance, and biology. However, event extraction in scientific domain is still insufficiently supported by comprehensive datasets and tailored methods. Compared with other domains, scientific domain has two characteristics: (1) denser nuggets and events, and (2) more complex information forms. To solve the above problem, considering these two characteristics, we first construct SciEvents, a large-scale multi-event document-level dataset with a schema tailored for scientific domain. It consists of 2,508 documents and 24,381 events under multi-stage manual annotation and quality control. Then, we propose EXCEEDS, an end-to-end scientific event extraction framework by encoding dense nuggets into a grid matrix and simplifying complex event extraction as a nugget-based grid modeling task. Experiments on SciEvents demonstrate state-of-the-art performances of EXCEEDS. Both the SciEvents dataset and the EXCEEDS framework are released publicly to facilitate future research.

[TACL] Inferring Scientific Cross-Document Coreference and Hierarchy with Definition-Augmented Relational Reasoning ▶ *Underline Video*

Lior Forer and Tom Hope

We address the fundamental task of inferring cross-document coreference and hierarchy in scientific texts, which has important applications in knowledge graph construction, search, recommendation, and discovery. LLMs can struggle when faced with many long-tail technical concepts with nuanced variations. We present a novel method which generates context-dependent definitions of concept mentions by retrieving full-text literature, and uses the definitions to enhance detection of cross-document relations. We further generate relational definitions, which describe how two concept mentions are related or different, and design an efficient reranking approach to address the combinatorial explosion involved in inferring links across papers. In both fine-tuning and in-context learning settings we achieve large gains in performance. We provide analysis of generated definitions, shedding light on the relational reasoning ability of Large Language Models (LLMs) over fine-grained scientific concepts.

[Long] MoCa: Modality-aware Continual Pre-training Makes Better Bidirectional Multimodal Embeddings ▶ *Underline Video*

Haonan Chen, Hong Liu, Yuping Luo, Liang Wang, Nan Yang, Furu Wei and Zhicheng Dou

Multimodal embedding models, built upon causal Vision Language Models (VLMs), have shown promise in various tasks. However, current approaches face three limitations: causal attention in VLM backbones is suboptimal for embedding tasks; scalability issues due to reliance on high-quality labeled paired data for contrastive learning; and limited diversity in training objectives and data. To address these issues, we propose MoCa, a two-stage framework for transforming pre-trained VLMs into bidirectional multimodal embedding models. The first stage, Modality-aware Continual Pre-training, introduces a joint reconstruction objective that simultaneously denoises interleaved texts and images, enhancing bidirectional context-aware reasoning. The second stage, Heterogeneous Contrastive Fine-tuning, leverages diverse, semantically rich multimodal data beyond simple image-caption pairs to enhance generalization and alignment. Our method addresses the stated limitations by introducing bidirectional attention through continual pre-training, scaling effectively with massive unlabeled datasets via joint reconstruction objectives, and utilizing diverse multimodal data for enhanced representation robustness. Experiments demonstrate that MoCa consistently improves performance across MMEB and ViDoRe-v2 benchmarks, achieving new state-of-the-arts, and exhibits strong scalability with both model size and training data on MMEB. We have released the model weights and data on our project page https://haon-chen.github.io/MoCa/.

[Long] ReasonEmbed: Enhanced Text Embeddings for Reasoning-Intensive Document Retrieval ▶ *Underline Video*

Jianlyu Chen, Junwei Lan, Chaofan Li, Defu Lian and Zheng Liu

In this paper, we introduce ReasonEmbed, a novel text embedding model developed for reasoning-intensive document retrieval. Our work includes three key technical contributions. First, we propose ReMixer, a new data synthesis method that overcomes the triviality problem prevalent in previous synthetic datasets, enabling large-scale production of 82K high-quality training samples. Second, we design Redapter, a self-adaptive learning algorithm that dynamically adjusts training each sample's weight based on its reasoning intensity. This allows the model to effectively capture the complex semantic relationships between queries and documents. Third, we implement ReasonEmbed across multiple backbones of varying sizes, all of which achieve superior performance on reasoning-intensive retrieval tasks. Notably, our ReasonEmbed-Qwen3-8B model offers a record-high nDCG@10 score of 38.1 on the BRIGHT benchmark, which significantly outperforms existing text embedding models. We will fully open-source our created resource in ReasonEmbed to push forward the research advancement in this field.

[TACL] Text-to-SQL Task-oriented Dialogue Ontology Construction ▶ *Underline Video*

Renato Vukovic, Carel van Niekerk, Michael Heck, Benjamin Ruppik, Hsien-Chin Lin, Shutong Feng, Nurul Lubis and Milica Gasic

Large language models (LLMs) are widely used as general-purpose knowledge sources, but they rely on parametric knowledge, limiting explainability and trustworthiness. In task-oriented dialogue (TOD) systems, this separation is explicit, using an external database structured by an explicit ontology to ensure explainability and controllability. However, building such ontologies requires manual labels or supervised training. We introduce TeQoDO: a Text-to-SQL task-oriented Dialogue Ontology construction method. Here, an LLM autonomously builds a TOD ontology from scratch without supervision using its inherent SQL programming capabilities combined with dialogue theory provided in the prompt. We show that TeQoDO outperforms transfer learning approaches, and its constructed ontology is competitive on a downstream dialogue state tracking task. Ablation studies demonstrate the key role of dialogue theory. TeQoDO also scales to allow construction of much

larger ontologies, which we investigate on a Wikipedia and ArXiv dataset. We view this as a step towards broader application of ontologies to increase LLM explainability.

Oral Session A: Interpretability and Analysis of Models for NLP 1

11:00-12:30 (Harbor D)

[Long] ADVICE: Answer-Dependent Verbalized Confidence Estimation ▶ *Underline Video*

KiJung Seo, Sehun Lim and Taeuk Kim

Recent progress in large language models (LLMs) has enabled them to communicate their confidence in natural language, improving transparency and reliability. However, this expressiveness is often accompanied by systematic overconfidence, whose underlying causes remain poorly understood. In this work, we analyze the dynamics of verbalized confidence estimation and identify answer-independence-the failure to condition confidence on the models own answer-as a primary driver of this behavior. To address this, we introduce ADVICE (Answer-Dependent VerbalIzed Confidence Estimation), a fine-tuning framework that promotes answer-grounded confidence estimation. Extensive experiments show that ADVICE substantially improves confidence calibration, while exhibiting strong generalization to unseen settings without degrading task performance.We further demonstrate that these gains stem from enhanced answer dependence, shedding light on the origins of overconfidence and enabling trustworthy confidence verbalization.

[Long] CRISP: Persistent Concept Unlearning via Sparse Autoencoders ▶ *Underline Video*

Tomer Ashuach, Dana Arad, Aaron Mueller, Martin Tutek and Yonatan Belinkov

As large language models (LLMs) are increasingly deployed in real-world applications, the need to selectively remove unwanted knowledge while preserving model utility has become paramount. Recent work has explored sparse autoencoders (SAEs) to perform precise interventions on monosemantic features. However, most SAE-based methods operate at inference time, which does not create persistent changes in the model's parameters. Such interventions can be bypassed or reversed by malicious actors with parameter access. We introduce CRISP, a parameter-efficient method for persistent concept unlearning using SAEs. CRISP automatically identifies salient SAE features across multiple layers and suppresses their activations. We experiment with two LLMs and show that our method outperforms prior approaches on safety-critical unlearning tasks from the WMDP benchmark, successfully removing harmful knowledge while preserving general and in-domain capabilities. Feature-level analysis reveals that CRISP achieves semantically coherent separation between target and benign concepts, allowing precise suppression of the target features.

[Long] Exploring Attention Attractors in Large Language Models ▶ *Underline Video*

Ziheng Wang, Zihao Yue, Wenxuan Wang and Qin Jin

This paper explores attention attractors, tokens that draw significantly high attention, in large language models. We analyze them from three perspectives: (1) Functionality: We demonstrate their role in aggregating information from preceding contexts to facilitate future predictions. (2) Distribution: Through layer-wise and token-wise analysis, we reveal that attention attractors are widely distributed across layers but predominantly originate from low-semantic words like "_the". (3) Mechanism: We demonstrate the correlation between attention weights allocated to tokens with their specific activation dimension values. We hope these findings provide new insights into the attention mechanisms of large language models and inspire further exploration.

[Long] Knowledge without Wisdom: Measuring Misalignment between LLMs and Intended Impact ▶ *Underline Video*

Michael Hardy and Yunsung Kim

LLMs increasingly excel on AI benchmarks, but doing so does not guarantee validity for downstream tasks. This study contrasts LLM alignment on benchmarks, downstream tasks, and, importantly the intended impact of those tasks. We evaluate the performance of leading LLMs (i.e., generative pre-trained base models) on difficult-to-verify tasks of the teaching and learning of schoolchildren. Across all LLMs, inter-model behaviors on disparate tasks correlate higher than they do with expert human behaviors on target tasks. These biases shared across LLMs are poorly aligned with downstream measures of teaching quality and often negatively aligned with the intended impact of student learning outcomes. Further, we find multi-model ensembles, both unanimous model voting and expert-weighting by benchmark performance, further exacerbate misalignment with learning. We measure that selection of LLM and/or prompting strategy only reliably accounts for 15% of all measured misalignment error and that variation in misalignment error is shared across LLMs, suggesting that common pretraining accounts for much of the misalignment in these tasks. We demonstrate methods for robustly measuring alignment of complex tasks and provide unique insights into practical applications of LLMs in high-noise contexts.

[Long] Mind the (DH) Gap! A Contrast in Risky Choices Between Reasoning and Conversational LLMs ▶ *Underline Video*

Luise Ge, Yongyan Zhang and Yevgeniy Vorobeychik

The use of large language models either as decision support systems, or in agentic workflows, is rapidly transforming the digital ecosystem. However, the understanding of LLM decision-making under uncertainty remains limited. We initiate a comparative study of LLM risky choices along two dimensions: (1) prospect representation (explicit vs. experience-based) and (2) decision rationale (explanation). Our

study, which involves 20 frontier and open LLMs, is complemented by a matched human subjects experiment, which provides one reference point, while an expected payoff maximizing rational agent model provides another. We find that LLMs cluster into two categories: reasoning models (RMs) and conversational models (CMs). RMs tend towards rational behavior, are insensitive to the order of prospects, gain/loss framing, and explanations, and behave similarly whether prospects are explicit or presented via experience history. CMs are significantly less rational, slightly more human-like, sensitive to prospect ordering, framing, and explanation, and exhibit a large description-history gap. Paired comparisons of open LLMs suggest that a key factor differentiating RMs and CMs is training for mathematical reasoning.

[Long] Quantifying Aleatoric Uncertainty of In-Context Learning for Robust Measure of LLM Prediction Confidence ▶ *Underline Video*

Jinseok Chung, Minkyoung Song, Hyunji Jung and Namhoon Lee

In-Context Learning (ICL) allows large language models to adapt to new tasks from a few demonstrations, but its reliability remains a concern: predictions are highly sensitive to both prompt design and the model's ability to understand the context, obscuring whether failures arise from data properties or model limitations. Uncertainty decomposition, separating aleatoric from epistemic sources, is particularly crucial in this setting, yet existing methods, designed for standard generation tasks, fail to capture the unique dynamics of ICL. To address this, we introduce a concept of self-function vectors, built upon Bayesian views and the mechanistic interpretability of ICL. These vectors leverage internal model representations to model the latent concept learned during in-context prompting, thereby enabling a direct estimation of aleatoric uncertainty within a Bayesian framework and circumventing the reliance on brittle input or decoding manipulations. Given the lack of established benchmarks and suitable evaluation protocols, we also propose the first and rigorous evaluation framework, initially grounded in synthetic tasks for conceptual development and subsequently extended to real-world datasets, in which aleatoric and epistemic uncertainty can be quantified separately under controlled settings. Moreover, we show it can be used as a practical tool for trustworthy-related applications, such as hallucination detection. Our findings pave a new direction for connecting the quantitative view of uncertainty with the mechanistic understanding of model behavior.

[Long] Schoenfeld's Anatomy of Mathematical Reasoning by Language Models ▶ *Underline Video*

Ming Li

Large language models increasingly expose reasoning traces, yet their underlying cognitive structure and steps remain difficult to identify and analyze beyond surface-level statistics. We adopt Schoenfeld's Episode Theory as an inductive, intermediate-scale lens and introduce ThinkARM (Anatomy of Reasoning in Models), a scalable framework that explicitly abstracts reasoning traces into functional reasoning steps such as Analysis, Explore, Implement, Verify, etc. When applied to mathematical problem solving by diverse models, this abstraction reveals reproducible thinking dynamics and structural differences between reasoning and non-reasoning models, which are not apparent from token-level views. We further present two diagnostic case studies showing that exploration functions as a critical branching step associated with correctness, and that efficiency-oriented methods selectively suppress evaluative feedback steps rather than uniformly shortening responses. Together, our results demonstrate that episode-level representations make reasoning steps explicit, enabling systematic analysis of how reasoning is structured, stabilized, and altered in modern language models.

[Long] SCOUT: Selective Coupling via Optimal Unbalanced Transport for Interpretable Text Classification ▶ *Underline Video*

Junhao Jia, Hanwen Zheng, Yueyi Wu, Huangwei Chen, Haishuai Wang, Jiajun Bu and Lei Wu

Natural language data is inherently noisy, yet standard interpretable models often rely on scalar similarities that obscure the true evidentiary basis of a prediction. This limitation is particularly detrimental to prototype-based classification, where traditional full-alignment mechanisms force non-informative background segments to match informative prototypes, yielding unstable or misleading explanations. To mitigate this, we present SCOUT, a novel paradigm that grounds prototype reasoning in the selective correspondence of discriminative fragments. Concretely, we represent each document as a discrete distribution over span embeddings and employ differentiable Unbalanced Optimal Transport (UOT) to align them with class-specific prototypes. Unlike standard methods, this mechanism enables the model to focus strictly on decisive evidence while leaving irrelevant noise unmatched via geometric mass suppression. To ensure verifiability, we anchor prototype supports to readable training spans, establishing a transparent bridge between input segments and stored knowledge. Comprehensive experiments on seven benchmarks demonstrate that SCOUT yields prototypes focused on semantically significant spans, significantly outperforming traditional rationale extraction and post-hoc attribution methods in terms of faithfulness and stability.

[Long] Sense and Sensitivity: Examining the Influence of Semantic Recall on Long Context Code Understanding ▶ *Underline Video*

Adam torek, Mukur Gupta, Samira Hajizadeh, Prashast Srivastava and Suman Jana

Large language models (LLMs) are increasingly deployed for understanding large codebases, but whether they understand operational semantics of long code context or rely on pattern matching shortcuts remains unclear. We distinguish between lexical recall (retrieving code verbatim) and semantic recall (understanding operational semantics). Evaluating 10 state-of-the-art LLMs, we find that while frontier models achieve near-perfect, position-independent lexical recall, semantic recall degrades severely when code is centrally positioned in long contexts. We introduce semantic recall sensitivity to measure whether tasks require understanding of code's operational semantics vs. permit pattern matching shortcuts. Through a novel counterfactual measurement method, we show that models rely heavily on pattern matching shortcuts to solve existing code understanding benchmarks. We propose a new task SemTrace, which achieves high semantic recall sensitivity through unpredictable operations; LLMs' accuracy exhibits severe positional effects, with median accuracy drops of 92.73% versus CRUXEval's 53.36% as the relevant code snippet approaches the middle of the input

code context. Our findings suggest current evaluations substantially underestimate semantic recall failures in long context code understanding.

Oral Session A: Language Models 1

11:00-12:30 (Harbor H-I)

[Long] C2: Scalable Rubric-Augmented Reward Modeling from Binary Preferences ▶ *Underline Video*

Akira Kawabata and Saku Sugawara

Rubric-augmented verification guides reward models with explicit evaluation criteria, yielding more reliable judgments than single-model verification. However, most existing methods require costly rubric annotations, limiting scalability. Moreover, we find that rubric generation is vulnerable to a failure of cooperation; low-quality rubrics actively mislead reward models rather than help. Inspired by the principle of cooperative communication, we propose Cooperative yet Critical reward modeling (C2), a framework that significantly improves reward model judgments by having the reward model critically collaborate with a rubric generator trained solely from binary preferences. In C2, we synthesize helpful and misleading rubric pairs by measuring how each rubric shifts the reward model toward or away from the correct preference. Using these contrastive pairs, we train a cooperative rubric generator to propose helpful rubrics, and a critical verifier to assess rubric validity before making its judgment, following only rubrics it deems helpful at inference time. C2 outperforms reasoning reward models trained on the same binary preferences, with gains of up to 6.5 points on RM-Bench and 6.0 points length-controlled win rate on AlpacaEval 2.0. Without external rubric annotations, C2 enables an 8B reward model to match performance achieved with rubrics from a 4x larger model. Overall, our work demonstrates that eliciting deliberate cooperation in rubric-augmented verification makes reward models more trustworthy in a scalable way.

[Long] Every Token Counts: Generalizing 16M Ultra-Long Context in Large Language Models ▶ *Underline Video*

Xiang Hu, Zhanchao Zhou, Ruiqi Liang, Zehuan Li, Wei Wu and Jianguo Li

This work explores efficient ultra-long context modeling. We posit that an effective solution requires three fundamental properties: sparsity, random-access flexibility, and length generalization. To achieve this, we leverage Hierarchical Sparse Attention (HSA), a novel attention mechanism that satisfies all three properties. We integrate HSA into the Transformer architecture to develop HSA-UltraLong, an 8B-parameter Mixture-of-Experts (MoE) model trained on over 8 trillion tokens. We rigorously evaluate the model across tasks with both in-domain and out-of-domain context lengths to validate its capabilities. Our model demonstrates comparable performance to full-attention baselines on in-domain sequence lengths. Crucially, it achieves over 90% accuracy on most in-context retrieval tasks with contexts up to 512 times the pre-training context length. This work reports our findings and remaining issues throughout the experiments, offering insights for future research in ultra-long context modeling.

[Long] Protecting Language Models Against Unauthorized Distillation through Trace Rewriting ▶ *Underline Video*

Xinhang Ma, William Yeoh, Ning Zhang and Yevgeniy Vorobeychik

Knowledge distillation is a widely adopted technique for transferring capabilities from LLMs to smaller, more efficient student models. However, unauthorized use of knowledge distillation takes unfair advantage of the considerable effort and cost put into developing frontier models. We investigate methods for modifying teacher-generated reasoning traces to achieve two objectives that deter unauthorized distillation: (1) anti-distillation, or degrading the training usefulness of query responses, and (2) API watermarking, which embeds verifiable signatures in student models. We introduce several approaches for dynamically rewriting a teacher's reasoning outputs while preserving answer correctness and semantic coherence. Two of these leverage the rewriting capabilities of LLMs, while others use gradient-based techniques. Our experiments show that a simple instruction-based rewriting approach achieves a strong anti-distillation effect while maintaining or even improving teacher performance. Furthermore, we show that our rewriting approach also enables embedding watermarks that can be reliably detected with essentially no false alarms. Our code is available at https://github.com/xhOwenMa/trace-rewriting.

[Long] Revisiting the Reliability of Language Models in Instruction-Following ▶ *Underline Video*

Jianshuo Dong, Yutong Zhang, Liu Yan, Zhenyu Zhong, Tao Wei, Chao Zhang and Han Qiu

Advanced LLMs have achieved near-ceiling instruction-following accuracy on benchmarks such as IFEval. However, these impressive scores do not necessarily translate to reliable services in real-world use, where users often vary their phrasing, contextual framing, and task formulations. In this paper, we study nuance-oriented reliability: whether models exhibit consistent competence across cousin prompts that convey analogous user intents but with subtle nuances. To quantify this, we introduce a new metric, reliable@k, and develop an automated pipeline that generates high-quality cousin prompts via data augmentation. Building upon this, we construct IFEval++ for systematic evaluation. Across 20 proprietary and 26 open-source LLMs, we find that current models exhibit substantial insufficiency in nuance-oriented reliability—their performance can drop by up to 61.8% with nuanced prompt modifications. What's more, we characterize it and explore three potential improvement recipes. Our findings highlight nuance-oriented reliability as a crucial yet underexplored next step toward more dependable and trustworthy LLM behavior. Our code and benchmark are accessible: https://github.com/jianshuod/IFEval-pp.

[Long] SLoRA: Balancing Plasticity and Forgetting in Large Language Models for Continual Learning ▶ *Underline Video*

Lina Yang, Yusheng Liao, Yanfeng Wang and Yu Wang
Large language models (LLMs) have achieved remarkable success across diverse tasks through large-scale pretraining. However, they remain prone to catastrophic forgetting in continual learning. To the best of our knowledge, this is the first work to identify noise accumulation in LoRA updates as a key cause of forgetting in continual learning. A preliminary two-task experiment demonstrates that removing less important components of the second tasks LoRA parameters improves performance on the first task, suggesting that later updates introduce noisy interference. Building on this insight, we propose Subspace-Denoised Low-Rank Adaptation (SLoRA), a simple and effective framework that filters noisy components from LoRA updates via subspace similarity with the base model. SLoRA is a regularization-free method without accessing data or gradients from previous tasks or modifying the training process. It offers two variants, SLoRA-Pre and SLoRA-Post, for online and offline continual learning, respectively. Extensive experiments across tasks and models validate the effectiveness of SLoRA. It improves final accuracy by up to 12%, reduces forgetting by 29%, and filters out over 30% of LoRA parameters identified as noisy. Our code is available at https://github.com/alina1031/SLoRA.

[Long] SPPO: Sequence-Level PPO for Long-Horizon Reasoning Tasks ▶ *Underline Video*
Tianyi Wang, Yixia Li, Long Li, Yibiao Chen, Shaohan Huang, Yun Chen, Peng Li, Yang Liu and Guanhua Chen
Proximal Policy Optimization (PPO) is central to aligning Large Language Models (LLMs) in reasoning tasks with verifiable rewards. However, standard token-level PPO struggles in this setting due to the instability of temporal credit assignment over long Chain-of-Thought (CoT) horizons and the prohibitive memory cost of the value model. While critic-free alternatives like GRPO mitigate these issues, they incur significant computational overhead by requiring multiple samples for baseline estimation, severely limiting training throughput. In this paper, we introduce Sequence-Level PPO (SPPO), a scalable algorithm that harmonizes the sample efficiency of PPO with the stability of outcome-based updates. SPPO reformulates the reasoning process as a Sequence-Level Contextual Bandit problem, employing a decoupled scalar value function to derive low-variance advantage signals without multi-sampling. Extensive experiments on mathematical benchmarks demonstrate that SPPO significantly surpasses standard PPO and matches the performance of computation-heavy group-based methods, offering a resource-efficient framework for aligning reasoning LLMs.

[Long] The Reasoning Trap: How Enhancing LLM Reasoning Amplifies Tool Hallucination ▶ *Underline Video*
Chenlong Yin, Zeyang Sha, shiwen cui, Changhua Meng and Zechao Li
Enhancing the reasoning capabilities of Large Language Models (LLMs) is a key strategy for building agents that "think then act". However, recent observations, like OpenAI's o3, suggest a paradox: stronger reasoning often coincides with increased hallucination, yet no prior work has systematically examined whether reasoning enhancement itself causes tool hallucination. We address this gap with the central question: Does strengthening reasoning increase tool hallucination? To answer this, we introduce SimpleToolHalluBench, a diagnostic benchmark measuring tool hallucination in two failure modes: (i) no tool available, and (ii) only distractor tools available. Through controlled experiments, we establish three key findings. First, we demonstrate a causal relationship: progressively enhancing reasoning through RL increases tool hallucination proportionally with task performance gains. Second, this effect transcends overfitting, training on non-tool tasks (e.g., mathematics) still amplifies subsequent tool hallucination. Third, the effect is method-agnostic, appearing when reasoning is instilled via supervised fine-tuning and when it is merely elicited at inference by switching from direct answers to step-by-step thinking. We also evaluate mitigation strategies including Prompt Engineering and Direct Preference Optimization (DPO), revealing a fundamental reliability-capability trade-off: reducing hallucination consistently degrades utility. Mechanistically, Reasoning RL disproportionately collapses tool-reliability-related representations, and hallucinations surface as amplified divergences concentrated in late-layer residual streams. These findings reveal that current reasoning enhancement methods inherently amplify tool hallucination, highlighting the need for new training objectives that jointly optimize for capability and reliability. Our implementation is provided at https://github.com/albert-y1n/Reasoning_Trap.

[Long] Unlocking the Potential of Diffusion Language Models through Template Infilling ▶ *Underline Video*
Junhoo Lee, Seungyeon Kim and Nojun Kwak
Diffusion Language Models (DLMs) have emerged as a promising alternative to Autoregressive Language Models, yet their inference strategies largely rely on prefix-based prompting inherited from the autoregressive paradigm. In this paper, we propose Template Infilling (TI), a conditioning methodology tailored for DLMs. Unlike conventional prefix prompting, TI distributes structural anchors across the target response, establishing a global template before infilling masked segments. This enables structured conditioning that leverages the bidirectional generation process of DLMs. We evaluate TI on diverse benchmarks, including mathematical reasoning, code generation, and trip planning, achieving consistent improvements of 9.40%p over baseline prompting strategies. Furthermore, TI naturally supports multi-token generation settings, providing practical speed advantages while maintaining generation quality and robustness. Overall, our results highlight a DLM-specific conditioning paradigm for structured generation, suggesting a promising direction for inference methods tailored to diffusion-based language models.

[Long] Writing-RL: Advancing Long-form Writing via Adaptive Curriculum Reinforcement Learning ▶ *Underline Video*
Xuanyu Lei, Chenliang Li, Yuning Wu, Kaiming Liu, Weizhou Shen, Peng Li, Ming Yan, Fei Huang, Ya-Qin Zhang and Yang Liu
Recent advances in Large Language Models (LLMs) have enabled strong performance in long-form writing, but current training paradigms remain limited: Supervised Fine-Tuning (SFT) remains constrained by data saturation and performance ceilings, while Reinforcement Learning with Verifiable Reward (RLVR), though successful in verifiable domains like math and code, cannot be directly migrated to

open-ended long-form writing due to a lack of ground-truths. To further advance long-form writing, we present Writing-RL: an Adaptive Curriculum Reinforcement Learning framework to advance long-form writing capabilities beyond SFT. The framework consists of three key components: Margin-aware Data Selection strategy that prioritizes samples with high learning potential, Pairwise Comparison Reward mechanism that provides discriminative learning signals in the absence of verifiable rewards, and Dynamic Reference Scheduling approach, which plays a critical role by adaptively adjusting task difficulty based on evolving model performance. Experiments on 7B-scale writer models show that Writing-RL effectively improves long-form writing performance over strong SFT baselines. Furthermore, we observe that models trained with long-output RL generalize surprisingly well to long-input reasoning tasks, potentially offering a promising perspective for rethinking long-context training.

Oral Session A: Multilinguality and Language Diversity 1

11:00-12:30 (Harbor A)

[Long] A Data-Efficient Path to Multilingual LLMs: Language Expansion via Post-training PARAM Integration into Upcycled MoE ▶ *Underline Video*

Hao Zhou, Tianhao Li, Zhijun Wang, Shuaijie She, Linjuan Wu, Hao-Ran Wei, Baosong Yang, Jiajun Chen and Shujian Huang

Expanding Large Language Models(LLMs) to new languages is a costly endeavor, demanding extensive Continued Pre-Training(CPT) and data-intensive alignment. While recent data-free merging techniques attempt to bypass alignment by fusing a multilingual CPT-enhanced model with its instruct counterpart, they are plagued by a critical trade-off: mitigating parameter conflicts to preserve original abilities inevitably dilutes new language acquisition, and vice-versa. To resolve this conflict, we introduce, which upcycles a dense model into a Mixture-of-Experts(MoE) architecture, allocating different experts to different languages. Alignment ability is then transferred by grafting a MoE-expanded parameter delta(_instruct) to the CPT-enhanced base model, bypassing the complex alignment phase. Experiments demonstrate 's superiority even against baselines with similar FLOPs or number of parameters; it improves performance on expanded languages while effectively preserving original capabilities. We further show our approach is highly applicable across different models and Post-training deltas.

[Long] Combining Distantly Supervised Models with In Context Learning for Monolingual and Cross-Lingual Relation Extraction ▶ *Underline Video*

Vipul Kumar Rathore, Malik Hammad Faisal, Parag Singla and Mausam Mausam

Distantly Supervised Relation Extraction (DSRE) remains a long-standing challenge in NLP, where models must learn from noisy bag-level annotations while making sentence-level predictions. While existing state-of-the-art (SoTA) DSRE models rely on task-specific training, their integration with in-context learning (ICL) using large language models (LLMs) remains underexplored. A key challenge is that the LLM may not learn relation semantics correctly, due to noisy annotation. In response, we propose HYDRE – HYbrid Distantly Supervised Relation Extraction framework. It first uses a trained DSRE model to identify the top-k candidate relations for a given test sentence, then uses a novel dynamic exemplar retrieval strategy that extracts reliable, sentence-level exemplars from training data, which are then provided in LLM prompt for outputting the final relation(s). We further extend HYDRE to cross-lingual settings for RE in low-resource languages. Using available English DSRE training data, we evaluate all methods on English as well as a newly curated benchmark covering four diverse low-resource Indic languages - Oriya, Santali, Manipuri, and Tulu. HYDRE achieves up to 20 F1 point gains in English and, on average, 17 F1 points on Indic languages over prior SoTA DSRE models and naive prompting baselines. Detailed ablations exhibit HYDRE's efficacy compared to other prompting strategies.

[Long] From Curated Data to Scalable Models: Continual Pre-training of Dense and MoE Large Language Models for Tibetan ▶ *Underline Video*

Lei Yang, Leiyu Pan, Bojian Xiong, Renren Jin, Shaowei Zhang, Yue chen, Ling Shi, Jiang Zhou, Junru Wu, Zhen Wang, Jianxiang Peng, Juesi Xiao, Tianyu Dong, Zhuowen Han, Zhuo Chen, Yuqi Ren and Deyi Xiong

Large language models (LLMs) have achieved remarkable success across a wide range of natural language processing tasks, yet their performance remains heavily biased toward high-resource languages. Tibetan, despite its cultural significance and large speaker population, is still substantially underrepresented. In this work, we present a comprehensive pipeline for advancing Tibetan language modeling through large-scale data curation and continual pre-training. We construct a 72 GB high-quality Tibetan corpus, the largest to date, and adapt Qwen2.5-7B through balanced multilingual continual pre-training with Tibetan, Chinese, and English, followed by multilingual instruction tuning. To further scale capacity efficiently, we extend the dense model to a 50B-A10B Mixture-of-Experts architecture. Due to the absence of standardized Tibetan benchmarks, we build multiple evaluation datasets via high-quality translation and human verification. Experimental results show that both dense and MoE models consistently outperform existing open-source and Tibetan-focused models of similar scale across diverse tasks. Our work advances Tibetan-centric LLM research and provides transferable insights for extending LLMs to other low-resource languages. We will release the model weights, evaluation benchmarks, and detailed data processing documentation in the follow-up.

[Long] Grammar as Control: Modular Language Generation for the Long Tail ▶ *Underline Video*
Ndapa Nakashole
Large language models (LLMs) can, in principle, bootstrap language technologies for long-tail languages due to their pattern recognition capabilities. Yet in practice, without structured guidance, they produce narrow, unrepresentative samples that fail to cover the morphosyntactic space of typologically underrepresented languages. We propose Modular Typology-Informed Generation (mTIG), a prompting framework that transforms descriptive grammars into explicit control mechanisms that guide LLMs to generate typologically balanced synthetic data for downstream training. mTIG decomposes grammars into modular grammar slices, each targeting a specific morphosyntactic phenomenon (e.g., passive voice, causative morphology). Across three low-resource languages, mTIG improves typological entropy by up to 19% and yields a "student-beats-teacher" effect, where distilled models outperform the source LLM by up to +20 chrF in machine translation. These findings show that grammar-as-control can construct training corpora wherever formal linguistic descriptions exist.

[Long] How Can Synthetic Data Improve Multilingual Language Model Pretraining? A Data Quality Perspective ▶ *Underline Video*
Tongyao Zhu, Qian Liu, Chang Ma, Jinghan Zhang, Longxu Dou, Junxian He and Shiqi Chen
Low-resource languages challenge multilingual LLMs due to limited high-quality training data, leading to weaker performance on complex reasoning and knowledge tasks. To address this, we propose improving training data quality through data synthesis, moving beyond simple resource scaling. First, we introduce SynTrans, which translates high-quality, knowledge-rich English data into low-resource languages during pre-training to inject world knowledge, though at the cost of semantic fluency. To overcome low-quality data issues while maintaining fluency, we also propose SynRank. SynRank leverages synthetic data as positive samples to train a classifier that ranks and filters noisy real-world data, enabling the extraction of high-quality subsets without expensive human cleaning. Experiments show SynRank matches handcrafted rule-based filtering by human experts and significantly improves knowledge-intensive task performance at the same filtering rate. Remarkably, higher filtering rates even improve performance with less data, demonstrating the efficiency and effectiveness of our method, surpassing expert filtering. Lastly, we introduce DA-QwenScore, a training-free metric that evaluates corpus quality by normalizing model loss with diversity measures, further enhancing evaluation efficiency. Our insights into knowledge injection could advance low-resource multilingual LLM development.

[Long] Lost in Translation, and Found: Detecting and Interpreting Translation Effects ▶ *Underline Video*
Shira Wein, Anna Serbina, Jiyuan Ji, Nathan Wolf, Jason DeGraaff, Prajakta Kini and Maria Leonor Pacheco
Translationese refers to the statistical patterns that distinguish translated texts from original texts, which are often subtle and imperceptible to human readers. When translated texts appear in either training or testing data, these patterns can negatively affect model performance or warp model evaluation. We approach the task of discerning whether a text was originally written in English or translated into English by fine-tuning contemporary foundation models at distinct item lengths and achieve state-of-the-art performance (94% Macro F1). Given that these linguistic cues are subtle and often imperceptible to humans, we analyze the features which enable our model's high performance. Employing a suite of interpretability-based techniques, we find that: (1) our high accuracy is enabled by a collection of linguistic features, a number of which correspond with linguistic theories of translationese, and (2) pretrained neural models are adept at picking up these features without any fine-tuning.

[Long] MSMO-ABSA: Multi-Scale and Multi-Objective Optimization for Cross-Lingual Aspect-Based Sentiment Analysis ▶ *Underline Video*
Chengyan Wu, Bolei Ma, Ningyuan Deng, He Yanqing, Yun Xue and Liu Xiaoyong
Aspect-based sentiment analysis (ABSA) garnered growing research interest in multilingual contexts in the past. However, the majority of the studies lack more robust feature alignment and finer aspect-level alignment. In this paper, we propose a novel framework, MSMO: Multi-Scale and Multi-Objective optimization for cross-lingual ABSA. During multi-scale alignment, we achieve cross-lingual sentence-level and aspect-level alignment, aligning features of aspect terms in different contextual environments. Specifically, we introduce code-switched bilingual sentences into the language discriminator and consistency training modules to enhance the model's robustness. During multi-objective optimization, we design two optimization objectives: supervised training and consistency training, aiming to enhance cross-lingual semantic alignment. To further improve model performance, we incorporate distilled knowledge of the target language into the model. Results show that MSMO significantly enhances cross-lingual ABSA by achieving state-of-the-art performance across multiple languages and models.

[Long] SAHM: A Benchmark for Arabic Financial and Shariah-Compliant Reasoning ▶ *Underline Video*
Rania Elbadry, Sarfraz Ahmad, Ahmed Heakl, Dani Bouch, Momina Ahsan, Muhra AlMahri, Marwa Elsaid khalil, Yuxia Wang, Salem Lahlou, Sophia Ananiadou, Veselin Stoyanov, Jimin Huang, Xueqing Peng, Preslav Nakov and Zhuohan Xie
English financial NLP has progressed rapidly through benchmarks for sentiment, document understanding, and financial question answering, while Arabic financial NLP remains comparatively under-explored despite strong practical demand for trustworthy finance and Islamic-finance assistants. We introduce SAHM, a document-grounded benchmark and instruction-tuning dataset for Arabic financial NLP and Shariah-compliant reasoning. SAHM contains 14,380 expert-verified instances spanning seven tasks: AAOIFI standards QA, fatwa-based QA/MCQ, accounting and business exams, financial sentiment analysis, extractive summarization, and event-cause reasoning, curated from authentic regulatory, juristic, and corporate sources. We evaluate 19 strong open and proprietary LLMs using task-specific metrics and rubric-based scoring for open-ended outputs, and find that Arabic fluency does not reliably translate to evidence-grounded financial reasoning: models are substantially stronger on recognition-style tasks than on generation and causal reasoning, with the largest gaps on event-cause reasoning. We release the benchmark, evaluation framework, and an instruction-tuned model to support future research

on trustworthy Arabic financial NLP.

[Long] Think Natively: Unlocking Multilingual Reasoning with Consistency-Enhanced Reinforcement Learning ▶ *Underline Video*
Xue Zhang, Yunlong Liang, Fandong Meng, Songming Zhang, Kaiyu Huang, Yufeng Chen, Xu Jinan and Jie Zhou
Large Reasoning Models (LRMs) have achieved remarkable performance on complex reasoning tasks by adopting the "think-then-answer" paradigm, which enhances both accuracy and interpretability. However, current LRMs exhibit two critical limitations when processing non-English languages: (1) They often struggle to maintain input-output language consistency; (2) They generally perform poorly with wrong reasoning paths and lower answer accuracy compared to English. These limitations significantly compromise the interpretability of reasoning processes and degrade the user experience for non-English speakers, hindering the global deployment of LRMs. To address these limitations, we propose M-Thinker, which is trained by the GRPO algorithm that involves a Language Consistency (LC) reward and a novel Cross-lingual Thinking Alignment (CTA) reward. Specifically, the LC reward defines a strict constraint on the language consistency between the input, thought, and answer. Besides, the CTA reward compares the model's non-English reasoning paths with its English reasoning path to transfer its own reasoning capability from English to non-English languages. Through an iterative RL procedure, our M-Thinker-1.5B/4B/7B models not only achieve nearly 100% language consistency and superior performance on two multilingual benchmarks (MMATH and PolyMath), but also exhibit excellent generalization on out-of-domain languages.

Oral Session A: NLP Applications 1

11:00-12:30 (Harbor E-F)

[Long] Beyond the Final Actor: Modeling the Dual Roles of Creator and Editor for Fine-Grained LLM-Generated Text Detection ▶ *Underline Video*
Yang Li, Qiang Sheng, Zhengjia Wang, Yehan Yang, Danding Wang and Juan Cao
The misuse of large language models (LLMs) requires precise detection of synthetic text. Existing works mainly follow binary or ternary classification settings, which can only distinguish pure human/LLM text or collaborative text at best. This remains insufficient for the nuanced regulation, as the LLM-polished human text and humanized LLM text often trigger different policy consequences. In this paper, we explore fine-grained LLM-generated text detection under a rigorous four-class setting. To handle such complexities, we propose RACE (Rhetorical Analysis for Creator-Editor Modeling), a fine-grained detection method that characterizes the distinct signatures of creator and editor. Specifically, RACE utilizes Rhetorical Structure Theory (RST) to construct a logic graph for the creator's foundation while extracting Elementary Discourse Unit (EDU)-level features for the editor's style. Experiments show that RACE outperforms 12 baselines in identifying fine-grained types with low false alarms, offering a policy-aligned solution for LLM regulation.

[Long] CHIMERA: A Knowledge Base of Scientific Idea Recombinations for Research Analysis and Ideation ▶ *Underline Video*
Noy Sternlicht and Tom Hope
A hallmark of human innovation is recombination—the creation of novel ideas by integrating elements from existing concepts and mechanisms. In this work, we introduce CHIMERA, the first large-scale Knowledge Base (KB) of recombination examples automatically mined from the scientific literature. CHIMERA enables empirical analysis of how scientists recombine concepts and draw inspiration from different areas, and enables training models that propose cross-disciplinary research directions. To construct this KB, we define a new information extraction task: identifying recombination instances in papers. We curate an expert-annotated dataset and use it to fine-tune an LLM-based extraction model, which we apply to a broad corpus of AI papers. We also demonstrate generalization to a biological domain. We showcase the utility of CHIMERA through two applications. First, we analyze patterns of recombination across AI subfields. Second, we train a scientific hypothesis generation model using the KB, showing that it can propose directions that researchers rate as inspiring.

[Long] Double: Breaking the Acceleration Limit via Double Retrieval Speculative Parallelism ▶ *Underline Video*
Yuhao Shen, Tianyu Liu, Junyi Shen, Jinyang Wu, Quan Kong, Huan Li and Cong Wang
Parallel Speculative Decoding (PSD) accelerates traditional Speculative Decoding (SD) by overlapping draft generation with verification. However, it remains hampered by two fundamental challenges: (1) a theoretical speedup ceiling dictated by the speed ratio between the draft and target models, and (2) high computational waste and pipeline stall due to mid-sequence token rejections of early errors. To address these limitations, we introduce Double (Double Retrieval Speculative Parallelism). By bridging the gap between SD and PSD, our framework resolves the Retrieval Precision-Efficiency Dilemma through a novel synchronous mechanism. Specifically, we enable the draft model to execute iterative retrieval speculations to break the theoretical speedup limits; to alleviate rejections without rollback, the target model performs authoritative retrieval to generate multi-token guidance. Double is entirely training-free and lossless. Extensive experiments demonstrate state-of-the-art speedup of 5.3x on LLaMA3.3-70B and 2.8x on Qwen3-32B, significantly outperforming the advanced method EAGLE-3 that requires extensive model training. Our code is available at https://github.com/Sylvan820/Double1.

[Long] Focusing Condition: Inference-Time Self-Contrastive Steering Elicits Better Conditional Text Embeddings in LLMs ▶ *Underline Video*

Zifeng Cheng, Lingyun Qian, Zhiwei Jiang, Cong Wang, Yafeng Yin, Fei Shen, Ao Zhou and Qing Gu
Extracting conditional text embeddings from large language models (LLMs) is a promising paradigm, as it requires neither additional data nor fine-tuning. Existing methods incorporate conditions into prompts to guide LLMs to focus on specific aspects and elicit conditional text embeddings. However, relying solely on prompts often fails to produce high-quality conditional text embeddings, as they remain entangled with general text embeddings, ultimately degrading their quality. To this end, we propose an inference-time, plug-and-play Self-Contrastive Steering (SCS) method that constructs unconditional general text embeddings and uses them to refine conditional text embeddings, making them more focused on the target condition. Specifically, we modify the attention mask and positional encodings to mask the condition, thereby obtaining unconditional text embeddings and intervening in the multi-head self-attention computation process. Notably, our method is highly efficient, requiring only a single additional multi-head self-attention computation at inference time. Extensive experiments on clustering, Semantic Textual Similarity, and triplet alignment datasets demonstrate that our method can seamlessly improve the performance of existing prompt-based methods across different LLMs in a training-free and plug-and-play manner.

[Long] GATE: Graph-based Adaptive Tool Evolution Across Diverse Tasks ▶ *Underline Video*
Jianwen Luo, Yiming Huang, Jinxiang Meng, Fangyu Lei, Shizhu He, Xiao Liu, Shanshan Jiang, Bin Dong, Jun Zhao and Kang Liu
Large Language Models (LLMs) have shown great promise in tool-making, yet existing frameworks often struggle to efficiently construct reliable toolsets and are limited to single-task settings. To address these challenges, we propose GATE (Graph-based Adaptive Tool Evolution), an adaptive framework that dynamically constructs and evolves a hierarchical graph of reusable tools across multiple scenarios. We evaluate GATE on open-ended tasks (Minecraft), agent-based tasks (TextCraft, DABench), and code generation tasks (MATH, Date, TabMWP). Our results show that GATE achieves up to 4.3x faster milestone completion in Minecraft compared to the previous state-of-the-art method, and provides an average improvement of 9.23% over existing tool-making methods in code generation tasks and 10.03% in agent tasks. Further analysis shows that GATE exhibits strong adaptive evolution capabilities, effectively balancing tool quantity, complexity, and functionality while maintaining high efficiency. Code and data are available at https://github.com/ayanami2003/GATE.

[Long] PUPPET: Neural-Symbolic Standardized Patients for Mental Health ▶ *Underline Video*
Chen Xu, Yu ji, Zhenyu Lv, Yang Yi, Yizhe Yang, Luyao Ji, Chaoyi Chen, Xianyang Wang, Tian Lan, Zhihua Wang, Juan Wang, Xunde Dong, Fuze Tian, Qunxi Dong and Bin Hu
The critical therapist shortage demands scalable training solutions. Standardized Patients, the gold standard, are scarce and costly. Current LLM-based approaches focus on patient simulation for conversational realism but lack pedagogical rigor as Virtual Standardized Patients, lacking faithful reactions to clinical errors and explainable feedback. To bridge this gap, we propose PUPPET, the first neural-symbolic Virtual Standardized Patient governed by an OBSERVE-THINK-BEHAVE architecture. PUPPET externalizes LLM reasoning into a symbolic system where experts implant causal associations between intervention logic (propositional logic) and patient mental states (state machine). This allows PUPPET to behave coherently with controllable and explainable psychological dynamics: intervention logic (OBSERVE) to state transition (THINK) to response (BEHAVE). Our PUPPET-TRAINER further leverages this chain to educate trainees about intervention consequences, standardizing and scaling mental health training. Experiments across three clinical scenarios confirm that PUPPET outperforms baselines in clinical faithfulness and pedagogical value.

[Long] Saber: Efficient Sampling with Adaptive Acceleration and Backtracking Enhanced Remasking for Diffusion Language Model in Code Generation ▶ *Underline Video*
Yihong Dong, Zhaoyu Ma, Xue Jiang, Zhiyuan Fan, Jiaru Qian, Yongmin Li, Jianha Xiao, Zhi Jin and Ge Li
Diffusion language models (DLMs) are emerging as a compelling alternative to the dominant autoregressive paradigm, offering inherent advantages in parallel generation and bidirectional context modeling. However, for the tasks with strict structural constraints such as code generation, DLMs face a critical trade-off between inference speed and output quality, where accelerating generation by reducing sampling steps often leads to catastrophic performance collapse. We find that the fundamental reasons are: 1) the generation difficulty is uneven in the structured sequence decoding steps, making DLM's static acceleration strategy suboptimal; 2) the context of tokens generated by DLM evolves continuously, causing early high-confidence predictions to turn into irreversible errors. In this paper, we introduce efficient Sampling with Adaptive acceleration and Backtracking Enhanced Remasking (i.e., Saber), a novel training-free sampling algorithm for DLMs that the first to improve both inference speed and output quality in code generation. Saber dynamically adjusts the number of tokens unmasked per step based on the model's evolving confidence, and utilizes a backtracking mechanism to revert tokens whose confidence drops as new context emerges, with its effectiveness supported by theoretical analysis. Extensive experiments on multiple mainstream code generation benchmarks show that Saber boosts Pass@1 accuracy by an average of 1.9% over mainstream DLM sampling methods, while achieving an average 251.4% inference speedup. By leveraging the inherent advantages of DLMs, our work significantly narrows the performance gap with autoregressive models in code generation.

[Long] Table Question Answering in the Era of Large Language Models: A Comprehensive Survey of Tasks, Methods, and Evaluation ▶ *Underline Video*
Wei Zhou, Bolei Ma, Annemarie Friedrich and Mohsen Mesgar
Table Question Answering (TQA) aims to answer natural language questions using tabular data, often accompanied by additional contexts, such as passages. The task spans diverse settings, varying in table representation, question/answer complexity, modality involved, and domain. While recent advances in large language models (LLMs) have driven substantial progress for TQA, the field still lacks a systematic

organization and understanding of task formulations, core challenges, and methodological trends, particularly in light of emerging research directions, such as reinforcement learning. This survey addresses this gap by providing a comprehensive and structured overview of TQA research using LLMs. We introduce various task setups and summarize available benchmarks based on task features. We categorize current modeling strategies according to the challenges they target, and analyze their strengths and limitations. Furthermore, we highlight underexplored but timely topics that have not been systematically covered in prior surveys. By unifying disparate research threads and identifying open problems, our survey offers a consolidated foundation for the TQA community, enabling a deeper understanding of the state of the art and guiding future developments in this rapidly evolving area.

[Long] VeriTaS: The First Dynamic Benchmark for Multimodal Automated Fact-Checking ▶ *Underline Video*

Mark Rothermel, Marcus Kornmann, Marcus Rohrbach and Anna Rohrbach

The growing scale of online misinformation urgently demands Automated Fact-Checking (AFC). Existing benchmarks for evaluating AFC systems, however, are largely limited in terms of task scope, modalities, domain, language diversity, realism, or coverage of misinformation types. Critically, they are static, thus subject to data leakage as their claims enter the pretraining corpora of LLMs. As a result, benchmark performance no longer reliably reflects the actual ability to verify claims. We introduce Verified Theses and Statements (VeriTaS), the first dynamic benchmark for multimodal AFC, designed to remain robust under ongoing large-scale pretraining of foundation models. VeriTaS currently comprises 25,000 real-world claims from 104 professional fact-checking organizations across 54 languages, covering textual and audiovisual content. Claims are added quarterly via a fully automated seven-stage pipeline that normalizes claim formulation, retrieves original media, and maps heterogeneous expert verdicts to a novel, standardized, and disentangled scoring scheme with textual justifications. Through human evaluation, we demonstrate that the automated annotations closely match human judgments. We commit to updating VeriTaS in the future, establishing a leakage-resistant benchmark, supporting meaningful AFC evaluation in the era of rapidly evolving foundation models. The code and data are publicly available under https://veritas.mai.informatik.tu-darmstadt.de.

Oral Session A: Resources and Evaluation 1

11:00-12:30 (Harbor B-C)

[Long] Afri-MCQA: Multimodal Cultural Question Answering for African Languages ▶ *Underline Video*

Atnafu Lambebo TonjaSrija Anand, Emilio Villa-Cueva, Israel Abebe Azime, Jesujoba Oluwadara Alabi, Muhidin A Mohamed, Debela Desalegn YadetaNegasi Haile Abadi, Abigail OppongNnaemeka Casmir Obiefuna, Idris Abdulmumin, Naome A Etori, Eric Peter Wairagala, Kanda Patrick Tshinu, Imanigirimbabazi Emmanuel, Gabofetswe Malema, Alham Fikri Aji, David Ifeoluwa Adelani and Thamar Solorio

Africa is home to over one-third of the world's languages, yet remains severely underrepresented in multimodal AI research. We introduce Afri-MCQA, the first Multilingual Cultural Question-Answering benchmark containing 7.5k Q&A pairs across 15 African languages from 12 countries. The benchmark offers parallel text and speech modalities and was entirely created by native speakers. We find that models show poor performance across evaluated cultures, with near-zero accuracy on open-ended VQA when queried through native language or speech. To test linguistic competence, we include control experiments meant to assess this specific aspect separate from cultural knowledge, and we observe significant performance gaps between native languages and English for both text and speech. These findings underscore the pressing need for speech-first approaches, culturally grounded pretraining, and cross-lingual cultural transfer. We release Afri-MCQA to support more inclusive multimodal AI development.

[Long] ChiKhaPo: A Large-Scale Multilingual Benchmark for Evaluating Lexical Comprehension and Generation in Large Language Models ▶ *Underline Video*

Emily Chang and Niyati Bafna

Existing benchmarks for large language models (LLMs) are largely restricted to high- or mid-resource languages, and often evaluate performance on higher-order tasks in reasoning and generation. However, plenty of evidence points to the fact that LLMs lack basic linguistic competence in the vast majority of the worlds 3800+ written languages. We introduce ChiKhaPo, consisting of eight subtasks of varying difficulty designed to evaluate the lexical comprehension and generation abilities of generative models. ChiKhaPo draws on existing lexicons, monolingual data, and bitext, and provides coverage for 2700+ languages for two word-translation-based subtasks, surpassing any existing benchmark in terms of language coverage. We further show that six SOTA models struggle on our benchmark, and discuss the factors contributing to performance scores, including language family, language resourcedness, task, and comprehension versus generation directions. With ChiKhaPo, we hope to enable and encourage the massively multilingual benchmarking of LLMs.

[Long] CogToM: A Comprehensive Theory of Mind Benchmark inspired by Human Cognition for Large Language Models ▶ *Underline Video*

Haibo Tong, Zeyang Yue, Feifei Zhao, Erliang Lin, Lu Jia, Ruolin Chen, Yinqian Sun, Qian Zhang and Yi Zeng

Whether Large Language Models (LLMs) truly possess human-like Theory of Mind (ToM) capabilities has garnered increasing attention. However, existing benchmarks remain largely restricted to narrow paradigms like false belief tasks, failing to capture the full spectrum of human cognitive mechanisms. We introduce CogToM, a comprehensive, theoretically grounded benchmark comprising over 8000

bilingual instances across 46 paradigms, validated by 49 human annotators. A systematic evaluation of 22 representative models, including frontier models like GPT-5.1 and Qwen3-Max, reveals significant performance heterogeneities and highlights persistent bottlenecks in specific dimensions. Further analysis based on human cognitive patterns suggests potential divergences between LLM and human cognitive structures. CogToM offers a robust instrument and perspective for investigating the evolving cognitive boundaries of LLMs. We release our code and data at https://github.com/Beijing-AISI/CogToM.

[Long] DeepFact: Co-Evolving Benchmarks and Agents for Deep Research Factuality ▶ *Underline Video*

Yukun Huang, Leonardo F. R. Ribeiro, Momchil Hardalov, Bhuwan Dhingra, Markus Dreyer and Venkatesh Saligrama

Search-augmented LLM agents can produce deep research reports (DRRs), but verifying claim-level factuality remains challenging. Existing fact-checkers usually target general-domain atomic claims, and there is no benchmark to test whether such verifiers transfer to DRRs. Yet building such a benchmark for DRR fact-checkers is itself difficult because it requires expert judgments over cognitively demanding, domain-specific claims. In a controlled study with PhD-level specialists, unassisted experts achieve only 60.8% accuracy on hidden known-answer claims. We therefore propose evolving benchmarking via Audit-then-Score (AtS), in which labels and rationales remain revisable: when a verifier disagrees with the current benchmark, it submits evidence; an auditor adjudicates the dispute; and accepted revisions update the benchmark before scoring. After three additional AtS rounds, expert accuracy rises to 90.9%, showing that experts are better auditors than one-shot labelers. We instantiate AtS as DeepFactBench, a versioned DRR factuality benchmark with auditable rationales, and introduce DeepFactEval, a claim-level verifier. On the frozen DeepFactBench release, DeepFactEval achieves 83.4% accuracy, outperforming the best prior deep-research and traditional fact-checkers by 14.3 and 24.9 points, respectively, and transferring well to external factuality datasets.

[Long] Fine-tuning vs. In-context Learning in Large Language Models: A Formal Language Learning Perspective ▶ *Underline Video*

Bishwamittra Ghosh, Soumi Das, Till Speicher, Qinyuan Wu, Mohammad Aflah Khan, Deepak Garg, Krishna P. Gummadi and Evimaria Terzi

Large language models (LLMs) operate in two fundamental learning modes - fine-tuning (FT) and in-context learning (ICL) - raising key questions about which mode yields greater language proficiency and whether they differ in their inductive biases. Prior studies comparing FT and ICL have yielded mixed and inconclusive results due to inconsistent experimental setups. To enable a rigorous comparison, we propose a formal language learning task - offering precise language boundaries, controlled string sampling, and no data contamination - and introduce a discriminative test for language proficiency, where an LLM succeeds if it assigns higher generation probability to in-language strings than to out-of-language strings. Empirically, we find that: (a) FT has greater language proficiency than ICL on in-distribution generalization, but both perform equally well on out-of-distribution generalization. (b) Their inductive biases, measured by the correlation in string generation probabilities, are similar when both modes partially learn the language but diverge at higher proficiency levels. (c) Unlike FT, ICL performance differs substantially across models of varying sizes and families and is sensitive to the token vocabulary of the language. Thus, our work demonstrates the promise of formal languages as a controlled testbed for evaluating LLMs, behaviors that are difficult to isolate in natural language datasets. Our source code is available at https://github.com/bishwamittra/formallm.

[Long] HalluCitation Matters: Revealing the Impact of Hallucinated References with 300 Hallucinated Papers in ACL Conferences ▶ *Underline Video*

Yusuke Sakai, Hidetaka Kamigaito and Taro Watanabe

Recently, we have often observed hallucinated citations or references that do not correspond to any existing work in papers under review, preprints, or published papers. Such hallucinated citations pose a serious concern to scientific reliability. When they appear in accepted papers, they may also negatively affect the credibility of conferences. In this study, we refer to hallucinated citations as “HalluCitation” and systematically investigate their prevalence and impact. We analyze all papers published at ACL, NAACL, and EMNLP in 2024 and 2025, including main conference, Findings, and workshop papers. Our analysis reveals that nearly 300 papers contain at least one HalluCitation, most of which were published in 2025. Notably, half of these papers were identified at EMNLP 2025, the most recent conference, indicating that this issue is rapidly increasing. Moreover, more than 100 such papers were accepted as main conference and Findings papers at EMNLP 2025, affecting the credibility.

[Long] ImplicitMemBench: Measuring Unconscious Behavioral Adaptation in Large Language Models ▶ *Underline Video*

Chonghan Qin, Xiachong Feng, Weitao Ma, Xiaocheng Feng and Lingpeng Kong

Existing memory benchmarks for LLM agents evaluate explicit recall of facts, yet overlook implicit memory where experience becomes automated behavior without conscious retrieval. This gap is critical: effective assistants must automatically apply learned procedures or avoid failed actions without explicit reminders. We introduce ImplicitMemBench, the first systematic benchmark evaluating implicit memory through three cognitively grounded constructs drawn from standard cognitive-science accounts of non-declarative memory: Procedural Memory (one-shot skill acquisition after interference), Priming (theme-driven bias via paired experimental/control instances), and Classical Conditioning (Conditioned Stimulus–Unconditioned Stimulus (CS–US) associations shaping first decisions). Our 300-item suite employs a unified Learning/Priming-Interfere-Test protocol with first-attempt scoring. Evaluation of 17 models reveals severe limitations: no model exceeds 66% overall, with top performers DeepSeek-R1 (65.3%), Qwen3-32B (64.1%), and GPT-5 (63.0%) far below human baselines. Analysis uncovers dramatic asymmetries (inhibition 17.6% vs. preference 75.0%) and universal bottlenecks requiring architectural innovations beyond parameter scaling. ImplicitMemBench reframes evaluation from "what agents recall" to "what they automatically enact".

[Long] Synthetic Data Generation for Training Diversified Commonsense Reasoning Models ▶ *Underline Video*
Tianhui Zhang, Bei Peng and Danushka Bollegala
Conversational agents are required to respond to their users not only with high quality (i.e. commonsense bearing) responses, but also considering multiple plausible alternative scenarios, reflecting the diversity in their responses. Despite the growing need to train diverse commonsense generators, the progress of this line of work has been significantly hindered by the lack of large-scale high-quality diverse commonsense training datasets. Due to the high annotation costs, existing Generative Commonsense Reasoning (GCR) datasets are created using a small number of human annotators, covering only a narrow set of commonsense scenarios. To address this training resource gap, we propose a two-stage method to create the first-ever synthetic dataset CommonSyn for diversified (GCR). The model fine-tuned on our synthetic data jointly increase both generation diversity and quality compared with vanilla models and the model fine-tuned on human-crafted dataset across different size Large Language Models (LLMs)

[Long] Uni-MMMU: A Massive Multi-discipline Multimodal Unified Benchmark ▶ *Underline Video*
Kai Zou, Ziqi Huang, Yuhao Dong, Shulin Tian, Dian Zheng, Hongbo Liu, Jingwen He, Bin Liu, Yu Qiao and Ziwei Liu
Unified multimodal models aim to jointly enable visual understanding and generation, yet current benchmarks rarely examine their true integration. Existing evaluations either treat the two abilities in isolation or overlook tasks that inherently couple them. To address this gap, we present Uni-MMMU, a comprehensive and discipline-aware benchmark that systematically unfolds the bidirectional synergy between generation and understanding across eight reasoning-centric domains, including science, coding, mathematics, and puzzles. Each task is bidirectionally coupled, demanding models to (i) leverage conceptual understanding to guide precise visual synthesis, or (ii) utilize generation as a cognitive scaffold for analytical reasoning. Uni-MMMU incorporates verifiable intermediate reasoning steps, unique ground truths, and a reproducible scoring protocol for both textual and visual outputs. Through extensive evaluation of state-of-the-art unified, generation-only, and understanding-only models, we reveal substantial performance disparities and cross-modal dependencies, offering new insights into when and how these abilities reinforce one another, and establishing a reliable foundation for advancing unified models.

Oral Session A: Student Research Workshop 1

11:00-12:30 (Gas Lamp A-B)

[SRW] BANGLASOCIALBENCH: A Benchmark for Evaluating Sociopragmatic and Cultural Alignment of LLMs in Bangladeshi Social Interaction ▶ *Underline Video*
Tanvir Ahmed Sijan, S. M Golam Rifat, Pankaj Chowdhury Partha, Md. Tanjeed Islam and Md. Musfique Anwar
Large Language Models have demonstrated strong multilingual fluency, yet fluency alone does not guarantee socially appropriate language use. In high-context languages, communicative competence requires sensitivity to social hierarchy, relational roles, and interactional norms that are encoded directly in everyday language. Bangla exemplifies this challenge through its three-tiered pronominal system, kinship-based addressing, and culturally embedded social customs. We introduce BANGLASOCIALBENCH, the first benchmark designed to evaluate sociopragmatic competence in Bangla through context-dependent language use rather than factual recall. The benchmark spans three domains: Bangla Address Terms, Kinship Reasoning, and Social Customs, and consists of 1,719 culturally grounded instances written and verified by native Bangla speakers. We evaluate twelve contemporary LLMs in a zero-shot setting and observe systematic patterns of cultural misalignment. Models frequently default to overly formal address forms, fail to recognize multiple socially acceptable address pronouns, and conflate kinship terminology across religious contexts. Our findings show that sociopragmatic failures are often structured and non-random, revealing persistent limitations in how current LLMs infer and apply culturally appropriate language use in realistic Bangladeshi social interactions.

[SRW] Disentangling the Effects of Unlearning in Measuring Parametric Faithfulness of Chain-of-Thought ▶ *Underline Video*
Ryo Mitsuhashi; Gaku Morio; Ayana Niwa; Masahiro Kaneko; Kentaro Inui; Terufumi Morishita; Yuta Koreeda; Yasuhiro Sogawa;
Chain-of-Thought (CoT) in large language models (LLMs) has been widely debated in terms of whether it faithfully reflects an internal reasoning process of models. Parametric faithfulness is a recently proposed metric that uses unlearning to assess whether a model encodes parametric beliefs corresponding to a reasoning chain. This paper refines this metric by accounting for the unintended artifacts of unlearning. We introduce control tasks that unlearn irrelevant knowledge and word-shuffled content and show that these control tasks can measure parametric faithfulness, suggesting the non-negligible effect of unlearning. We also found that control tasks help explain the significant variations in parametric faithfulness observed across different model sizes and CoT lengths. We conclude that the effects of unlearning should be considered when measuring parametric faithfulness.

[SRW] Evaluation of Multilingual Ability to Use Spatial Deictic Expressions in Vision-Language Models ▶ *Underline Video*
Kaito Watanabe; Taisei Yamamoto; Tomoki Doi; Hitomi Yanaka;
One of the expected abilities of vision-language models (VLMs) is spatial reasoning ability based on a given text and image. To evaluate the spatial reasoning abilities of VLMs, we focus on the use of spatial deictic expressions, which are defined as spatial expressions whose referent is determined by their situational context, such as this and that. To handle spatial deictic expressions, VLMs must jointly reason over

language and visual space, grounding context-dependent references in the image's spatial structure. In addition, selecting appropriate spatial deictic expressions across languages requires VLMs to understand the language-specific spatial distinctions encoded by these expressions. In this paper, we develop a benchmark to evaluate the multilingual ability of VLMs to use spatial deictic expressions in four languages. Our experiments using this benchmark reveal that the tested models use demonstratives in a manner different from that of humans, particularly in selecting the appropriate demonstratives based on the distance from the object.

[SRW] LLM-based Literal Example Generation for Japanese Multiword Expressions ▶ ***Underline Video***
Mio Ohashi; Hajime Kiyama; Zhidong Ling; Mamoru Komachi
We investigate whether large language models (LLMs) can generate literal usage examples for Japanese multiword expressions (MWEs), whose literal readings are structurally low-frequency in available corpora. Prior work on MWEs has largely focused on detecting idiomatic usages in context, leaving literal usages underrepresented particularly for Japanese MWEs whose literal readings are rare and structurally diverse. Because literal readings are rarely attested in corpora, we design a lexicon-grounded setup that uses corpus non-literal usages as contrastive cues for controlled prompting. We evaluate the generated sentences using automatic literalness judgments and human literalness judgments, together with manual inspection. Our results show that providing contrastive non-literal information stabilizes literal generation and improves quality compared with prompts that include only literal information or no hints. In addition, we conduct an LLM-based understanding test that compares model predictions of literal and idiomatic plausibility with human judgments. The results indicate that the model aligns more closely with human judgments for idiomatic interpretations than for literal ones, highlighting the relative difficulty of modeling literal readings of MWEs. The study demonstrates that LLMs can complement existing resources by supplying frequency-independent literal examples and offers a controlled framework for examining contextual meaning understanding of Japanese MWEs.

[SRW] Neural KWIC: Inducing Contextualized Word Embeddings from KWIC Concordance Examples ▶ ***Underline Video***
Mao Shimada; Hajime Kiyama; Zhidong Ling; Mamoru Komachi; Toshinobu Ogiso; Hiroya Takamura; Daichi Mochihashi
KWIC (Keyword-in-Context) retrieves a target keyword with a fixed-length local context window and is widely used in corpus-based linguistic analysis. However, interpreting large collections of KWIC instances remains challenging. To address this issue, we propose Neural KWIC, a method that induces contextualized word embeddings from KWIC-based local contexts and enables their visualization in embedding space. The method learns a transformation function to approximate contextualized embeddings produced by a pretrained language model, thereby distilling representational knowledge acquired through large-scale pretraining. Once trained, inference requires only a single mapping, enabling efficient large-scale computation. Quantitative and qualitative analyses demonstrate that Neural KWIC achieves competitive alignment with the pretrained language model while substantially reducing inference cost and facilitating interpretable large-scale analysis of KWIC instances.

[SRW] The Silence of the Facts: Popularity as a Barrier to Machine Unlearning ▶ ***Underline Video***
Anna Borisiuk; Andrey Savchenko; Alexander Panchenko; Elena Tutubalina
Machine Unlearning is a valuable ability of LLMs, enabling the removal of unsafe, outdated, or private information. Existing unlearning methods, however, are often evaluated under the assumption that all facts are equally challenging to forget. Controllable knowledge removal is essential for reliable NLP systems. In this paper, we investigate whether fact popularity influences the efficiency of LLM unlearning. To answer this question, we build UNLamb benchmark designed to systematically investigate this relationship. It consists of 11.6k question-answer pairs derived from real-world knowledge in Wikidata, explicitly partitioned into rare and popular facts. Using this benchmark, we perform a comprehensive evaluation of state-of-the-art unlearning algorithms on a set of models of different sizes. We conduct a comprehensive analysis of four unlearning methods across three validation sets and two LLMs. We show that larger models struggle more to forget popular entities, often damaging related knowledge in the process. In contrast, it is much easier to remove rare facts without side effects.

Poster Session A (Coronado)

11:00-12:30 (Coronado)

[Findings] Affectron: Emotional Speech Synthesis with Affective and Contextually Aligned Nonverbal Vocalizations ▶ ***Underline Video***
Deok-Hyeon Cho, Hyung-Seok Oh, Seung-Bin Kim and Seong-Whan Lee
Nonverbal vocalizations (NVs), such as laughter and sighs, are central to the expression of affective cues in emotional speech synthesis. However, learning diverse and contextually aligned NVs remains challenging in open settings due to limited NV data and the lack of explicit supervision. Motivated by this challenge, we propose Affectron as a framework for affective and contextually aligned NV generation. Built on a small-scale open and decoupled corpus, Affectron introduces an NV-augmented training strategy that expands the distribution of NV types and insertion locations. We further incorporate NV structural masking into a speech backbone pre-trained on purely verbal speech to enable diverse and natural NV synthesis. Experimental results demonstrate that Affectron produces more expressive and diverse NVs than baseline systems while preserving the naturalness of the verbal speech stream.

[Findings] AI, Take the Wheel: What Drives Delegation and Trust in Human-Computer Cooperative Question Answering? ▶ *Underline Video*

Maharshi Gor, Yoo Yeon Sung, Yu Hou, Eve Fleisig, Zhu Irene Ying, Tianyi Zhou and Jordan Lee Boyd-Graber

AI systems are fallible, and humans can make mistakes in deciding whether to trust AI over their own judgment. Thus, improving human-AI collaboration requires that we understand when, why, and how humans decide to rely on AI. We study two reliance decisions: delegating a task to AI without seeing its output (whether AI is used) and evaluating AI suggestions to decide whether to adopt them how AI output shapes final decisions). Both matter for effective collaboration, yet prior work lacks naturalistic experiments capturing both patterns for the same users. We address this gap by studying collaborative human–AI teams competing in a question-answering game in which humans can choose when and how to work with AI agents to win. Our 24 matches pair 23 expert humans with 16 AI agents, capturing 387 delegation and 1440 adoption decisions. While human–AI collaboration performs better than either AI or humans alone, humans make suboptimal collaboration decisions, both under-relying on correct AI suggestions (3.7% of opportunities missed) and over-relying when AI misleads them (1.5%). Both parties contribute wrong answers: reported model confidence is near chance when humans and AI disagree, while confirmation bias drives higher under-reliance (60.7%) when an AI suggestion agrees with humans' initial incorrect answer.

[Long] An Existence Proof for Neural Language Models That Can Explain Garden-Path Effects via Surprisal ▶ *Underline Video*

Ryo Yoshida, Shinnosuke Isono, Taiga Someya, Yohei Oseki and Tatsuki Kuribayashi

Surprisal theory hypothesizes that the difficulty of human sentence processing increases linearly with surprisal, the negative log-probability of a word given its context. Computational psycholinguistics has tested this hypothesis using language models (LMs) as proxies for human prediction. While surprisal derived from recent neural LMs generally captures human processing difficulty on naturalistic corpora that predominantly consist of simple sentences, it severely underestimates processing difficulty on sentences that require syntactic disambiguation (garden-path effects). This leads to the claim that the processing difficulty of such sentences cannot be reduced to surprisal, although it remains possible that neural LMs simply differ from humans in next-word prediction. In this paper, we investigate whether it is truly impossible to construct a neural LM that can explain garden-path effects via surprisal. Specifically, instead of evaluating off-the-shelf neural LMs, we fine-tune these LMs on garden-path sentences so as to better align surprisal-based reading-time estimates with actual human reading times. Our results show that fine-tuned LMs do not overfit and successfully capture human reading slowdowns on held-out garden-path items; they even improve predictive power for human reading times on naturalistic corpora and preserve their general LM capabilities. These results provide an existence proof for a neural LM that can explain both garden-path effects and naturalistic reading times via surprisal, but also raise a theoretical question: what kind of evidence can truly falsify surprisal theory?

[Long] An Information-Theoretic Foundation for the Subregular Hierarchy ▶ *Underline Video*

Hung Mai, Khanh Nguyen, Luong Doan, Ngan Duong, Thao Duong and Tuan Do

The Subregular Hypothesis posits that phonological patterns in natural languages occupy a restricted region of the formal language hierarchy, yet the cognitive basis for this restriction remains unclear. We propose an information-theoretic characterization: Strictly Local languages, when formalized as shifts of finite type, are exactly those admitting stationary Markov sources, which exhibit zero conditional mutual information between distant positions given intervening symbols. We prove that certain non-subregular patterns such as first-last assimilation admit no such Markov realization, explaining their unlearnability. Empirical validation on English phonotactics versus Finnish, Turkish, and Hungarian vowel harmony confirms that MI profiles statistically distinguish SL-like from TSL-like patterns ($p < 0.001$, $r = 0.84$). This work bridges formal language theory and information theory, offering a unified framework for understanding computational restrictions on natural language phonology.

[Findings] Arab Voices: Mapping Standard and Dialectal Arabic Speech Technology ▶ *Underline Video*

Peter Sullivan, AbdelRahim A. Elmadany, Alcides Alcoba Inciarte and Muhammad Abdul-Mageed

Dialectal Arabic datasets embody a range of domain, dialect, and quality. To better understand the landscape of these datasets, we perform a computational analysis of the 'dialectness' and a set of measures of audio quality. This analysis of the training splits of dialectal Arabic datasets, provides a valuable complement to existing literature surveys of dialectal Arabic. To further address inconsistencies between datasets, we also introduce Arab Voices, a standardized framework for supporting Automatic Speech Recognition in dialectal Arabic. This framework provide access to 31 datasets covering 14 dialects, to better address the limited data availability encountered in dialectal Arabic speech processing. Our benchmark further provides a current evaluation of SOTA tools as well as modern multimodal LLMs at dialectal Arabic ASR.

[Findings] ArgBench: Benchmarking LLMs on Computational Argumentation Tasks ▶ *Underline Video*

Yamen Ajjour, Carlotta Quensel, Nedim Lipka and Henning Wachsmuth

Argumentation skills are an essential toolkit for large language models (LLMs). These skills are crucial in various use cases, including self-reflection, debating collaboratively for diverse answers, and countering hate speech. In this paper, we create the first benchmark for a standardized evaluation of LLM-based approaches to computational argumentation, encompassing 33 datasets from previous work in unified form. Using the benchmark, we evaluate the generalizability of five LLM families across 46 computational argumentation tasks that cover mining arguments, assessing perspectives, assessing argument quality, reasoning about arguments, and generating arguments. On the benchmark, we conduct an extensive systematic analysis of the contribution of few-shot examples, reasoning steps, model size, and training skills to the performance of LLMs on the computational argumentation tasks in the benchmark.

[Long] ATIR: Towards Audio-Text Interleaved Contextual Retrieval ▶ *Underline Video*

Tong Zhao, Chenghao Zhang, Yutao Zhu and Zhicheng Dou

Audio carries richer information than text, including emotion, speaker traits, and environmental context, while also enabling lower-latency processing compared to speech-to-text pipelines. However, recent multimodal information retrieval research has predominantly focused on images, largely overlooking audio, especially in the setting of interleaved audio-text contextual retrieval. In this work, we introduce the Audio-Text Interleaved contextual Retrieval (ATIR) task, where queries can alternate between audio and text modalities. We construct an ATIR benchmark by integrating several Automatic Speech Recognition (ASR), QA, and retrieval datasets, ultimately unifying four types of contextual retrieval tasks. This benchmark substantially addresses the limitations of existing audio retrieval datasets in semantic retrieval. To study this task, we evaluate several off-the-shelf retrievers and train our ATIR model based on a Multimodal Large Language Model (MLLM). We further propose a novel token compression mechanism, which is orthogonal to existing compression methods, to mitigate the challenge of excessive audio tokens in MLLM-based ATIR models. Experimental results show that our ATIR model achieves significant improvements over strong baselines.

[Findings] Benchmarking Agentic Newswriting via Journalistic Workflows ▶ *Underline Video*

Yen-Che Chien, Kuang-Da Wang, Wei-Yao Wang and Wen-Chih Peng

Recent advances in autonomous digital agents from industry (e.g., Manus AI and Geminis research mode) highlight their potential for structured tasks through autonomous decision-making and task decomposition, but it remains unclear how well such systems support real-world information-intensive workflows. We study this question in journalism, where newswriting requires iterative planning, contextual reasoning, and active discovery of missing background to produce a coherent article. We introduce NEWSAGENT, a benchmark for evaluating how agents search raw materials, select relevant information, and iteratively revise drafts through core journalistic functions. Given a writing instruction and partial firsthand materials, agents must identify narrative perspectives, issue keyword-based queries, retrieve historical context, and generate complete news articles. Unlike typical summarization or retrieval tasks, essential context is not directly available and must be actively discovered, reflecting real-world reporting constraints. NEWSAGENT consists of 6k human-verified examples derived from real news. We evaluate open- and closed-sourced LLMs with commonly-used agentic frameworks on NEWSAGENT, which shows that agents are capable of retrieving relevant facts but struggling with planning and narrative integration. We believe that NEWSAGENT serves a realistic testbed for iterating and evaluating agent capabilities in terms of web data manipulation to real-world productivity. The benchmark resources are publicly available at https://github.com/wywyWang/CoachAI-Projects.

[Long] BioTool: A Comprehensive Tool-Calling Dataset for Enhancing Biomedical Capabilities of Large Language Models ▶ *Underline Video*

Xin Gao, Ruiyi Zhang, Meixi Du, Peijia Qin and Pengtao Xie

Despite the success of large language models (LLMs) on general-purpose tasks, their performance in highly specialized domains such as biomedicine remains unsatisfactory. A key limitation is the inability of LLMs to effectively leverage biomedical tools, which clinical experts and biomedical researchers rely on extensively in daily workflows. While recent general-domain tool-calling datasets have substantially improved the capabilities of LLM agents, existing efforts in the biomedical domain largely rely on in-context learning and restrict models to a small set of tools. To address this gap, we introduce BioTool, a comprehensive biomedical tool-calling dataset designed for fine-tuning LLMs. BioTool comprises 34 frequently used tools collected from the NCBI, Ensembl, and UniProt databases, along with 7,040 high-quality, human-verified query-API call pairs spanning variation, genomics, proteomics, evolution, and general biology. Fine-tuning a 4-billion-parameter LLM on BioTool yields substantial improvements in biomedical tool-calling performance, outperforming cutting-edge commercial LLMs such as GPT-5.1. Furthermore, human expert evaluations demonstrate that integrating a BioTool-fine-tuned tool caller significantly improves downstream answer quality compared to the same LLM without tool usage, highlighting the effectiveness of BioTool in enhancing the biomedical capabilities of LLMs. The full dataset and evaluation code are available at https://github.com/gxx27/BioTool.

[Short] Calibrated? Not for Everyone: How Sexual Orientation and Religious Markers Distort LLM Accuracy and Confidence in Medical QA ▶ *Underline Video*

Alberto Testoni and Iacer Calixto

Safe clinical deployment of Large Language Models (LLMs) requires not only high accuracy but also robust uncertainty calibration to ensure models defer to clinicians when appropriate. Our paper investigates how social descriptors of a patient (specifically sexual orientation and religious affiliation) distort these uncertainty signals and model accuracy. Evaluating nine general-purpose and biomedical LLMs on 2,364 medical questions and their counterfactual variants, we demonstrate that identity markers cause a "calibration crisis". Homosexual markers consistently trigger performance drops, and intersectional identities produce idiosyncratic, non-additive harms to calibration. Moreover, a clinician-validated case study in an open-ended generation setting confirms that these failures are not an artifact of the multiple-choice format. Our results demonstrate that the presence of social identity cues does not merely shift predictions; it affects the reliability of confidence signals, posing a significant risk to equitable care and safe deployment in confidence-based clinical workflows.

[Long] Can You Make It Sound Like You? Post-Editing LLM-Generated Text for Personal Style ▶ *Underline Video*

Connor Baumler, Calvin Bao, Huy Nghiem, Xinchen Yang, Marine Carpuat and Hal Daumé III

Despite the growing use of large language models (LLMs) for writing tasks, users may hesitate to rely on LLMs when personal style is

important. Post-editing LLM-generated drafts or translations is a common collaborative writing strategy, but it remains unclear whether users can effectively reshape LLM-generated text to reflect their personal style. We conduct a pre-registered online study (n=81) in which participants post-edit LLM-generated drafts for writing tasks where personal style matters to them. Using embedding-based style similarity metrics, we find that post-editing increases stylistic similarity to participants' unassisted writing and reduces similarity to fully LLM-generated output. However, post-edited text still remains stylistically closer in style to LLM text than to participants' unassisted control text, and it exhibits reduced stylistic diversity compared to unassisted human text. We find a gap between perceived stylistic authenticity and model-measured stylistic similarity, with post-edited text often perceived as representative of participants' personal style despite remaining detectable LLM stylistic traces.

[Findings] CliniCAST: Benchmarking Acoustic Grounding and Text Dominance in Medical Triage ▶ *Underline Video*

Kyusik Kim, Hyunwoo Yoo, Jaehoon Choi, Kitae Kim, Gail Rosen and Bongwon Suh

Recent Large Audio-Language Models (LALMs) integrate acoustic capabilities into reasoning, yet whether they reliably ground clinical judgments in audible evidence remains unproven. We introduce CliniCAST (Clinical Controlled Acoustic Synthetic Triage), a controlled benchmark that disentangles clinically meaningful acoustic cues from lexical content and speaker demographics. CliniCAST comprises 5,856 synthetic samples across 12 disease conditions: 4,800 audio samples forming 2,400 tagged–untagged pairs for five-level emergency triage, and 1,056 audio–text inconsistent samples in which reassuring speech is paired with high-risk acoustic cues. Evaluating a diverse suite of audio-capable foundation models, we find that LALMs exhibit fragile acoustic grounding and a pronounced "text dominance" failure mode: reassuring lexical content suppresses response to audible distress signals even under safety-critical conditions. Age and gender interactions are weak across conditions, indicating that the primary failure mode is insufficient cross-modal integration rather than demographic bias. These results suggest current LALMs are not yet robust enough for high-stakes medical triage, and motivate training objectives that explicitly enforce reliance on clinically grounded audible evidence.

[Findings] Cross-lingual Matryoshka Representation Learning across Speech and Text ▶ *Underline Video*

Yaya Sy, Christophe Cerisara and Irina Illina

Speakers of under-represented languages face both a language barrier, as most online knowledge is in a few dominant languages, and a modality barrier, since information is largely text-based while many languages are primarily oral. We address this for French-Wolof by training the first bilingual speech-text Matryoshka embedding model, enabling efficient retrieval of French text from Wolof speech queries without relying on a costly ASR-translation pipelines. We introduce large-scale data curation pipelines and new benchmarks, compare modeling strategies, and show that modality fusion within a frozen text Matryoshka model performs best. Although trained only for retrieval, the model generalizes well to other tasks, such as speech intent detection, indicating the learning of general semantic representations. Finally, we analyze cost-accuracy trade-offs across Matryoshka dimensions and ranks, showing that information is concentrated only in a few components, suggesting potential for efficiency improvements.

[Findings] Decide less, communicate more: On the construct validity of end-to-end fact-checking in medicine ▶ *Underline Video*

Sebastian Joseph, Lily Chen, Barry Wei, Michael Mackert, Iain J. Marshall, Paul Pu Liang, Ramez Kouzy, Byron C. Wallace and Junyi Jessy Li

Technological progress has led to concrete advancements in tasks that were regarded as challenging, such as automatic fact-checking. Interest in adopting these systems for public health and medicine has grown due to the high-stakes nature of medical decisions and challenges in critically appraising a vast and diverse medical literature. Evidence-based medicine connects to every individual, and yet the nature of it is highly technical, rendering the medical literacy of majority users inadequate to sufficiently navigate the domain. Such problems with medical communication ripen the ground for end-to-end fact-checking agents: check a claim against current medical literature and return with an evidence-backed verdict. And yet, such systems remain largely unused. In this position paper, developed with expert input, we present the first study examining how clinical experts verify real claims from social media by synthesizing medical evidence. In searching for this upper-bound, we reveal fundamental challenges in end-to-end fact-checking when applied to medicine: Difficulties connecting claims in the wild to scientific evidence in the form of clinical trials; ambiguities in underspecified claims mixed with mismatched intentions; and inherently subjective veracity labels. We argue that fact-checking should be approached as an interactive communication problem, rather than an end-to-end process.

[Findings] Decoupling the Effect of Chain-of-Thought Reasoning: A Human Label Variation Perspective ▶ *Underline Video*

Beiduo Chen, Tiancheng Hu, Caiqi Zhang, Robert Litschko, Anna Korhonen and Barbara Plank

Reasoning-tuned LLMs utilizing long Chain-of-Thought (CoT) excel at single-answer tasks, yet their ability to model Human Label Variation, which requires capturing probabilistic ambiguity rather than resolving it, remains underexplored. We investigate this through systematic disentanglement experiments on distribution-based tasks, employing Cross-CoT experiments to isolate the effect of reasoning text from intrinsic model priors. We observe a distinct "decoupled mechanism": while CoT improves distributional alignment, final accuracy is dictated by CoT content (99% variance contribution), whereas distributional ranking is governed by model priors (over 80%). Step-wise analysis further shows that while CoTs influence on accuracy grows monotonically during the reasoning process, distributional structure is largely determined by LLMs intrinsic priors. These findings suggest that long CoT serves as a decisive LLM decision-maker for the top option but fails to function as a granular distribution calibrator for ambiguous tasks.

[Findings] Detecting Hallucinations in SpeechLLMs at Inference Time Using Attention Maps ▶ *Underline Video*

Jonas Waldendorf, Bashar Awwad Shiekh Hasan and Evgenii Tsymbalov

Hallucinations in Speech Large Language Models (SpeechLLMs) pose significant risks, yet existing detection methods typically rely on gold-standard outputs that are costly or impractical to obtain. Moreover, hallucination detection methods developed for text-based LLMs do not directly capture audio-specific signals. We investigate four attention-derived metrics: AudioRatio, AudioConsistency, AudioEntropy, and TextEntropy, designed to capture pathological attention patterns associated with hallucination, and train lightweight logistic regression classifiers on these features for efficient inference-time detection. Across automatic speech recognition and speech-to-text translation tasks, evaluations on Qwen-2-Audio and Voxtral-3B show that our approach outperforms uncertainty-based and prior attention-based baselines on in-domain data, achieving improvements of up to +0.23 PR-AUC, and generalises to out-of-domain ASR settings. We further find that strong performance can be achieved with approximately 100 attention heads, improving out-of-domain generalisation compared to using all heads. While effectiveness is model-dependent and task-specific training is required, our results demonstrate that attention patterns provide a valuable tool for hallucination detection in SpeechLLMs

[Findings] Dialectical Structured Reasoning for Explainable Multimodal Fake News Detection ▶ *Underline Video*

Ruichao Yang, Yufan Bian, Wei Gao, Bo-Wen Zhang, Jing Ma, Hongzhan Lin, Ziyang Luo, Xiaobin Zhu and Xu-Cheng Yin

Current multimodal fake news detectors predominantly function as opaque classifiers, offering limited deductive transparency and little insight into how conflicting evidence is reconciled. To address this limitation, we propose Dialectical Structured Reasoning (DSR), a framework modeling fake news detection as an explicit dialectical process over multimodal social context. DSR instantiates two opposing agents: a Verifier, which constructs evidence paths supporting semantic consistency, and a Debunker, which actively explores exposing logical or factual contradictions. Then a differentiable Judge agent adjudicates between these competing perspectives by integrating local evidence with global parametric knowledge. Experiments on three benchmarks demonstrate that DSR achieves state-of-the-art performance while producing transparent, dialectically grounded explanations that closely mirror human reasoning process.

[Long] Discovering Properties of Inflectional Morphology in Neural Emergent Communication ▶ *Underline Video*

Miles Gilberti, Shane Storks and Huteng Dai

Emergent communication (EmCom) with deep neural network-based agents promises to yield insights into the nature of human language, but remains focused primarily on a few subfield-specific goals and metrics that prioritize communication schemes which represent attributes with unique characters one-to-one and compose them syntactically. We thus reinterpret a common EmCom setting, the attribute-value reconstruction game, by imposing a small-vocabulary constraint to simulate double articulation, and formulating a novel setting analogous to naturalistic inflectional morphology (enabling meaningful comparison to natural language communication schemes). We develop new metrics and explore variations of this game motivated by real properties of inflectional morphology: concatenativity and fusion. Through our experiments, we discover that simulated phonological constraints encourage concatenative morphology, and emergent languages replicate the tendency of natural languages to fuse grammatical attributes.

[Long] Do LLM Agents Mirror Socio-Cognitive Effects in Power-Asymmetric Conversations? ▶ *Underline Video*

Anvesh Rao Vijjini, Sagar Manjunath and Snigdha Chaturvedi

Power differences shape human communication through well-documented socio-cognitive effects, including language coordination, pronoun usage, authority bias, and harmful compliance. We examine whether large language models (LLMs) exhibit similar behaviors when assigned high- or low-status personas. Using personas from diverse professions, we simulate multi-turn, power-asymmetric dialogues (e.g., principal-teacher, justice-lawyer) and measure (i) linguistic coordination, (ii) pronoun usage, (iii) persuasion success, and (iv) compliance with unsafe requests. Our results show that LLMs show key socio-cognitive effects of power, albeit with nuances and variability, linking simulated interactions to both desirable and unsafe behaviors.

[Long] Dual Alignment Between Language Model Layers and Human Sentence Processing ▶ *Underline Video*

Tatsuki Kuribayashi, Alex Warstadt, Yohei Oseki and Ethan Gotlieb Wilcox

A recent study~kuribayashi2025largelanguagemodelshumanlike has shown that human sentence processing behavior, typically measured on syntactically unchallenging constructions, can be effectively modeled using surprisal from early layers of large language models (LLMs). This raises the question of whether such advantages of internal layers extend to more syntactically challenging constructions, where surprisal has been reported to underestimate human cognitive effort. In this paper, we begin by exploring internal layers that better estimate human cognitive effort observed in syntactic ambiguity processing in English. Our experiments show that, in contrast to naturalistic reading, later layers better estimate such a cognitive effort, but still underestimate the human data. This dual alignment sheds light on different modes of sentence processing in humans and LMs: naturalistic reading employs a somewhat weak prediction akin to earlier layers of LMs, while syntactically challenging processing requires more fully-contextualized representations, better modeled by later layers of LMs. Motivated by these findings, we also explore several probability-update measures using shallow and deep layers of LMs, showing a complementary advantage to single-layer's surprisal in reading time modeling.

[Findings] DualAlign: Generating Clinically Grounded Synthetic Data ▶ *Underline Video*

Rumeng Li, XWang XWang and hong yu

Synthetic clinical data are essential for advancing AI in healthcare, given strict privacy constraints on electronic health records (EHRs),

the scarcity of annotated data for rare or slowly progressing conditions, and demographic biases in observational cohorts. Large language models (LLMs) can generate fluent clinical text, but ensuring that such outputs are both clinically grounded and useful for downstream modeling remains challenging. We present DualAlign, a disease-agnostic framework for generating privacy-preserving, clinically faithful synthetic EHR narratives. DualAlign improves generation fidelity through two complementary alignment mechanisms: persona alignment, which conditions generation on patient demographics and risk factors, and symptom-trajectory alignment, which grounds narratives in empirically observed longitudinal symptom patterns. Using Alzheimers disease (AD) as a case study, DualAlign produces context-aware, symptom-rich sentences that more closely reflect real-world clinical documentation. Augmenting limited gold-standard data with DualAlign substantially improves AD symptom classification, outperforming both gold-only training and unconstrained synthetic baselines. Overall, DualAlign provides a generalizable approach for generating high-utility synthetic clinical text in chronic and progressive diseases, reducing annotation burden while enabling scalable and privacy-conscious clinical NLP research.

[Findings] Efficient and Effective Internal Memory Retrieval for LLM-Based Healthcare Prediction ▶ *Underline Video*
Mingchen Li, Jiatan Huang, Zonghai Yao and hong yu
Large language models (LLMs) hold significant promise for healthcare, yet their reliability in high-stakes clinical settings is often compromised by hallucinations and a lack of granular medical context. While Retrieval-Augmented Generation (RAG) can mitigate these issues, standard supervised pipelines require computationally intensive searches over massive external knowledge bases, leading to high latency that is impractical for time-sensitive care. To address this, we introduce Keys-to-Knowledge (K2K), a novel framework that replaces external retrieval with internal, key-based knowledge access. By encoding essential clinical information directly into the models parameter space, K2K enables rapid retrieval from internal key-value memory without inference-time overhead. We further enhance retrieval quality through activation-guided probe construction and cross-attention reranking. Experimental results demonstrate that K2K achieves state-of-the-art performance across four benchmark healthcare outcome prediction datasets.

[Long] emg2speech: synthesizing speech from electromyography using self-supervised speech models ▶ *Underline Video*
Harshavardhana T Gowda, Daniel C Comstock and Lee M. Miller
We present a neuromuscular speech interface that translates electromyographic (EMG) signals recorded from orofacial muscles during speech articulation directly into audio. We find that self-supervised speech (S3) representations are strongly linearly related to the electrical power of muscle activity: a simple linear mapping predicts EMG power from S3 representations with a correlation of r = 0.85. In addition, EMG power vectors associated with distinct articulatory gestures form structured, separable clusters. Together, these observations suggest that S3 models implicitly encode articulatory mechanisms, as reflected in EMG activity. Leveraging this structure, we map EMG signals into the S3 representation space and synthesize speech, enabling end-to-end EMG-to-speech generation without explicit articulatory modeling or vocoder training. We demonstrate this system with a participant with amyotrophic lateral sclerosis (ALS), converting orofacial EMG recorded while she silently articulated speech into audio.

[Findings] Faithfulness vs. Safety: Evaluating LLM Behavior Under Counterfactual Medical Evidence ▶ *Underline Video*
Kaijie Mo, Siddhartha Venkatayogi, Chantal Shaib, Ramez Kouzy, Wei Xu, Byron C Wallace and Junyi Jessy Li
In high-stakes domains like medicine, it may be generally desirable for models to faithfully adhere to the context provided. But what happens if the context does not align with model priors or safety protocols? In this paper, we investigate how LLMs behave and reason when presented with counterfactual (or even adversarial) medical evidence. We first construct MedCounterFact, a counterfactual medical QA dataset that requires the models to answer clinical comparison questions (i.e., judge the efficacy of certain treatments, with evidence consisting of randomized controlled trials provided as context). In MedCounterFact, real-world medical interventions within the questions and evidence are systematically replaced with four types of counterfactual stimuli, ranging from unknown words to toxic substances. Our evaluation across multiple frontier LLMs on MedCounterFact reveals that in the presence of counterfactual evidence, existing models overwhelmingly accept such "evidence" at face value even when it is dangerous or implausible, and provide confident and uncaveated answers. While it may be prudent to draw a boundary between faithfulness and safety, our findings suggest that models arguably overemphasize the former.

[Long] FC-TTS: Style and Timbre Control in Zero-Shot Text-to-Speech with Disentangled Speech Representations ▶ *Underline Video*
Yoonhyung Lee, Hyunsin Park, Jinhwan Park and Jinkyu Lee
Recent advances in zero-shot text-to-speech (TTS) have enabled accurate imitation of reference speech in terms of both speaking style and speaker timbre. However, achieving disentangled control over these aspects from separate references remains a challenging task. Several studies have proposed disentangled speech representations that decompose speech into interpretable attributes (e.g., timbre, prosody, and content), providing a promising foundation for TTS with attribute control from separate references. Yet, how to effectively integrate such representations into TTS systems to achieve independent and precise control remains underexplored. In this paper, we present FC-TTS, a zero-shot TTS framework that enables disentangled control of style and timbre by conditioning on two distinct reference utterances. Unlike existing systems that inherit limitations from those pre-trained disentangled representations, FC-TTS introduces key design strategies, including architectural choices, training framework, and auxiliary training objectives, which improve the reliability of attribute separation and dual-reference control. Experiments show that FC-TTS achieves high-fidelity synthesis and competitive zero-shot naturalness, while uniquely supporting consistent and independent manipulation of style and timbre. Audio samples are available at https://qualcomm-ai-research.github.io/fc-tts

[Findings] Feedback Adaptation for Retrieval-Augmented Generation ▶ *Underline Video*

Jihwan Bang, Seunghan Yang, Kyuhong Shim, Simyung Chang, Juntae Lee and Sungha Choi

Retrieval-Augmented Generation (RAG) systems are typically evaluated under static assumptions, despite being frequently corrected through user or expert feedback in deployment. Existing evaluation protocols focus on overall accuracy and fail to capture how systems adapt after feedback is introduced. We introduce feedback adaptation as a problem setting for RAG systems, which asks how effectively and how quickly corrective feedback propagates to future queries. To make this behavior measurable, we propose two evaluation axes: correction lag, which captures the delay between feedback provision and behavioral change, and post-feedback performance, which measures reliability on semantically related queries after feedback. Using these metrics, we show that training-based approaches exhibit a trade-off between delayed correction and reliable adaptation. We further propose PatchRAG, a minimal inference-time instantiation that incorporates feedback without retraining, demonstrating immediate correction and strong post-feedback generalization under the proposed evaluation. Our results highlight feedback adaptation as a previously overlooked dimension of RAG system behavior in interactive settings.

[Findings] Filling in the Mechanisms: How do LMs Learn Filler-Gap Dependencies under Developmental Constraints? ▶ *Underline Video*

Atrey Desai and Sathvik Nair

For humans, filler-gap dependencies require a shared representation across different syntactic constructions. Although causal analyses suggest this may also be true for LLMs (Boguraev et al., 2025), it is still unclear if such a representation also exists for language models trained on developmentally feasible quantities of data. We applied Distributed Alignment Search (DAS, Geiger et al. (2024)) to checkpoints of a language model from the BabyLM challenge (Warstadt et al., 2023), to evaluate whether representations of filler-gap dependencies transfer between wh-questions and topicalization, which greatly vary in terms of their input frequency. Our results suggest shared, yet item-sensitive mechanisms may develop with limited training data. More importantly, LMs still require far more data than humans to learn comparable generalizations, highlighting the need for language-specific biases in models of language acquisition.

[Findings] From Fallback to Frontline: When Can LLMs be Superior Annotators of Human Perspectives? ▶ *Underline Video*

Hasan Amin, Harry Yizhou Tian, Xiaoni Duan, Chien-Ju Ho, Rajiv Khanna and Ming Yin

Although large language models (LLMs) are increasingly used as annotators at scale, they are typically treated as a pragmatic fallback rather than a faithful estimator of human perspectives. This work challenges that presumption. By framing perspective-taking as the estimation of a latent group-level judgment, we characterize the conditions under which modern LLMs can outperform human annotators, including in-group humans, when predicting aggregate subgroup opinions on subjective tasks, and show that these conditions are common in practice. This advantage arises from structural properties of LLMs as estimators, including low variance and reduced coupling between representation and processing biases, rather than any claim of lived experience. Our analysis identifies clear regimes where LLMs act as statistically superior frontline estimators, as well as principled limits where human judgment remains essential. These findings reposition LLMs from a cost-saving compromise to a principled tool for estimating collective human perspectives.

[Findings] GerAV: Towards New Heights in German Authorship Verification using Fine-Tuned LLMs on a New Benchmark ▶ *Underline Video*

Lotta Kiefer, Christoph Leiter, Sotaro Takeshita, Elena Schmidt and Steffen Eger

Authorship verification (AV) is the task of determining whether two texts were written by the same author and has been studied extensively, predominantly for English data. In contrast, large-scale benchmarks and systematic evaluations for other languages remain scarce. We address this gap by introducing GerAV, a comprehensive benchmark for German AV comprising over 400k labeled text pairs. GerAV is built from Twitter and Reddit data, with the Reddit part further divided into in-domain and cross-domain message-based subsets, as well as a profile-based subset. This design enables controlled analysis of the effects of data source, topical domain, and text length. Using the provided training splits, we conduct a systematic evaluation of strong baselines and state-of-the-art models and find that our best approach, a fine-tuned large language model, outperforms recent baselines by up to 0.09 absolute F1 score and surpasses GPT-5 in a zero-shot setting by 0.08. We further observe a trade-off between specialization and generalization: models trained on specific data types perform best under matching conditions but generalize less well across data regimes, a limitation that can be mitigated by combining training sources. Overall, GerAV provides a challenging and versatile benchmark for advancing research on German and cross-domain AV.

[Long] GUI0: Self-Evolving Foundational GUI Agents in Super App Ecosystems ▶ *Underline Video*

Xinyi Wang, Wei Dai, Kyle Qiao, Ke Wang, Peng Chen, Gang Cao, Kang Qin, Zhongpu Wang, Xiaode Zhang, Yanming Liu, Jihao Gu, Jingtao Xu and Gong Zhi

Automated interaction with graphical user interfaces (GUIs) is central to General Artificial Intelligence yet remains challenging within Super App ecosystems, characterized by non-standard rendering and absent accessibility metadata. While GUI agents often rely on explicit accessibility trees or static imitation, they are less explored for dynamic environments marked by sparse feedback and implicit visual cues. We present GUI0, a framework synergizing autonomous data synthesis with dual-agent co-evolution. GUI0 establishes a domain-aware foundation model via synthesized corpora and employs curriculum-driven reinforcement learning, where a curriculum agent generates boundary tasks to optimize an actor agent. Empirical results demonstrate three key advantages: (1) State-of-the-art performance on the SuperAPP benchmark, outperforming Gemini-2.5-Pro and Claude-4-Sonnet; (2) universal efficacy across diverse base models, consistently yielding substantial improvements on both Qwen2.5-VL and GUI-Owl variants; and (3) robust zero-shot generalization to standard GUIs

(e.g., +62.7% on ScreenSpot Pro).

[Findings] HypEHR: Hyperbolic Modeling of Electronic Health Records for Efficient Question Answering ▶ *Underline Video*
Yuyu Liu, Sarang Rajendra Patil, Mengjia Xu and Tengfei Ma
Electronic health record (EHR) question answering is often handled by LLM-based pipelines that are costly to deploy and do not explicitly leverage the hierarchical structure of clinical data. Motivated by evidence that medical ontologies and patient trajectories exhibit hyperbolic geometry, we propose HypEHR, a compact Lorentzian model that embeds codes, visits, and questions in hyperbolic space and answers queries via geometry-consistent cross-attention with type-specific pointer heads. HypEHR is pretrained with next-visit diagnosis prediction and hierarchy-aware regularization to align representations with the ICD ontology. On two MIMIC-IV-based EHR-QA benchmarks, HypEHR approaches LLM-based methods while using far fewer parameters. Our code is publicly available at https://github.com/yuyuliu11037/HypEHR.

[Findings] Interpreting Style Representations via Style-Eliciting Prompts ▶ *Underline Video*
Junghwan Kim and David Jurgens
Style representation learning is a powerful tool for authorship analysis and modeling writing style, yet the latent nature of learned representations makes them difficult to interpret. Recent work has attempted to explain these representations by generating natural language descriptions with large language models (LLMs) conditioned on input text. However, such descriptions are often prone to the LLM's biases and hallucinations, and they lack an explicit objective and practical utility. In this work, we propose a novel framework for interpreting style representations through style-eliciting prompts: natural language instructions designed to steer LLMs to generate text that reflects specific stylistic attributes. We curate 1,010 distinct style features spanning 26 stylistic categories and construct a dataset by prompting an LLM to generate text conditioned on these features. Using this data, we train a decoder to generate a style prompt from the style representation of the generated text. We evaluate our approach on three tasks: (1) recovering original style prompts from generated text, (2) generating text in the same style using the recovered prompts, and (3) steering LLM outputs to match the style of human-written texts. Experiments demonstrate that our method consistently outperforms strong baselines that directly prompt LLMs with target text, achieving superior performance in both style description and style imitation. These results highlight style-eliciting prompts can provide a practical and interpretable interface to stylistic information encoded in style representations.

[Long] JanusMM: A Benchmark for Self-Deprecation Understanding in Real-World Multimodal Conversations ▶ *Underline Video*
Xinyi Xu, Bingguang Hao, Yongyi Xiong, Zimo Chen, Xinchen Liu, Hongxin Guo, xuelong Wang, Silin Zhou and Shihan Dou
Self-deprecation is a prevalent communicative strategy in human society, often using image-text interplay to express emotions and intentions. Despite self-deprecation is widespread in real-world conversations, the ability of multimodal large language models (MLLMs) to understand it remains underexplored. To fill this gap, we introduce JanusMM, the first benchmark designed to evaluate MLLMs' understanding of self-deprecation in real-world conversations. JanusMM contains 2,016 bilingual memes from three types of social interactions and provides a dual-task evaluation framework with six new metrics. The first task assesses MLLMs' abilities in self-deprecation recognition and reasoning, while the second task evaluates the consistency of their understanding by simulating the perspectives of the initiator and responder. We evaluate ten frontier MLLMs and find that they exhibit weak recognition and reasoning abilities, with their understanding of self-deprecation remaining inconsistent across both perspectives.

[Long] Language Statistics and False Belief Reasoning: Evidence from 41 Open-Weight LMs ▶ *Underline Video*
Sean Trott, Samuel Taylor, Cameron Jones, James Michaelov and Pamela Rivière
Research on mental state reasoning in language models (LMs) has the potential to inform theories of human social cognition—such as the theory that mental state reasoning emerges in part from language exposure, and our understanding of LMs themselves. Yet much published work on LMs relies on a relatively small sample of closed-source LMs, limiting our ability to rigorously test psychological theories and evaluate LM capacities. Here, we replicate and extend published work on the false belief task by assessing LM mental state reasoning behavior across 41 open-weight models (from distinct model families). We find sensitivity to implied knowledge states in 34% of the LMs tested; however, consistent with prior work, none fully "explain away" the effect in humans. Larger LMs show increased sensitivity and also exhibit higher psychometric predictive power. Finally, we use LM behavior to generate and test a novel hypothesis about human cognition: both humans and LMs show a bias towards attributing false beliefs when knowledge states are cued using a non-factive verb ("John thinks...") than when cued indirectly ("John looks in the..."). Unlike the primary effect of knowledge states, where human sensitivity exceeds that of LMs, the magnitude of the human knowledge cue effect falls squarely within the distribution of LM effect sizes—suggesting that distributional statistics of language can in principle account for the latter but not the former in humans. These results demonstrate the value of using larger samples of open-weight LMs to test theories of human cognition and evaluate LM capacities.

[Long] LCMA-SRT: Language-Conditional Mixture-of-Experts Adapters for Joint Multilingual Speech Recognition and Translation ▶ *Underline Video*
Nanjie Li, Xiaoyong Guo, Hao Huang, Xu Haihua and Wei Shi
Neural transducers offer an alignment-free framework for speech-to-text modeling, and hierarchical transducer architectures further improve multilingual joint automatic speech recognition (ASR) and speech translation (ST) by stacking a translation-focused encoder on top of an ASR encoder. However, extending hierarchical transducers to multilingual many-to-many settings remains challenging: fully shared

models often suffer from negative transfer and unstable target-language generation, while training separate models for each direction is computationally prohibitive. We propose LCMA-SRT (Language-Conditional Mixture-of-Experts Adapters for Speech Recognition and Translation), which augments a hierarchical transducer with language-conditional Mixture-of-Experts (MoE) adapters. A source-conditioned MoE adapter (SRC-MoE) uses source-language embeddings to reduce cross-language interference and improve multilingual ASR. A target-conditioned MoE adapter (TGT-MoE) uses the desired target language to reduce cross-target interference and stabilize target-language generation in many-to-many ST. Experiments on Europarl-ST (9 languages, 72 directions) show that LCMA-SRT improves both ASR and ST within a single joint model, reducing average WER and improving BLEU and COMET over strong hierarchical transducer baselines. We release our code and models at https://github.com/linanjie0820/LCMA-SRT.

[Long] LENS: LLM-Enabled Narrative Synthesis for Mental Health by Aligning Multimodal Sensing with Language Models ▶ *Underline Video*

Wenxuan Xu, Arvind Pillai, Subigya Nepal, Amanda Collins, Daniel Mackin, Michael Heinz, Tess Griffin, Nicholas Jacobson and Andrew Campbell

Multimodal health sensing offers rich behavioral signals for assessing mental health, yet translating these numerical time-series measurements into natural language remains challenging. Current LLMs cannot natively ingest long-duration sensor streams, and paired sensor-text datasets are scarce. To address these challenges, we introduce LENS, a framework that aligns multimodal sensing data with language models to generate clinically grounded mental-health narratives. LENS first constructs a large-scale dataset by transforming Ecological Momentary Assessment (EMA) responses related to depression and anxiety symptoms into natural-language descriptions, yielding over 100,000 sensor-text QA pairs from 258 participants. To enable native time-series integration, we train a patch-level encoder that projects raw sensor signals directly into an LLMs representation space. Our results show that LENS outperforms strong baselines on standard NLP metrics and task-specific measures of symptom-severity accuracy. A user study with 13 mental-health professionals further indicates that LENS-produced narratives are comprehensive and clinically meaningful. Ultimately, our approach advances LLMs as interfaces for health sensing, providing a scalable path toward models that can reason over raw behavioral signals and support downstream clinical decision-making.

[Findings] LVLM-Aware Multimodal Retrieval for RAG-Based Medical Diagnosis with General-Purpose Models ▶ *Underline Video*

Nir Mazor and Tom Hope

Retrieving visual and textual information from medical literature and hospital records can enhance diagnostic accuracy for clinical image interpretation. However, multimodal retrieval-augmented diagnosis is highly challenging. We explore a lightweight mechanism for enhancing diagnostic performance of retrieval-augmented LVLMs. We train an LVLM-aware multimodal retriever, such that the retriever learns to return images and texts that guide the LVLM toward correct predictions. In our low-resource setting, we perform only lightweight fine-tuning with small amounts of data, and use only general-purpose backbone models, achieving competitive results in clinical classification and VQA tasks compared to medically pre-trained models with extensive training. In a novel analysis, we highlight a previously unexplored class of errors that we term inconsistent retrieval predictions: cases where different top-retrieved images yield different predictions for the same target. We find that these cases are challenging for all models, even for non-retrieval models, and that our retrieval optimization mechanism significantly improves these cases over standard RAG. However, our analysis also sheds light on gaps in the ability of LVLMs to utilize retrieved information for clinical predictions.

[Short] MARCH: Multi-Agent Radiology Clinical Hierarchy for CT Report Generation ▶ *Underline Video*

Yi Lin, Yihao Ding, Yonghui Wu and Yifan Peng

Automated 3D radiology report generation often suffers from clinical hallucinations and a lack of the iterative verification found in human practice. While recent Vision-Language Models (VLMs) have advanced the field, they typically operate as monolithic "black-box" systems without the collaborative oversight characteristic of clinical workflows. To address these challenges, we propose MARCH (Multi-Agent Radiology Clinical Hierarchy), a multi-agent framework that emulates the professional hierarchy of radiology departments and assigns specialized roles to distinct agents. MARCH utilizes a Resident Agent for initial drafting with multi-scale CT feature extraction, multiple Fellow Agents for retrieval-augmented revision, and an Attending Agent that orchestrates an iterative, stance-based consensus discourse to resolve diagnostic discrepancies. On the RadGenome-ChestCT dataset, MARCH significantly outperforms state-of-the-art baselines in both clinical fidelity and linguistic accuracy. Our work demonstrates that modeling human-like organizational structures enhances the reliability of AI in high-stakes medical domains.

[Findings] MedCPI: A Construct-Personalize-Integrate Framework for KG-enhanced Clinical Prediction ▶ *Underline Video*

Hang Wang, Hang Dong and Lu Liu

Electronic health records (EHRs) provide longitudinal evidence for clinical prediction, but EHR data are sparse, incomplete, and heterogeneous, which can limit robustness. Medical knowledge graphs (MKGs) have therefore been incorporated to support KG-enhanced clinical prediction by linking heterogeneous EHR codes to shared medical concepts via structured relations. However, existing KG-enhanced approaches remain limited in two aspects: (i) task-specific knowledge selection when extracting knowledge from a large multi-source MKG; and (ii) patient-level personalization and knowledge integration, where personalization is often weakly controlled and knowledge integration is not sufficiently aligned with longitudinal patient trajectories. To address these issues, we propose MedCPI, a unified Construct-Personalize-Integrate framework. MedCPI first performs task-guided schema induction and KG normalization to build a task-specific Concept MKG as a denoised knowledge pool, then constructs controlled patient-level PKGs via local expansion and short path

search, and finally integrates PKG representations with time-aware EHR representations via cross-attention for prediction. Experiments on MIMIC-III and MIMIC-IV across four clinical prediction tasks show consistent improvements over strong EHR-only and KG-enhanced baselines. Ablations and additional analyses further validate the contribution of each stage and illustrate how MedCPI utilizes structured medical knowledge.

[Findings] MEDSYN: Benchmarking Multi-EviDence SYNthesis in Complex Clinical Cases for Multimodal Large Language Models ▶ *Underline Video*

Boqi Chen, Xudong Liu, jiachuan peng, Marianne Frey-Marti, Kyle Lam, Bang Zheng, Lin Li and Jianing Qiu

Multimodal large language models (MLLMs) have shown great potential in medical applications, yet existing benchmarks inadequately capture real-world clinical complexity. We introduce MEDSYN, a multilingual, multimodal benchmark of highly complex clinical cases with up to 7 distinct visual clinical evidence (CE) types per case. Mirroring clinical workflow, we evaluate 18 MLLMs on differential diagnosis (DDx) generation and final diagnosis (FDx) selection. While frontier models often match or even outperform human experts on DDx generation, all MLLMs exhibit a much larger DDx–FDx performance gap compared to expert clinicians, indicating a failure mode in synthesis of heterogeneous CE types. Ablations attribute this failure to (i) overreliance on less discriminative textual CE (e.g., medical history) and (ii) a cross-modal CE utilization gap. We introduce Evidence Sensitivity to quantify the latter and show that a smaller gap correlates with higher diagnostic accuracy. Finally, we demonstrate how it can be used to guide interventions to improve model performance. https://github.com/jianing-lab/MEDSYN.

[Findings] Model in Distress: Sentiment Analysis on French Synthetic Social Media ▶ *Underline Video*

Pierre-Carl Langlais, Pavel Chizhov, Yannick Detrois, Carlos Rosas Hinostroza, Ivan P. Yamshchikov and Bastien Perroy

Automated analysis of customer feedback on social media is hindered by three challenges: the high cost of annotated training data, the scarcity of evaluation sets, especially in multilingual settings, and privacy concerns that prevent data sharing and reproducibility. We address these issues by developing a generalizable synthetic data generation pipeline applied to a case study on customer distress detection in French public transportation. Our approach utilizes backtranslation with fine-tuned models to generate 1.7 million synthetic tweets from a small seed corpus, complemented by synthetic reasoning traces. We train 600M-parameter reasoners with English and French reasoning that achieve 77-79% accuracy on human-annotated evaluation data, matching or exceeding SOTA proprietary LLMs and specialized encoders. Beyond reducing annotation costs, our pipeline preserves privacy by eliminating the exposure of sensitive user data. Our methodology can be adopted for other use cases and languages.

[Long] On the Effect of Hyperparameters in Language Modeling for Computational Linguistics ▶ *Underline Video*

Ruoxi Ning, Yongpeng Zhu, Qingcheng Zeng, Tatsuki Kuribayashi and Freda Shi

Training language models and examining their linguistic behaviors have been a common protocol in computational linguistics for studying linguistic phenomena and modeling human language processing. However, work in this area is often limited to proof-of-concept demonstrations with arbitrary model configurations, without considering hyperparameter sensitivity, an important source of variation in model performance. In this work, we replicate three prior studies (Chang and Bergen, 2022; Hu et al., 2020b; Kuribayashi et al., 2024) with hyperparameters varied within a practical range, and show that modest hyperparameter changes can alter some qualitative conclusions about models' linguistic abilities and even reverse the ranking of model performance. Our results highlight the risk that prior work may have reflected optimization artifacts rather than the genuine inductive biases of model classes, and that hyperparameter sensitivity should receive more attention as a factor that can meaningfully influence model behavior. We suggest future work to report the variation of performance across the configuration space to enhance the reliability and generalizability of conclusions. Code: https://github.com/compling-wat/tune-linguistic-lms.

[Findings] On the Fallacy of Global Token Perplexity in Spoken Language Model Evaluation ▶ *Underline Video*

Chan-Jan Hsu, Liang-Hsuan Tseng, Yi-Cheng Lin, Yen-Chun Kuo, Ju-Chieh Chou, Kai-Wei Chang, Hung-yi Lee and Carlos Busso

Generative spoken language models pretrained on large-scale raw audio can continue a speech prompt with appropriate content while preserving attributes like speaker and emotion, serving as foundation models for spoken dialogue. In prior literature, these models are often evaluated using "global token perplexity", which directly applies the text perplexity formulation to speech tokens. However, this practice overlooks fundamental differences between speech and text modalities, possibly leading to an underestimation of the speech characteristics. In this work, we propose a variety of likelihood- and generative-based evaluation methods that serve in place of naive global token perplexity. We demonstrate that the proposed evaluations more faithfully reflect perceived generation quality, as evidenced by stronger correlations with human-rated mean opinion scores (MOS). When assessed under the new metrics, the relative performance landscape of spoken language models is reshaped, revealing a significantly reduced gap between the best-performing model and the human topline. Together, these results suggest that appropriate evaluation is critical for accurately assessing progress in spoken language modeling.

[Long] On the Proper Treatment of Units in Surprisal Theory ▶ *Underline Video*

Samuel Kiegeland, Vésteinn Snæbjarnarson, Tim Vieira and Ryan Cotterell

Surprisal theory links human processing effort to the predictability of an upcoming linguistic unit, but empirical work often leaves the notion of a unit underspecified. In practice, experimental stimuli are segmented into linguistically motivated units (e.g., words), while pretrained language models assign probability mass to a fixed token alphabet that typically does not align with those units. As a result, surprisal-based

predictors depend implicitly on ad hoc procedures that conflate two distinct modeling choices: the definition of the unit of analysis and the choice of regions of interest over which predictions are evaluated. In this paper, we disentangle these choices and give a unified framework for reasoning about surprisal over arbitrary unit inventories. We argue that surprisal-based analyses should make these choices explicit and treat tokenization as an implementation detail rather than a scientific primitive.

[Findings] Optimizing Conversational Quality in Spoken Dialogue Systems with Reinforcement Learning from AI Feedback ► *Underline Video*

Siddhant Arora, Jinchuan Tian, Jiatong Shi, Hayato Futami, Yosuke Kashiwagi, Emiru Tsunoo and Shinji Watanabe

Reinforcement learning from human or AI feedback (RLHF/RLAIF) for speech-in/speech-out dialogue systems (SDS) remains underexplored, with prior work largely limited to single semantic rewards applied at the utterance level. Such setups overlook the multi-dimensional and multi-modal nature of conversational quality, which encompasses semantic coherence, audio naturalness, speaker consistency, emotion alignment, and turn-taking behavior. Moreover, they are fundamentally mismatched with duplex spoken dialogue systems that generate responses incrementally, where agents must make decisions based on partial utterances. We address these limitations with the first multi-reward RLAIF framework for SDS, combining semantic, audio-quality, and emotion-consistency rewards. To align utterance-level preferences with incremental, blockwise decoding in duplex models, we apply turn-level preference sampling and aggregate per-block log-probabilities within a single DPO objective. We present the first systematic study of preference learning for improving SDS quality in both multi-turn Chain-of-Thought and blockwise duplex models, and release a multi-reward DPO dataset to support reproducible research. Experiments show that single-reward RLAIF selectively improves its targeted metric, while joint multi-reward training yields consistent gains across semantic quality and audio naturalness. These results highlight the importance of holistic, multi-reward alignment for practical conversational SDS.

[Findings] Peering Behind the Shield: Guardrail Identification in Large Language Models ► *Underline Video*

Ziqing Yang, Yixin Wu, Rui Wen, Michael Backes and Yang Zhang;

With the rapid adoption of large language models (LLMs), conversational AI agents have become widely deployed across real-world applications. To enhance safety, these agents are often equipped with guardrails that moderate harmful content. Identifying the guardrails in an agent thus becomes critical for adversaries to understand the system and design guard-specific attacks. In this work, we introduce AP-Test, a novel approach that leverages guard-specific adversarial prompts to detect the identity of guardrails deployed in black-box AI agents. Our method addresses key challenges in this task, including the influence of safety-aligned LLMs and other guardrails, as well as a lack of principled decision-making strategies. AP-Test employs two complementary testing strategies, input and output guard tests, and a new metric, match score, to enable robust identification. Experiments across diverse agents and four open-source guardrails demonstrate that AP-Test achieves perfect classification accuracy in multiple scenarios. Ablation studies further highlight the necessity of our proposed components. Our findings reveal a practical path toward guardrail identification in real-world AI systems.

[Findings] PersonaAgent: Bridging Memory and Action for Personalized LLM Agents ► *Underline Video*

Weizhi Zhang, Xinyang Zhang1, Chenwei Zhang1, Liangwei Yang, Jingbo Shang, Zhepei Wei, Henry Peng Zou, Zijie Huang, Zhengyang Wang, Yifan Gao, Xiaoman Pan, Lian Xiong, Jingguo Liu, Philip S. Yu and Xian Li

Large Language Model (LLM) empowered agents have recently emerged as advanced paradigms that exhibit impressive capabilities in a wide range of domains and tasks. Despite their potential, current LLM agents often adopt a one-size-fits-all approach, lacking the flexibility to respond to users varying needs and preferences. This limitation motivates us to develop PersonaAgent, the first personalized LLM agent framework designed to address versatile personalization tasks. Specifically, PersonaAgent integrates two complementary components: a personalized memory module that includes episodic and semantic memory mechanisms; a personalized action module that enables the agent to perform tool actions tailored to the user. At the core, the persona (defined as unique system prompt for each user) functions as an intermediary: it leverages insights from personalized memory to control agent actions, while the outcomes of these actions in turn refine the memory. Based on the framework, we propose a test-time user-preference alignment strategy that simulate the latest n interactions to optimize the persona prompt, ensuring real-time user preference alignment through textual loss feedback between simulated and ground-truth responses. Experimental evaluations demonstrate that PersonaAgent significantly outperforms other baseline methods by not only personalizing the action space effectively but also scaling during test-time real-world applications. These results underscore the feasibility and potential of our approach in delivering tailored, dynamic user experiences.

[Findings] PlanRAG-Audio: Planning and Retrieval Augmented Generation for Long-form Audio Understanding ► *Underline Video*

Masao Someki, Chien-yu Huang, Siddhant Arora, Samuele Cornell, Markus Müller, Nathan Susanj, Rupak Vignesh Swaminathan, Grant Strimel, Jing Liu and Shinji Watanabe

Long-form audio understanding poses significant challenges for large audio language models (LALMs) due to the extreme length of audio sequences and the need to reason over heterogeneous acoustic cues distributed over time, such as speech content, speaker identity, emotion, and sound events. To address these challenges, we propose PlanRAG-Audio, a planning-based retrieval-augmented generation framework for scalable long-form audio understanding. Rather than having audio LALMs process entire recordings directly, PlanRAG-Audio explicitly plans which modalities and temporal spans are required for a given query, and retrieves only query-relevant information from a structured text and audio database. This retrieval planning enables effective reasoning over complex, cross-domain audio queries while substantially reducing the input length passed to the large language models. Experiments across a wide range of speech/audio retrieval demonstrate that

PlanRAG-Audio improves reasoning accuracy and stabilizes performance as audio duration increases by decoupling inference cost from raw audio length.

[Short] Privacy-preserving Prosody Representation Learning ▶ *Underline Video*

Kevin Everson and Mari Ostendorf

Speech representations that capture prosodic information can be useful for both understanding and generation. However, speaker characteristics are reflected in acoustic-prosodic features (e.g., pitch). To address privacy concerns from the leakage of identity information, we propose a new self-supervised approach to learning prosody representations that incorporates speaker disentanglement strategies. We evaluate our encoder on three tasks to probe representation capabilities, including pitch reconstruction and detection of different prosodic events. Our encoder outperforms raw prosody and HuBERT-base baselines, achieving strong speaker disentanglement without adverse impact on prosody-related downstream tasks.

[Long] Reasoning Is Not All You Need: Examining LLMs for Multi-Turn Mental Health Conversations ▶ *Underline Video*

Mohit Chandra, Siddharth Sriraman, Harneet Singh Khanuja, Yiqiao Jin and Munmun De Choudhury

Limited access to mental healthcare, extended wait times, and increasing capabilities of Large Language Models (LLMs) has led individuals to turn to LLMs for fulfilling their mental health needs. However, examining the multi-turn mental health conversation capabilities of LLMs remains under-explored. Existing evaluation frameworks typically focus on diagnostic accuracy and win-rates and often overlook alignment with patient-specific goals, values, and personalities required for meaningful conversations. To address this, we introduce MedAgent, a novel framework for synthetically generating realistic, multi-turn mental health sensemaking conversations and use it to create the Mental Health Sensemaking Dialogue (MHSD) dataset, comprising over 2,200 patient-LLM conversations. Additionally, we present MultiSenseEval, a holistic framework to evaluate the multi-turn conversation abilities of LLMs in healthcare settings using human-centric criteria. Our findings reveal that frontier reasoning models yield below-par performance for patient-centric communication and struggle at precise ("hard") diagnostic capabilities with average accuracy of ~31%. Additionally, we observed variation in model performance based on patient's persona and performance drop with increasing turns in the conversation. Our work provides a comprehensive synthetic data generation framework, a dataset and evaluation framework for assessing LLMs in multi-turn mental health conversations.

[Findings] Scaling Collaborative Effort with Agents ▶ *Underline Video*

Shannon Shen, Valerie Chen, Ken Gu, Alexis Ross, Zixian Ma, Jillian Ross, Alex Gu, Chenglei Si, Wayne Chi, Andi Peng, Jocelyn Shen, Ameet Talwalkar, Sherry Wu and David Sontag

Current evaluations of agents remain centered around one-shot task completion, failing to account for the inherently iterative and collaborative nature of many real-world problems, where human goals are often underspecified and evolve. We argue for a shift from building and assessing task completion agents to developing collaborative agents, assessed not only by the quality of their final outputs but by how well they engage with and enhance human effort throughout the problem-solving process. To support this shift, we introduce collaborative effort scaling, a framework that captures how an agent's utility grows with increasing user involvement. Through case studies and simulated evaluations, we show that state-of-the-art agents often underperform in multi-turn, real-world scenarios, revealing a missing ingredient in agent design: the ability to sustain engagement and scaffold user understanding. Collaborative effort scaling offers a lens for diagnosing agent behavior and guiding development toward more effective interactions.

[Findings] Scaling is Not All You Need: Clinical-Oriented Reinforcement Learning Makes Parameter-Efficient Clinical Reasoning ▶ *Underline Video*

Chi Liu, Yan Shu, Mengzhuo Chen, Hongming Piao, Zhijian Duan, Derek Li and Bryan Dai

While large language models show promise in medical applications, achieving expert-level clinical reasoning efficiently remains challenging due to the need for massive amounts of manually labeled data and large-scale models. To address this challenge, we propose Clinical-Oriented Reinforcement Learning (CORL), the first fully open-source, end-to-end reinforcement learning training pipeline in the clinical reasoning domain, incorporating a Reasoning-Oriented Data Strategy (RODS) based on topological synthesis, CoT cold-start, and two-stage reinforcement learning. Through CORL, we trained the Fleming-R1 series of models. Among them, Fleming-R1-7B significantly outperforms models of comparable size while approaching or even surpassing certain 32B and 72B models. Fleming-R1-32B achieves near-parity with GPT-4o and outperforms the strongest open-source alternatives up to 671B in MedXpertQA. This demonstrates that in clinical reasoning field, a meticulously designed training pipeline holds greater importance than scaling model size alone. Data and Models are available at https://github.com/UbiquantAI/Fleming-R1 and https://huggingface.co/collections/IQuestLab/fleming.

[Long] Shanks: Simultaneous Hearing and Thinking for Spoken Language Models ▶ *Underline Video*

Cheng-Han Chiang, Xiaofei Wang, Linjie Li, Chung-Ching Lin, Kevin Lin, Shujie LIU, Zhendong Wang, Zhengyuan Yang, Hung-yi Lee and Lijuan Wang

Current large language models (LLMs) and spoken language models (SLMs) begin thinking and taking actions only after the user has finished their turn. This prevents the model from interacting with the user during the user's turn and can lead to high response latency when the model is thinking. To address this issue, we draw inspiration from the think while listening behavior of humans. In this paper, we propose SHANKS, a general inference framework that enables SLMs to generate unspoken chain-of-thought reasoning while listening to user input. SHANKS streams input speech in fixed-duration chunks and, as soon as a chunk is received, generates unspoken reasoning based on all

previous speech and reasoning, while the user continues speaking. SHANKS uses unspoken reasoning to determine whether to interrupt the user and make tool calls to complete the task. We demonstrate that SHANKS enhances real-time user-SLM interaction in two scenarios: (1) SHANKS can listen to the user's speech and interrupt when the user makes a mistake. (2) In a tool-augmented dialogue scenario, SHANKS can complete 56.9% of the tool calls before the user ends their turn. Overall, SHANKS is a step toward models that keep thinking throughout the conversation, not only after a turn ends. Demos can be found on the project page: https://d223302.github.io/SHANKS/.

[Long] SpeechMedAssist: Efficiently and Effectively Adapting Speech Language Models for Medical Consultation ▶ *Underline Video*
Sirry Chen, Jieyi Wang, Wei Chen and Zhongyu Wei
Medical consultations are intrinsically speech-centric. However, most prior works focus on long-text-based interactions, which are cumbersome and patient-unfriendly. Recent advances in speech language models (SpeechLMs) have enabled more natural speech-based interaction, yet the scarcity of medical speech data and the inefficiency of directly fine-tuning on speech data jointly hinder the adoption of SpeechLMs in medical consultation. In this paper, we propose SpeechMedAssist, a SpeechLM natively capable of conducting speech-based multi-turn interactions with patients. By exploiting the architectural properties of SpeechLMs, we decouple the conventional one-stage training into a two-stage paradigm consisting of (1) Knowledge & Capability Injection via Text and (2) Modality Re-alignment with Limited Speech Data, thereby reducing the requirement for medical speech data to only 10k synthesized samples. To evaluate SpeechLMs for medical consultation scenarios, we design a benchmark comprising both single-turn question answering and multi-turn simulated interactions. Experimental results show that our model outperforms all baselines in both effectiveness and robustness in most evaluation settings.

[Long] Still Between Us? Evaluating and Improving Voice Assistant Robustness to Third-Party Interruptions ▶ *Underline Video*
Dongwook Lee, Eunwoo Song, Che Hyun Lee, Heeseung Kim and Sungroh Yoon
While recent Spoken Language Models (SLMs) have been actively deployed in real-world scenarios, they lack the capability to discern Third-Party Interruptions (TPI) from the primary users ongoing flow, leaving them vulnerable to contextual failures. To bridge this gap, we introduce TPI-Train, a dataset of 88K instances designed with speaker-aware hard negatives to enforce acoustic cue prioritization for interruption handling, and TPI-Bench, a comprehensive evaluation framework designed to rigorously measure the interruption-handling strategy and precise speaker discrimination in deceptive contexts. Experiments demonstrate that our dataset design mitigates semantic shortcut learning, a critical pitfall where models exploit semantic context while neglecting acoustic signals essential for discerning speaker changes. We believe our work establishes a foundational resource for overcoming text-dominated unimodal reliance in SLMs, paving the way for more robust multi-party spoken interaction. The code for the framework is publicly available at https://tpi-va.github.io

[Findings] StoryMI: Steerable Multi-Agent Therapeutic Dialogue Generation ▶ *Underline Video*
Qingyu Meng, Min Chen, Dingming Liu, Yifan Mo, Yue Su, Xin Sun, Koen Hindriks and Jiahuan Pei
Large language models (LLMs) can generate fluent dialogue, but prior works lack situational grounding, dynamic strategy control, and evaluation aligned with clinical standards in motivational interviewing (MI). We introduce StoryMI, a multi-LLM agent framework for controllable MI dialogue generation, where questionnaire-based client profiles are expanded into situational stories that provide narrative context for the dialogue. Therapist and client agents generate MI-coded utterances guided by MI codes selected by the interaction agent, while an interaction agent dynamically coordinates exchanges to control MI strategies during a multi-turn conversation. We propose a two-level evaluation protocol: lexical metrics and MI-specific measures of macro-level counseling strategies, alongside LLM-as-judge and human expert assessments. We construct a dataset of 6K simulated MI dialogues grounded in 1K questionnaire-story pairs, covering 12 MI codes and 13 symptom domains, and benchmark six open- and closed-source LLMs. Our results show that situational grounding and macro-level control can improve MI adherence and clinical plausibility, demonstrating the effectiveness of a structured multi-agent workflow for psychotherapy dialogue generation. We provide code and data for reproducibility.

[Findings] StressTest: Can YOUR Speech LM Handle the Stress? ▶ *Underline Video*
Iddo Yosha, Gallil Maimon and Yossi Adi
Sentence stress refers to emphasis on words within a spoken utterance to highlight or contrast an idea. It is often used to imply an underlying intention not explicitly stated. Recent speech-aware language models (SLMs) have enabled direct audio processing, allowing models to access the full richness of speech to perform audio reasoning tasks such as spoken question answering. Despite the crucial role of sentence stress in shaping meaning and intent, it remains largely overlooked in evaluation and development of SLMs. We address this gap by introducing StressTest, a benchmark designed to evaluate models' ability to distinguish between meanings of speech based on the stress pattern. We evaluate leading SLMs, and find that despite their overall capabilities, they perform poorly on such tasks. Hence, we propose a novel data generation pipeline, and create Stress-17k, a training set that simulates change of meaning implied by stress variation. Results suggest, that our finetuned model, StresSLM, generalizes well to real recordings and notably outperforms existing SLMs on sentence stress reasoning and detection. Models, code, data, samples - https://pages.cs.huji.ac.il/adiyoss-lab/stresstest.

[Findings] Timesteps of Mamba Align with Human Reading Times ▶ *Underline Video*
Yuji Yamamoto, Shinnosuke Isono, Yoshinobu Kawahara and Sho Yokoi
This study demonstrates an alignment of per-word processing time in a popular state-space language model Mamba and human readers. In Mamba, the recurrent state transition at each layer conceptually takes some duration of time, the discretization timestep _t, determined dynamically in response to the input. Using a naturalistic reading dataset, we show that the per-word timestep from Mamba is a powerful

predictor of human reading times, comparable to strong baselines such as word frequency and GPT-2 surprisal and significant even when they are controlled for. We further suggest, through formal analysis of Mamba's architecture and internal dynamics, that Mamba can serve as a new, valuable lens to look at human real-time language processing with ever-updated memory, because it allows us to look at how each module (layer) weighs short- and long-term information retention, and how noise may interact with dynamic, continuous memory representation. Code is available via an (anonymized) link.

[Findings] Towards semantic reliable clinical QA: Query pipeline optimization for cancer patient question answering systems ▶ *Underline Video*

MaoLin He, Rena Wei Gao, Mike Conway and Brian E. Chapman

Large Language Models (LLMs) show promise in medical Question-Answering (QA) but suffer from hallucinations that jeopardize patient safety. While Retrieval-Augmented Generation (RAG) mitigates this by grounding outputs in external evidence, existing pipelines struggle with the complex, rapidly evolving nature of oncology. We present CoMeta, a three-level controllable metadata-aware framework optimized for Cancer Patient QA (CPQA). We introduce Clinical Hybrid Semantic-Symbolic Document Retrieval (CHSDR), which synergizes real-time Boolean search via NCBI E-Utilities with semantic retrieval to overcome metadata blindness. Additionally, we propose Semantic Enhanced Overlap Segmentation (SEOS) to prevent contextual fragmentation. Our results demonstrate that CHSDR significantly improves retrieval performance, CoMeta improved the answer accuracy of Claude-3-haiku by 5.24% over chain-of-thought prompting and about 3% over a naive RAG setup. This study highlights the importance of domain-specific query optimization in realizing the full potential of RAG and provides a robust framework for building more reliable CPQA systems.

[Findings] Uncertainty-Calibrated Elastic Alignment for Multimodal Sentiment Analysis with Missing Modalities ▶ *Underline Video*

Kang He, Yuzhe Ding, Rao Fu, Yukang Feng, Kaipeng Zhang, Yiming Liu, Fei Li, Chong Teng and Donghong Ji

Multimodal sentiment analysis (MSA) in real-world scenarios is often challenged by dynamically missing modalities. Existing methods predominantly rely on deterministic imputation and rigid alignment, which compels the model to overfit noise in ambiguous regions while neglecting the decision shift induced by modality inertia. To address these issues, we propose a novel uncertainty-calibrated elastic alignment framework, termed EASE. Specifically, we employ probabilistic imputation to capture cross-modal ambiguity and leverage the estimated uncertainty to drive elastic alignment, thereby adaptively relaxing constraints in ambiguous regions to avoid rigid fitting. Meanwhile, we introduce cross-view predictive consistency constraints to unify discriminative logic across different modality views, stabilizing the decision boundary under modality degradation. Extensive experiments demonstrate that EASE consistently outperforms existing state-of-the-art baselines across multiple benchmarks, exhibiting exceptional robustness particularly under high missing-rate scenarios.

[Long] UniVocal: Unified Speech-Singing Code-Switching Synthesis ▶ *Underline Video*

YuFei Shi, Qian Chen, Wen Wang, Xiangang Li, Zhen-Hua Ling and Yang Ai

We propose UniVocal, a unified framework that implicitly infers vocal modes from text context to pioneer Speech-Singing Code-Switching (SCS) Synthesis, a task where transitions are autonomously driven by textual semantics, akin to seamless human language blending. Unlike single-mode generation or systems relying on switching-control tags, our proposed UniVocal implicitly infers vocal modes solely from text context. To achieve this, we employ a data-efficient two-stage curriculum learning strategy that progressively trains a competitive TTS system to acquire the desired SCS capability. Addressing data scarcity, we introduce a scalable pipeline to synthesize diverse code-switching data that is both semantically and acoustically natural, alongside a new multi-scenario benchmark, SCSBench. To address limitations of semantic tokenizers in capturing acoustic details, we also introduce refined cent token and Chain-of-Thought (CoT) generation for planning prosody before content generation, effectively enhancing empathetic speech generation and singing melody. Experimental results demonstrate that UniVocal achieves state-of-the-art performance on SCSBench while maintaining competitive performance on regular speech and singing tasks. Audio samples are available at https://project-univocal-demo.github.io/demo/. The code and dataset are released at https://github.com/FunAudioLLM/FunResearch/tree/main/UniVocal.

[Long] Vocabulary Shapes Cross-Lingual Variation of Word-Order Learnability in Language Models ▶ *Underline Video*

Jonas Mayer Martins, Jaap Jumelet, Viola Priesemann and Lisa Beinborn

Why do some languages like Czech permit free word order, while others like English do not? We address this question by pretraining transformer language models on a spectrum of synthetic word-order variants of natural languages. We observe that greater word-order irregularity consistently raises model surprisal, indicating reduced learnability. Sentence reversal, however, affects learnability only weakly. A coarse distinction of free- (e.g., Czech and Finnish) and fixed-word-order languages (e.g., English and French) does not explain cross-lingual variation. Instead, the structure of the word and subword vocabulary strongly predicts the model surprisal. Overall, vocabulary structure emerges as a key driver of computational word-order learnability across languages.

[Long] VoxMind: An End-to-End Agentic Spoken Dialogue System ▶ *Underline Video*

Tianle Liang, Yifu Chen, Shengpeng Ji, Yijun Chen, Zhiyang Jia, Jingyu Lu, Fan Zhuo, Xueyi Pu, Yangzhuo Li and Zhou Zhao

Recent end-to-end spoken dialogue models enable natural interaction. However, as user demands become increasingly complex, models that rely solely on conversational abilities often struggle to cope. Incorporating agentic capabilities is therefore essential: by enabling tool use, these models can extend their knowledge boundaries and better solve real-world tasks. Yet, existing research has largely concentrated on core perception and generation, with comparatively limited exploration of such tool-augmented extensions. To bridge this gap, we present

VoxMind, an integrated framework designed to equip end-to-end spoken dialogue models with comprehensive agentic abilities. Leveraging our curated 470-hour AgentChat dataset, we incorporate a "Think-before-Speak" mechanism, enabling the model to internalize structured reasoning as a critical prerequisite for planning and response generation. Furthermore, to mitigate latency bottlenecks caused by large-scale tool integration, we propose a Multi-Agent Dynamic Tool Management architecture. By asynchronously delegating retrieval tasks to an auxiliary agent aligned with the main models reasoning trajectory, this system effectively decouples inference latency from toolset size. Experimental results confirm that VoxMind achieves significant improvements in agent performance: compared with strong baselines, the task completion rate increases from 34.88% to 74.57%, outperforming Gemini-2.5-Pro on spoken agent tasks while preserving general conversational quality. The source code and associated data are publicly available at https://github.com/MM-Speech/VoxMind.

[Long] What Makes a Good Curriculum? Disentangling the Effects of Data Ordering on LLM Mathematical Reasoning ▶ ***Underline Video***

Yaning Jia, Chunhui Zhang, Xingjian Diao, Xiangchi Yuan, Zhongyu Ouyang, Chiyu Ma and Soroush Vosoughi

Curriculum learning (CL), which orders training data from easy to hard, has become a popular strategy for improving reasoning in large language models (LLMs). Yet prior work employs disparate difficulty metrics and training setups, leaving open fundamental questions: When does curriculum help? Which direction, forward or reverse, is better? And does the answer depend on what we measure? We address these questions through a unified offline evaluation framework that decomposes curriculum difficulty into five complementary dimensions: Problem Difficulty, Model Surprisal, Confidence Margin, Predictive Uncertainty, and Decision Variability. Through controlled post-training experiments on mathematical reasoning benchmarks with Llama3.1-8B, Mistral-7B, and Gemma3-4B, we find that: (i) no curriculum strategy dominates universally, the relative effectiveness of forward versus reverse CL depends jointly on model capability and task complexity; (ii) even within a single metric, samples at different difficulty levels produce distinct gains depending on task demands; and (iii) Task-aligned curricula focus on shaping the models final representations and generalization, whereas inner-state curricula modulate internal states such as confidence and uncertainty. Our findings challenge the notion of a universal curriculum strategy and offer actionable guidance across model and task regimes, with some metrics indicating that prioritizing decision-uncertain samples can further enhance learning outcomes.

[Findings] Who's Asking? Simulating Role-Based Questions for Conversational AI Evaluation ▶ ***Underline Video***

Navreet Kaur, Hoda Ayad, Hayoung Jung, Shravika Mittal, Munmun De Choudhury and Tanu Mitra

Language model users often embed personal and social context in their questions. The askers role, implicit in how the question is framed, creates specific needs for an appropriate response. However, most evaluations, while capturing the models capability to respond, often ignore who is asking. This gap is especially critical in stigmatized domains such as opioid use disorder (OUD), where accounting for users contexts is essential to provide accessible, stigma-free responses. We propose CORUS (COmmunity-driven Roles for User-centric Question Simulation), a framework for simulating role-based questions. Drawing on role theory and posts from an online OUD recovery community (r/OpiatesRecovery), we first build a taxonomy of asker roles, patients, caregivers, practitioners. Next, we use it to simulate 15,321 questions that embed each roles goals, behaviors, and experiences. Our evaluations show that these questions are both highly believable and comparable to real-world data. When used to evaluate five LLMs, for the same question but differing roles, we find systematic differences: vulnerable roles, such as patients and caregivers, elicit more supportive responses (+17%) and reduced knowledge content (-19%) in comparison to practitioners. Our work demonstrates how implicitly signaling a users role shapes model responses, and provides a methodology for role-informed evaluation of conversational AI.

Poster Session A (Grand Hall)

11:00-12:30 (Grand Hall)

[Long] A BERTology View of LLM Orchestrations: Token- and Layer-Selective Probes for Efficient Single-Pass Classification ▶ ***Underline Video***

Gonzalo Ariel Meyoyan and Luciano Del Corro

Production LLM systems often rely on separate models for safety and other classification-heavy steps, increasing latency, VRAM footprint, and operational complexity. We instead reuse computation already paid for by the serving LLM: we train lightweight probes on its hidden states and predict labels in the same forward pass used for generation. We frame classification as representation selection over the full tokenxlayer hidden-state tensor, rather than committing to a fixed token or fixed layer (e.g., first-token logits or final-layer pooling). To implement this, we introduce a two-stage aggregator that (i) summarizes tokens within each layer and (ii) aggregates across layer summaries to form a single representation for classification. We instantiate this template with direct pooling, a 100K-parameter scoring-attention gate, and a downcast multi-head self-attention (MHA) probe with up to 35M trainable parameters. Across safety and sentiment benchmarks our probes improve over logit-only reuse (e.g., MULI) and are competitive with substantially larger task-specific baselines, while preserving near-serving latency and avoiding the VRAM and latency costs of a separate guard-model pipeline. Multi-backbone experiments on dense and mixture-of-experts architectures (Llama-3.2-3B, GPT-OSS-20B, Qwen3-30B-A3B) confirm that these findings generalize beyond a single model family.

[Long] Action Boundary Blindness: When LLM Agents Cannot Tell Where One Action Ends and Another Begins ▶ *Underline Video*
Wang zhangyi, Bingnan Yu, JIEXIANG XU and Li Zongze
Large language model (LLM) agents excel at multi-step tasks yet frequently exhibit Action Boundary Blindness—the inability to correctly determine action granularity, scope, and completeness. Grounded in Event Segmentation Theory from cognitive science, we formalize three violation types: granularity confusion, scope creep, and boundary ambiguity. We propose four automatic metrics—Action Boundary Score (ABS), Granularity Alignment Rate (GAR), Scope Violation Rate (SVR), and Boundary-Aware Success Rate (BASR)—requiring no human annotation. Experiments on 1,655 tasks across six benchmarks (-bench, WebArena, ALFWorld, TheAgentCompany, OSWorld) with seven LLMs reveal that: (1) the best model achieves only 0.424 ABS; (2) using a multi-label attribution framework validated by inter-annotator agreement (= 0.78), boundary blindness is the primary failure mode in 37.2% of failures (25.8% as sole cause; 55.9% total involvement including contributing factors); (3) under-action dominates at 48.4%; (4) BASR is consistently 4 points lower than traditional success rate, exposing "lucky successes." Critically, Explicit Boundary Prompting (EBP) improves ABS by 0.08–0.13 across all models, demonstrating that boundary blindness is better characterized as an elicitation gap rather than a fundamental capability limitation—LLMs possess latent boundary perception not activated by default. This finding has implications for alignment and instruction tuning. We validate metrics through state-based cross-validation and human audit, estimating 22% false positive rate from valid alternative paths, with model rankings remaining stable (Spearman = 1.0).

[Findings] Activation Reward Models for Few-Shot Model Alignment ▶ *Underline Video*
Tianning Chai, Chancharik Mitra, Brandon Huang, Gautam Rajendrakumar Gare, Zhiqiu Lin, Assaf Arbelle, Leonid Karlinsky, Rogerio Feris, Trevor Darrell, Deva Ramanan and Roei Herzig
Aligning Large Language Models (LLMs) and Large Multimodal Models (LMMs) to human preferences is crucial for improving their real-world behavior. A common approach is to use reward models that enable reinforcement-learning post-training. However, traditional reward modeling requires finetuning on large preference datasets, limiting adaptability to new preferences. We introduce Activation Reward Models (Activation RMs), the first mechanistic interpretability approach that steers LLM activations to align with few-shot preference data without finetuning. Our method combines activation denoising and output token likelihood scoring, achieving state-of-the-art performance on standard reward modeling benchmarks, surpassing zero-shot, few-shot, and voting-based baselines. We further demonstrate that Activation RMs mitigate reward hacking behaviors and remain robust to noisy exemplars and spurious reward signals. To evaluate this, we propose PreferenceHack, a novel few-shot benchmark testing reward models on reward hacking in a paired preference format, where Activation RMs achieve state-of-the-art performance, surpassing GPT-4o.

[Findings] Aligning with Your Own Voice: Self-Corrected Preference Learning for Hallucination Mitigation in LVLMs ▶ *Underline Video*
Byeonggeuk Lim, JungMin Yun, JuneHyoung Kwon, Kyeonghyun Kim and YoungBin Kim
Large Vision-Language Models (LVLMs) frequently suffer from hallucinations. Existing preference learning-based approaches largely rely on proprietary models to construct preference datasets. We identify that this reliance introduces a distributional mismatch between the proprietary and target models that hinders efficient alignment. To address this, we propose Alignment via VErified Self-correction DPO (AVES-DPO), a framework that aligns LVLMs using in-distribution data derived from the model's intrinsic knowledge. Our approach employs a consensus-based verification mechanism to diagnose diverse hallucinations and guides the model to self-correct, thereby generating preference pairs strictly compatible with its internal distribution. Extensive experiments demonstrate that AVES-DPO surpasses existing baselines in hallucination mitigation while requiring only 5.2k samples.

[Findings] Alignment Data Map for Efficient Preference Data Selection and Diagnosis ▶ *Underline Video*
Seohyeong Lee, Eunwon Kim, Hwaran Lee and Buru Chang
Human preference data is essential for aligning large language models (LLMs) with human values, but collecting such data is often costly and inefficient-motivating the need for efficient data selection methods that reduce annotation costs while preserving alignment effectiveness. To address this issue, we propose Alignment Data Map, a data analysis tool for identifying and selecting effective preference data. We first evaluate alignment scores of the preference data by LLM-as-a-judge, explicit reward model, and reference-based approaches. The Alignment Data Map considers both response quality and inter-response variability based on the alignment scores. From our experimental findings, training on only 33% of samples that exhibit high-quality and low-variability, achieves comparable or superior alignment performance on MT-Bench, Evol-Instruct, and AlpacaEval, compared to training with the full dataset. In addition, Alignment Data Map detects potential label misannotations by analyzing correlations between annotated labels and alignment scores, improving annotation accuracy. The implementation is available at https://github.com/01choco/Alignment-Data-Map.

[Long] Android Coach: Improve Online Agentic Training Efficiency with Single State Multiple Actions ▶ *Underline Video*
Guo Gan and Cong Chen
Online reinforcement learning (RL) serves as an effective method for enhancing the capabilities of Android agents. However, guiding agents to learn through online interaction is prohibitively expensive due to the high latency of emulators and the sample inefficiency of existing RL algorithms. We identify a fundamental limitation in current approaches: the Single State Single Action paradigm, which updates the policy with one-to-one state-action pairs from online one-way rollouts without fully exploring each costly emulator state. In this paper, we propose

Android Coach, a novel framework that shifts the training paradigm to Single State Multiple Actions, allowing the agent to sample and utilize multiple actions for a single online state. We enable this without additional emulator overhead by learning a critic that estimates action values. To ensure the critic serves as a reliable coach, we integrate a process reward model and introduce a group-wise advantage estimator based on the averaged critic outputs. Extensive experiments demonstrate the effectiveness and efficiency of Android Coach: it achieves 7.5% and 8.3% success rate improvements on AndroidLab and AndroidWorld over UI-TARS-1.5-7B, and attains 1.4x higher training efficiency than Single State Single Action methods PPO and GRPO at matched success rates.

[Long] Attention Weights as an Indicator: Analyzing and Improving Document Utilization in Retrieval-Augmented Generation ▶ *Underline Video*

Jing Jin, Yuhan Song, Wen Luo and Houfeng Wang

The generation of Retrieval-Augmented Generation (RAG) models is affected by factors such as the quality and order of input documents, indicating that their ability to utilize documents remains underdeveloped. This ability encompasses not only identifying useful documents from inputs but also minimizing positional bias and filtering irrelevant documents. To achieve this, key challenges include the model's internal estimation of document importance and positional bias. In this paper, we conduct a comprehensive study on the properties of attention weights, examining the impact of factors like aggregation methods, document quality, document position, token type, and so on. Based on our findings, we propose strategies to enhance document utilization from three perspectives: document ranking, placement, and filtering. Comprehensive experiments show that our method outperforms baselines and improves document utilization effectiveness in a training-free manner.

[Long] Audio Jailbreak: An Open Comprehensive Benchmark for Jailbreaking Large Audio-Language Models ▶ *Underline Video*

Zirui Song, Qian Jiang, Mingxuan Cui, Mingzhe Li, Lang Gao, Zeyu Zhang, Zixiang Xu, Yanbo Wang, Guanxian Ouyang, Zhenhao Chen and Xiuying Chen

The rise of Large Audio-Language Models (LAMs) brings both potential and risks, as their audio outputs may contain harmful or unethical content. However, current research lacks a systematic, quantitative evaluation of LAM safety, especially against jailbreak attacks, which are challenging due to the temporal and semantic nature of speech. To bridge this gap, we introduce AJailBench, the first benchmark specifically designed to evaluate jailbreak vulnerabilities in LAMs. We begin by constructing -Base, a dataset of 1,495 adversarial audio prompts spanning 10 policy-violating categories. Using this dataset, we evaluate several state-of-the-art LAMs and reveal that none exhibit consistent robustness across attacks. To further strengthen jailbreak testing and simulate more realistic attack conditions, we propose a method to generate dynamic adversarial variants. Our Audio Perturbation Toolkit (APT) applies targeted distortions across time, frequency, and amplitude domains. To preserve the original jailbreak intent, we enforce a semantic consistency constraint and employ Bayesian optimization to efficiently search for perturbations that are both subtle and highly effective. This results in AJailBench-APT+, an extended dataset of optimized adversarial audio samples. Our findings demonstrate that even small, semantically preserved perturbations can significantly reduce the safety performance of leading LAMs, underscoring the need for more robust and semantically aware defense mechanisms. We release AJailBench to facilitate future research: https://anonymous.4open.science/r/AudioJailbreak-4262/

[Findings] AUDITA: A New Dataset to Audit Humans vs. AI Skill at Audio QA ▶ *Underline Video*

Tasnim Kabir, Dmytro Kurdydyk, Aadi Palnitkar, Liam Dorn, Ahmed Haj Ahmed and Jordan Lee Boyd-Graber

Existing audio question answering benchmarks largely emphasize sound event classification or caption-grounded queries, often enabling models to succeed through shortcut strategies, short-duration cues, lexical priors, dataset-specific biases, or even bypassing audio via metadata and captions rather than genuine reasoning Thus, we present AUDITA (Audio Understanding from Diverse Internet Trivia Authors), a large-scale, real-world benchmark to rigorously evaluate audio reasoning beyond surface-level acoustic recognition. AUDITA comprises carefully curated, human-authored trivia questions grounded in real-world audio, designed to stress robust auditory reasoning through challenging distractors and long-range temporal dependencies, using probing queries that cannot be answered from isolated text or sound cues alone. Human average accuracy of 32.13% shows both the challenge of the task while demonstrating meaningful comprehension of the audio. In stark contrast, state-of-the- art audio question answering models perform poorly, with average accuracy below 8.86%. Beyond raw accuracy, we apply Item Response Theory (IRT) to estimate latent proficiency, question difficulty, and expose systematic deficiencies of the models and data.

[Long] AutoRAN: Automated Hijacking of Safety Reasoning in Large Reasoning Models ▶ *Underline Video*

Jiacheng Liang, Tanqiu Jiang, Yuhui Wang, Rongyi Zhu, Fenglong Ma and Ting Wang

This paper presents AutoRAN, the first framework to automate the hijacking of internal safety reasoning in large reasoning models (LRMs). At its core, AutoRAN pioneers an execution simulation paradigm that leverages a weaker but less-aligned model to simulate execution reasoning for initial hijacking attempts and iteratively refine attacks by exploiting reasoning patterns leaked through the target LRM's refusals. This approach steers the target model to bypass its own safety guardrails and elaborate on harmful instructions. We evaluate AutoRAN against state-of-the-art LRMs, including GPT-o3/o4-mini and Gemini-2.5-Flash, across multiple benchmarks (AdvBench, HarmBench, and StrongReject). Results show that AutoRAN achieves approaching 100% success rate within one or few turns, effectively neutralizing reasoning-based defenses even when evaluated by robustly aligned external models. This work reveals that the transparency of the reasoning process itself creates a critical and exploitable attack surface, highlighting the urgent need for new defenses that protect models' reasoning traces rather than merely their final outputs.

[Long] Bayesian Social Deduction with Graph-Informed Language Models ▶ *Underline Video*

Shahab Rahimirad, Guven Gergerli, Lucia Romero, Angela Qian, Matthew Lyle Olson, Simon Stepputtis and Joseph Campbell

Social reasoning—inferring unobservable beliefs and intentions from partial observations of other agents—remains a challenging task for large language models (LLMs). We evaluate the limits of current reasoning language models in the social deduction game Avalon and find that while the largest models demonstrate strong performance, they require extensive test-time inference and degrade sharply when distilled to smaller, real-time-capable variants. To address this, we introduce a hybrid reasoning framework that externalizes belief inference to a structured probabilistic model, while using an LLM for language understanding and interaction. Our approach achieves competitive performance with much larger models in Agent-Agent play and, notably, is the first language agent to defeat human players in a controlled study—achieving a 67% win rate and receiving higher qualitative ratings than both reasoning baselines and human teammates. We release code, models, and a dataset to support future work on social reasoning in LLM agents.

[Findings] Belief in Authority: Impact of Authority in Multi-Agent Evaluation Framework ▶ *Underline Video*

Junhyuk Choi, Jeongyoun Kwon, Heeju Kim, Haeun Cho, Hayeong Jung, Sehee Min and Bugeun Kim

Multi-agent systems utilizing large language models often assign authoritative roles to improve performance, yet the impact of authority bias on agent interactions remains underexplored. We present the first systematic analysis of role-based authority bias in free-form multi-agent evaluation using ChatEval. Applying French and Raven's power-based theory, we classify authoritative roles into legitimate, referent, and expert types and analyze their influence across 12-turn conversations. Experiments with GPT-4o and DeepSeek R1 reveal that Expert and Referent power roles exert stronger influence than Legitimate power roles. Crucially, authority bias emerges not through active conformity by general agents, but through authoritative roles consistently maintaining their positions while general agents demonstrate flexibility. Furthermore, authority influence requires clear position statements, as neutral responses fail to generate bias. These findings provide key insights for designing multi-agent frameworks with asymmetric interaction patterns.

[Long] Beyond Code Pairs: Dialogue-Based Data Generation for LLM Code Translation ▶ *Underline Video*

Le Chen, Nuo Xu, Winson Chen, Bin Lei, Pei-Hung Lin, Dunzhi Zhou, Rajeev Thakur, Caiwen Ding, Ali Jannesari and Chunhua Liao

Large language models (LLMs) have shown remarkable capabilities in code translation, yet their performance deteriorates in low-resource programming domains such as Fortran and emerging frameworks like CUDA, where high-quality parallel data are scarce. We present an automated dataset generation pipeline featuring a dual-LLM Questioner-Solver design that incorporates external knowledge from compilers and runtime feedback. Beyond traditional source-target code pair datasets, our approach additionally generates (1) verified translations with unit tests for assessing functional consistency, and (2) multi-turn dialogues that capture the reasoning process behind translation refinement. Applied to Fortran to C++ and C++ to CUDA, the pipeline yields 3.64k and 3.93k dialogues, respectively. Fine-tuning on this data yields dramatic improvements in functional correctness, boosting unit test success rates by over 56% on the challenging C++-to-CUDA task. We show that the generated data enables a 7B open-weight model to significantly outperform larger proprietary systems on key metrics like compilation success.

[Findings] Beyond Monolithic Rewards: Hybrid Multi-Aspect Reward Optimization ▶ *Underline Video*

Radha Gulhane and Sathish Reddy Indurthi

Reinforcement learning optimization policies have traditionally relied on a single reward mechanism, most commonly a model-based reward. Such monolithic rewards often lack confidence calibration across domain-specific tasks and fail to capture diverse aspects of model responses. This approach requires extensive data annotation and reward model training, which is particularly challenging for multimodal models. In this work, we propose and provide a thorough study of hybrid reward and multi-aspect reward modeling. For accuracy and confidence calibration, we introduce a hybrid reward modeling framework that integrates complementary reward paradigms: model-based rewards, in which a learned reward model predicts scalar or vector scores, and rule-based reward, in which domain-specific heuristics provide explicit correctness signals with confidence. Beyond accuracy, we further incorporate multi-aspect rewards to enforce instruction adherence and introduce a generalized length-penalty reward to stabilize training and improve performance. Experiments demonstrate that this approach significantly enhances reasoning capabilities: our best-performing 3B model achieves an average improvement of ~9.5% across multimodal benchmarks, with a notable ~16% gain in mathematical reasoning tasks.

[Findings] Beyond Reasoning Gains: Mitigating General-Capability Forgetting in Large Reasoning Models ▶ *Underline Video*

Hoang Phan, Xianjun Yang, Yuanshun Yao, Jingyu Zhang, Shengjie Bi, Xiaocheng Tang, Madian Khabsa, Lijuan Liu and Deren Lei

Reinforcement learning with verifiable rewards (RLVR) has delivered impressive gains in mathematical and multimodal reasoning and has become a standard post-training paradigm for contemporary language and vision-language models. However, the RLVR recipe introduces a significant risk of capability regression, where models forget foundational skills after prolonged training without employing regularization strategies. We empirically confirm this concern, observing that open-source reasoning models suffer performance degradation on core capabilities such as perception and faithfulness. While imposing regularization terms like KL divergence can help prevent deviation from the base model, these terms are calculated on the current task, thus they do not guarantee broader knowledge. Meanwhile, commonly used experience replay across heterogeneous domains makes it nontrivial to decide how much training focus each objective should receive. To address this, we propose RECAP, a replay strategy with dynamic objective reweighting for general knowledge preservation. Our reweighting mechanism adapts in an online manner using short-horizon signals of convergence and instability, shifting the post-training focus away

from saturated objectives and toward underperforming or volatile ones. Our method is end-to-end and readily applicable to existing RLVR pipelines without training additional models or heavy tuning. Extensive experiments on benchmarks based on Qwen2.5-VL-3B and Qwen2.5-VL-7B demonstrate the effectiveness of our method, which not only preserves general capabilities but also improves reasoning by enabling more flexible trade-offs among in-task rewards.

[Findings] Beyond Single Plots: A Benchmark for Question Answering on Multi-Charts ▶ *Underline Video*

Azher Ahmed Efat, Seok Hwan Song and Wallapak Tavanapong

Charts are widely used to present complex information. Deriving meaningful insights in real-world contexts often requires interpreting multiple related charts together. Research on understanding multi-chart images has not been extensively explored. We introduce PolyChartQA, a mid-scale dataset specifically designed for question answering over multi-chart images. PolyChartQA comprises 534 multi-chart images (with a total of 2,297 sub-charts) sourced from peer-reviewed computer science research publications and 2,694 QA pairs. We evaluate the performance of nine state-of-the-art Multimodal Language Models (MLMs) on PolyChartQA across question type, difficulty, question source, and key structural characteristics of multi-charts. Our results show a 27.4% LLM-based accuracy (L-Accuracy) drop on human-authored questions compared to MLM-generated questions, and a 5.39% L-accuracy gain with our proposed prompting method.

[Findings] Beyond Static Toolsets: Self-Evolving LLM Tool Agents via Continual Documentation Adaptation ▶ *Underline Video*

Bin Wu, Edgar Meij and Emine Yilmaz

Large language models (LLMs) increasingly act as tool-using agents, and existing methods for evaluating and optimizing tool usage by LLMs typically assume a static tool environment with fixed APIs and documentation. In practice, toolsets evolve as tools are added, changed, or deprecated, introducing instability for agents that must retain prior competence while adapting to new capabilities. We formalize this challenge as the stability–adaptation dilemma. To address it, we propose ContDa, a continual documentation adaptation framework that provides a generalizable solution to this problem, enabling LLM agents to self-evolve by updating tool documentation. ContDa combines relation-guided exploration, which leverages functionally related existing tools as anchors to probe and identify new tool capabilities, with relation-aware adjustment that organizes overlapping tools and explicitly encodes usage preferences and fallback options among them. We then introduce complementary metrics that disentangle performance from stability and adaptation. Experiments across three evolution patterns on dynamic extensions of StableToolBench and RestBench show that ContDa consistently improves average performance by enhancing the discovery of new capabilities while incurring only limited loss of previously solved tasks, demonstrating documentation adaptation as an effective and lightweight mechanism for robust tool use in evolving environments. Our code is available at https://github.com/Bingo-W/ContDa.

[Long] Beyond the Last Frame: Process-aware Evaluation for Generative Video Reasoning ▶ *Underline Video*

Yifan Li, YuKai Gu, Yingqian Min, Zikang Liu, Yifan Du, Kun Zhou, Min Yang, Xin Zhao and Minghui Qiu

Recent breakthroughs in video generation have demonstrated an emerging capability termed Chain-of-Frames (CoF) reasoning, where models resolve complex tasks through the generation of continuous frames. While these models show promise for Generative Video Reasoning (GVR), existing evaluation frameworks often rely on single-frame assessments, which can lead to outcome-hacking, where a model reaches a correct conclusion through an erroneous process. To address this, we propose a process-aware evaluation paradigm. We introduce VIPER, a comprehensive benchmark spanning 16 tasks across temporal, structural, symbolic, spatial, physics, and planning reasoning. Furthermore, we propose Process-outcome Consistency (POC@r), a new metric that utilizes VLM-as-Judge with a hierarchical rubric to evaluate both the validity of the intermediate steps and the final result. Our experiments reveal that state-of-the-art video models achieve POC@1.0 only about 20% and exhibit a significant outcome-hacking. We further explore the impact of test-time scaling and sampling robustness, highlighting a substantial gap between current video generation and true generalized visual reasoning. Our benchmark are released at https://github.com/RUCAIBox/VIPER.

[Long] Beyond Value Benchmarks: Measuring Value-Structure Alignment in Large Language Models via Symmetric Q-Sorts ▶ *Underline Video*

Jingting Zheng, Yuqi Ren, Linhao Yu, Yongqi Leng and Deyi Xiong

Large Language Models (LLMs) are increasingly deployed in contexts requiring complex moral reasoning and value trade-offs. However, existing evaluations typically rely on item-level behavioral metrics, which fail to capture how models structurally prioritize competing values as a cohesive system. To address this, we propose a symmetric human-LLM evaluation framework, grounded in Q methodology, to measure value-structure alignment. Under our protocol, humans and models sort an identical 140-item moral statement set into a shared nine-column forced distribution; for LLMs, we elicit strict rankings and deterministically map them to Q-sort buckets. Using a human reference sample (N=35), we establish a stable three-factor reference geometry specific to this instrument and sample. We evaluate 12 LLMs across four model families via 240 replicated Q-sorts at two temperature settings, quantifying structural alignment via Procrustes similarity () and RSA-based Spearman correlation (). Our results reveal significant cross-family heterogeneity, model-specific sensitivity to generation stochasticity and localized misalignment, which demonstrate that favorable global scores can obscure underlying regional distortions. While rank- and bucket-based analyses remain highly consistent, prompt phrasing introduces notable variance. Ultimately, assessing value-structure alignment provides a crucial structural complement to traditional itemwise moral benchmarks.

[Long] Biomedical Question Answering via Multi-Level Summarization on a Local Knowledge Graph ▶ *Underline Video*

Lingxiao Guan, Yuanhao Huang and Jie Liu

In Question Answering (QA), Retrieval Augmented Generation (RAG) has revolutionized performance in various domains. However, how to effectively capture multi-document relationships remains an open question. This is particularly critical for biomedical tasks due to their reliance on information spread across multiple documents. In this work, we propose a novel method CLAIMS, which utilizes propositional claims to construct a local knowledge graph from retrieved documents. Summaries are then derived via layerwise summarization from the knowledge graph to contextualize a small language model to perform QA. The structured summaries effectively capture explicit and implicit relationships between entities in the documents, thus having a more comprehensive context to provide to LLMs. CLAIMS achieved comparable or superior performance over RAG baselines on several biomedical QA benchmarks. We also evaluated its generalizability and each individual step of our approach with a targeted set of metrics, demonstrating its effectiveness.

[Findings] Bridging the Culture Gap: A Framework for LLM-Driven Socio-Cultural Localization of Math Word Problems in Low-Resource Languages ► *Underline Video*

Israel Abebe Azime, Tadesse Destaw Belay, Dietrich Klakow, Philipp Slusallek and Anshuman Chhabra

Large language models (LLMs) have demonstrated significant capabilities in solving mathematical problems expressed in natural language. However, multilingual and culturally-grounded mathematical reasoning in low-resource languages lags behind English due to the scarcity of socio-cultural task datasets that reflect accurate native entities such as person names, organization names, and currencies. Existing multilingual benchmarks are predominantly produced via translation and typically retain English-centric entities, owing to the high cost associated with human annotator-based localization. Moreover, automated localization tools are limited, and hence, truly localized datasets remain scarce. To bridge this gap, we introduce a framework for LLM-driven cultural localization of math word problems that automatically constructs datasets with native names, organizations, and currencies from existing sources. We find that translated benchmarks can obscure true multilingual math ability under appropriate socio-cultural contexts. Through extensive experiments, we also show that our framework can help mitigate English-centric entity bias and improve robustness when native entities are introduced across various languages.

[Long] Bringing Real-World Relations into Video Generation with Graph-Structured Knowledge ► *Underline Video*

Joonhyung Park, Jaeyun Song, Sihwan Park and Eunho Yang

Recent proprietary video generation models have demonstrated remarkable proficiency in synthesizing highly realistic videos from textual instructions. Most open-source text-to-video models, however, still struggle to accurately simulate real-world physics and dynamic entity interactions. Existing approaches rely on scaling laws and large-scale, high-quality video datasets to implicitly learn physical dynamics, yet this paradigm is constrained by prohibitive costs and the burdensome demands of data curation. Motivated by this, we propose a novel framework that integrates graph-structured temporal knowledge into video latent diffusion models to enhance compositional generation and interaction fidelity. Our framework constructs video scene graphs specifically designed to capture entity relationships, temporal dynamics, and global scene context. These graph-structured representations guide the generation process through cross-attention mechanisms. Additionally, we introduce Graph-Aligned Denoising Loss (GADL), a training objective that ensures adherence to conditioned graphs by incorporating node modification tasks within the denoising process, leveraging synchronized edited video-graph pairs. Comprehensive evaluations demonstrate that incorporating graph-structured knowledge significantly enhances compositionality and the accurate portrayal of real-world interactions in generated videos.

[Findings] BrowseConf: Confidence-Guided Test-Time Scaling for Web Agents ► *Underline Video*

Litu Ou, Kuan Li, Huifeng Yin, Liwen Zhang, Zhongwang Zhang, Xixi Wu, Rui Ye, Zile Qiao, Yong Jiang, Pengjun Xie, Fei Huang and Jingren Zhou

Confidence in LLMs is a useful indicator of model uncertainty and answer reliability. Existing work mainly focused on single-turn scenarios, while research on confidence in complex multi-turn interactions is limited. In this paper, we investigate whether LLM-based search agents have the ability to communicate their own confidence through verbalized confidence scores after long sequences of actions, a significantly more challenging task compared to outputting confidence in a single interaction. Experimenting on open-source agentic models, we first find that models exhibit much higher task accuracy at high confidence while having near-zero accuracy when confidence is low. Based on this observation, we propose Test-Time Scaling (TTS) methods that use confidence scores to determine answer quality, encourage the model to try again until reaching a satisfactory confidence level. Results show that our proposed methods significantly reduce token consumption while demonstrating competitive performance compared to baseline fixed budget TTS methods.

[Long] CAMEC: Complexity-Aware Multi-Expert Collaboration for Reliable Chinese Medical Question Answering ► *Underline Video*

Yukang Wu, Xiyuan Jia, Jiayi Wu, HongChen Yu, Yuhan Qiu and Wu Guohua

Large language models (LLMs) are promising for medical question answering (QA) but remain unreliable in Chinese clinical settings due to hallucinations, weak factual grounding, and difficulty handling clinically complex cases. We propose CAMEC (Complexity-Aware Multi-Expert Collaboration), a framework that combines hierarchical medical adaptation with complexity-aware expert routing for reliable Chinese medical QA. We adopt a three-stage LoRA-based supervised fine-tuning pipeline for domain adaptation, instruction following, and clinical reasoning. At inference, CAMEC routes each query by predicted complexity and selectively recruits three experts: an internal chain-of-thought (CoT) expert, a retrieval-augmented expert over a dense medical vector database, and a knowledge graph (KG) expert over a structured medical knowledge base. An LLM-as-a-Judge module evaluates and critiques expert reports, iteratively refining them into a consensus answer. Experiments on four Chinese medical benchmarks show that CAMEC consistently outperforms strong general and

medical LLM baselines, achieving 78.86% (CMExam), 84.15% (MedQA-CN), 78.51% (CMMLU-Med), and 74.40% (CMB-exam), with consistent absolute improvements over the previous state-of-the-art HuatuoGPT-o1-7B across all benchmarks. The complexity-aware router reduces expert invocations and inference cost, making CAMEC both highly effective and computationally efficient.

[Findings] CausalDetox: Causal Head Selection and Intervention for Language Model Detoxification ► *Underline Video*

Yian Wang, Yuen Chen, Agam Goyal and Hari Sundaram

Large language models (LLMs) frequently generate toxic content, posing significant risks for safe deployment. Current mitigation strategies often degrade generation quality or require costly human annotation. We propose CausalDetox, a framework that identifies and intervenes on the specific attention heads causally responsible for toxic generation. Using the Probability of Necessity and Sufficiency (PNS), we isolate a minimal set of heads that are necessary and sufficient for toxicity. We utilize these components via two complementary strategies: (1) Local Inference-Time Intervention, which constructs dynamic, input-specific steering vectors for context-aware detoxification, and (2) PNS-Guided Fine-Tuning, which permanently unlearns toxic representations. We also introduceParaTox, a novel benchmark of aligned toxic/non-toxic sentence pairs enabling controlled counterfactual evaluation. Experiments on ToxiGen, ImplicitHate, and ParaDetox show that CausalDetox achieves up to 5.34% greater toxicity reduction compared to baselines while preserving linguistic fluency, and offers a 7x speedup in head selection.

[Findings] CLewR: Curriculum Learning with Restarts for Machine Translation Preference Learning ► *Underline Video*

Alexandra Dragomir, Florin Brad and Radu Tudor Ionescu

Large language models (LLMs) have demonstrated competitive performance in zero-shot multilingual machine translation (MT). Some follow-up works further improved MT performance via preference optimization, but they leave a key aspect largely underexplored: the order in which data samples are given during training. We address this topic by integrating curriculum learning into various state-of-the-art preference optimization algorithms to boost MT performance. We introduce a novel curriculum learning strategy with restarts (CLewR), which reiterates easy-to-hard curriculum multiple times during training to effectively mitigate the catastrophic forgetting of easy examples. We demonstrate consistent gains across several model families (Gemma2, Qwen2.5, Llama3.1) and preference optimization techniques. We publicly release our code at https://github.com/alexandra-dragomir/CLewR.

[Findings] CodeWiki: Evaluating AIs Ability to Generate Holistic Documentation for Large-Scale Codebases ► *Underline Video*

Anh Nguyen Hoang, Minh Le-Anh, Bach Le and Nghi D. Q. Bui

Comprehensive software documentation is crucial yet costly to produce. Despite recent advances in large language models (LLMs), generating holistic, architecture-aware documentation at the repository level remains challenging due to complex and evolving codebases that exceed LLM context limits. Existing automated methods struggle to capture rich semantic dependencies and architectural structure. We present CodeWiki, a unified framework for automated repository-level documentation across seven mainstream programming languages. CodeWiki combines top-down hierarchical decomposition with a divide-and-conquer agent system to preserve architectural context and scale documentation generation, and a bottom-up synthesis that integrates textual descriptions with visual artifacts such as architecture and data-flow diagrams. We also introduce CodeWikiBench, a benchmark with hierarchical rubrics and LLM-based evaluation protocols. Experiments show that CodeWiki achieves a 68.79% quality score with proprietary models, outperforming the closed-source DeepWiki baseline by 4.73%, with especially strong gains on scripting languages. CodeWiki is released as open source to support future research.

[Long] COMPASS: A Framework for Evaluating Organization-Specific Policy Alignment in LLMs ► *Underline Video*

Dasol Choi, DongGeon Lee, Brigitta Jesica Kartono, Helena Berndt, Taeyoun Kwon, Joonwon Jang, Haon Park, Hwanjo Yu and Minsuk Kahng

As large language models are deployed in high-stakes enterprise applications, from healthcare to finance, ensuring adherence to organization-specific policies has become essential. Yet existing safety evaluations focus exclusively on universal harms. We present COMPASS (Company/Organization Policy Alignment Assessment), the first systematic framework for evaluating whether LLMs comply with organizational allowlist and denylist policies. We apply COMPASS to eight diverse industry scenarios, generating and validating 5,920 queries that test both routine compliance and adversarial robustness through strategically designed edge cases. Evaluating seven state-of-the-art models, we uncover a fundamental asymmetry: models reliably handle legitimate requests (>95% accuracy) but catastrophically fail at enforcing prohibitions, refusing only 13-40% of adversarial denylist violations. These results demonstrate that current LLMs lack the robustness required for policy-critical deployments, establishing COMPASS as an essential evaluation framework for organizational AI safety.

[Long] Compiling Activation Steering into Weights via Null-Space Constraints for Stealthy Backdoors ► *Underline Video*

Rui Yin, Tianxu Han, Naen Xu, Changjiang Li, Ping He, Chunyi Zhou, Jun Wang, Zhihui Fu, Tianyu Du, Jinbao Li and Shouling Ji

Safety-aligned large language models (LLMs) are increasingly deployed in real-world pipelines, yet this deployment also enlarges the supply-chain attack surface: adversaries can distribute backdoored checkpoints that behave normally under standard evaluation but jailbreak when a hidden trigger is present. Recent post-hoc weight-editing methods offer an efficient approach to injecting such backdoors by directly modifying model weights to map a trigger to an attacker-specified response. However, existing methods typically optimize a token-level mapping that forces an affirmative prefix (e.g., “Sure”), which does not guarantee sustained harmful output—the model may begin with apparent agreement yet revert to safety-aligned refusal within a few decoding steps. We address this reliability gap by shifting the backdoor objective from surface tokens to internal representations. We extract a steering vector that captures the difference between

compliant and refusal behaviors, and compile it into a persistent weight modification that activates only when the trigger is present. To preserve stealthiness and benign utility, we impose a null-space constraint so that the injected edit remains dormant on clean inputs. The method is efficient, requiring only a small set of examples and admitting a closed-form solution. Across multiple safety-aligned LLMs and jailbreak benchmarks, our method achieves high triggered attack success while maintaining non-triggered safety and general utility.

[Long] CoreGaze: Core Subgraph-Driven Visual Gaze Diffusion for Training-Free Referring Multimodal Large Language Models ▶ *Underline Video*

Xiaoyang Yi, Jing Chen, Yuru Bao and Jian Zhang

Referring multimodal large language models enable users to ground queries to specific image regions via spatial prompts, supporting fine-grained referring dialogue. However, existing methods rely on extensive fine-tuning to mitigate attention distraction, which incurs high computational costs and limits adaptability. Without sufficient training data, irrelevant regions in single images easily divert model focus, leading to redundant outputs or hallucinations. To address this, we propose CoreGaze, a training-free framework that simulates human visual gaze diffusion for fine-grained comprehension. First, CoreGaze constructs a sparse semantic graph from visual tokens, modeling region-wise affinities via thresholded similarity. It then maps the user's visual prompt to a core subgraph with amplified initial influence, which drives a degree-normalized diffusion process using restart-equipped random walks to propagate relevance to contextual neighborhoods. This process prunes irrelevant tokens while preserving user-indicated targets and semantically linked context, distilling a focused yet comprehensive subgraph. Finally, CoreGaze fuses this subgraph with prompt tokens in the frozen large language model decoder, facilitating fine-grained referring generation. Experimental results show that CoreGaze achieves outstanding performance in multiple referring dialogue tasks, showcasing its effectiveness.

[Findings] CURaTE: Continual Unlearning in Real Time with Ensured Preservation of LLM Knowledge ▶ *Underline Video*

Seyun Bae, Seokhan Lee and Eunho Yang

The inability to filter out in advance all potentially problematic data from the pre-training of large language models has given rise to the need for methods for unlearning specific pieces of knowledge after training. Existing techniques overlook the need for continuous and immediate action, causing them to suffer from degraded utility as updates accumulate and protracted exposure of sensitive information. To address these issues, we propose Continual Unlearning in Real Time with Ensured Preservation of LLM Knowledge (CURaTE). Our method begins by training a sentence embedding model on a dataset designed to enable the formation of sharp decision boundaries for determining whether a given input prompt corresponds to any stored forget requests. The similarity of a given input to the forget requests is then used to determine whether to answer or return a refusal response. We show that even with such a simple approach, not only does CURaTE achieve more effective forgetting than existing methods, but by avoiding modification of the language model parameters, it also maintains near perfect knowledge preservation over any number of updates and is the only method capable of continual unlearning in real-time.

[Long] d-TreeRPO: Towards More Reliable Policy Optimization for Diffusion Language Models ▶ *Underline Video*

Leyi Pan, Shuchang Tao, Yunpeng Zhai, Zheyu Fu, Liancheng Fang, Minghua He, Lingzhe Zhang, Zhaoyang Liu, Bolin Ding, Aiwei Liu and Lijie Wen

Reinforcement learning (RL) is pivotal for enhancing the reasoning capabilities of diffusion large language models (dLLMs). However, existing dLLM policy optimization methods suffer from two critical reliability bottlenecks: (1) reward sparsity, arising from coarse or unverifiable signals that impede accurate advantage calculation; and (2) their probability estimates do not account for the gap to the unbiased expectation over all decoding orders, which are intractable to compute. To mitigate these issues, we propose d-TreeRPO, a reliable RL framework for dLLMs that leverages tree-structured rollouts and bottom-up advantage computation based on verifiable outcome rewards to provide fine-grained and verifiable step-wise reward signals. Furthermore, we provide a theoretical proof demonstrating that increasing prediction confidence effectively minimizes the gap between unbiased expected prediction probabilities and its single-step forward pass estimate. Guided by this analysis, we introduce a time-scheduled self-distillation loss during training that enhances prediction confidence in later training stages, thereby enabling more accurate probability estimation and better performance. Experiments demonstrate that d-TreeRPO outperforms existing baselines and achieves significant improvements across multiple reasoning benchmarks. Specifically, it achieves +86.2% on Sudoku, +51.6% on Countdown, +4.5% on GSM8K, and +5.3% on Math500 compared to the base model.

[Long] Decisive: Guiding User Decisions with Optimal Preference Elicitation from Unstructured Documents ▶ *Underline Video*

Akriti Jain, Anish Mulay, Divyansh Verma, Aishani Pandey, Pritika Ramu and Aparna Garimella

Decision-making is a cognitively intensive task that requires synthesizing relevant information from multiple unstructured sources, weighing competing factors, and incorporating subjective user preferences. Existing methods, including large language models and traditional decision-support systems, fall short: they often overwhelm users with information or fail to capture nuanced preferences accurately. We present Decisive, an interactive decision-making framework that combines document-grounded reasoning with Bayesian preference inference. Our approach grounds decisions in an objective option-scoring matrix extracted from source documents, while actively learning a user's latent preference vector through targeted elicitation. Users answer pairwise tradeoff questions adaptively selected to maximize information gain over the final decision. This process converges efficiently, minimizing user effort while ensuring recommendations remain transparent and personalized. Through extensive experiments, we demonstrate that our approach significantly outperforms both general-purpose LLMs and existing decision-making frameworks achieving up to 20% improvement in decision accuracy over strong baselines across domains.

[Findings] Decomposing Unitization and Typing for Efficient and Consistent Span-Bound Concept Annotation ▶ *Underline Video*
Nupoor Gandhi, Michael Bada and Emma Strubell
In specialized domains that require expert annotators and high inter-annotator agreement, high-quality datasets with span-bound semantic concept annotations remain expensive to develop. Substantial resources are typically spent on unitizing, the task of identifying precise span boundaries for entity mentions. Unitizing is a significant source of inter-annotator disagreement, a poor use of expensive domain expertise, and very time-consuming. We propose a lighter annotation procedure that concentrates manual efforts on typed position annotations, marking positions in the text that overlap with mentions of each entity type, abstracting away span boundary decisions. With as few as 100-200 example sentences, we train span boundary detection models to unitize typed position annotations. Through evaluation over three datasets: CRAFT (biomedical), GENIA (molecular biology), and POLIANNA (climate/energy policy text), we demonstrate that (1) annotating typed positions in the text instead of full concept annotation is a more efficient use of time in low-resource settings, and (2) model-inferred span boundaries result in higher agreement at both the annotator training and corpus annotation phases, without sacrificing utility.

[Short] Dialogue is the Plan: From Interface to Joint Action in Agentic AI ▶ *Underline Video*
Mert Inan, Malihe Alikhani and Anthony Sicilia
Large Language Model agents can seemingly plan and act, yet their language use is often treated primarily as an interface for instructing actions and reporting results. We argue that this framing is one important cause of recurrent coordination failures in human-facing and multiagent settings, including ungrounded assumptions, silent goal misalignment, brittle protocol adherence, and failures to maintain or update shared dialogue state over time, a limitation previously linked to the absence of explicit common ground tracking in collaborative systems (Geib et al., 2022). Drawing from classical dialogue system research on joint action, common ground, grounding, repair, and incremental processing, we re-frame dialogue as part of the planning loop itself (rather than its output). We distill this re-framing into concrete implications for agentic architecture and evaluation, including explicit representations of shared commitments, clarification as a first class action available to the policy, and process metrics that approximate grounding behavior, repair, and commitment formation rather than task completion alone. We lastly discuss how dialogue-centered requirements can inform standards and governance for safe deployment of agentic systems.

[Long] Discourse Coherence and Response-Guided Context Rewriting for Multi-Party Dialogue Generation ▶ *Underline Video*
Zhiyu Cao, Peifeng Li and Qiaoming Zhu
Previous research on multi-party dialogue generation has predominantly leveraged structural information inherent in dialogues to directly inform the generation process. However, the prevalence of colloquial expressions and incomplete utterances in dialogues often impedes comprehension and weakens the fidelity of dialogue structure representations, which is particularly pronounced in multi-party dialogues. In this work, we propose a novel framework DRCR (Discourse coherence and Response-guided Context Rewriting) to improve multi-party dialogue generation through dialogue context rewriting. Specifically, DRCR employs two complementary feedback signals, discourse coherence and response quality, to construct preference data for both context rewriting and response generation. Moreover, we propose a dynamic self-evolution learning method that allows the rewriter and responder to continuously enhance their capabilities through mutual interaction in an iterative training loop. Comprehensive experiments conducted on four multi-party dialogue datasets substantiate the effectiveness of DRCR.

[Findings] Discovering the Gems in Early Layers: Accelerating Long-Context LLMs with 1000x Input Token Reduction ▶ *Underline Video*
Zhenmei Shi, Yifei Ming, Xuan-Phi Nguyen, Yingyu Liang and Shafiq Joty
Large Language Models (LLMs) have demonstrated remarkable capabilities in handling long context inputs, but this comes at the cost of increased computational resources and latency. Our research introduces a novel approach for the long context bottleneck to accelerate LLM inference and reduce GPU memory consumption. We show that LLMs can identify relevant tokens in the early layers prior to generating query responses. Leveraging this insight, we propose an algorithm that uses early layers of an LLM as filters to select and compress input tokens, significantly reducing the context length for subsequent processing. Our method, GemFilter, demonstrates substantial improvements in both speed and memory efficiency compared to existing techniques, such as standard attention and SnapKV/H2O. Notably, it achieves a 2.4X speedup and 30% reduction in GPU memory usage compared to SOTA methods. When evaluated on the Needle in a Haystack task, GemFilter significantly outperforms standard attention and SnapKV, while demonstrating comparable performance on the LongBench challenge. GemFilter is simple, training-free, and broadly applicable across different LLMs. Moreover, it provides interpretability by allowing humans to inspect the selected input sequence. Our findings provide practical benefits for deploying LLMs and deepen our understanding of their internal mechanisms, paving the way for further optimizations in LLM design and inference. Our code is available at https://github.com/SalesforceAIResearch/GemFilter.

[Findings] Distilling LLM Reasoning into Dense Encoders: Bridging the Accuracy-Efficiency Gap in Recommendation ▶ *Underline Video*
Donghee Han, Daeyoung Roh, A Young Kim, Hwanjun Song and Mun Yong Yi
Large Language Models (LLMs) have shown remarkable potential in recommendation systems but suffer from prohibitive inference latency. Existing distillation approaches typically target Small Language Models (SLMs) or Conventional Recommendation Models (CRMs), face a critical trade-off between computational cost and semantic reasoning capacity. To bridge this accuracy-efficiency gap, we introduce

Reasoning-to-Encoder Distillation (R2END), a framework that establishes a text encoder as the optimal student architecture for scalable recommendation. Unlike methods that mimic token generation, R2END compresses the teachers reasoning into a dense vector space via a semantic alignment objective, effectively capturing user-item dynamics. Extensive experiments on four datasets demonstrate that R2END not only outperforms state-of-the-art baselines but also achieves drastically reduced latency, offering a sweet spot for recommendation.

[Findings] Distilling Long-CoT Reasoning through Collaborative Step-wise Multi-Teacher Decoding ► *Underline Video*

Taewon Yun, Jisu Shin, Jeonghwan Choi, Seunghwan Bang and Hwanjun Song

Distilling large reasoning models (LRMs) has become essential for making their Long-CoT reasoning capabilities practical, as full-scale inference remains computationally prohibitive. Existing curation-based approaches, which select complete reasoning traces post-hoc, overlook the collaborative potential of heterogeneous teachers and fail to adapt exploration dynamically, often leading to redundant sampling and missed opportunities for complementary reasoning. To address this, we introduce CoRD, a collaborative multi-teacher decoding framework that performs step-wise reasoning synthesis guided by predictive perplexity-based scoring and beam search. This approach enables heterogeneous LRMs to jointly construct coherent reasoning trajectories while maintaining diverse, high-potential hypotheses efficiently. Experiments show that CoRD generates higher-quality reasoning data and achieves student performance approaching teacher-level results, demonstrating that fine-grained collaboration among diverse LRMs yields structured, efficient, and robust reasoning distillation. The dataset and model are available at https://github.com/DISL-Lab/CoRD

[Findings] Distilling the Essence: Efficient Reasoning Distillation via Sequence Truncation ► *Underline Video*

Wei-Rui Chen, Vignesh Kothapalli, Ata Fatahibaarzi, Hejian Sang, Shao Tang, Qingquan Song, Zhipeng Wang and Muhammad Abdul-Mageed

Distilling the capabilities from a large reasoning model (LRM) to a smaller student model often involves training on substantial amounts of reasoning data. However, knowledge distillation (KD) over lengthy sequences with prompt (P), chain-of-thought (CoT), and answer (A) sections makes the process computationally expensive. In this work, we investigate how the allocation of supervision across different sections (P, CoT, A) affects student performance. Our analysis shows that selective KD over only the CoT tokens can be effective when the prompt and answer information is encompassed by it. Building on this insight, we establish a truncation protocol to quantify computation-quality tradeoffs as a function of sequence length. We observe that beyond a specific length, longer training sequences provide marginal returns for downstream performance but require substantially higher memory and FLOPs. To this end, training on only the first 50% of tokens of every training sequence can retain, on average, approx 91% of full-sequence performance on math benchmarks while reducing training time, memory usage, and FLOPs by about 50% each. Codes are available at https://github.com/weiruichen01/distilling-the-essence.

[Findings] Distorted or Fabricated? A Survey on Hallucination in Video LLMs ► *Underline Video*

Yiyang Huang, Yitian Zhang, Yizhou Wang, Mingyuan Zhang, Liang Shi, Huimin Zeng and Yun Fu

Despite significant progress in video-language modeling, hallucinations remain a persistent challenge in Video Large Language Models (Vid-LLMs), referring to outputs that appear plausible yet contradict the content of the input video. This survey presents a comprehensive analysis of hallucinations in Vid-LLMs and introduces a systematic taxonomy that categorizes them into two core types: dynamic distortion and content fabrication, each comprising two subtypes with representative cases. Building on this taxonomy, we review recent advances in the evaluation and mitigation of hallucinations, covering key benchmarks, metrics, and intervention strategies. We further analyze the root causes of dynamic distortion and content fabrication, which often result from limited capacity for temporal representation and insufficient visual grounding. These insights inform several promising directions for future work, including the development of motion-aware visual encoders and the integration of counterfactual learning techniques. This survey consolidates scattered progress to foster a systematic understanding of hallucinations in Vid-LLMs, laying the groundwork for building robust and reliable video-language systems.

[Long] Doc-V^*: Coarse-to-Fine Interactive Visual Reasoning for Multi-Page Document VQA ► *Underline Video*

Yuanlei Zheng, Pei Fu, Hang Li, Ziyang Wang, Yuyi Zhang, Wenyu Ruan, Xiaojin Zhang, Zhongyu Wei, Zhenbo Luo, Jian Luan, Wei Chen and Xiang Bai

Multi-page Document Visual Question Answering requires reasoning over semantics, layouts, and visual elements in long, visually dense documents. Existing OCR-free methods face a trade-off between capacity and precision: end-to-end models scale poorly with document length, while visual retrieval-based pipelines are brittle and passive. We propose Doc-V^, an OCR-free agentic framework that casts multi-page DocVQA as sequential evidence aggregation. Doc-V^ begins with a thumbnail overview, then actively navigates via semantic retrieval and targeted page fetching, and aggregates evidence in a structured working memory for grounded reasoning. Trained by imitation learning from expert trajectories and further optimized with Group Relative Policy Optimization, Doc-V^ balances answer accuracy with evidence-seeking efficiency. Across five benchmarks, Doc-V^ outperforms open-source baselines and approaches proprietary models, improving out-of-domain performance by up to 47.9% over RAG baseline. Other results reveal effective evidence aggregation with selective attention, not increased input pages.

[Findings] DRPG (Decompose, Retrieve, Plan, Generate): An Agentic Framework for Academic Rebuttal ► *Underline Video*

Peixuan Han, YingJie Yu, Jingjun Xu and Jiaxuan You

Despite the growing adoption of large language models (LLMs) in scientific research workflows, automated support for academic rebuttal, a crucial step in academic communication and peer review, remains largely underexplored. Existing approaches typically rely on off-the-shelf

LLMs or simple pipelines, which struggle with long-context understanding and often fail to produce targeted and persuasive responses. In this paper, we propose DRPG, an agentic framework for automatic academic rebuttal generation that operates through four steps: Decompose reviews into atomic concerns, Retrieve relevant evidence from the paper, Plan rebuttal strategies, and Generate responses accordingly. Notably, the Planner in DRPG reaches over 98% accuracy in identifying the most feasible rebuttal direction. Experiments on data from top-tier conferences demonstrate that DRPG significantly outperforms existing rebuttal pipelines and achieves performance beyond the average human level using only an 8B model. Our analysis further demonstrates the effectiveness of the planner design and its value in providing multi-perspective and explainable suggestions. We also showed that DRPG works well in a more complex multi-round setting. These results highlight the effectiveness of DRPG and its potential to provide high-quality rebuttal content and support the scaling of academic discussions.

[Findings] Efficient Inference for Large Vision-Language Models: Bottlenecks, Techniques, and Prospects ▶ *Underline Video*

Jun Zhang, Yicheng Ji, Feiyang Ren, Yihang Li, Bowen Zeng, Zonghao Chen, Ke Chen, Lidan Shou, Gang Chen and Huan Li

Large Vision-Language Models (LVLMs) enable sophisticated reasoning over images and videos, yet their inference is hindered by a systemic efficiency barrier known as visual token dominance. This overhead is driven by a multi-regime interplay between high-resolution feature extraction, quadratic attention scaling, and memory bandwidth constraints. We present a systematic taxonomy of efficiency techniques structured around the inference lifecycle, consisting of encoding, prefilling, and decoding. Unlike prior reviews focused on isolated optimizations, we analyze the end-to-end pipeline to reveal how upstream decisions dictate downstream bottlenecks, covering compute-bound visual encoding, the intensive prefilling of massive contexts, and the ”visual memory wall” in bandwidth-bound decoding. By decoupling the efficiency landscape into the axes of shaping information density, managing long-context attention, and overcoming memory limits, this work provides a structured analysis of how isolated optimizations compose to navigate the trade-off between visual fidelity and system efficiency. The survey concludes by outlining four future frontiers supported by pilot empirical insights, including hybrid compression based on functional unit sensitivity, modality-aware decoding with relaxed verification, progressive state management for streaming continuity, and stage-disaggregated serving through hardware-algorithm co-design. The submitted software contains a snapshot of our literature repository, which is designed to be maintained as a living resource for the community.

[Findings] EOP-LLM: Energy Oriented Pruning for Large Language Models ▶ *Underline Video*

Guan Huang and Tao Shu

In deployed large language models (LLMs), inference energy consumption has grown rapidly and has emerged as a key bottleneck in large-scale deployment, yet most existing inference efficiency methods focus on reducing FLOPs or latency, rather than explicitly modeling or enforcing end-to-end inference energy constraints. We propose EOP-LLM, an energy-oriented dynamic pruning framework that enables LLM inference under explicit per-sequence energy budgets. EOP-LLM combines a device-calibrated energy proxy with lightweight token and feed-forward (FFN) selectors, coordinated through a global dual variable, to dynamically allocate computation while preserving model quality. Extensive experiments on LLaMA 3.2 (1B/3B) and LLaMA 3.1 (8B) demonstrate that EOP-LLM consistently outperforms state-of-the-art dynamic pruning baselines under matched energy budgets, while strictly adhering to per-sequence energy constraints.

[Findings] Evaluating Scene-based In-Situ Item Labeling for Immersive Conversational Recommendation ▶ *Underline Video*

Jiazhou Liang, Yifan Simon Liu, David Guo, Yilun Jiang, Minqi Sun and Scott Sanner

The growing ubiquity of Extended Reality (XR) is driving Conversational Recommendation Systems (CRS) toward visually immersive experiences. We formalize this paradigm as Immersive CRS (ICRS), where recommended items are highlighted directly in the users scene-based visual environment and augmented with in-situ labels. While item recommendation has been widely studied, the problem of how to select and evaluate which information to present as immersive labels remains an open problem. To this end, we introduce a principled categorization of information needs into explicit intent satisfaction and proactive information needs and use these to define novel evaluation metrics for item label selection. We benchmark IR-, LLM-, and VLM-based methods across three datasets and ICRS scenarios: fashion, movie recommendation, and retail shopping. Our evaluation reveals three important limitations of existing methods: (1) they fail to leverage scenario-specific information modalities (e.g., visual cues for fashion, metadata for retail), (2) they present redundant information that is visually inferable, and (3) they poorly anticipate users proactive information needs from explicit dialogue alone. In summary, this work provides both a novel evaluation paradigm for in-situ item labeling in ICRS and highlights key challenges for future work.

[Long] Evo-PI: Aligning Medical Reasoning via Evolving Principle-Guided Supervision ▶ *Underline Video*

Xianda Zheng, Huan Gao, Meng-Fen Chiang, Michael Witbrock, Kaiqi Zhao and Shangyang Li

Despite recent progress, the reasoning capabilities of large multimodal language models (MLLMs) remain fundamentally constrained by static supervision, where fixed prompts, rules, or reward models provide non-adaptive guidance throughout training. Such static signals are often sufficient to enforce output formats, but fail to shape the underlying reasoning process, leading to brittle generalization and performance saturation in complex decision-making tasks. We propose Evo-PI, a principle-centric learning framework that treats reasoning principles as explicit, language-based supervision signals that can be generated, evaluated, and iteratively evolved. Instead of relying on fixed rewards, Evo-PI enables a co-evolutionary loop in which principles guide model reasoning, while model behaviors in turn refine the principles that supervise them. This dynamic alignment mechanism allows supervision to progressively adapt to the models reasoning deficiencies. We instantiate Evo-PI in medical visual question answering as a high-stakes testbed requiring structured visual-textual reasoning. Across eight benchmarks and multiple model backbones, Evo-PI consistently improves reasoning accuracy, achieving gains of up to 24.6%. Our results

suggest that evolving principle-guided supervision offers a scalable and general paradigm for training expert-aligned reasoning in multimodal language models.

[Long] EvoRoute: Experience-Driven Self-Routing LLM Agent Systems ▶ *Underline Video*

Guibin Zhang, Haiyang Yu, Kaiming Yang, Bingli Wu, Fei Huang, Yongbin Li and Shuicheng YAN

Complex agentic AI systems, powered by a coordinated ensemble of Large Language Models (LLMs), tool and memory modules, have demonstrated remarkable capabilities on intricate, multi-turn tasks. However, this success is shadowed by prohibitive economic costs and severe latency, exposing a critical, yet underexplored, trade-off. We formalize this challenge as the Agent System Trilemma: the inherent tension among achieving state-of-the-art performance, minimizing monetary cost, and ensuring rapid task completion. To dismantle this trilemma, we introduce EvoRoute, a self-evolving model routing paradigm that transcends static, pre-defined model assignments. Leveraging an ever-expanding knowledge base of prior experience, EvoRoute dynamically selects Pareto-optimal LLM backbones at each step, balancing accuracy, efficiency, and resource use, while continually refining its own selection policy through environment feedback. Experiments on challenging agentic benchmarks such as GAIA and BrowseComp+ demonstrate that EvoRoute, when integrated into off-the-shelf agentic systems, not only sustains or enhances system performance but also reduces execution cost by up to 80% and latency by over 70%.

[Long] Experience-driven Multi-turn Reinforcement Learning for GUI Agents ▶ *Underline Video*

Zhengxi Lu, Jiabo Ye, Fei Tang, Yongliang Shen, Haiyang Xu, Ziwei Zheng, Weiming Lu, Ming Yan, Fei Huang, Jun Xiao and Yueting Zhuang

GUI agents have demonstrated remarkable progress in automating complex user interface interactions. However, training such agents for long-horizon tasks remains challenging. Single-turn reinforcement learning conditions on expert histories during training but self-generated histories during deployment, causing distribution mismatch. Online multi-turn methods eliminate this gap via environment interaction but suffer from sparse rewards and prohibitive costs. We propose Experience-driven Multi-turn Policy Optimization (EMPO), which leverages expert trajectories as environment experiences for on-policy multi-turn training. The agent constructs self-generated history throughout rollouts; when actions match expert experiences, the trajectory provides valid state transitions, and a Patch Module recovers mismatched steps to maintain on-policy rollouts. EMPO further incorporates discounted future rewards and dual-level advantage estimation to capture long-horizon dependencies. We also propose AndroidControl-Real, an evaluation metric strongly correlated with real-world performance (R^2=0.934). With only 1K public trajectories as RL experiences, our method achieves substantial gains over the base model (e.g., +12.0% on AndroidWorld and +23.8% on AITW) and achieves competitive performance against strong baselines such as UI-TARS-7B and GPT-4o, demonstrating better generalization than prior single-turn RL approaches. Code available: https://anonymous.4open.science/r/UI-S1-0DAF.

[Findings] F-Actor: Controllable Conversational Behavior in Full-Duplex Models ▶ *Underline Video*

Maike Züfle, Ondrej Klejch, Nicholas Sanders, Jan Niehues, Alexandra Birch and Tsz Kin Lam

Spoken conversational systems require more than accurate speech generation to have human-like conversations: to feel natural and engaging, they must produce conversational behaviour that adapts dynamically to the context. Current spoken conversational systems, however, rarely allow such customization, limiting their naturalness and usability. In this work, we present the first open, instruction-following full-duplex conversational speech model that can be trained efficiently under typical academic resource constraints. By keeping the audio encoder frozen and finetuning only the language model, our model requires just 2,000 hours of data, without relying on large-scale pretraining or multi-stage optimization. The model can follow explicit instructions to control speaker voice, conversation topic, conversational behaviour (e.g., backchanneling and interruptions), and dialogue initiation. We propose a single-stage training protocol and systematically analyze design choices. Both the model and training code is released to enable reproducible research on controllable full-duplex speech systems.

[Findings] Faithful Persona Steering under Incongruity via Dual-Stream Refinement ▶ *Underline Video*

Yu-An ChuJen-Ren PongChia-Yao YehMeng-Fen Chiang

Standard LLM personalization typically frames identity as a static retrieval task, overlooking the inherent incongruity of human personas, where stable traits coexist with atypical, context-specific stances. Existing methods struggle to reconcile these dimensions: prompting succumbs to context drift over long sequences, while fine-tuning often suppresses idiosyncratic "quirks" in favor of generic distributional patterns. To bridge this gap, we present QuirkyMind, a framework that disentangles identity definition from its expression. First, Traits Anchoring constructs a dual-stream latent state, fusing a sentence-level summary for semantic stability with a token-level sequence for generative control. This state is stabilized via In-Context Narrative Refinement using an alternating objective: a discriminative InfoNCE loss anchors the persona in representation space to prevent drift, while a generative cross-entropy loss ensures faithful verbalization. Finally, Persona Steered Generalization transfers the refined state to downstream tasks via parameter-efficient adapters. Empirical evaluations on Persona-Steered QA and Narrative Inference demonstrate that QuirkyMind mitigates drift, consolidating persona knowledge without erasing authentic incongruities.

[Findings] Faithfulness-Aware Uncertainty Quantification for Fact-Checking the Output of Retrieval-Augmented Generation ▶ *Underline Video*

Ekaterina Fadeeva, Aleksandr Rubashevskii, Dzianis Piatrashyn, Roman Vashurin, Shehzaad Dhuliawala, Artem Shelmanov, Timothy Baldwin, Preslav Nakov, Mrinmaya Sachan and Maxim Panov

Large Language Models (LLMs) enhanced with knowledge retrieval, an approach known as Retrieval-Augmented Generation (RAG), have achieved strong performance in open-domain question answering. However, RAG remains prone to hallucinations: factually incorrect outputs

may arise from inaccuracies in the model's internal knowledge and the retrieved context. Existing approaches to mitigating hallucinations often conflate factuality with faithfulness to the retrieved evidence, incorrectly labeling factually correct statements as hallucinations if they are not explicitly supported by the retrieval. In this paper, we introduce FRANQ (Faithfulness-aware Retrieval-Augmented UNcertainty Quantification), a new method for hallucination detection in RAG outputs. FRANQ applies distinct uncertainty quantification techniques to estimate factuality, conditioning on whether a statement is faithful to the retrieved context. To evaluate FRANQ and competing uncertainty quantification methods, we construct a new long-form question answering dataset annotated for both factuality and faithfulness, combining automated labeling with manual validation of challenging cases. Extensive experiments across multiple datasets, tasks, and LLMs show that FRANQ achieves more accurate detection of factual errors in RAG-generated responses compared to existing uncertainty quantification and hallucination detection approaches.

[Findings] Fast and Effective On-Policy Distillation from Reasoning Prefixes ▶ *Underline Video*

Dongxu Zhang, Zhichao Yang, Sepehr Janghorbani, Jun Han, Andrew Ressler II, Qian Qian, Gregory D Lyng, Sanjit Singh Batra and Robert E. Tillman

On-policy distillation (OPD), which samples trajectories from the student model and supervises them with a teacher at the token-level, avoids relying solely on verifiable terminal rewards and can yield better generalization than off-policy distillation. However, OPD requires expensive on-the-fly sampling of the student policy during training, which substantially increases training cost, especially for responses with long reasoning traces. Our initial analysis shows that, during OPD, training signals are stronger in the prefix of each output reasoning trace, and that even a short teacher-generated prefix can significantly help the student produce the correct answer. Motivated by these observations, we propose a simple yet effective modification of OPD: we apply the distillation objective only to prefixes of student-generated outputs and terminate each sampling early during distillation. Experiments on a suite of AI-for-Math and out-of-domain reasoning benchmarks show that on-policy prefix distillation matches the performance of full OPD in long reasoning outputs while reducing training FLOP by 2x-40x.

[Findings] Fico: Evaluating Vision-Language Models under Visual Fidelity and Compression at Scale ▶ *Underline Video*

Jianhong Tu, Nicholas Crispino, Kyle Montgomery, Chenguang Wang and Dawn Song

Visual text compression is an emerging paradigm for rendering text as images for processing by vision-language models (VLMs), enabling higher information density per context token. However, the robustness of VLMs under dense, text-based visual inputs remains unevaluated. We introduce Fico, a benchmark designed to assess VLM robustness across seven controlled variants of visual fidelity and information density. Fico spans documents of 8k to 64k tokens and includes three tasks of increasing semantic granularity: optical character recognition (OCR), needle-in-a-haystack (NIAH) retrieval, and visual question answering (VQA). Evaluating 13 general-purpose VLMs and 3 OCR-specialized models reveals three consistent trends: performance drops sharply under increased density or reduced resolution; cross-task transfer between OCR, NIAH, and VQA is limited; and VQA is comparatively robust because low-level details are lost before high-level semantics. By exposing failure modes that remain invisible under conventional VLM evaluations, Fico establishes a rigorous test-bed for visual text compression.

[Long] FIGMA: Towards FIne-Grained Music retrievAl ▶ *Underline Video*

Nishit Anand, Ashish Seth, Sreyan Ghosh, Dinesh Manocha and Ramani Duraiswami

Retrieving music using natural language descriptions has improved with contrastive audio-text models such as CLAP, but current systems remain limited to coarse semantic queries. When descriptions specify fine-grained musical attributes such as tempo, key, chord progression, or rhythmic structure, existing models often fail to retrieve the correct audio. We show that this limitation stems from the contrastive learning objective itself: despite being trained on long captions, CLAP-based models effectively utilize only the first few tokens, discarding much of the information encoded in detailed prompts. Then, we propose FIGMA (Fine-Grained Music Retrieval), a multi-view contrastive architecture that addresses this limitation by jointly optimizing global audio-text alignment and frame-level, token-wise alignment. This design enables FIGMA to capture both high-level semantic context and fine-grained musical attributes within a unified representation space. Moreover, we formalize the task of Fine-Grained Music Retrieval and construct Fine-Grained Music Caption dataset (FGMCaps), a large-scale dataset of 380K music-caption pairs for training along with a 10K test set, both annotated with tempo, key, chord progression, beat count, as well as genre and mood. Extensive experiments demonstrate that FIGMA consistently outperforms existing CLAP-based music retrieval models across multiple music retrieval benchmarks, including out-of-domain evaluations, with relative improvements of up to 73.3%.

[Findings] FLEXITOKENS: Flexible Tokenization for Evolving Language Models ▶ *Underline Video*

Abraham Toluwase Owodunni, Orevaoghene Ahia and Sachin Kumar

Adapting language models to new data distributions by simple finetuning is challenging. This is due to the rigidity of their subword tokenizers, which typically remain unchanged during adaptation. This inflexibility often leads to inefficient tokenization, causing overfragmentation of text in out-of-distribution domains, unseen languages, or scripts. In this work, we develop byte-level LMs with learnable tokenizers to make tokenization adaptive. Our models include a submodule that learns to predict boundaries given the input byte sequence, encoding it into variable-length segments. Most tokenizer-free methods train this boundary predictor using an auxiliary loss that enforces a fixed compression rate across the training corpus, introducing a new kind of rigidity. We propose FLEXITOKENS, a simplified training objective that enables significantly greater flexibility during adaptation. Evaluating across multiple multilingual benchmarks, morphologically diverse tasks, and domains, we demonstrate that FLEXITOKENS consistently reduces token over-fragmentation and achieves up to 10% point improvements on token classification and generative tasks compared to BPE and other gradient-based tok-

enizer baselines. We validate our findings using models of varying sizes, and our method demonstrates consistent improvements across scales.

[Long] Focus-dLLM: Accelerating Long-Context Diffusion LLM Inference via Confidence-Guided Context Focusing ▶ *Underline Video*

Lingkun Long, Yushi Huang, Shihao Bai, Ruihao Gong, Jun Zhang, Ao Zhou and Jianlei Yang

Diffusion Large Language Models (dLLMs) deliver strong long-context processing capability in a non-autoregressive decoding paradigm. However, the considerable computational cost of bidirectional full attention limits the inference efficiency. Although sparse attention is promising, existing methods remain ineffective. This stems from the need to estimate attention importance for tokens yet to be decoded, while the unmasked token positions are unknown during diffusion. In this paper, we present Focus-dLLM, a novel training-free attention sparsification framework tailored for accurate and efficient long-context dLLM inference. Based on the finding that token confidence strongly correlates across adjacent steps, we first design a past confidence-guided indicator to predict unmasked regions. Built upon this, we propose a sink-aware pruning strategy to accurately estimate and remove redundant attention computation, while preserving highly influential attention sinks. To further reduce overhead, this strategy reuses identified sink locations across layers, leveraging the observed cross-layer consistency. Experimental results show that our method offers more than 29x lossless speedup under 32K context length.

[Long] From "Aha Moments" to Controllable Thinking: Toward Meta-Cognitive Reasoning in LRMs via Decoupled Reasoning and Control ▶ *Underline Video*

Rui Ha, Rui Pu, Chaozhuo Li, Li Sun and Sen Su

Large Reasoning Models (LRMs) can exhibit step-by-step reasoning, reflection, and backtracking, but these behaviors are often unregulated, leading to overthinking. As a result, LRMs continue generating redundant reasoning even after reaching high-confidence conclusions. This increases inference cost and latency, limiting practical deployment. The root cause is the absence of an intrinsic mechanism to monitor the reasoning state and decide when to continue, backtrack, or stop. We propose MERA, a meta-cognitive reasoning framework that decouples reasoning from control to enable independent optimization of control strategies. MERA constructs high-quality reasoning-control supervision data via a takeover-based pipeline, and transforms long-horizon traces into structured reasoning-control alternating sequences for training. The model is trained with supervised fine-tuning to internalize the structured separation, and further optimized with Control-Segment Policy Optimization (CSPO), which combines segment-wise GRPO with control masking to focus learning on control segments. Experiments across reasoning benchmarks show that MERA improves both efficiency and accuracy.

[Findings] From Coarse to Fine: Benchmarking and Reward Modeling for Writing-Centric Generation Tasks ▶ *Underline Video*

Qingyu Ren, Tianjun Pan, Xingzhou Chen and Xuhong Wang

Large language models have achieved remarkable progress in text generation but still struggle with generative writing tasks. In terms of evaluation, existing evaluation benchmarks include few requirement types and writing reward models are not evaluated. In terms of training, existing studies often enhance writing ability through reinforcement learning with verifiable rewards (RLVR). However, existing reward model training remains coarse-grained. To address these issues, we introduce WšBench, a comprehensive evaluation benchmark, and WRL, a fine-grained training framework. WšBench covers five task categories and seven requirement types, enabling systematic evaluation of both writing and writing reward models by measuring the correlation between reward rankings and golden rankings. WRL constructs positive and negative samples by dropping instruction requirements to construct positive and negative examples, allowing more precise reward model training. Experiments show that our models achieve substantial improvements on various writing benchmarks and exhibit strong generalization. We will release our code and data to support future research.

[Long] From Competition to Synergy: Unlocking Reinforcement Learning for Subject-Driven Image Generation ▶ *Underline Video*

Ziwei Huang, Ying Shu, Fanghao Fanghao, Quanyu Long, Wenya Wang, Qiushi Guo, Tiezheng Ge and Leilei Gan

Subject-driven image generation models face a fundamental trade-off between identity preservation (fidelity) and prompt adherence (editability). While online reinforcement learning (RL), specifically GPRO, offers a promising solution, we find that a naive application of GRPO leads to competitive degradation, as the simple linear aggregation of rewards with static weights causes conflicting gradient signals and a misalignment with the temporal dynamics of the diffusion process. To overcome these limitations, we propose Customized-GRPO, a novel framework featuring two key innovations: (i) Synergy-Aware Reward Shaping (SARS), a non-linear mechanism that explicitly penalizes conflicted reward signals and amplifies synergistic ones, providing a sharper and more decisive gradient. (ii) Time-Aware Dynamic Weighting (TDW), which aligns the optimization pressure with the model's temporal dynamics by prioritizing prompt-following in the early, identity preservation in the later. Extensive experiments demonstrate that our method significantly outperforms naive GRPO baselines, successfully mitigating competitive degradation. Our model achieves a superior balance, generating images that both preserve key identity features and accurately adhere to complex textual prompts.

[Short] From Narrow Unlearning to Emergent Misalignment in LLMs ▶ *Underline Video*

Erum Mushtaq, Anil Ramakrishna, Satyapriya Krishna, Sattvik Sahai, Prasoon Goyal, Kai-Wei Chang, Tao Zhang and Rahul Gupta

Recent work has shown that fine-tuning on insecure code data can trigger an emergent misalignment (EMA) phenomenon, where models generate malicious responses even to prompts unrelated to the original insecure code-writing task. Such cross-domain generalization of harmful behavior underscores the need for a deeper understanding of the algorithms, tasks, and datasets that induce emergent misalignment. In this work, we extend this study by demonstrating that emergent misalignment can also arise from narrow refusal unlearning in specific

domains. We perform refusal unlearning on Cybersecurity and Safety concept, and evaluate EMA by monitoring refusal scores across seven responsible AI (RAI) domains, Cybersecurity, Safety, Toxicity, Bias, Sensitive Content, Medical/Legal, and Privacy. Our work shows that narrow domain unlearning can yield compliance responses for the targeted concept, however, it may also propagate EMA to unrelated domains. Among the two intervened concepts, Cybersecurity and Safety, we find that the safety concept can have larger EMA impact, i.e, causing lower refusal scores, across other unrelated domains such as bias. We observe this effect consistently across two model families, Mistral-7b-0.3v, and Qwen-7b-2.5. Further, we show that refusal unlearning augmented with cross-entropy loss function on a small set of retain data from the affected domains can largely, if not fully, restore alignment across the impacted domains while having lower refusal rate on the concept we perform unlearning on. To investigate the underlying causes of EMA, we analyze concept entanglements at the representation level via concept vectors. Our analysis reveals that concepts with higher representation similarity in earlier layers are more susceptible to EMA after intervention when the refusal stream is altered through targeted refusal unlearning.

[Findings] From Scenes to Elements: Multi-Granularity Evidence Retrieval for Verifiable Multimodal RAG ▶ *Underline Video*

Guanhua Chen, Chuyue Huang, Yutong Yao, Shudong Liu, Xueqing Song, Lidia S. Chao and Derek F. Wong

Multimodal Retrieval-Augmented Generation (RAG) systems retrieve evidence at coarse granularities (entire images or scenes), creating a mismatch with fine-grained user queries and making failures unverifiable. We introduce GranuVistaVQA, a multimodal benchmark featuring real-world landmarks with element-level annotations across multiple viewpoints, capturing the partial observation challenge where individual images contain only subsets of entities. We further propose GranuRAG, a multi-granularity framework that treats visual elements as first-class retrieval units through three stages: element-level detection and classification, multi-granularity cross-modal alignment for evidence retrieval, and attribution-constrained generation. By grounding retrieval at the element level rather than relying on implicit attention, our approach enables transparent error diagnosis. Experiments demonstrate that GranuRAG achieves up to 29.2% improvement over six strong baselines for this task.

[Findings] From Style to Story: A Curriculum Learning Approach for Imitative Novel Generation ▶ *Underline Video*

Xueran Han, Yuhan Liu, Mingzhe Li, Wei Liu, Sen Hu, Rui Yan, zhiqiang xu and Xiuying Chen

Great novels create immersive worlds with rich character arcs, well-structured plots, and nuanced writing styles. However, current novel generation methods often rely on brief, simplistic story outlines and generate details using plain, generic language. To bridge this gap, we introduce the task of Imitative Novel Generation, which requires the generated novels to imitate the distinctive features of the original work, including understanding character profiles and world views, predicting plausible plot developments, and writing concrete details using vivid, expressive language. To achieve this, we propose WriterAgent, a novel generation system designed to master the core aspects of literary imitative. WriterAgent is trained through a curriculum learning paradigm, progressing from low-level stylistic mastery to high-level narrative coherence. Its key tasks include language style learning, character modeling, plot planning, and stylish writing, ensuring comprehensive narrative control. To support this, WriterAgent leverages the WriterLoRA framework, an extension of LoRA with hierarchical and cumulative task-specific modules, each specializing in a different narrative aspect. We evaluate WriterAgent on multilingual classics like Harry Potter and Dream of the Red Chamber, demonstrating its superiority over baselines in capturing the target authors settings, character dynamics, and writing style to produce coherent, faithful narratives. We hope this work inspires literary creativity in NLP: WriterAgent.

[Long] GameplayQA: A Benchmarking Framework for Decision-Dense POV-Synced Multi-Video Understanding of 3D Virtual Agents ▶ *Underline Video*

Yunzhe Wang, Runhui Xu, Kexin Zheng, Tianyi Zhang, Jayavibhav Kogundi, Soham Hans and Volkan Ustun

Multimodal LLMs are increasingly deployed as perceptual backbones for autonomous agents in 3D environments, from robotics to virtual worlds. These applications require agents to perceive rapid state changes, attribute actions to the correct entities, and reason about concurrent multi-agent behaviors from a first-person perspective, capabilities that existing benchmarks do not adequately evaluate. We introduce GameplayQA, a framework for evaluating agentic-centric perception and reasoning through video understanding. Specifically, we densely annotate multiplayer 3D gameplay videos at 1.22 labels/second, with time-synced, concurrent captions of states, actions, and events structured around a triadic system of Self, Other Agents, and the World, a natural decomposition for multi-agent environments. From these annotations, we refined 2.4K diagnostic QA pairs organized into three levels of cognitive complexity, accompanied by a structured distractor taxonomy that enables fine-grained analysis of where models hallucinate. Evaluation of frontier MLLMs reveals a substantial gap from human performance, with common failures in temporal and cross-video grounding, agent-role attribution, and handling the decision density of the game. We hope GameplayQA stimulates future research at the intersection of embodied AI, agentic perception, and world modeling.

[Long] GMoE: Global Mixture of Experts with Logit Propagation ▶ *Underline Video*

Geonwoo Hong and Taehwan Kim

Sparse Mixture of Experts (SMoE) architectures reduce computational cost by activating only a subset of experts per token, yet they often retain large memory footprints and exhibit significant redundancy, both within and across layers. We propose GMoE, a sparse MoE architecture designed to explicitly address these inefficiencies. Instead of maintaining separate expert sets for each layer, GMoE uses Global Experts shared across all layers and adds a Local Expert per layer for layer-specific adaptation. This architecture reuses Global Experts across layers, thereby mitigating inter-layer redundancy while substantially reducing model parameters. In addition, we introduce a Global Router with a GRU-based recurrent component shared across layers and layer-specific routing heads that propagate routing logits across layers. This routing mechanism couples routing decisions across layers, progressively refines routing paths, and helps mitigate intra-layer

redundancy. Across diverse language modeling corpora and downstream benchmarks, GMoE remains competitive while using substantially fewer parameters. Routing path analyses and an ablation study show that GMoE reduces cross-layer routing concentration and increases path diversity, with the Global Experts, the Local Expert, and the Global Router all contributing to the gains. The code is available at https://github.com/GEONWOOHONG/GMoE.

[Long] GROKE: Vision-Free Navigation Instruction Evaluation via Graph Reasoning on OpenStreetMap ▶ *Underline Video*

Farzad Shami, Subhrasankha Dey, Nico Van de Weghe and Henrikki Tenkanen

The evaluation of navigation instructions remains a persistent challenge in Vision-and-Language Navigation (VLN) research. Traditional reference-based metrics such as BLEU and ROUGE fail to capture the functional utility of spatial directives, specifically whether an instruction successfully guides a navigator to the intended destination. Although existing VLN agents could serve as evaluators, their reliance on high-fidelity visual simulators introduces licensing constraints and computational costs, and perception errors further confound linguistic quality assessment. This paper introduces GROKE (Graph-based Reasoning over OSM Knowledge for instruction Evaluation), a vision-free training-free hierarchical LLM-based framework for evaluating navigation instructions using OpenStreetMap data. Through systematic ablation studies, we demonstrate that structured JSON and textual formats for spatial information substantially outperform grid-based and visual graph representations. Our hierarchical architecture combines sub-instruction planning with topological graph navigation, reducing navigation error by 68.5% compared to heuristic and sampling baselines on the Map2Seq dataset. The agent's execution success, trajectory fidelity, and decision patterns serve as proxy metrics for functional navigability given OSM-visible landmarks and topology, establishing a scalable and interpretable evaluation paradigm without visual dependencies.

[Long] HERMES: KV Cache as Hierarchical Memory for Efficient Streaming Video Understanding ▶ *Underline Video*

Haowei Zhang, Shudong Yang, Jinlan Fu, See-Kiong Ng and Xipeng Qiu

Recent advancements in Multimodal Large Language Models (MLLMs) have demonstrated significant improvement in offline video understanding. However, extending these capabilities to streaming video inputs, remains challenging, as existing models struggle to simultaneously maintain stable understanding performance, real-time responses, and low GPU memory overhead. To address this challenge, we propose HERMES, a novel training-free architecture for real-time and accurate understanding of video streams. Based on a mechanistic attention investigation, we conceptualize KV cache as a hierarchical memory framework that encapsulates video information across multiple granularities. During inference, HERMES reuses a compact KV cache, enabling efficient streaming understanding under resource constraints. Notably, HERMES requires no auxiliary computations upon the arrival of user queries, thereby guaranteeing real-time responses for continuous video stream interactions. HERMES achieves 10x faster TTFT compared to prior SOTA. Even when reducing video tokens by up to 68% compared with uniform sampling, HERMES achieves superior or comparable accuracy across all benchmarks, with up to 11.4% gains on streaming datasets.

[Findings] Hidden States as Early Signals: Step-level Trace Evaluation and Pruning for Efficient Test-Time Scaling ▶ *Underline Video*

Zhixiang Liang, Beichen Huang, Zheng Wang and Minjia Zhang

Large Language Models (LLMs) can enhance reasoning capabilities through test-time scaling by generating multiple traces. However, the combination of lengthy reasoning traces with multiple sampling introduces substantial computation and high end-to-end latency. Prior work on accelerating this process has relied on similarity-based or confidence-based pruning, but these signals do not reliably indicate trace quality. To address these limitations, we propose STEP: Step-level Trace Evaluation and Pruning, a novel pruning framework that evaluates reasoning steps using hidden states and dynamically prunes unpromising traces during generation. We train a lightweight step scorer to estimate trace quality, and design a GPU memory-aware pruning strategy that triggers pruning as the GPU memory is saturated by KV cache to reduce end-to-end latency. Experiments across challenging reasoning benchmarks demonstrate that STEP reduces end-to-end inference latency by 45%-70% on average compared to self-consistency while also improving reasoning accuracy.

[Findings] Hybrid Self-evolving Structured Memory for Computer-Use Agents ▶ *Underline Video*

Sibo Zhu, Wenyi Wu, Kun Zhou, Stephen Wang and Biwei Huang

The remarkable progress of vision-language models (VLMs) has enabled computer-use agents to interact with computers in a human-like manner. Yet real-world computer-use tasks remain difficult due to long-horizon workflows, diverse interfaces, and frequent intermediate errors. Prior work equips agents with external memory built from large collections of trajectories, but relies on flat retrieval over discrete summaries or continuous embeddings, falling short of the structured organization and self-evolving characteristics of human memory. Inspired by the brain, we propose Hybrid Self-evolving Structured Memory (HyMEM), a graph-based memory that couples discrete high-level symbolic nodes with continuous trajectory embeddings. HyMEM maintains a graph structure to support multi-hop retrieval, self-evolution via node update operations, and on-the-fly working-memory refreshing during inference. Extensive experiments show that HyMEM consistently improves open-source computer-use agents, enabling 7B/8B backbones to match or surpass strong closed-source models; notably, it boosts Qwen2.5-VL-7B by +22.5% and outperforms Gemini2.5-Pro-Vision and GPT-4o.

[Findings] Improved Policy Optimization for Mixture-of-Experts Models: Importance Sampling and Rewarding from an Expert-Centric Perspective ▶ *Underline Video*

Yining Qian, Jinpeng Li, Fei Mi, Lifeng Shang and Xiang Zhang

Reinforcement learning (RL) has demonstrated considerable promise in enhancing large language models. However, its application to

Mixture-of-Experts (MoE) architectures is frequently hindered by training instability, primarily stemming from token-level misalignment in expert assignments between current and behavior policies. Existing approaches often oscillate between overly coarse sequence-level importance sampling, which ignores token-specific discrepancies, and restrictive expert-selection constraints that suppress beneficial policy exploration. To bridge this gap, we propose Expert Relative Policy Optimization (ERPO), which introduces expert-level importance sampling. By grouping tokens according to their routing assignments, ERPO mitigates the high variance of token-level importance sampling while overcoming the token-agnostic limitations of sequence-level methods. Furthermore, ERPO leverages this expert-centric granularity to introduce an Expert-Selection Entropy Reward, which dynamically adjusts routing uncertainty based on task-specific feedback. This unique mechanism ensures a rigorous alignment between reward signals and policy updates, a capability inherently unattainable by traditional importance sampling methods. Experimental results demonstrate that ERPO significantly outperforms strong baselines across multiple reasoning tasks, highlighting the efficacy of tailoring RL objectives to the structural inductive biases of MoE models.

[Findings] InfiAgent: An Infinite-Horizon Framework for General-Purpose Autonomous Agents ▶ *Underline Video*

Chenglin Yu, Yuchen Wang, Songmiao Wang, Hongxia Yang and Ming Li

LLM agents can reason and use tools, but they often break down on long-horizon tasks due to unbounded context growth and accumulated errors. Common remedies such as context compression or retrieval-augmented prompting introduce trade-offs between information fidelity and reasoning stability. We present InfiAgent, a general-purpose framework that keeps the agents reasoning context strictly bounded regardless of task duration by externalizing persistent state into a file-centric state abstraction. At each step, the agent reconstructs context from a workspace state snapshot plus a fixed window of recent actions. Experiments on DeepResearch and an 80-paper literature review task show that, without task-specific fine-tuning, InfiAgent with a 20B open-source model is competitive with larger proprietary systems and maintains substantially higher long-horizon coverage than context-centric baselines. These results support explicit state externalization as a practical foundation for stable long-horizon agents.

[Findings] Inject to Heal: Alleviating hallucination in LVLMs via Context Embedding Injection ▶ *Underline Video*

Mehrdad Fazli, Bowen Wei and Ziwei Zhu

Hallucinations, generating responses inconsistent with the visual input, remain a critical limitation of large vision-language models (LVLMs), especially in open-ended tasks such as image captioning and visual reasoning. In this work, we probe the layer-wise generation dynamics that drive hallucinations and propose a training-free mitigation strategy. Employing the Logit Lens, we examine how LVLMs construct next-token distributions across decoder layers, uncovering a pronounced commitment-depth gap: truthful tokens accumulate probability mass on their final candidates earlier than hallucinatory ones. Drawing on this discovery, we introduce Context Embedding injection (CEI), a lightweight method that harnesses the hidden state of the last input token, the context embedding, as a grounding signal to maintain visual fidelity throughout decoding and curb hallucinations. Evaluated on the CHAIR, AMBER, and MMHal-Bench benchmarks (with a maximum token length of 512), CEI outperforms state-of-the-art baselines across three LVLMs, with its dynamic variant yielding the lowest overall hallucination rates. By integrating novel mechanistic insights with a scalable intervention, this work advances the mitigation of hallucinations in LVLMs. Data and code are available at https://github.com/mehrdadfazli/CEI.

[Findings] JTPRO: A Joint Tool-Prompt Reflective Optimization Framework for Language Agents ▶ *Underline Video*

Sandip Ghoshal, Anshul Mittal, Jyotika Singh, Miguel Ballesteros, Weiyi Sun, Fang Tu, Shailender Singh, Yassine Benajiba, Fahad Shah, Sujeeth Bharadwaj, Sujith Ravi and Dan Roth;

Large language model (LLM) agents augmented with external tools often struggle as number of tools grow large and become domain-specific. In such settings, ambiguous tool descriptions and under-specified agent instructions frequently lead to tool mis-selection and incorrect slot/value instantiation. We hypothesize that this is due to two root causes: generic, one-size-fits-all prompts that ignore tool-specific nuances, and underspecified tool schemas that lack clear guidance on when and how to use each tool and how to format its parameters. We introduce Joint Tool-Prompt Reflective Optimization (JTPRO), a framework for improving tool-calling reliability in trace-supervised settings by iteratively using rollout-driven reflection to co-optimize global instructions and per-tool schema/argument descriptions for accurate tool selection and argument instantiation in large tool inventories. JTPRO is designed to preserve only tool-local cues needed for correct disambiguation and slot filling. We evaluate JTPRO across multi-tool benchmarks, which account for different number of tools using three metrics: Tool Selection Accuracy (TSA), Slot Filling Accuracy(SFA), and Overall Success Rate(OSR) (correct tool + correct slots + correct values). JTPRO consistently outperforms strong baselines, including CoT-style agents, and reflective prompt optimizers such as GEPA by 5%-20% (relative) on OSR. Ablations show that joint optimization of instructions and tool schemas is more effective and robust than optimizing either component in isolation.

[Long] JW-SVD: Bridging the Cross-Modal Mismatch in Post-Training MLLM Compression ▶ *Underline Video*

Runchao Li, Yao Fu, Mu Sheng, Xianxuan Long, Haotian Yu and Kenneth Loparo

Post-training compression of Multimodal LLMs faces a fundamental geometric conflict: parameter subspaces optimized for text often suppress orthogonal visual features. We demonstrate that standard SVD fails to resolve this cross-modal mismatch, causing catastrophic visual degradation. To bridge this gap, we introduce Joint-Whitening SVD (JW-SVD), a dual-objective framework that aligns vision and language manifolds via a Joint Covariance basis, preserving features critical to both. Additionally, we propose Global Spectrum-Aware Truncation to dynamically transfer parameter budget from the redundant Vision Tower to the sensitive Backbone. Experiments on Qwen2.5-VL and Llama-3-Next confirm that JW-SVD demonstrates superior retention of both text and image capabilities. In addition,

it resolves the modality trade-off: it recovers over 30% of perceptual performance lost by baselines while maintaining parity in textual reasoning, enabling robust multimodal performance even at extreme compression rates.

[Long] KCVR: Knowledge-Centric Video Reconstruction for Structured Pedagogical Summarization via Dynamic Graph Planning ▶ *Underline Video*

Jingjiang Liu, Jia Zhu, Hanghui Guo, Weijie Shi, Yue Cui, Xiaokang Jin, Yilin Wang, Qingyu Niu, Jiawei Shen, Guoqing Ma, Yidan Liang, Shimin Di and Jiajie Xu

Existing video summarization methods mainly compress content for gist browsing, but they often break the prerequisite logic in instructional videos and induce logical inversions (e.g., conclusions before premises). We formalize this problem as Structure-Pedagogical Reconstruction (SPR). SPR raises two challenges: (1) Structure Hallucination, where retrieved knowledge is topologically valid but not evidence-grounded by the blackboard; and (2) Logical Inversion, where soft prompt-level graph injection fails to enforce prerequisite order during decoding. To address these challenges, we propose Knowledge-Centric Video Reconstruction (KCVR), a Plan-then-Generate neuro-symbolic framework that decouples epistemic planning from content generation. KCVR prunes a Dual-Layer Epistemic Graph into a minimal video-supported plan, then realizes the plan with visually anchored attention and topology-constrained decoding. We additionally release EduStruct, a 10-discipline benchmark for SPR and structure-centric evaluation. Experiments show that KCVR outperforms strong end-to-end baselines on Knowledge Progression Consistency and Learning Objective Coverage. Our code and data are available at https://github.com/mark1001-ljj/video_sum.

[Long] Knowledge-driven Augmentation and Retrieval for Integrative Temporal Adaptation ▶ *Underline Video*

Weisi Liu, Guangzeng Han and Xiaolei Huang

Time introduces fundamental challenges in model development and deployment: models are usually trained on historical data while deployed on future data where semantic distributions and domain knowledge may evolve. Unfortunately, existing studies either overlook temporal shifts or hardly capture rich shifting patterns of both semantic and knowledge. We develop Knowledge-driven Augmentation and Retrieval for Integrative Temporal Adaptation (KARITA) to capture diverse temporal shifts (e.g., uncertainty and feature shift), construct and integrate rich knowledge sources (e.g., medical ontology like MeSH), and leverage shifting insights for selecting-retrieval augmented learning. We evaluate KARITA on classification tasks across multiple domains, clinical, legal, and scientific corpora, demonstrating consistent improvements across all domain with temporal adaptation. Our results show that knowledge integration can be more critical and effective in temporal augmentation and learning.

[Findings] LAD: Learning Advantage Distribution for Reasoning ▶ *Underline Video*

Wendi Li and Sharon Li

Current reinforcement learning objectives for large-model reasoning primarily focus on maximizing expected rewards. This paradigm can lead to overfitting to dominant reward signals, while neglecting alternative yet valid reasoning trajectories, thereby limiting diversity and exploration. To address this issue, we introduce Learning Advantage Distributions (LAD), a distribution-matching framework that replaces advantage maximization with learning the advantage-induced distribution. By establishing the equivalence between the optimal policy update and an advantage-based target distribution, we derive a practical LAD objective formulated as minimizing an f-divergence between the policy-induced and advantage-induced distributions. This yields a gradient update that increases likelihood for high-advantage responses while suppressing over-confident probability growth, preventing collapse without requiring auxiliary entropy regularization. LAD incurs no extra training cost compared to GRPO and scales naturally to LLM post-training. In a controlled bandit setting, LAD faithfully recovers the multimodal advantage distribution, validating the theoretical formulation. Experiments on math and code reasoning tasks across several LLM backbones show that LAD reliably improves both accuracy and generative diversity.

[Findings] Large Language Models Can Help Mitigate Barren Plateaus in Quantum Neural Networks ▶ *Underline Video*

Jun Zhuang and Chaowen Guan

In the era of noisy intermediate-scale quantum (NISQ) computing, Quantum Neural Networks (QNNs) have emerged as a promising approach for various applications, yet their training is often hindered by barren plateaus (BPs), where gradient variance vanishes exponentially as the qubit size increases. Most initialization-based mitigation strategies rely heavily on pre-designed static parameter distributions, thereby lacking adaptability to diverse model sizes or data conditions. To address these limitations, we propose AdaInit, a foundational framework that leverages large language models with the submartingale property to iteratively synthesize initial parameters for QNNs that yield non-negligible gradient variance, thereby mitigating BPs. Unlike conventional one-shot initialization methods, AdaInit adaptively explores the parameter space by incorporating dataset characteristics and gradient feedback, with theoretical guarantees of convergence to finding a set of effective initial parameters for QNNs. We provide rigorous theoretical analyses of the submartingale-based process and empirically validate that AdaInit consistently outperforms existing initialization methods in maintaining higher gradient variance across various QNN scales. We believe this work may initiate a new avenue to mitigate BPs.

[Long] LASA: Language-Agnostic Semantic Alignment at the Semantic Bottleneck for LLM Safety ▶ *Underline Video*

Junxiao Yang, Haoran Liu, Jinzhe Tu, Jiale Cheng, Zhexin Zhang, Shiyao Cui, Jiaqi Weng, Jialing Tao, Hui Xue, Hongning Wang, Han Qiu and Minlie Huang

Large language models (LLMs) have demonstrated better safety performance in high-resource languages than in low-resource languages. We

attribute this issue as a mismatch gap between language-agnostic semantic understanding ability and language dominant safety alignment biased toward high-resource languages. Based on above insights, we empirically identify the semantic bottleneck in LLMs: intermediate layers in which the geometry of model representations is governed primarily by shared semantic content rather than language identity. Then, we propose Language-Agnostic Semantic Alignment (LASA), which anchors safety alignment directly in semantic bottlenecks. Experiments show that LASA substantially improves safety across all languages: average attack success rate (ASR) drops from 24.7% to 2.8% on LLaMA-3.1-8B-Instruct and remains within 3-4% across Qwen2.5 and Qwen3 Instruct models (7B-32B). Besides, our analysis and method offer a representation-level perspective on LLM safety, suggesting that safety alignment requires anchoring safety understanding not in surface text, but in the model's language-agnostic semantic space.

[Long] LazyEviction: Lagged KV Eviction with Attention Pattern Observation for Efficient Long Reasoning ► *Underline Video*

Haoyue Zhang, Hualei Zhang, Xiaosong Ma, Jie Zhang and Song Guo

Large Language Models (LLMs) exhibit enhanced capabilities by Chain-of-Thought reasoning. However, the extended reasoning sequences introduce significant GPU memory overhead due to increased key-value (KV) cache. Existing KV cache compression methods mitigate memory bottlenecks but struggle in long reasoning tasks. In this paper, we analyze attention patterns in reasoning tasks and reveal a Token Importance Recurrence phenomenon: a large proportion of tokens regain high attention after multiple decoding steps, which is failed to capture by existing works and may lead to unpredictable eviction on such periodically critical tokens. To address this, we propose LazyEviction, an observation window-based lagged eviction framework retaining latent recurring tokens by prioritized eviction based on tokens' recurrence patterns. Extensive experiments demonstrate that LazyEviction reduces KV cache by 50%~70% while maintaining comparable accuracy, outperforming existing KV cache baselines. Our implementation code can be found at https://github.com/Halo-949/LazyEviction.

[Findings] Learning to Retrieve User History and Generate User Profiles for Personalized Persuasiveness Prediction ► *Underline Video*

Sejun Park, Yoonah Park, Jongwon Lim and Yohan Jo

Estimating the persuasiveness of messages is critical in various applications, from recommender systems to safety assessment of LLMs. While it is imperative to consider the target persuadee's characteristics, such as their values, experiences, and reasoning styles, there is currently no established systematic framework to optimize leveraging a persuadee's past activities (e.g., conversations) to the benefit of a persuasiveness prediction model. To address this problem, we propose a context-aware user profiling framework with two trainable components: a query generator that generates optimal queries to retrieve persuasion-relevant records from a user's history, and a profiler that summarizes these records into a profile to effectively inform the persuasiveness prediction model. Our evaluation on the ChangeMyView Reddit dataset shows consistent improvements over existing methods across multiple predictor models, raising F1 from 33% to 47% on Llama-3.3-70B-Instruct. Further analysis shows that effective user profiles are context-dependent and predictor-specific, rather than relying on static attributes or surface-level similarity. Together, these results highlight the importance of task-oriented, context-dependent user profiling for personalized persuasiveness prediction.

[Long] Locate and Explain: Joint Multimodal Emotion Cause Extraction and Summarization in Conversation ► *Underline Video*

Jikun Wan, Chen Gong and Guohong Fu

Multimodal emotion cause analysis in conversation aims to identify the causes of emotions by leveraging multimodal information. Existing studies mainly formulate this problem as either utterance-level emotion cause extraction, which provides clear cause localization but limited explanation, or multimodal emotion cause generation, which offers fine-grained explanations but lacks explicit traceability to source utterances. Moreover, existing datasets rely heavily on human judgment and lack well-defined structured theoretical guidance, leading to subjective and inconsistent annotations. To address these issues, we introduce joint Multimodal Emotion Cause Extraction and Summarization in conversation (MECES), a new task that simultaneously extracts emotion cause utterances and generates cause summaries, enabling both precise localization and interpretable explanations of emotion cause. We further construct a MECES dataset guided by the Activating Events-Beliefs-Consequences theory from psychology. This dataset consists of 5,787 emotion utterances annotated with causes, comprising 12,231 emotion-cause pairs and 6,040 cause summaries. We also propose an effective end-to-end joint learning approach for MECES task, establishing strong benchmark results for this newly introduced task and dataset.

[Findings] Logit Arithmetic Elicits Long Reasoning Capabilities Without Training ► *Underline Video*

Yunxiang Zhang, Muhammad Khalifa, Lechen Zhang, Xin Liu, Ayoung Lee, Xinliang Frederick Zhang, Farima Fatahi Bayat and Lu Wang

Large reasoning models exhibit long chain-of-thought reasoning with complex strategies such as backtracking and self-verification. Yet, these capabilities typically require resource-intensive post-training. We investigate whether such behaviors can be elicited in large models without any gradient updates. To this end, we propose a decoding-time approach, ThinkLogit, which utilizes logit arithmetic to transfer these capabilities from a substantially smaller reasoning guider to a large non-reasoning target. We further show that we can boost performance by training the guider to correct the target's errors using preference optimization over mixed model outputs, a setup we refer to as ThinkLogit-DPO. We evaluate these methods across six reasoning benchmarks spanning math, science, and coding domains using the Qwen2.5-32B guided by R1-Distill-Qwen-1.5B, a model 21x smaller. Our experiments demonstrate that ThinkLogit and ThinkLogit-DPO achieve a relative improvement of 21.5% and 24.2%, respectively, over the target model. Moreover, ThinkLogit remains effective even when the guider and target come from different model families. Crucially, our method requires zero training for the large model and would incur minimal inference overhead when logits are computed in parallel, presenting a practical solution for enabling long reasoning at scale.

[Findings] LongCLI-Bench: A Preliminary Benchmark and Study for Long-horizon Agentic Programming in Command-Line Interfaces ▶ *Underline Video*

Yukang Feng, Jianwen Sun, Zelai Yang, Jiaxin Ai, Chuanhao Li, Zizhen Li, Fanrui Zhang, Kang He, Rui Ma, Jifan Lin, Jie Sun, Yang Xiao, Sizhuo Zhou, Wenxiao Wu, Yiming Liu, Pengfei Liu, Shenglin Zhang and Kaipeng Zhang

Recent advances in AI-assisted programming have empowered agents to execute complex workflows via command-line interfaces, however, existing benchmarks are limited by short task horizons, data contamination from GitHub scraping, and a lack of fine-grained evaluation metrics, fail to rigorously evaluate the long-horizon planning and execution capabilities essential for realistic software engineering. To address these gaps, we introduce LongCLI-Bench, a comprehensive benchmark designed to evaluate agentic capabilities across long-horizon, realistic, sequential engineering tasks. We curated 20 high-quality, long-horizon tasks from over 1,000 computer science assignments and real-world workflows, covering four engineering categories: from scratch, feature addition, bug fixing, and refactoring. LongCLI-Bench employs a dual-set testing protocol, which measures requirement fulfillment (fail(->)pass) and regression avoidance (pass(->)pass), and incorporates step-level scoring to pinpoint execution failures. Extensive experiments reveal that even state-of-the-art agents achieve pass rates below 20% in LongCLI-Bench. Step-level analysis further indicates that the majority of tasks stall at less than 30% completion, highlighting that critical failures often occur in the early stages. Although self-correction offers marginal gains, human-agent collaboration through plan injection and interactive guidance yields significantly higher improvements. These results highlight that future research must emphasize the development of synergistic human-agent workflows alongside advances in agents planning and execution capabilities to overcome key challenges in long-horizon task performance.

[Long] MAGIC: Deep Geometric Evolution with Structural Consensus for Temporal Knowledge Graph Reasoning ▶ *Underline Video*

Chengao Liu, Yuan Li, Yingze Wang and Jianbin Jiao

Temporal Knowledge Graph (TKG) reasoning remains challenging to characterize with conventional flat representations due to its intrinsic heterogeneous structure. Existing multi-geometry approaches face two key bottlenecks: 1) the Riemannian depth barrier driven by numerical instability, which restricts models to shallow architectures; and 2) gate collapse, where adaptive fusion mechanisms suffer from gradient starvation and degenerate into single-geometry solutions. To this end, we propose MAGIC (Multi-geometry Annealing Graph Interaction with Consensus). Our framework introduces a Tangent-Residual Engine in multi-geometric spaces, which enables the first stable 8-layer geometric evolution and reveals a phenomenon termed Geometric Annealing, where manifold curvature spontaneously evolves from semantic flatness in shallow layers to structural complexity in deeper layers. We further design an explicit reasoning module with structural consensus, leveraging geometric invariants and structural priors to regulate gradient flow, prevent collapse, and ensure robust synergy across Hyperbolic, Spherical, and Euclidean spaces. Experiments show that MAGIC achieves state-of-the-art performance in TKG reasoning, improving MRR by up to 2.9 points.

[Findings] Making MLLMs Blind: Adversarial Smuggling Attacks in MLLM Content Moderation ▶ *Underline Video*

Zhiheng Li, Zongyang Ma, Yuntong Pan, Ziqi Zhang, Xiaolei Lv, Bo Li, Jun Gao, Jianing Zhang, Chunfeng Yuan, Bing Li and Weiming Hu

Multimodal Large Language Models (MLLMs) are increasingly being deployed as automated content moderators. Within this landscape, we uncover a critical threat: Adversarial Smuggling Attacks. Unlike adversarial perturbations (for misclassification) and adversarial jailbreaks (for harmful output generation), adversarial smuggling exploits the Human-AI capability gap. It encodes harmful content into human-readable visual formats that remain AI-unreadable, thereby evading automated detection and enabling the dissemination of harmful content. We classify smuggling attacks into two pathways: (1) Perceptual Blindness, disrupting text recognition; and (2) Reasoning Blockade, inhibiting semantic understanding despite successful text recognition. To evaluate this threat, we constructed SmuggleBench, the first comprehensive benchmark comprising 1,700 adversarial smuggling attack instances. Evaluations on SmuggleBench reveal that both proprietary (e.g., GPT-5) and open-source (e.g., Qwen3-VL) SOTA models are vulnerable to this threat, producing Attack Success Rates (ASR) exceeding 90%. By analyzing the vulnerability through the lenses of perception and reasoning, we identify three root causes: the limited capabilities of vision encoders, the robustness gap in OCR, and the scarcity of domain-specific adversarial examples. We conduct a preliminary exploration of mitigation strategies, investigating the potential of test-time scaling (via CoT) and adversarial training (via SFT) to mitigate this threat.

[Long] MemCoRL: Alternating Co-Optimization of Memory Retrieval and Utilization via Collaborative Reinforcement Learning ▶ *Underline Video*

Yuewen Liu, Peng Xu, Muxi Diao, Anyi Zhang, Yang Li and Yutong Zhang

Large Language Models (LLMs) are inherently constrained by their fixed-length context windows, which limits LLMs' ability to retain and utilize information across long-term interactions. To address this limitation, recent work has proposed external memory modules for LLMs. Using memory modules typically involves two stages: evidence retrieval and memory utilization. While prior work focuses on the architecture of memory modules and the retrieval stage, the equally critical memory utilization stage remains underexplored. Building on this, we propose MemCoRL, a two-stage alternating co-optimization reinforcement learning method. Stage 1 optimizes evidence retrieval using citation feedback and semantic accuracy from utilization as rewards. Stage 2 optimizes utilization with rewards combining semantic similarity and lexical overlap. Iterative co-optimization establishes a positive feedback loop: better retrieval improves memory utilization, which in turn refines retrieval rewards. Experimental results show our approach outperforms the leading baselines on both lexical overlap and semantic similarity metrics, confirming the co-optimization in memory retrieval and memory utilization.

[Findings] MemoBrain: Executive Memory as an Agentic Brain for Reasoning ▶ *Underline Video*

Hongjin Qian, Zhao Cao and Zheng Liu

Complex reasoning in tool-augmented agent frameworks is inherently long-horizon, causing reasoning traces and transient tool artifacts to accumulate and strain the bounded working context of large language models. Without explicit memory mechanisms, such accumulation disrupts logical continuity and undermines task alignment. This positions memory not as an auxiliary efficiency concern, but as a core component for sustaining coherent, goal-directed reasoning over long horizons. We propose MemoBrain, an executive memory model for tool-augmented agents that constructs a dependency-aware memory over reasoning steps, capturing salient intermediate states and their logical relations. Operating as a co-pilot alongside the reasoning agent, MemoBrain organizes reasoning progress without blocking execution and actively manages the working context. Specifically, it prunes invalid steps, folds completed sub-trajectories, and preserves a compact, high-salience reasoning backbone under a fixed context budget. Together, these mechanisms enable explicit cognitive control over reasoning trajectories rather than passive context accumulation. We evaluate MemoBrain on challenging long-horizon benchmarks, including GAIA, WebWalker, and BrowseComp-Plus, demonstrating consistent improvements over strong baselines.

[Long] METRO: Towards Strategy Induction from Expert Dialogue Transcripts for Non-collaborative Dialogues ▶ *Underline Video*

Haofu Yang, Jiaji Liu, Chen Huang, Faguo Wu, Wenqiang Lei and See-Kiong Ng

Developing non-collaborative dialogue agents traditionally requires the manual, unscalable codification of expert strategies. We propose METRO, a method that leverages large language models to autonomously induce both strategy actions and planning logic directly from raw transcripts. METRO formalizes expert knowledge into a Strategy Forest, a hierarchical structure that captures both short-term responses (nodes) and long-term strategic foresight (branches). Experimental results across two benchmarks show that METRO demonstrates promising performance, outperforming existing methods by an average of 9%-10%. Our further analysis not only reveals the success behind METRO (strategic behavioral diversity and foresight), but also demonstrates its robust cross-task transferability. This offers new insights into building non-collaborative agents in a cost-effective and scalable way. Our code is available at https://github.com/Humphrey-0125/METRO.

[Findings] Mina: A Multilingual LLM-Powered Legal Assistant Agent for Empowering Access to Justice in Bangladesh ▶ *Underline Video*

Azmine Toushik Wasi, Wahid Faisal, Mst Rafia Islam and Md Rizwan Parvez

Bangladesh's low-income population faces major barriers to affordable legal advice due to complex legal language, procedural opacity, and high costs. Existing AI legal assistants lack Bengali-language support and jurisdiction-specific adaptation, limiting their effectiveness. To address this, we developed Mina, a multilingual LLM-based legal assistant tailored for the Bangladeshi context. It employs multilingual embeddings and a RAG-based chain-of-tools framework for retrieval, reasoning, translation, and document generation, delivering context-aware legal drafts, citations, and plain-language explanations via an interactive chat interface. Evaluated by law faculty from leading Bangladeshi universities across all stages of the 2022 and 2023 Bangladesh Bar Council examinations, Mina achieved scores of 75-80% in the preliminary MCQs, written, and simulated viva voce components. These results matched or surpassed average human performance, demonstrating strong clarity, contextual understanding, and sound legal reasoning, while operating at approximately 0.1-0.6% of the cost of human lawyers. These results confirm its potential as a low-cost, multilingual AI assistant that automates key legal tasks and scales access to justice, offering a real-world details on building domain-specific, low-resource systems and addressing challenges of multilingual adaptation, efficiency, and sustainable public-service AI deployment.

[Long] Mitigating Structural Knowledge Collapse in Domain-Specific LLMs via Morpheme-Aware KV-Aggregation ▶ *Underline Video*

Yuxuan Si, Zheqi Lv, Chengxi Zang, Zhengyu Chen and Fei Wu

Standard tokenizers over-fragment domain terms, disrupting morpheme semantics. We characterize this representational misalignment as Structural Knowledge Collapse (SKC), where attention mechanisms fail to reconstruct coherent concepts from fragmented inputs. While existing input-centric solutions like vocabulary expansion address this, they necessitate expensive embedding retraining and neglect internal attention compositionality. To this end, we introduce Morpheme-aware KV-aggregation Attention (MorphKA), a lightweight adapter that dynamically consolidates fragments without tokenizer changes. Bypassing tokenizer retraining, MorphKA employs a dual-phase strategy, Input-Level Morpheme Aggregation (IMA) and Context-Aware KV-Aggregation (AMRF), to stabilize morpheme spans and synthesize higher-order concepts. Experiments on medical and legal benchmarks show MorphKA outperforms vocabulary adaptation baselines by 3.2–4.6%, reaching 7.9% on high-fragmentation terms. Moreover, MorphKA reduces catastrophic interference on general capabilities by 18–22% with ~80% fewer parameters than embedding retraining approaches.

[Long] Model-Based Imaginative Planning for Embodied Agents ▶ *Underline Video*

Junru Song, Hengzhe Jin, Yucong Huang, Tingsong Jiang, Weien Zhou, Feifei Wang, Yang Yang, Ying Wen and Wen Yao

Reasoning and planning critically rely on a predictive dynamics model. In symbolic domains such as mathematics and code, large language models (LLMs) internalize transition rules during pretraining, allowing reinforcement learning or test-time scaling to effectively elicit and generalize their reasoning ability. Embodied decision making is fundamentally different: agents must reason from sparse visual evidence under partial observability, while coping with environment-specific dynamics and affordances not captured by language priors. Here we propose IMPLEMENT, a model-based reasoning framework that enables frozen LLMs to perform imaginative planning. A lightweight

world model converts raw pixels into object-centric symbolic states amenable to language-based reasoning, and predicts their evolution under hypothetical actions. To address partial observability, we perform Monte Carlo state prediction via temperature sampling, enabling decision evaluation over multiple plausible futures. To support adaptation to unseen environments, we integrate Meta In-Context Learning, conditioning the world model on interaction history to continuously refine its predictions. At inference time, the LLM and world model form a tight co-reasoning loop: the LLM proposes candidate actions, the world model simulates future trajectories, and the LLM refines its decisions, effectively inducing an online policy iteration scheme. Extensive experiments in ALFWorld demonstrate consistent advantages over finetuning-based and strong test-time scaling approaches, validating IMPLEMENT as an effective framework for grounding language agents in visual embodied environments.

[Findings] MT-OSC: Path for LLMs that Get Lost in Multi-Turn Conversation ▶ *Underline Video*

Jyotika Singh, Fang Tu, Miguel Ballesteros, Weiyi Sun, Sandip Ghoshal, Michelle Yuan, Yassine Benajiba, Sujith Ravi and Dan Roth

Large language models (LLMs) suffer significant performance degradation when user instructions and context are distributed over multiple conversational turns, yet multi-turn (MT) interactions dominate chat interfaces. The routine approach of appending full chat history to prompts rapidly exhausts context windows, leading to increased latency, higher computational costs, and diminishing returns as conversations extend. We introduce MT-OSC, a One-off Sequential Condensation framework that efficiently and automatically condenses chat history in the background without disrupting the user experience. MT-OSC employs a Condenser Agent that uses a few-shot inference-based Condenser and a lightweight Decider to selectively retain essential information, reducing token counts by up to 72% in 10-turn dialogues. Evaluated across 13 state-of-the-art LLMs and diverse multi-turn benchmarks, MT-OSC consistently narrows the multi-turn performance gap, yielding improved or preserved accuracy across datasets while remaining robust to distractors and irrelevant turns. Our results establish MT-OSC as a scalable solution for multi-turn chats, enabling richer context within constrained input spaces, reducing latency and operational cost, while balancing performance.

[Long] MTAVG-Bench: A Diagnostic Benchmark for Multi-Talker Dialogue-Centric Audio-Video Generation ▶ *Underline Video*

Yanghao Zhou, Haitian Li, Rexar Lin, Heyan Huang, Jinxing Zhou, Changsen Yuan, Tian Lan, Ziqin Zhou, Yudong Li, Jiajun Xu, Jingyun Liao, YiMing Cheng, Xuefeng Chen, Xian-Ling Mao and Yousheng Feng

Recent advances in text-to-audio-video (T2AV) generation have enabled models to synthesize audio-visual videos with multi-participant dialogues. However, existing evaluation benchmarks remain largely designed for human-recorded videos or single-speaker settings. As a result, structural failures in generated multi-talker dialogue videos, such as identity drift, unnatural turn transitions, and audio-visual misalignment, cannot be effectively diagnosed. To address this issue, we introduce MTAVG-Bench, a failure-driven diagnostic benchmark for multi-talker dialogue-centric audio-video generation. MTAVG-Bench is built via a semi-automatic pipeline, where 1.8k videos are generated using mainstream T2AV models with carefully designed prompts, yielding 2.4k manually annotated QA pairs for fine-grained failure diagnosis. The benchmark evaluates multi-speaker dialogue generation at four levels: audio-visual signal fidelity, temporal attribute consistency, social interaction, and cinematic expression. Built on a hierarchical failure taxonomy and a targeted QA protocol, MTAVG-Bench is primarily designed to evaluate whether proprietary and open-source omni-models can reliably identify failure modes in multi-speaker T2AV outputs. We benchmark 12 proprietary and open-source omni-models on MTAVG-Bench, with Gemini 3 Pro achieving the strongest overall performance, while leading open-source models remain competitive in signal fidelity and consistency. Overall, MTAVG-Bench enables fine-grained failure analysis for rigorous model comparison and targeted video generation refinement.

[Findings] Multimodal Chemical Structure-Text Coreference in Intellectual Property via Rule-guided Reinforcement Learning ▶ *Underline Video*

Hanmeng Zhong, Wentao Wu, Linqing Chen and Peng Zhou

Navigating biopharmaceutical intellectual property necessitates precisely associating visual chemical structures with their textual referents across lengthy documents. Despite its critical role in drug discovery, this multimodal coreference task remains underexplored. It presents unique challenges, including handling Markush structures and distinguishing the atom-level differences between adjacent structures. To bridge this gap, we define the multimodal Chemical Structure-Text coreference and introduce CheST, the first dataset explicitly designed for the task. Furthermore, to satisfy the strict logical consistency in the task, we propose RULER, a RULE-guided multimodal Reinforcement learning framework built upon an SFT cold start. RULER utilizes rule-driven reward functions operationalizing multidimensional consistencies, acting as a domain-specific "verifier" to obtain the correct domain knowledge. Experimental results demonstrate that RULER achieves a 40% improvement over the strongest baseline–Gemini-2.5-Pro, demonstrating the superior efficacy.

[Findings] Music Audio-Visual Question Answering Requires Specialized Multimodal Designs ▶ *Underline Video*

Wenhao You, Xingjian Diao, Wenjun Huang, Chunhui Zhang, Keyi Kong, Weiyi Wu, Chiyu Ma, Zhongyu Ouyang, Tingxuan Wu, Ming Cheng, Soroush Vosoughi and Jiang Gui

While recent Multimodal Large Language Models exhibit impressive capabilities for general multimodal tasks, specialized domains like music necessitate tailored approaches. Music Audio-Visual Question Answering (Music AVQA) particularly underscores this, presenting unique challenges with its continuous, densely layered audio-visual content, intricate temporal dynamics, and the critical need for domain-specific knowledge. Through a systematic analysis of Music AVQA datasets and methods, this paper identifies that specialized input processing, architectures incorporating dedicated spatial-temporal designs, and music-specific modeling strategies are critical for success in this domain. Our study provides valuable insights for researchers by highlighting effective design patterns empirically linked to strong

performance, proposing concrete future directions for incorporating musical priors, and aiming to establish a robust foundation for advancing multimodal musical understanding. We aim to encourage further research in this area and provide a GitHub repository of relevant works: https://github.com/WenhaoYou1/Survey4MusicAVQA.

[Long] Musical Score Understanding Benchmark: Evaluating Large Language Models' Comprehension of Complete Musical Scores ▶ *Underline Video*

Congren Dai

Understanding complete musical scores entails integrated reasoning over pitch, rhythm, harmony, and large-scale structure, yet the ability of Large Language Models and Vision–Language Models to interpret full musical notation remains insufficiently examined. We introduce Musical Score Understanding Benchmark (MSU-Bench), a human-curated benchmark for score-level musical understanding across textual (ABC notation) and visual (PDF) modalities. MSU-Bench contains 1,800 generative question–answer pairs from works by Bach, Beethoven, Chopin, Debussy, and others, organised into four levels of increasing difficulty, ranging from onset information to texture and form. Evaluations of more than fifteen state-of-the-art models, in both zero-shot and fine-tuned settings, reveal pronounced modality gaps, unstable level-wise performance, and challenges in maintaining multilevel correctness. Fine-tuning substantially improves results across modalities while preserving general knowledge, positioning MSU-Bench as a robust foundation for future research in multimodal reasoning. The benchmark and code are available at https://github.com/Congren-Dai/MSU-Bench.

[Long] MVP: Enhancing Video Large Language Models via Self-supervised Masked Video Prediction ▶ *Underline Video*

Xiaokun Sun, Zezhong Wu, Zewen Ding and Linli Xu

Reinforcement learning based post-training paradigms for Video Large Language Models (VideoLLMs) have achieved significant success by optimizing for visual-semantic tasks such as captioning or VideoQA. However, while these approaches effectively enhance perception abilities, they primarily target holistic content understanding, often lacking explicit supervision for intrinsic temporal coherence and inter-frame correlations. This tendency limits the models' ability to capture intricate dynamics and fine-grained visual causality. To explicitly bridge this gap, we propose a novel post-training objective: Masked Video Prediction (MVP). By requiring the model to reconstruct a masked continuous segment from a set of challenging distractors, MVP forces the model to attend to the sequential logic and temporal context of events. To support scalable training, we introduce a scalable data synthesis pipeline capable of transforming arbitrary video corpora into MVP training samples, and further employ Group Relative Policy Optimization (GRPO) with a fine-grained reward function to enhance the model's understanding of video context and temporal properties. Comprehensive evaluations demonstrate that MVP enhances video reasoning capabilities by directly reinforcing temporal reasoning and causal understanding.

[Findings] Navigating the Alignment-Calibration Trade-off: A Pareto-Superior Frontier via Model Merging ▶ *Underline Video*

Tiancheng Hu, Benjamin Minixhofer and Nigel Collier

The "alignment tax" of post-training is typically framed as a drop in task accuracy. We show it also involves a severe loss of calibration, making models overconfident, less reliable, and model outputs less diverse. We demonstrate that this trade-off can be navigated effectively via a simple post-hoc intervention: interpolating between a model's weights before and after alignment. Crucially, this is not a strict trade-off. We find that the process consistently reveals Pareto-optimal interpolations, models that improve accuracy beyond both parents while substantially recovering the calibration lost during alignment. Our work demonstrates that simple model merging provides a computationally efficient method for mitigating the full scope of the alignment tax, yielding models that are more capable and more reliable.

[Long] Not All Directions Matter: Towards Structured and Task-Aware Low-Rank Model Adaptation ▶ *Underline Video*

Xi Xiao, Chenrui Ma, Yunbei Zhang, Chen Liu, Zhuxuanzi Wang, Yanshu Li, Lin Zhao, Guosheng Hu, Tianyang Wang and Hao Xu

Low-Rank Adaptation (LoRA) has become a cornerstone of parameter-efficient fine-tuning (PEFT). Yet, its efficacy is hampered by two fundamental limitations: semantic drift, arising from treating all update directions with equal importance, and structural incoherence, due to adapting layers independently, resulting in uncoordinated and suboptimal updates. To address these issues, we propose StructLoRA, a framework that tackles both limitations through a principled dual-component design: (1) an Information Bottleneck-guided filter that prunes task-irrelevant directions to mitigate semantic drift, and (2) a lightweight, training-only graph-based coordinator that enforces inter-layer consistency to resolve structural incoherence. Extensive experiments across large language models, vision language models, and vision models (including LLaMA, LLaVA, and ViT) demonstrate that StructLoRA consistently establishes a new state of the art, outperforming not only vanilla LoRA but also advanced dynamic rank allocation and sparsity-based methods. Notably, the gains are particularly pronounced in challenging low-rank and low-data regimes. Crucially, since the proposed modules operate only during training, StructLoRA improves performance with zero additional inference cost, shifting the focus of PEFT from mere parameter compression to a more holistic optimization of information quality and structural integrity.

[Long] ODUTQA-MDC: A Task for Open-Domain Underspecified Tabular QA with Multi-turn Dialogue-based Clarification ▶ *Underline Video*

Zhensheng Wang, Zhanteng Lin, Wenmian Yang, Kun Zhou, Yiquan Zhang and Weijia Jia

The advancement of large language models (LLMs) has enhanced tabular question answering (Tabular QA), yet they struggle with open-domain queries exhibiting underspecified or uncertain expressions. To address this, we introduce the ODUTQA-MDC task and the first comprehensive benchmark to tackle it. This benchmark includes: (1) a large-scale ODUTQA dataset with 209 tables and 25,105 QA

pairs; (2) a fine-grained labeling scheme for detailed evaluation; and (3) a dynamic clarification interface that simulates user feedback for interactive assessment. We also propose MAIC-TQA, a multi-agent framework that excels at detecting ambiguities, clarifying them through dialogue, and refining answers. Experiments validate our benchmark and framework, establishing them as a key resource for advancing conversational, underspecification-aware Tabular QA research.

[Long] On the (In-)Security of the Shuffling Defense in the Transformer Secure Inference ▶ *Underline Video*

Zhengyi Li, Yakai Wang, Jingwen Leng, Kang Yang, Yu Yu, Jiaping Gui, Yu Feng, Ning Liu and Minyi Guo

For Transformer models, cryptographically secure inference ensures that the client learns only the final output, while the server learns nothing about the clients input. However, securely computing nonlinear layers remains a major efficiency bottleneck due to the substantial communication rounds and data transmission required. To address this issue, prior works reveal intermediate activations to the client, allowing nonlinear operations to be computed in plaintext. Although this approach significantly improves efficiency, exposing activations enables adversaries to extract model weights. To mitigate this risk, existing works employ a shuffling defense that reveals only randomly permuted activations to the client. In this work, we show that the shuffling defense is not as robust as previously claimed. We propose an attack that aligns differently shuffled activations to a common permutation and subsequently exploits them to extract model weights. Experiments on Pythia-70m and GPT-2 demonstrate that the proposed attack can align shuffled activations with mean squared errors ranging from 10^-9 to 10^-6. With a query cost of approximately 1, the adversary can recover model weights with L1-norm differences ranging from 10^-4 to 10^-2 compared to the oracle weights.

[Long] On-policy Reinforcement Fine-tuning with Offline reward for Multi-step Embodied Planning ▶ *Underline Video*

Di Wu, Jiaxin Fan, Chloe Gu, Guanbo Wang, Wei Yin, Wenhao Li and Bo Jin

Embodied planning requires agents to make coherent multi-step decisions based on dynamic visual observations and verbal goals. While recent vision-language models (VLMs) excel at static perception tasks, they struggle in interactive environments. Reinforcement learning (RL) offers a natural way to address this limitation, yet online RL approaches suffer from costly interaction and sparse rewards in embodied settings. This paper introduces ORBIT, an On-policy Reinforcement fine-tuning (RFT) framework with offline rewards for EmBodIed Task Planning, that preserves the generalization benefits of RFT while addressing the challenges of costly interaction and sparse rewards, supported by solid theoretical guarantees. Our approach is evaluated on EmbodiedBench, a recent benchmark for interactive embodied tasks, covering both in-domain and out-of-domain scenarios. Experimental results show that ORBIT achieves SOTA performance on EB-ALFRED, outperforming all closed-source and online-RL-based methods, while being substantially more efficient in training speed and computational cost, remaining robust to sub-optimal expert trajectories, and exhibiting strong generalization to unseen environments. We released all code and data at https://github.com/mail-taii/Reinforced-Reasoning-for-Embodied-Planning.

[Short] One-step Nonautoregressive Natural Language Generation with Shortcut Flow Matching Models ▶ *Underline Video*

Jdrzej Warczyski, Ondrej Dusek and Mateusz Lango

While having a significant potential for parallel processing in theory, diffusion-based non-autoregressive text generation remains inefficient due to the need for multiple denoising steps. Performance degrades sharply if a low number of steps is used, such as in flow matching. To enable accurate one-step generation, we propose a novel shortcut flow-matching model that learns to directly predict multi-step denoising outcomes in a single step. Experiments conducted on three datasets demonstrate consistent improvements over classic flow-matching, with BLEU scores more than doubling on two datasets. We also tested five different ways of extending shortcut models with commonly used techniques.

[Findings] OpenPhone: Mobile Agentic Foundation Models ▶ *Underline Video*

Yangqin Jiang and Chao Huang

With the advancement of multimodal large language models (MLLMs), building GUI agent systems has become an increasingly promising direction, especially for mobile platforms, given their rich app ecosystems and intuitive touch interactions. Yet mobile GUI agents face a critical dilemma: truly on-device models (4B or smaller) lack sufficient performance, while capable models (starting from 7B) are either too large for mobile deployment or prohibitively costly (e.g., cloud-only closed-source MLLMs). To resolve this, we propose OpenPhone, a mobile GUI agent system that leverages device-cloud collaboration to tap the cost-efficiency of on-device models and the high capability of cloud models, while avoiding their drawbacks. Specifically, OpenPhone enhances Qwen2.5-VL-3B via two-stage SFT to GRPO training on synthetic GUI data for strong decision-making, integrates an efficient long-reasoning mechanism to utilize historical interactions under tight resources, and defaults to on-device execution, only escalating challenging subtasks to the cloud via real-time complexity assessment. Experiments on the online AndroidLab benchmark and diverse apps show OpenPhone matches or nears larger models, with a significant reduction in cloud costs.

[Long] Optimizing Length Compression in Large Reasoning Models ▶ *Underline Video*

Zhengxiang Cheng, Dongping Chen, Mingyang Fu and Tianyi Zhou

Large Reasoning Models (LRMs) have achieved remarkable success, yet they often suffer from producing unnecessary and verbose reasoning chains. We identify a core aspect of this issue as "invalid thinking"— models tend to repeatedly double-check their work after having derived the correct answer. To address this specific inefficiency, we move beyond the general principles of Efficacy and Efficiency to propose two new, fine-grained principles: Brevity, which advocates for eliminating redundancy, and Sufficiency, which

ensures critical reasoning steps are preserved. Guided by these principles, we introduce LC-R1, a post-training method based on Group Relative Policy Optimization (GRPO). LC-R1 employs a novel combination of a Length Reward for overall conciseness and a Compress Reward that is specifically designed to remove the invalid portion of the thinking process. Extensive experiments on multiple reasoning benchmarks demonstrate that LC-R1 achieves a significant reduction in sequence length (\{}~50%) with only a marginal (\{}~2%) drop in accuracy, achieving a favorable trade-off point on the Pareto frontier that prioritizes high compression. Our analysis further validates the robustness of LC-R1 and provides valuable insights for developing more powerful yet computationally efficient LRMs.

[Long] Paper Circle: An Open-source Multi-agent Research Discovery and Analysis Framework ▶ *Underline Video*

Komal Kumar, Aman Chadha, Salman Khan, Fahad Shahbaz Khan and Hisham Cholakkal

The rapid growth of scientific literature has made it increasingly difficult for researchers to efficiently discover, evaluate, and synthesize relevant work. Recent advances in multi-agent large language models (LLMs) have demonstrated strong potential for understanding user intent and are being trained to utilize various tools. In this paper, we introduce Paper Circle, a multi-agent research discovery and analysis system designed to reduce the effort required to find, assess, organize, and understand academic literature. The system comprises two complementary pipelines: (1) a Discovery Pipeline that integrates offline and online retrieval from multiple sources, multi-criteria scoring, diversity-aware ranking, and structured outputs; and (2) an Analysis Pipeline that transforms individual papers into structured knowledge graphs with typed nodes (e.g., concepts, methods, experiments, and figures) and edges, enabling graph-aware question answering and coverage verification. Both pipelines are implemented within a coder LLM-based multi-agent orchestration framework and produce fully reproducible, synchronized outputs (JSON, CSV, BibTeX, Markdown, and HTML) at each agent step. This paper describes the system architecture, agent roles, retrieval and scoring methods, knowledge graph schema, and evaluation interfaces that together form the Paper Circle research workflow. We benchmark Paper Circle on both paper retrieval and paper review generation, reporting hit rate, MRR, and Recall@K. Results show consistent improvements with stronger agent models. We have publicly released the website:https://papercircle.vercel.app/ and code: https://github.com/MAXNORM8650/papercircle.

[Findings] Parallel-SFT: Improving Zero-Shot Cross-Programming-Language Transfer for Code RL ▶ *Underline Video*

Zhaofeng Wu, Shiqi Wang, Boya Peng, Anuj Kumar Goyal, Melanie Kambadur, Sebastian Ruder, Yoon Kim and Chloe Bi

Modern language models demonstrate impressive coding capabilities in common programming languages (PLs), such as C++ and Python, but their performance in lower-resource PLs is often limited by training data availability. In principle, however, most programming skills are universal across PLs, so the capability acquired in one PL should transfer to others. In this work, we propose the task of zero-shot cross-programming-language transfer for code RL. We find that, for Llama-3.1, RL training for code generation in a source PL fails to improve, and sometimes even degrades, the performance on other target PLs. To address this, we hypothesize that effective RL transfer requires a generalizable SFT initialization before RL. We thus propose Parallel-SFT, an SFT strategy that incorporates "parallel programs"—functionally equivalent code implemented in multiple PLs—into the data mixture. We demonstrate that this improves transferability: when we subsequently perform RL on our Parallel-SFT model, we observe better generalization to unseen PLs. Analysis of the model internal representations reveals that Parallel-SFT leads to a more functionality-centric latent space, where equivalent programs across PLs are more tightly clustered, which we hypothesize to contribute to the improved transferability.

[Long] Paramanu: Compact and Competitive Monolingual Language Models for Low-Resource Morphologically Rich Indian Languages ▶ *Underline Video*

Mitodru Niyogi, Eric Gaussier and Arnab Bhattacharya

Multilingual large language models (LLMs) are expensive to pretrain and often suffer from imbalances across languages and datasets, English-centric bias, tokenizer oversegmentation for morphologically rich low-resource languages, and the curse of multilinguality. We introduce PARAMANU, a family of Indian language-only autoregressive language models trained from scratch on open-source language-specific data for the five most spoken Indian languages: Bangla (Bengali), Hindi, Marathi, Tamil, and Telugu. All models are designed for affordability and are trained on a single GPU with a budget under $1,000, allowing under-resourced researchers to build competitive language models. To address low-resource challenges, we develop morphology-aligned, low-fertility tokenizers, and propose an interpolation-based method for token position indices in RoPE based scaling to train longer sequences efficiently. We also create instruction-tuning datasets in Bangla that are then translated to the other four languages. Despite their small size (108M-367M parameters), Paramanu achieves a strong performance-efficiency tradeoff and outperforms most larger multilingual models up to 8B across all five languages. The models and datasets are available at: https://huggingface.co/collections/mitodru/paramanu.

[Long] Piece of Table: A Divide-and-Conquer Approach for Selecting Subtables in Table Question Answering ▶ *Underline Video*

Wonjin Lee, Kyumin Kim, Sungjae Lee, Jihun Lee and Kwang In Kim

Applying language models (LMs) to tables is challenging due to the mismatch between the two-dimensional structure of tables and the one-dimensional inputs expected by LMs. This mismatch forces linearization, making LMs particularly sensitive to irrelevant cells. Subtable selection mitigates this challenge by isolating question-relevant content prior to answer generation. However, existing approaches either rely on independent row or column selection, failing to capture cross-row and cross-column dependencies, or attempt global reasoning and face challenges similar to holistic table QA under noisy contexts. We propose PieTa (Piece of Table), a divide-and-conquer subtable selection framework that progressively aggregates locally selected evidence without requiring explicit global reasoning. PieTa uses an iterative, window-based multi-resolution process to construct compact subtables that capture global dependencies while limiting LM exposure to

irrelevant content. Extensive experiments demonstrate that PieTa consistently outperforms prior subtable-based and holistic table QA approaches.

[Findings] PolyAudio: Advancing Multi-Audio Reasoning in Large Audio Language Models with Interleaved Multi-Audio Contexts ▶ *Underline Video*

Sonal Kumar, Sreyan Ghosh, Yueqian Lin, S Sakshi, Ashish Seth, Yiran Chen, Ramani Duraiswami and Dinesh Manocha

Large Audio Language Models have shown impressive performance on single-clip audio language tasks such as automatic speech recognition, captioning, and sound event recognition. Yet, their ability to reason over interleaved multi-audio contexts-where answering a query requires relating information across multiple audio clips-remains limited. We present PolyAudio, a LALM built on Audio Flamingo 3 that targets multi-audio understanding via instruction tuning rather than massive-scale pre-training, and PolyAudio-Instruct, a high-quality instruction-tuning dataset consisting of 1.3M+ QA pairs, spanning over 14 task subsets to empower multi-audio understanding and reasoning. PolyAudio uses an explicit interleaved representation with clip indexing to encourage faithful grounding and reduce ambiguity in multi-clip references. We evaluate PolyAudio on a diverse suite of multi-audio benchmarks alongside standard single-audio tasks. PolyAudio achieves strong performance on multi-audio reasoning, outperforming competitive baselines that are also often limited to reasoning over up-to 2 audio clips, while preserving robust single-clip performance. Overall, our results suggest that precise, academic-scale multi-audio instruction tuning can unlock advanced cross-clip reasoning capabilities, enabling more capable audio-centric assistants.

[Findings] ProFit: Leveraging High-Value Signals in SFT via Probability-Guided Token Selection ▶ *Underline Video*

Tao Liu, Taiqiang Wu, Runming Yang, Shaoning Sun, Junjie Wang and Yujiu Yang

Supervised fine-tuning (SFT) is a fundamental post-training strategy to align Large Language Models (LLMs) with human intent. However, traditional SFT often ignores the one-to-many nature of language by forcing alignment with a single reference answer, leading to the model overfitting to non-core expressions. Although our empirical analysis suggests that introducing multiple reference answers can mitigate this issue, the prohibitive data and computational costs necessitate a strategic shift: prioritizing the mitigation of single-reference overfitting over the costly pursuit of answer diversity. To achieve this, we reveal the intrinsic connection between token probability and semantic importance: high-probability tokens carry the core logical framework, while low-probability tokens are mostly replaceable expressions. Based on this insight, we propose ProFit, which selectively masks low-probability tokens to prevent surface-level overfitting. Extensive experiments confirm that ProFit consistently outperforms traditional SFT baselines on general reasoning and mathematical benchmarks

[Findings] PROGRAM: Programmatic Retrieval Optimization with Generative Reasoning and Augmented Multi-queries ▶ *Underline Video*

Gun Il Kim, Jungkyu Shin, Jong Wook Kim and Beakcheol Jang

Current retrieval-augmented generation (RAG) methods struggle with complex multi-hop reasoning, relying on unstructured semantic matching that lacks the logical structure needed to systematically guide retrieval. We introduce Programmatic Retrieval Optimization with Generative Reasoning and Augmented Multi-queries (PROGRAM), a novel framework that elevates retrieval to structured, program-guided reasoning. PROGRAM treats retrieval as execution of specific program types, such as logical, temporal, causal, and so forth, through three stages of 'Program-Type Selection' with dual-metric optimization, 'Iterative Active Program Pruning' with evidence accumulation, and 'Final Answer Generation' with reranking. Evaluated on five benchmarks including HotPotQA, 2WikiMultihopQA, ARC-Challenge, MMLU-Pro, and MedQA with various LLMs, PROGRAM achieves state-of-the-art performance with up to 24% relative improvement on HotPotQA and 13.2% on MedQA over strong baselines including FLARE, ProbTree and Self-RAG.

[Long] Protecting Bystander Privacy via Selective Hearing in Audio LLMs ▶ *Underline Video*

Xiao Zhan, Guangzhi Sun, Jose Such and Philip C. Woodland

Audio Large language models (LLMs) are increasingly deployed in the real world, where they inevitably capture speech from unintended nearby bystanders, raising privacy risks that existing benchmarks and defences did not consider. We introduce SH-Bench, the first benchmark designed to evaluate selective hearing: a model's ability to attend to an intended main speaker while refusing to process or reveal information about incidental bystander speech. SH-Bench contains 3,968 multi-speaker audio mixtures, including both real-world and synthetic scenarios, paired with 77k multiple-choice questions that probe models under general and selective operating modes. In addition, we propose Selective Efficacy (SE), a novel metric capturing both multi-speaker comprehension and bystander-privacy protection. Our evaluation of state-of-the-art open-source and proprietary LLMs reveals substantial bystander privacy leakage, with strong audio understanding failing to translate into selective protection of bystander privacy. To mitigate this gap, we also present Bystander Privacy Fine-Tuning (BPFT), a novel training pipeline that teaches models to refuse bystander-related queries without degrading main-speaker comprehension. We show that BPFT yields substantial gains, achieving an absolute 47% higher bystander accuracy under selective mode and an absolute 16% higher SE compared to Gemini 2.5 Pro, which is the best audio LLM without BPFT. Together, SH-Bench and BPFT provide the first systematic framework for measuring and improving bystander privacy in audio LLMs.

[Findings] Purging the Gray Zone: Latent-Geometric Denoising for Precise Knowledge Boundary Awareness ▶ *Underline Video*

Hao An, Yibin Lou, Jiayi Guo and Yang Xu

Large language models (LLMs) often exhibit hallucinations due to their inability to accurately perceive their own knowledge boundaries. Existing abstention fine-tuning methods typically partition datasets directly based on response accuracy, causing models to suffer from

severe label noise near the decision boundaries and consequently exhibit high rates of abstentions or hallucinations. This paper adopts a latent space representation perspective, revealing a "gray zone" near the decision hyperplane where internal belief ambiguity constitutes the core performance bottleneck. Based on this insight, we propose the GeoDe (Geometric Denoising) framework for abstention fine-tuning. This method constructs a truth hyperplane using linear probes and performs "geometric denoising" by employing geometric distance as a confidence signal for abstention decisions. This approach filters out ambiguous boundary samples while retaining high-fidelity signals for fine-tuning. Experiments across multiple models (Llama3, Qwen3) and benchmark datasets (TriviaQA, NQ, SciQ, SimpleQA) demonstrate that GeoDe significantly enhances model truthfulness and demonstrates strong generalization in out-of-distribution (OOD) scenarios. Code is available at https://github.com/Notbesidemoon/GeoDe.

[Long] RAG-on-a-Diet: A Reinforcement Learning-Based Dynamic Resource Optimization Framework for RAG ► *Underline Video*

Hongwen Ding and Yizheng Zhao

Retrieval-Augmented Generation (RAG) has become the backbone of knowledge-intensive multi-hop question answering, yet routing every sub-query through a frontier model turns every hop into a cost multiplier and makes real-world deployment prohibitively expensive. Existing remedies either fix the retrieval schedule, route once at the query level, or lack a principled stopping rule, leaving a critical gap: no framework adapts, hop by hop, to how a trajectory actually unfolds. We introduce RAG-on-a-Diet, a lightweight reinforcement-learning agent that treats each reasoning hop as an independent decision and selects the smallest model (Qwen3-4B, Qwen3-30B, or DS-R1-671B) sufficient for it, guided by entity- and confidence-aware features. Trained via behavior cloning followed by PPO under a five-component cost-aware reward (final, cumulative, step-wise, cost, balance) and coupled with an explicit two-tier termination policy (5-hop cap plus a tau=0.3 confidence gate), the agent carves a Pareto-optimal efficiency frontier. On HotpotQA it cuts Monetary Inference Cost by 60.07% against IRCoT with only a 3.7% F1 drop; it matches Adaptive-RAG's F1 at 37.30% lower cost; and it attains up to 2.33x higher Quality-per-Monetary-Cost. Consistent gains on MuSiQue, 2WikiMultiHopQA, CRAG, and Bamboogle confirm strong out-of-distribution robustness, setting a new paradigm for fine-grained resource control in multi-hop RAG.

[Findings] Reading Between the Lines: The One-Sided Conversation Problem ► *Underline Video*

Victoria Ebert, Rishabh Singh, Tuochao Chen, Noah A. Smith and Shyamnath Gollakota

Conversational AI is constrained in many real-world settings where only one side of a dialogue can be recorded. We formalize the one-sided conversation problem (1SC): inferring and learning from only one side of a conversation. We study two tasks: (1) reconstructing the missing speaker's turns and (2) generating summaries from one-sided transcripts. Evaluating models on MultiWOZ, DailyDialog, SpokenWOZ and Candor with both human A/B testing and LLM-as-a-judge metrics, we find that additional context improves reconstruction, and while large models generate promising reconstructions with prompting, smaller models require finetuning. Further, high-quality summaries can be generated without reconstructing missing turns. We present 1SC as a novel challenge and report promising results that mark a step toward privacy-aware conversational AI.

[Findings] Reason in Chains, Learn in Trees: Self-Rectification and Grafting for Multi-turn Agent Policy Optimization ► *Underline Video*

Yu Li, Sizhe Tang and Tian Lan

Reinforcement learning for Large Language Model agents is often hindered by sparse rewards in multi-step reasoning tasks. Existing approaches like Group Relative Policy Optimization treat sampled trajectories as independent chains, assigning uniform credit to all steps in each chain and ignoring the existence of critical steps that may disproportionally impact reasoning outcome. In this paper, we propose T-STAR(Tree-structured Self-Taught Agent Rectification), a framework that recovers the latent correlated reward structure across seemingly independent trajectories. Specifically, we consolidate trajectories into a unified Cognitive Tree by identifying and merging functionally similar steps/nodes. It enables an Introspective Valuation mechanism that back-propagates trajectory-level rewards through the tree to obtain a new notion of variance-reduced relative advantage at step-level. Using the Cognitive Tree, we also develop In-Context Thought Grafting to synthesize corrective reasoning by contrasting successful and failed branches at critical divergence points/steps. Our proposed Surgical Policy Optimization then capitalizes on the rich policy gradient information concentrated at these critical points/steps through a Bradley-Terry type of surgical loss. Extensive experiments across embodied, interactive, reasoning, and planning benchmarks demonstrate that T-STAR achieves consistent improvements over strong baselines, with gains most pronounced on tasks requiring extended reasoning chains.

[Long] Reasoning over Precedents Alongside Statutes: Case-Augmented Deliberative Alignment for LLM Safety ► *Underline Video*

Can Jin, Rui Wu, Tong Che, Qixin Zhang, Hongwu Peng, Jiahui Zhao, Zhenting Wang, Wenqi Wei, Ligong Han, Zhao Zhang, Yuan Cao, Ruixiang Tang and Dimitris N. Metaxas

Ensuring that Large Language Models (LLMs) adhere to safety principles without refusing benign requests remains a significant challenge. While OpenAI introduces deliberative alignment (DA) to enhance the safety of its o-series models through reasoning over detailed "code-like" safety rules, the effectiveness of this approach in open-source LLMs, which typically lack advanced reasoning capabilities, is understudied. In this work, we systematically evaluate the impact of explicitly specifying extensive safety codes versus demonstrating them through illustrative cases. We find that referencing explicit codes inconsistently improves harmlessness and systematically degrades helpfulness, whereas training on case-augmented simple codes yields more robust and generalized safety behaviors. By guiding LLMs with case-augmented reasoning instead of extensive code-like safety rules, we avoid rigid adherence to narrowly enumerated rules and enable

broader adaptability. Building on these insights, we propose CADA, a case-augmented deliberative alignment method for LLMs utilizing reinforcement learning on self-generated safety reasoning chains. CADA effectively enhances harmlessness, improves robustness against attacks, and reduces over-refusal while preserving utility across diverse benchmarks, offering a practical alternative to rule-only DA for improving safety while maintaining helpfulness.

[Long] Reasoning Structure Matters for Safety Alignment of Reasoning Models ▶ *Underline Video*

Yeonjun In, Wonjoong Kim, Sangwu Park and Chanyoung Park

Large reasoning models (LRMs) achieve strong performance on complex reasoning tasks but often generate harmful responses to malicious user queries. This paper investigates the underlying cause of these safety risks and shows that the issue lies in the reasoning structure itself. Based on this insight, we claim that effective safety alignment can be achieved by altering the reasoning structure. We propose AltTrain, a simple yet effective post-training method that explicitly alters the reasoning structure of LRMs. AltTrain is both practical and generalizable, requiring no complex reinforcement learning (RL) training or reward design, only supervised fine-tuning (SFT) with a lightweight 1K training examples. Experiments across LRM backbones and model sizes demon strate strong safety alignment, along with robust generalization across reasoning, QA, summarization, and multilingual setting.

[Long] Reasoning While Asking: Transforming Reasoning Large Language Models from Passive Solvers to Proactive Inquirers ▶ *Underline Video*

Xin Chen, Feng Jiang, Yiqian Zhang, Hardy Chen, Shuo Yan, Wenya Xie, Min Yang and Shujian Huang

Reasoning-oriented Large Language Models (LLMs) have achieved remarkable progress with Chain-of-Thought (CoT) prompting, yet they remain fundamentally limited by a blind self-thinking paradigm: performing extensive internal reasoning even when critical information is missing or ambiguous. We propose Proactive Interactive Reasoning (PIR), a new reasoning paradigm that transforms LLMs from passive solvers into proactive inquirers that interleave reasoning with clarification. Unlike existing search- or tool-based frameworks that primarily address knowledge uncertainty by querying external environments, PIR targets premise- and intent-level uncertainty through direct interaction with the user. PIR is implemented via two core components: (1) an uncertainty-aware supervised fine-tuning procedure that equips models with interactive reasoning capability, and (2) a user-simulator-based policy optimization framework driven by a composite reward that aligns model behavior with user intent. Extensive experiments on mathematical reasoning, code generation, and document editing demonstrate that PIR consistently outperforms strong baselines, achieving up to 32.70% higher accuracy, 22.90% higher pass rate, and 41.36 BLEU improvement, while reducing nearly half of the reasoning computation and unnecessary interaction turns. Further reliability evaluations on factual knowledge, question answering, and missing-premise scenarios confirm the strong generalization and robustness of PIR.

[Findings] Reasoning with Memory: Adaptive Information Management for Retrieval-Augmented Generation ▶ *Underline Video*

Hieu Man, Ro-ee Tal, Abhishek Kumar, Jaejin Cho and Benjamin Hsu

Multi-hop reasoning remains a fundamental challenge for Retrieval-Augmented Generation (RAG) systems. Recent approaches—from adaptive retrieval to agentic pipelines—struggle to maintain coherent intermediate reasoning states as chains grow longer. We introduce State-Aware RAG, a framework that addresses this limitation through an explicit working memory that serves as a dynamic cognitive workspace for reasoning. Our modular architecture features a lightweight, trainable extractor that learns to actively filter, consolidate, and update this working memory via a novel Path-Outcome Dual Reward paradigm, which balances local coherence with global strategy. The retriever and generator remain frozen, enabling plug-and-play flexibility. Experiments on eight QA benchmarks demonstrate state-of-the-art results, on average achieving +8.6% over the best memory-augmented baseline and +9.3% over the best RL-enhanced baseline. Our architecture generalizes seamlessly to stronger generators and retrievers without retraining, establishing dynamic memory management as a critical yet underexplored dimension for advancing RAG systems.

[Long] ReflectRM: Boosting Generative Reward Models via Self-Reflection within a Unified Judgment Framework ▶ *Underline Video*

Kai Qin, Liangxin Liu, Yu Liang, Longzheng Wang, Yan Wang, Yueyang Zhang, Long Xia, Zhiyuan Sun, Houde Liu and Daiting Shi

Reward Models (RMs) are critical components in the Reinforcement Learning from Human Feedback (RLHF) pipeline, directly determining the alignment quality of Large Language Models (LLMs). Recently, Generative Reward Models (GRMs) have emerged as a superior paradigm, offering higher interpretability and stronger generalization than traditional scalar RMs. However, existing methods for GRMs focus primarily on outcome-level supervision, neglecting analytical process quality, which constrains their potential. To address this, we propose ReflectRM, a novel GRM that leverages self-reflection to assess analytical quality and enhance preference modeling. ReflectRM is trained under a unified generative framework for joint modeling of response preference and analysis preference. During inference, we use its self-reflection capability to identify the most reliable analysis, from which the final preference prediction is derived. Experiments across four benchmarks show that ReflectRM consistently improves performance, achieving an average accuracy gain of +3.7 on Qwen3-4B. Further experiments confirm that response preference and analysis preference are mutually reinforcing. Notably, ReflectRM substantially mitigates positional bias, yielding +10.2 improvement compared with leading GRMs and establishing itself as a more stable evaluator. Our code is available at https://github.com/yuliangCarmelo/ReflectRM.

[Findings] ReMedi: Reasoner for Medical Clinical Prediction ▶ *Underline Video*

Yushi Cao, Yiming Chen, Hongchao Jiang, Hung-yi Lee and Robby T. Tan

Predicting future clinical outcomes from electronic health records (EHR) remains challenging due to the complexity and heterogeneity

of patient data. LLMs have shown strong potential for such predictive tasks, yet existing approaches mainly focus on enhancing medical knowledge through distillation or RAG while relying on the model's internal ability to interpret contextual information. In this work, we present ReMedi (Reasoner for Medical Clinical Prediction), a framework for improving clinical outcome prediction from EHR. ReMedi generates rationale-answer pairs using a challenging sample regeneration mechanism for complex clinical questions, which leverages ground-truth answers as hints to enhance reasoning for further fine-tuning and preference tuning. ReMedi integrates ground-truth outcome guidance into the preference data construction loop, regenerating rationale-answer variants. By tuning on these rationale-answer pairs, the model improves its predictive performance. Experiments on multiple EHR prediction tasks demonstrate substantial gains of up to 19.9% over state-of-the-art baselines in terms of F1 score, underscoring ReMedi's effectiveness in real-world clinical prediction.

[Long] Resolving the Security-Auditability Dilemma with Auditable Latent Chain-of-Thought Alignment ▶ *Underline Video*

Guan Wang, Biyu Zhou, Xuehai Tang, Jizhong Han and Songlin Hu

To address the increasingly severe safety risk of large language models (LLMs), reasoning-based safety alignment methods have emerged. These methods overcome the limitations of 'shallow alignment' by exposing the model's Chain-of-Thought (CoT), enabling auditability of safety reasoning process through both training-phase supervision and post-generation verification. However, this transparency creates a critical vulnerability, a tension we define as the Security Auditability Dilemma: while explicit reasoning is a prerequisite for safety, its textual Auditable paradoxically transforms it into an optimization target for adaptive attackers and induces the model to unintentionally copy harmful content from its own reasoning context. To address this, we propose Auditable Latent CoT Alignment (ALCA), a framework that decouples internal reasoning from external output. ALCA shifts the safety deliberation process into a continuous latent space. This allows the safety reasoning process to guide the generation of harmless outputs, while eliminates the discrete textual surface that facilitates internal copying and adaptive attack. Yet, this process is not a black box. we introduce a restricted Self-Decoding mechanism that allows the model to reconstruct its latent reasoning into human-readable text for supervision under specific guidance. Extensive experiments show that ALCA achieves robustness alignment, reducing the success rate of adaptive jailbreak attacks by over 40% compared to strong baselines, while preserving performance. Our framework presents a path toward building LLMs that are both robustly secure and auditable.

[Long] Revealing the Seen, Imagining the Beyond: A Survey of Image-Grounded Chain-of-Thought Reasoning in Multimodal LLMs ▶ *Underline Video*

Qihua Dong, Yitian Zhang, Huimin Zeng, Yizhou Wang, Jianglin Lu, Kuo Yang and Yun Fu

Multimodal large language models (MLLMs) are making rapid strides in complex visual reasoning. This survey synthesizes the emerging paradigm of Image-Grounded Chain-of-Thought (IG-CoT), where models ground intermediate inferences by interleaving textual rationales with visual state updates. We formalize IG-CoT, present a method-centric taxonomy covering prompting, supervised fine-tuning, and reinforcement learning, and map these techniques to representative benchmarks. Our analysis identifies two domains where IG-CoT offers significant advantages: detail-oriented reasoning requiring meticulous perception, and imagined-world reasoning for simulating unseen states in games, geometry, and planning. We discuss the practical trade-offs of current methods regarding controllability, data, and compute. We conclude by highlighting key challenges (efficiency, data quality, and generative capabilities) and outlining promising future directions, including lightweight architectures, richer intermediate supervision, and method-aware evaluations that better assess faithfulness and long-horizon reasoning. We maintain a continuously updated paper list at https://github.com/dddraxxx/Awesome-Image-Grounded-CoT.

[Long] Revisiting Audio-language Pretraining for Learning General-purpose Audio Representation ▶ *Underline Video*

Wei-Cheng Tseng, Xuanru Zhou, Mingyue Huo, Yiwen Shao, Hao Zhang and Dong Yu

Audio-language pretraining (ALP) holds promise for learning general-purpose audio representation, yet remains underexplored. Crucially, there is no consensus on whether audio-language models can build effective general-purpose audio encoders, nor a systematic understanding of how pretraining objectives behave across diverse tasks and scales. We identify three key barriers: limited scale of audio-text corpora, limited coverage of audio attributes in existing caption corpora, and lack of systematic exploration and evaluation. To fill this gap, we present the first principled empirical study of ALP. We first introduce CaptionStew, a 10.7M caption dataset aggregating open-source audio-text corpora across multiple domains and captioning focuses. We then conduct the first comprehensive evaluation comparing contrastive and captioning objectives for learning audio representation across speech, music, and environmental sound tasks. Our results not only demonstrate that ALP yields competitive, transferable representations, but reveal critical trade-offs: contrastive learning offers superior data efficiency, while captioning exhibits better scalability. Furthermore, we find that the benefits of supervised initialization often diminish at larger scales, challenging common practices. By grounding these claims in empirical evidence, we establish a viable pathway toward general-purpose audio representation learning, guiding future research.

[Long] Reward Modeling for Scientific Writing Evaluation ▶ *Underline Video*

Furkan ahinuç, Subhabrata Dutta and Iryna Gurevych

Scientific writing is an expert-domain task that demands deep domain knowledge, task-specific requirements and reasoning capabilities that leverage the domain knowledge to satisfy the task specifications. While scientific text generation has been widely studied, its evaluation remains a challenging and open problem. It is critical to develop models that can be reliably deployed for evaluating diverse open-ended scientific writing tasks while adhering to their distinct requirements. However, existing LLM-based judges and reward models are primarily optimized for general-purpose benchmarks with fixed scoring rubrics and evaluation criteria. Consequently, they often fail to reason over sparse knowledge of scientific domains when interpreting task-dependent and multi-faceted criteria.

Moreover, fine-tuning for each individual task is costly and impractical for low-resource settings. To bridge these gaps, we propose cost-efficient, open-source reward models tailored for scientific writing evaluation. We introduce a two-stage training framework that initially optimizes scientific evaluation preferences and then refines reasoning capabilities. Our multi-aspect evaluation design and joint training across diverse tasks enable fine-grained assessment and robustness to dynamic criteria and scoring rubrics. Experimental analysis shows that our training regime strongly improves LLM-based scientific writing evaluation. Our models generalize effectively across tasks and to previously unseen scientific writing evaluation settings, allowing a single trained evaluator to be reused without task-specific retraining.

[Findings] Ro-SLM: Onboard Small Language Models for Robot Task Planning and Operation Code Generation ► *Underline Video*
Wenhao Wang, Yanyan Li, Long Jiao and Jiawei Yuan
Recent advances in large language models (LLMs) provide robots with contextual reasoning abilities to comprehend human instructions. Yet, current LLM-enabled robots typically depend on cloud-based models or high-performance computing infrastructure, which limit their deployment on robots under unreliable internet environments or with constrained computational resources, such as UAVs and small ground vehicles. Thus, deploying fine-tuned small language models (SLMs) that support onboard deployment offers a promising alternative. This paper introduces Ro-SLM, a framework that enables reliable SLM-driven robot operation by distilling LLMs' knowledge and reasoning. Ro-SLM starts from dataset synthesis by leveraging LLMs to generate diverse task instructions, produce corresponding ground truth code with minimal human assistance, and augment instructions into real-world application scenarios. Ro-SLM is then fine-tuned with the dataset, in which LLM serves as a reward function to guide the training. Extensive experiments on UAV operation tasks demonstrate that Ro-SLM improves the performance of SLM from being incapable of supporting robotic task planning and code generation to achieving performance that approaches LLM.

[Long] RSDA: Restoring Stale Data Affinity via Dynamic Renovation Strategy for Mitigating Data Scarcity ► *Underline Video*
Yidan Liang, Jia Zhu, Weijie Shi, Hanghui Guo, Yue Cui, Jiawei Shen, Guoqing Ma, Jingjiang Liu, Qingyu Niu, Yilin Wang, Shimin Di and Jiajie Xu
High-quality data is the cornerstone of advancing large language models. However, the field currently faces a critical dilemma: the supply of premium data is nearing depletion, while vast stale corpora remain underutilized. Our empirical analysis reveals that training models on such data directly often leads to performance degradation. We attribute this phenomenon to the data affinity gap, a misalignment stemming from the model's inability to effectively comprehend the data or inherent quality defects. To bridge this gap, we propose Restoring Stale Data Affinity (RSDA) framework. First, utilizing our proposed potential entropy metric, RSDA quantifies the latent value of samples to effectively identify stale data with higher renovation potential. Subsequently, the framework employs a dynamic renovation strategy selection mechanism to determine the optimal component-level strategy for each instance, transforming low-affinity stale samples into high-quality training data. Comprehensive experimental results demonstrate that RSDA effectively enhances data affinity, achieving performance improvements using less than 10% of the data volume, thereby underscoring that the latent potential of stale corpora remains largely untapped. The code is available at https://github.com/wenfiii/RSDA.

[Findings] SceneLM: 3D-Aware Language Models for Editable 3D Scene Synthesis ► *Underline Video*
Xingbo Yao, Xiaoyu Chen, Doudou ZHANG, Mingzhi Sheng, Boyuan Cao, Yingcong Chen and Hui Xiong
Synthesizing an editable 3D scene from a single RGB image is central to content creation, embodied-agent data generation, and AR/VR, yet remains challenging to achieve both high-fidelity reconstruction and convenient interactive editing. Existing geometry-based pipelines produce high-quality 3D results but are typically hard to refine without rerunning the full process, while LLM-driven procedural systems enable interactive tool use but are mostly text-driven and lack precise metric 3D understanding from images. We present SceneLM, a language-model-based framework that grounds 3D scene synthesis in visual evidence by recovering an executable metric 3D layout directly from a single image. Given an RGB image (and camera intrinsics when available), SceneLM outputs a JSON-form layout specifying each objects category, 3D center, size, and discretized yaw, and then deterministically executes this layout with a tool suite to instantiate, place, and edit objects for iterative refinement. To train metric layout recovery at scale, we curate five datasets covering diverse indoor, outdoor, and tabletop scenes and convert heterogeneous 3D annotations into a unified instruction-tuning format. To improve numerical stability and metric accuracy while preserving the text interface, we augment autoregressive JSON generation with a lightweight geometry prediction branch and dual supervision. Experiments show that SceneLM substantially improves single-image 3D layout estimation over strong open and proprietary MLLM baselines, and yields higher-quality end-to-end scene generation in geometric consistency, physical plausibility, semantic alignment, and realism.

[Findings] Sculpting the Vector Space: Towards Efficient Multi-Vector Visual Document Retrieval via Prune-then-Merge Framework ► *Underline Video*
Yibo Yan, Mingdong Ou, Yi Cao, Xin Zou, Jiahao Huo, Shuliang Liu, James Kwok and Xuming Hu
Visual Document Retrieval (VDR), which aims to retrieve relevant pages within vast corpora of visually-rich documents, is of significance in current multimodal retrieval applications. The state-of-the-art multi-vector paradigm excels in performance but suffers from prohibitive overhead, a problem that current efficiency methods like pruning and merging address imperfectly, creating a difficult trade-off between compression rate and feature fidelity. To overcome this dilemma, we introduce Prune-then-Merge, a novel two-stage framework that synergizes these complementary approaches. Our method first employs an adaptive pruning stage to filter out low-information patches, creating a refined, high-signal set of embeddings. Subsequently, a hierarchical merging stage compresses this pre-filtered set, effectively

summarizing semantic content without the noise-induced feature dilution seen in single-stage methods. Extensive experiments on 29 VDR datasets demonstrate that our framework consistently outperforms existing methods, significantly extending the near-lossless compression range and providing robust performance at high compression ratios.

[Findings] Self-Consistency from Only Two Samples: CoT-PoT Ensembling for Efficient LLM Reasoning ▶ *Underline Video*

Raman Saparkhan, Majd Hawasly, Md Rizwan Parvez and Mohammad Raza

Self-consistency (SC) is a popular technique for improving the reasoning accuracy of large language models by aggregating multiple sampled outputs, but it comes at a high computational cost due to extensive sampling. We introduce a hybrid ensembling approach that leverages the complementary strengths of two distinct modes of reasoning: Chain-of-Thought (CoT) and Program-of-Thought (PoT). We describe a general framework for combining these two forms of reasoning in self-consistency, as well as particular strategies for both full sampling and early-stopping. We show that CoT-PoT ensembling not only improves overall accuracy, but also drastically reduces the number of samples required for SC by a factor of 9.3x. In particular, the majority of tasks (78.6%) can be addressed with only two samples, which has not been possible with any prior SC methods.

[Findings] Semantic-Space Exploration and Exploitation in RLVR for LLM Reasoning ▶ *Underline Video*

Fanding Huang, Guanbo Huang, Xiao Fan, Yi He, Xiao Liang, Xiao Chen, Qinting Jiang, Faisal Nadeem Khan, Jingyan Jiang and Zhi Wang

Reinforcement Learning with Verifiable Rewards (RLVR) for LLM reasoning is often framed as balancing exploration and exploitation in action space, typically operationalized with token-level proxies (e.g., output entropy or confidence). We argue that this apparent trade-off is largely a measurement artifact: token-level statistics reflect next-token uncertainty rather than how reasoning progresses over multi-token semantic structures. We therefore study exploration and exploitation in the hidden-state space of response trajectories. We use Effective Rank (ER) to quantify representational exploration and introduce its temporal derivatives, Effective Rank Velocity (ERV) and Effective Rank Acceleration (ERA), to characterize exploitative refinement dynamics. Empirically and theoretically, ER and ERV exhibit near-zero correlation in semantic space, suggesting the two capacities can be improved simultaneously. Motivated by this, we propose Velocity-Exploiting Rank Learning (VERL), which shapes the RL advantage with an auxiliary signal derived from ER/ERV and uses the more stable ERA as a meta-control variable to adaptively balance the incentives. Across multiple base models, RL algorithms, and reasoning benchmarks, VERL yields consistent improvements, including large gains on challenging tasks (e.g., 21.4% in Gaokao 2024).

[Findings] SepSeq: A Training-Free Framework for Long Numerical Sequence Processing in LLMs ▶ *Underline Video*

Jie Sun, Yu Liu, Lu Han, Qiwen Deng, Xiang Shu, Yang Xiao, Lintao Ma, Xingyu Lu, Jun Zhou, Pengfei Liu, Jiancan Wu and Xiang Wang

While transformer-based Large Language Models (LLMs) theoretically support massive context windows, they suffer from severe performance degradation when processing long numerical sequences. We attribute this failure to the attention dispersion in the Softmax mechanism, which prevents the model from concentrating attention. To overcome this, we propose Separate Sequence (SepSeq), a training-free, plug-and-play framework to mitigate dispersion by strategically inserting separator tokens. Mechanistically, we demonstrate that separator tokens act as an attention anchor, recalibrating attention to focus on local segments while preserving global context. Extensive evaluations on 9 widely-adopted LLMs confirm the effectiveness of our approach: SepSeq yields an average relative accuracy improvement of 35.6% across diverse domains while reducing 16.4% inference token consumption.

[Long] SGPVT: Self-Generated Proximal Visual Tokens for Mitigating Proximal Collateral Damage in MLLM Unlearning ▶ *Underline Video*

Jiaqi Li, Zhijing Zhang, Jiahui Geng, Sheng Bi, Chuanyi Zhang, Fan Liu and Guilin Qi

Machine unlearning in multimodal large language models (MLLMs) aims to remove specific concepts while preserving overall utility. However, existing approaches focus primarily on general utility metrics, overlooking the preservation of semantically related concepts. We present the first systematic analysis of this proximal collateral damage, revealing that forgetting vulnerability correlates strongly with visual embedding similarity in a smooth gradient across the semantic space. Based on this insight, we propose a novel unlearning framework that introduces Self-Generated Proximal Visual Tokens (SGPVTs), which are synthetically perturbed visual representations around the target concept. Our method employs an adaptive cosine-band curriculum with a dual-stream objective: forgetting the target via gradient ascent while distilling knowledge from a frozen teacher model into proximal tokens to prevent degradation. Extensive experiments demonstrate that our approach significantly outperforms existing methods in preserving semantically related concepts while achieving effective target unlearning, eliminating the need for manual retention set curation. Our source code will be released in the near future.

[Findings] Shorten After Youre Right: Lazy Length Penalties for Reasoning RL ▶ *Underline Video*

Danlong Yuan, Tian Xie, Shaohan Huang, Huishuai Zhang, Zhuocheng Gong, Chong Luo, Furu Wei and Dongyan Zhao

Long-reasoning models achieve strong accuracy on complex reasoning tasks, but their extended reasoning trajectories incur substantial memory and latency costs. Several existing shortening methods rely on additional supervision or multi-stage post-training, which primarily reduces inference length and does not reduce the rollout tokens during on-policy reinforcement learning (RL). We instead target on-policy response shortening, aiming to improve both inference efficiency and RL training throughput. However, because on-policy RL couples optimization with exploration, naively penalizing length can destabilize training and suppress exploration. To impose length pressure safely, we propose a lazy length penalty integrated into the rule-based RL pipeline: it activates only on correct trajectories, only after training accuracy enters a stably improving regime, and only when responses exceed a tolerance band beyond the minimal correct length. Across four

settings, our method significantly reduces response length without extra training stages while maintaining or improving performance. In a logic reasoning setting, we achieve a 40% reduction in step-averaged response length alongside a 14-point gain in performance. For math problems, we reduce step-averaged response length by 33% while preserving performance.

[Long] Simple Agents, Biased Judges: Efficient Multi-Party Dialogue Generation & The Evaluation Gap ▶ *Underline Video*

Kunal Samanta, Faisal Shohan, Amine Trabelsi and Richard Khoury

Multi-party social dialogue remains underexplored in the literature, in part due to the difficulty and cost of evaluation. As a result, recent work on synthetic dialogue generation often relies on automated metrics and LLM-as-a-Judge frameworks, despite limited evidence that such judges reflect human preferences in social settings. In this work, we introduce a lightweight and controllable multi-party dialogue generation framework (MPOD) as an experimental instrument for studying generation and evaluation in social interaction. Using this framework, we conduct human evaluations of open-domain multi-party dialogue simulation and directly compare human judgments against state-of-the-art LLM judges. Across 319 pairwise comparisons, we observe near-random agreement between humans and automated judges (Cohen's ~ 0.11), driven by systematic behaviors including extreme tie aversion and strong sensitivity to assistant-style verbosity. Crucially, human-human inter-annotator agreement (= 0.29) is substantially higher than human-LLM agreement. To isolate the mechanism underlying this misalignment, we introduce a controlled Transplant Ablation, showing that LLM judges consistently prefer conversations containing a single proprietary, assistant-style agent. Additional stress tests show that judges prefer GPT-style conversations even when utterance order is randomly shuffled, indicating insensitivity to conversational structure and coherence. Our findings provide controlled evidence that current instruction-tuned LLM judges do not reliably reflect human preferences for naturalness, engagingness, and overall quality in multi-party social dialogue, calling into question their widespread use for validating synthetic conversational data.

[Long] Sparse-RL: Breaking the Memory Wall in LLM Reinforcement Learning via Stable Sparse Rollouts ▶ *Underline Video*

Sijia Luo, Xiaokang Zhang, Yuxuan Hu, Bohan Zhang, Ke Wang, Jinbo Su, Mengshu Sun, Lei Liang and Jing Zhang

Reinforcement Learning (RL) has become essential for eliciting complex reasoning capabilities in Large Language Models (LLMs). However, the substantial memory overhead of storing Key-Value (KV) caches during long-horizon rollouts acts as a critical bottleneck, often prohibiting efficient training on limited hardware. While existing KV compression techniques offer a remedy for inference, directly applying them to RL training induces a severe policy mismatch, leading to catastrophic performance collapse. To address this, we introduce Sparse-RL, which empowers stable RL training under sparse rollouts. We show that instability arises from a fundamental policy mismatch among the dense old policy, the sparse sampler policy, and the learner policy. To mitigate this issue, Sparse-RL incorporates Sparsity-Aware Rejection Sampling and Importance-based Reweighting to correct the off-policy bias introduced by compression-induced information loss. Experimental results show that Sparse-RL reduces rollout overhead compared to dense baselines while preserving the performance. Furthermore, Sparse-RL inherently implements sparsity-aware training, significantly enhancing model robustness during sparse inference deployment.

[Long] Spatial-Agent: Agentic Geo-spatial Reasoning with Scientific Core Concepts ▶ *Underline Video*

Riyang Bao, Cheng Yang, Dazhou Yu, Zhexiang Tang, Gengchen Mai and Liang Zhao

Geospatial reasoning is essential for real-world applications such as urban analytics, transportation planning, and disaster response. However, existing LLM-based agents often fail at genuine geospatial computation, relying instead on web search or pattern matching while hallucinating spatial relationships. We present Spatial-Agent, an AI agent grounded in foundational theories of spatial information science. Our approach formalizes geo-analytical question answering as a concept transformation problem, where natural-language questions are parsed into executable workflows represented as GeoFlow Graphs—directed acyclic graphs with nodes corresponding to spatial concepts and edges representing transformations. Drawing on spatial information theory, Spatial-Agent extracts spatial concepts, assigns functional roles with principled ordering constraints, and composes transformation sequences through template-based generation. Extensive experiments on MapEval-API and MapQA benchmarks demonstrate that Spatial-Agent significantly outperforms existing baselines including ReAct and Reflexion, while producing interpretable and executable geospatial workflows.

[Findings] SQLAgent: Learning to Explore Before Generating as a Data Engineer ▶ *Underline Video*

Wenjia Jiang, Yiwei Wang, Boyan Han, Joey Tianyi Zhou and Chi Zhang

Large Language Models have recently shown impressive capabilities in reasoning and code generation, making them promising tools for natural language interfaces to relational databases. However, existing approaches often fail to generalize in complex, real-world settings due to the highly database-specific nature of SQL reasoning, which requires deep familiarity with unique schemas, ambiguous semantics, and intricate join paths. To address this challenge, we introduce a novel two-stage LLM-based framework that decouples knowledge acquisition from query generation. In the Exploration Stage, the system autonomously constructs a database-specific knowledge base by navigating the schema with a Monte Carlo Tree Search-inspired strategy, generating triplets of schema fragments, executable queries, and natural language descriptions as usage examples. In the Deployment Stage, a dual-agent system leverages the collected knowledge as in-context examples to iteratively retrieve relevant information and generate accurate SQL queries in response to user questions. This design enables the agent to proactively familiarize itself with unseen databases and handle complex, multi-step reasoning. Extensive experiments on large-scale benchmarks demonstrate that our approach significantly improves accuracy over strong baselines, highlighting its effectiveness and generalizability.

[Long] SSA: Improving Performance With a Better Scoring Function ▶ *Underline Video*

Omar Naim, Swarnadeep Bhar, Jerome Bolte and Nicholas Asher

While transformer models exhibit strong in-context learning (ICL) abilities, they often fail to generalize under simple distribution shifts. We analyze these failures and identify Softmax, the scoring function in the attention mechanism, as a contributing factor. We propose Scaled Signed Averaging (SSA), a novel attention scoring function that mitigates these failures. SSA significantly improves performance on our ICL tasks and outperforms transformer models with Softmax on several NLP benchmarks and linguistic probing tasks, in both decoder-only and encoder-only architectures.

[Long] Stable Language Guidance for Vision-Language-Action Models ▶ *Underline Video*

Zhihao Zhan, Yuhao Chen, Jiaying Zhou, Qinhan Lyu, Hao Liu, Keze Wang, Liang Lin and Guangrun Wang

Vision-Language-Action (VLA) models have demonstrated impressive capabilities in generalized robotic control; however, they remain notoriously brittle to linguistic perturbations. We identify a critical "modality collapse" phenomenon where strong visual priors overwhelm sparse linguistic signals, causing agents to overfit to specific instruction phrasings while ignoring the underlying semantic intent. To address this, we propose Residual Semantic Steering (RSS), a probabilistic framework that disentangles physical affordance from semantic execution. RSS introduces two theoretical innovations: (1) Monte Carlo Syntactic Integration, which approximates the true semantic posterior via dense, LLM-driven distributional expansion, and (2) Residual Affordance Steering, a dual-stream decoding mechanism that explicitly isolates the causal influence of language by subtracting the visual affordance prior. Theoretical analysis suggests that RSS effectively maximizes the mutual information between action and intent while suppressing visual distractors. Empirical results across diverse manipulation benchmarks demonstrate that RSS achieves state-of-the-art robustness, maintaining performance even under adversarial linguistic perturbations.

[Findings] Stable On-Policy Distillation through Adaptive Target Reformulation ▶ *Underline Video*

Ijun Jang, Jewon Yeom, Juan Yeo, Hyunggu Lim and Taesup Kim

Knowledge distillation (KD) is a widely adopted technique for transferring capabilities from large language models to smaller student models. However, conventional supervised KD often suffers from a distribution mismatch between training and inference. While on-policy KD approaches attempt to mitigate this issue by learning directly from student-generated outputs, they frequently encounter training instabilities and noisy teacher feedback during early optimization stages. These challenges manifest as pathological gradients in forward KL objectives when students encounter unfamiliar tokens, or as a collapse in distributional diversity within reverse KL regimes. To address these limitations, we propose Veto, an objective-level reformulation that constructs a geometric target distribution in logit space to emphasize agreement between the teacher and the student. By introducing a tunable parameter beta, Veto serves as an Adaptive Gradient Veto that stabilizes optimization by suppressing harmful gradients on low-confidence tokens, while simultaneously acting as a Decisiveness Knob to balance reward-driven performance with output diversity. Extensive experiments across various reasoning and generation tasks demonstrate that Veto consistently outperforms supervised fine-tuning and existing on-policy baselines.

[Findings] Stress-Testing Emotional Support Models: Moving from Homogeneous to Diverse Help Seekers ▶ *Underline Video*

Chaewon Heo, Cheyon Jin and Yohan Jo;

As emotional support chatbots have recently gained significant traction across both research and industry, a common evaluation strategy has emerged: use help-seeker simulators to interact with supporter chatbots. However, current simulators suffer from two critical limitations: (1) they fail to capture the behavioral diversity of real-world seekers, often portraying them as overly cooperative, and (2) they lack the controllability required to simulate specific seeker profiles. To address these challenges, we present a controllable seeker simulator driven by nine psychological and linguistic features that underpin seeker behavior. Using authentic Reddit conversations, we train our model via a Mixture-of-Experts (MoE) architecture, which effectively differentiates diverse seeker behaviors into specialized parameter subspaces, thereby enhancing fine-grained controllability. Our simulator achieves superior profile adherence and behavioral diversity compared to existing approaches. Furthermore, evaluating 7 prominent supporter models with our system uncovers previously obscured performance degradations. These findings underscore the utility of our framework in providing a more faithful and stress-tested evaluation for emotional support chatbots.

[Findings] Superficial Success vs. Internal Breakdown: An Empirical Study of Generalization in Adaptive Multi-Agent Systems ▶ *Underline Video*

Namyeong So, Seokgyu Jang and Taeuk Kim

Adaptive multi-agent systems (MAS) are increasingly adopted as solutions to complex problems. However, their optimization for narrow task ranges leaves it unclear whether they can function as general-purpose systems. To fill this gap, we conduct an extensive empirical study on adaptive MAS, revealing two key findings: (1) they are prone to topological overfitting, defined as failures in domain transfer; and (2) they exhibit illusory coordination, where surface-level accuracy is high but underlying agent coordination deviates from ideal MAS behavior, raising concerns about their practical effectiveness. These observations highlight the urgent need to prioritize generalization in MAS development and motivate more thorough evaluation beyond correctness of the final answer.

[Findings] Task-Stratified Knowledge Scaling Laws for Post-Training Quantized Large Language Models ▶ *Underline Video*

Chenxi Zhou, Pengfei Cao, Jiang Li, Bohan Yu, Jinyu Ye, Jun Zhao and Kang Liu

Post-Training Quantization (PTQ) is a critical strategy for efficient large language models (LLMs) deployment. However, existing scaling laws primarily focus on general performance, overlooking crucial fine-grained factors and how quantization differentially impacts diverse

knowledge capabilities. To address this, we establish Task-Stratified Knowledge Scaling Laws. By stratifying capabilities into memorization, application, and reasoning, we develop a framework that unifies model size, bit-width, and fine-grained factors: group size and calibration set size. Validated on 293 diverse PTQ configurations, our framework demonstrates strong fit and cross-architecture consistency. It reveals distinct sensitivities across knowledge capabilities: reasoning is precision-critical, application is scale-responsive, and memorization is calibration-sensitive. We highlight that in low-bit scenarios, optimizing these fine-grained factors is essential for preventing performance collapse. These findings provide an empirically-backed foundation for designing knowledge-aware quantization strategies.

[Findings] The Agent's First Day: Benchmarking Learning, Exploration, and Scheduling in the Workplace Scenarios ▶ *Underline Video*

Daocheng Fu, Jianbiao Mei, Rong Wu, Xuemeng Yang, Jia Xu, Ding Wang, Pinlong Cai, Yong Liu, Licheng Wen and Botian Shi

The rapid evolution of Multi-modal Large Language Models (MLLMs) has advanced workflow automation; however, existing research mainly targets performance upper bounds in static environments, overlooking robustness for stochastic real-world deployment. We identify three key challenges: dynamic task scheduling, active exploration under uncertainty, and continuous learning from experience. To bridge this gap, we introduce TraineeBench, a dynamic evaluation environment that simulates a "trainee" agent continuously exploring a novel setting. Unlike traditional benchmarks, TraineeBench evaluates agents along three dimensions: (1) context-aware scheduling for streaming tasks with varying priorities; (2) prudent information acquisition to reduce hallucination via active exploration; and (3) continuous evolution by distilling generalized strategies from rule-based, dynamically generated tasks. Experiments show that cutting-edge agents have significant deficiencies in dynamic environments, especially in active exploration and continual learning. Our work establishes a framework for assessing agent reliability, shifting evaluation from static tests to realistic, production-oriented scenarios.

[Short] The Personalization Trap: How User Memory Alters Emotional Reasoning in LLMs ▶ *Underline Video*

Xi Fang, Weijie Xu, Yuchong Zhang, Scott Nickleach, Stephanie Eckman and Chandan K. Reddy

When an AI assistant remembers that Sarah is a single mother working two jobs, does it interpret her stress differently than if she were a wealthy executive? As personalized AI systems increasingly incorporate long-term user memory, understanding how this memory shapes emotional reasoning is critical. We investigate how user memory affects emotional intelligence in large language models (LLMs) by evaluating 15 models on human validated emotional intelligence tests. We find that identical scenarios paired with different user profiles produce systematically divergent emotional interpretations. Across validated user-independent emotional scenarios and diverse user profiles, systematic biases emerged in several high-performing LLMs where advantaged profiles received more accurate emotional interpretations. Moreover, LLMs demonstrate significant disparities across demographic factors in emotion understanding and supportive recommendations tasks, indicating that personalization mechanisms can embed social hierarchies into models emotional reasoning. These results highlight a key challenge for memory-enhanced AI: systems designed for personalization may inadvertently reinforce social inequalities.

[Long] The Pitfalls of KV Cache Compression ▶ *Underline Video*

Alex Chen, Renato Geh, Aditya Grover, Guy Van den Broeck and Daniel Mingyi Israel

KV cache compression promises increased throughput and efficiency with negligible loss in performance. While the gains in throughput are indisputable and recent literature has indeed shown minimal degradation on particular benchmarks, in general the consequences of compression in realistic scenarios such as multi-instruction prompting have been insufficiently studied. In this paper, we identify several pitfalls that practitioners should be aware of when deploying KV cache compressed LLMs. We evaluate five KV cache compression methods (StreamingLLM, SnapKV, TOVA, H2O, and K-Norm) on Llama3.1 8B and Qwen2.5 14B under multi-instruction prompting with IFEval. Importantly, we show that certain instructions degrade much more rapidly with compression, effectively causing them to be completely ignored by the LLM. As a practical example, we highlight system prompt leakage as a case study, empirically demonstrating the impact of compression on leakage and general instruction-following. We identify several factors that contribute to system prompt leakage: compression method, instruction order, and KV eviction bias. We then propose simple changes to KV cache eviction policies that can reduce the impact of these factors and improve the overall performance in multi-instruction tasks.

[Long] Think How to Think: Mitigating Overthinking with Autonomous Difficulty Cognition in Large Reasoning Models ▶ *Underline Video*

Yongjiang Liu, Haoxi Li, Xiaosong Ma, Jie ZHANG and Song Guo

Recent Large Reasoning Models (LRMs) excel at complex reasoning tasks but often suffer from overthinking, generating overly long and redundant reasoning trajectories. To explore its essence, our empirical analysis reveals that LRMs are primarily limited to recognizing task properties (i.e., difficulty levels) like humans before solving the problem, leading to a one-size-fits-all reasoning strategy. This observation motivates a fundamental question: Can we explicitly bootstrap such ability to alleviate overthinking in LRMs? To this end, we propose Think-How-to-Think (TH2T), a novel two-stage fine-tuning strategy that progressively inspires LRMs' difficulty cognition and redundancy cognition of LRMs. Specifically, we first inject Difficulty Dypnosis into output prefixes as cues for global, prospective reasoning strategy selection, stimulating the model's sharper sensitivity to task complexity and adaptive control of reasoning depth. Then, we incorporate Redundancy Hypnosis into in-progress reasoning steps, which serve as local, retrospective signals for behavior correction by identifying and eliminating superfluous reasoning detours. Experiments across 7B/14B/32B models demonstrate that TH2T significantly reduces inference costs by over 70% on easy tasks and 40% on complex ones without compromising performance. The resultant models exhibit a nascent ability for difficulty-aware reasoning, effectively mitigating behaviors like excessive reflection and looping, thereby paving the way for more

cognitively efficient LRMs.

[Findings] Think Outside the Policy: In-Context Steered Policy Optimization ▶ *Underline Video*
Hsiu-Yuan Huang, Chenming Tang, Weijie Liu, Clive Bai, Saiyong Yang and Yunfang Wu
Existing Reinforcement Learning from Verifiable Rewards (RLVR) methods, such as Group Relative Policy Optimization (GRPO), have achieved remarkable progress in improving the reasoning capabilities of Large Reasoning Models (LRMs). However, they exhibit limited exploration due to reliance on on-policy rollouts which are confined to the current policy's distribution, resulting in narrow trajectory diversity. Recent approaches attempt to expand policy coverage by incorporating trajectories generated from stronger expert models, yet this reliance increases computational cost and such advanced models are often inaccessible. To address these issues, we propose In-Context Steered Policy Optimization (ICPO), a unified framework that leverages the inherent in-context learning capability of LRMs to provide expert guidance using existing datasets. ICPO introduces mixed-policy GRPO with implicit expert forcing, which expands exploration beyond the current policy distribution without requiring advanced LRM trajectories. To further stabilize optimization, ICPO integrates expert region reject sampling to filter unreliable off-policy trajectories and annealed expert-bonus reward shaping to balance early expert guidance with later autonomous improvement. Results demonstrate that ICPO consistently enhances RLVR performance and training stability on mathematical reasoning benchmarks, revealing a scalable and effective RLVR paradigm for LRMs. Our code is available at https://github.com/Celine-hxy/ICPO.

[Findings] Think Smart, Not Hard: Difficulty Adaptive Reasoning for Large Audio Language Models ▶ *Underline Video*
Zhichao Sheng, Shilin Zhou, Chen Gong and Zhenghua Li
Large Audio Language Models (LALMs) employing the Chain-of-Thought paradigm have demonstrated remarkable reasoning capabilities. Though different problems naturally require varying depths of reasoning, existing methods often determine whether to perform reasoning, lacking fine-grained mechanisms to adapt reasoning length to problem complexity. As a result, LALMs often adopt a one-size-fits-all reasoning strategy, leading to redundant overthinking for simple tasks and insufficient reasoning for complex ones. In this paper, we conduct an in-depth analysis of LALM reasoning behavior and argue that effective and efficient reasoning should be adaptively aligned with task difficulty. To this end, we propose a difficulty-adaptive reasoning method for LALMs. Specifically, we introduce a reward function that dynamically links reasoning length to the model's perceived problem difficulty, encouraging shorter reasoning for easy tasks and longer reasoning for more complex ones. Extensive experiments on three datasets demonstrate that our method consistently improves performance while reducing average reasoning length by at least 50%, achieving higher efficiency without sacrificing accuracy.

[Findings] ThinkBrake: Efficient Reasoning via Log-Probability Margin Guided Decoding ▶ *Underline Video*
Sangjun Song, Minjae Oh, Seungkyu Lee, Sungmin Jo and Yohan Jo
Large Reasoning Models (LRMs) allocate substantial inference-time compute to Chain-of-Thought (CoT) reasoning, improving performance on mathematics, scientific QA, and tool usage. However, this introduces overthinking: LRMs often reach a correct intermediate solution, continue reasoning, and overwrite it with an incorrect answer. We first demonstrate that oracle stopping, where we inject </think> at every sentence boundary and select the best stopping point in hindsight, improves average accuracy by 8% while reducing thinking tokens by 72%, exposing substantial overthinking. Motivated by this finding, we propose ThinkBrake, which monitors the log-probability margin between the top continuation token and </think> at sentence boundaries, stopping reasoning when this margin narrows. ThinkBrake requires no training and achieves favorable accuracy-efficiency trade-offs across math, scientific QA, and tool usage benchmarks, reducing thinking token usage by up to 30%. Furthermore, we provide theoretical analysis showing that ThinkBrake is equivalent to test-time realignment with a reward bonus for the </think> token.

[Long] TInR: Exploring Tool-Internalized Reasoning in Large Language Models ▶ *Underline Video*
Qiancheng Xu, Yongqi Li, Fan Liu, Hongru WANG, Min Yang and Wenjie Li
Tool-Integrated Reasoning (TIR) has emerged as a promising direction by extending Large Language Models' (LLMs) capabilities with external tools during reasoning. Existing TIR methods typically rely on external tool documentation during reasoning. However, this leads to tool mastery difficulty, tool size constraints, and inference inefficiency. To mitigate these issues, we explore Tool-Internalized Reasoning (TInR), aiming at facilitating reasoning with tool knowledge internalized into LLMs. Achieving this goal presents notable requirements, including tool internalization and tool-reasoning coordination. To address them, we propose TInR-U, a tool-internalized reasoning framework for unified reasoning and tool usage. TInR-U is trained through a three-phase pipeline: 1) tool internalization with a bidirectional knowledge alignment strategy; 2) supervised fine-tuning warm-up using high-quality reasoning annotations, and 3) reinforcement learning with TInR-specific rewards. We comprehensively evaluate our method across in-domain and out-of-domain settings. Experiment results show that TInR-U achieves superior performance in both settings, highlighting its effectiveness and efficiency. The codes are attached in the supplementary file for review.

[Findings] Too Good to be Bad: On the Failure of LLMs to Role-Play Villains ▶ *Underline Video*
Zihao Yi, Qingxuan Jiang, Ruotian Ma, Xingyu Chen, Qu Yang, Mengru Wang, Fanghua Ye, Ying Shen, Zhaopeng Tu, Xiaolong Li and Liefeng Bo
Large Language Models (LLMs) are increasingly tasked with creative generation, including the simulation of fictional characters. However, their ability to portray non-prosocial, antagonistic personas remains largely unexamined. We hypothesize that the safety alignment of modern

LLMs creates a fundamental conflict with the task of authentically role-playing morally ambiguous or villainous characters. To investigate this, we introduce the Moral RolePlay benchmark, a new dataset featuring a four-level moral alignment scale and a balanced test set for rigorous evaluation. We task state-of-the-art LLMs with role-playing characters from moral paragons to pure villains. Our large-scale evaluation reveals a consistent, monotonic decline in role-playing fidelity as character morality decreases. We find that models struggle most with traits directly antithetical to safety principles, such as "Deceitful" and "Manipulative", often substituting nuanced malevolence with superficial aggression. Furthermore, we demonstrate that general chatbot proficiency is a poor predictor of villain role-playing ability, with highly safety-aligned models performing particularly poorly. Our work provides the first systematic evidence of this critical limitation, highlighting a key tension between model safety and creative fidelity. Our benchmark and findings pave the way for developing more nuanced, context-aware alignment methods.

[Findings] Towards LLM Agents for Earth Observation ▶ ***Underline Video***

Chia Hsiang Kao, Wenting Zhao, Cheryl Lam, Aarush Umap, Shreelekha Revankar, Samuel Speas, Snehal Bhagat, Rajeev Datta, Cheng Perng Phoo, Utkarsh Mall, Carl Vondrick, Kavita Bala and Bharath Hariharan

Earth Observation (EO) provides critical planetary data for environmental monitoring, disaster management, climate science, and other scientific domains. In this work we ask: Are AI systems ready for reliable Earth Observation? To answer this, we introduce UnivEARTH, a coding benchmark of 408 yes/no questions from NASA Earth Observatory articles across 7 various topics and over 15 satellite instruments and sources. Using Google Earth Engine API as a tool in a zero-shot setup, LLM agents achieve an accuracy of 40.0% where the code fails to run over 44% of the time. To better understand LLM agent behavior, we also analyze the impact of using the JavaScript API versus Python and the effect of providing documentation. Furthermore, we find that using a reflexion framework significantly reduces errors: Claude-4.5-Sonnet, Gemini-2.5-Pro, and GPT-5 accuracies rise to around 60%. However, these results remain only marginally above random chance. Taken together, our findings identify significant challenges to be solved before AI agents can automate earth observation, and suggest paths forward.

[Findings] Towards Self-Improving Error Diagnosis in Multi-Agent Systems ▶ ***Underline Video***

Jiazheng Li, Emine Yilmaz, Bei Chen and Thu Le

Large Language Model (LLM)-based Multi-Agent Systems (MAS) enable complex problem-solving but introduce significant debugging challenges, characterized by long interaction traces, inter-agent dependencies, and delayed error manifestation. Existing diagnostic approaches often rely on expensive expert annotation or "LLM-as-a-judge" paradigms, which struggle to pinpoint decisive error steps within extended contexts. In this paper, we introduce ErrorProbe, a self-improving framework for semantic failure attribution that identifies responsible agents and the originating error step. The framework operates via a three-stage pipeline: (1) operationalizing the MAS failure taxonomy to detect local anomalies, (2) performing symptom-driven backward tracing to prune irrelevant context, and (3) employing a specialized multi-agent team (Strategist, Investigator, Arbiter) to validate error hypotheses through tool-grounded execution. Crucially, ErrorProbe maintains a verified episodic memory that updates only when error patterns are confirmed by executable evidence, without the need for annotation. Experiments across the TracerTraj and Who&When benchmarks demonstrate that ErrorProbe significantly outperforms baselines, particularly in step-level localization, while the verified memory enables robust cross-domain transfer without retraining.

[Findings] Transition-Matrix Regularization for Next Dialogue Act Prediction in Counselling Conversations ▶ ***Underline Video***

Eric Rudolph, Philipp Steigerwald and Jens Albrecht

This paper studies how empirical dialogue-flow statistics can be incorporated into Next Dialogue Act Prediction (NDAP). A KL regularization term is proposed that aligns predicted act distributions with corpus-derived transition patterns. Evaluated on a 60-class German counselling taxonomy using 5-fold cross-validation, this improves macro-F1 by 9–42% relative depending on encoder and substantially improves dialogue-flow alignment. Cross-dataset validation on HOPE suggests that improvements transfer across languages and counselling domains. In systematic ablations across pretrained encoders and architectures, the findings indicate that transition regularization provides consistent gains and disproportionately benefits weaker baseline models. The results suggest that lightweight discourse-flow priors complement pretrained encoders, especially in fine-grained, data-sparse dialogue tasks.

[Findings] TravelBehaviorQA: A Benchmark Dataset for Behavioral Interpretation of GPS Trajectories ▶ ***Underline Video***

Dongyang Zhen, Niping Duan, Huan Zhou and Qingbin Cui

GPS trajectories encode rich behavioral information about how people move, organize activities, and form daily routines. Recent advances in large language models (LLMs) raise a natural question: can such models infer and summarize travel behavior directly from mobility traces? This paper introduces TravelBehaviorQA, a large-scale benchmark dataset that reframes trajectory analysis as a language-based behavioral understanding task. The dataset links raw GPS trajectories with human-grounded question-answering (QA) pairs that capture travel intensity, temporal structure, activity patterns, mode usage, and behavioral routines. Unlike prior mobility datasets focused on prediction or classification, TravelBehaviorQA emphasizes semantic interpretation through a unified mix of deterministic and open-ended questions. In this benchmark, we construct over 143k QA instances spanning users and years, and evaluate a broad range of state-of-the-art LLMs under controlled settings. Our results reveal substantial gaps between factual extraction and genuine behavioral reasoning, showing that model scale alone is insufficient and that trajectory representation is a primary bottleneck. TravelBehaviorQA exposes critical limitations of current models and establishes a rigorous benchmark for advancing language-based understanding of human mobility behavior.

[Findings] TT-SI: Self-Improving LLM Agents with Test-Time Training ▶ *Underline Video*

Emre Can Acikgoz, Cheng Qian, Heng Ji, Dilek Hakkani-Tür and Gokhan Tur

One paradigm of language model (LM) fine-tuning relies on creating large training datasets, under the assumption that high quantity and diversity will enable models to generalize to novel tasks after posttraining. In practice, gathering large sets of data is inefficient, and training on them is prohibitively expensive; worse, there is no guarantee that the resulting model will handle complex scenarios or generalize better. Moreover, existing techniques rarely assess whether a training sample provides novel information, resulting in unnecessary costs. In this work, we explore a new Test-Time Self-Improvement (TT-SI) algorithm to create more effective and generalizable agentic LMs on-the-fly. TT-SI can be summarized in three steps: (i) first it identifies the samples that model struggles with (self-awareness), (ii) then generates similar examples from detected uncertain samples (self-data augmentation), and (iii) uses these newly generated samples at test-time training (self-improvement). We further explore Test-Time Distillation (TT-D), which leverages a stronger supervisor for targeted data generation. Empirical evaluations across different agent benchmarks demonstrate that TT-SI improves the performance with +5.48% absolute accuracy gain on average across all benchmarks and surpasses other standard learning methods more efficiently. Our findings highlight the promise of TT-SI, demonstrating the potential of self-improvement algorithms at test-time as a new paradigm for building more capable agents toward self-evolution.

[Long] Uncertainty Quantification in LLM Agents: Foundations, Emerging Challenges, and Opportunities ▶ *Underline Video*

Changdae Oh, Seongheon Park, To Eun Kim, Jiatong Li, Wendi Li, Samuel Yeh, Sean Du, Hamed Hassani, Paul Bogdan, Dawn Song and Sharon Li

Uncertainty quantification (UQ) for large language models (LLMs) is a key building block for safety guardrails of daily LLM applications. Yet, even as LLM agents are increasingly deployed in highly complex tasks, most UQ research still centers on single-turn question-answering. We argue that UQ research must shift to realistic settings with interactive agents, and that a new principled framework for agent UQ is needed. This paper presents three pillars to build a solid ground for future agent UQ research: (1. Foundations) We present the first general formulation of agent UQ that subsumes broad classes of existing UQ setups; (2. Challenges) We identify four technical challenges specifically tied to agentic setups—selection of uncertainty estimator, uncertainty of heterogeneous entities, modeling uncertainty dynamics in interactive systems, and lack of fine-grained benchmarks—with numerical analysis on a real-world agent benchmark, ^2-bench; (3. Future Directions) We conclude with noting on the practical implications of agent UQ and remaining open problems as forward-looking discussion for future explorations.

[Findings] Unifying Inference-Time Planning Language Generation ▶ *Underline Video*

Prabhu Prakash Kagitha, Bo Sun, Ishan Desai, Andrew Zhu, Cassie Huang, Manling Li, Ziyang Li and Li Zhang

A line of work in planning uses LLM not to generate a plan, but to generate a formal representation in some planning language, which can be input into a symbolic solver to deterministically find a plan. While showing improved trust and promising performance, dozens of recent publications have proposed scattered methods on a variety of benchmarks under different experimental settings. We attempt to unify the inference-time LLM-as-formalizer methodology for classical planning by proposing a unifying organizational framework based on intermediate representations. We thus systematically evaluate more than a dozen pipelines that subsume most existing work, while proposing novel ones that involve syntactically similar but high-resource intermediate languages (such as a Python wrapper of PDDL). We provide recipes for planning language generation pipelines, draw a series of conclusions showing the efficacy of their various components, and evidence their robustness against problem complexity.

[Long] Unlocking Implicit Experience: Synthesizing Tool-Use Trajectories from Text ▶ *Underline Video*

Zhihao Xu, Rumei Li, Jiahuan Li, Rongxiang Weng, Jingang Wang, Xunliang Cai and Xiting Wang

Enabling Large Language Models (LLMs) to effectively utilize tools in multi-turn interactions is essential for building capable autonomous agents. However, acquiring diverse and realistic multi-turn tool-use data remains a significant challenge. In this work, we propose a novel text-based paradigm. We observe that textual corpora naturally contain rich, multi-step problem-solving experiences, which can serve as an untapped, scalable, and authentic data source for multi-turn tool-use tasks. Based on this insight, we introduce GEM, a data synthesis pipeline that enables the generation and extraction of multi-turn tool-use trajectories from text corpora through a four-stage process: relevance filtering, workflow & tool extraction, trajectory grounding, and complexity refinement. To reduce the computational cost, we further train a specialized Trajectory Synthesizer via supervised fine-tuning. This model distills the complex generation pipeline into an efficient, end-to-end trajectory generator. Experiments demonstrate that our GEM-32B achieve a 14.9% improvement on the BFCL V3 Multi-turn benchmark. Our models partially surpass the performance of models trained on -bench (Airline and Retail) in-domain data, highlighting the superior generalization capability derived from our text-based synthesis paradigm. Notably, our Trajectory Synthesizer matches the quality of the full pipeline while significantly reducing inference latency and costs.

[Findings] Valid != Necessary: Diagnosing Latent Inefficiency in Chain-of-Thought ▶ *Underline Video*

Daeyeop Lee and Hwanjo Yu

Chain-of-Thought (CoT) prompting has significantly advanced the reasoning capabilities of Large Language Models (LLMs), yet it often incurs substantial computational costs due to over-reasoning, the generation of redundant, verbose, or irrelevant steps. While existing reasoning step evaluators effectively detect logical fallacies and factual errors, our analysis reveals a critical blind spot: they fail to penalize valid but inefficient reasoning steps that inflate token usage without contributing to the solution. To systematically diagnose this limitation,

we introduce RIV-GSM8K, a diagnostic benchmark injected with five distinct types of inefficiencies, including circular reasoning and excessive decomposition. Diagnostic experiments reveal that state-of-the-art evaluators struggle to distinguish these inefficiencies from necessary reasoning. To address this gap, we propose CAID (Context-Aware Information Density), a training-free metric grounded in information theory that identifies low-utility steps. To validate the metrics practical utility, we apply it within PACE, a post-hoc compression strategy. Additional control experiments show that the gains of PACE are not explained by trivial pruning: compared with random step removal and PRM-based compression baselines, it preserves accuracy at substantially higher compression rates. Empirical results on GSM8K, StrategyQA, and ARC-Challenge demonstrate that PACE reduces token consumption by 31-53% while maintaining accuracy, confirming that CAID successfully distills informational froth from reasoning chains without compromising deductive validity.

[Long] VC-Inspector: Advancing Reference-free Evaluation of Video Captions with Factual Analysis ▶ ***Underline Video***

Shubhashis Roy Dipta, Tz-Ying Wu and Subarna Tripathi

We propose VC-Inspector, a lightweight, open-source large multimodal model (LMM) for reference-free evaluation of video captions, with a focus on factual accuracy. Unlike existing metrics that suffer from limited context handling, weak factuality assessment, or reliance on proprietary services, VC-Inspector offers a reproducible, fact-aware alternative that aligns closely with human judgments. To enable robust training and interpretable evaluation, we introduce a systematic approach for generating captions with controllable errors, paired with graded quality scores and explanatory annotations. Experiments show that VC-Inspector achieves state-of-the-art correlation with human judgments, generalizing across diverse domains (e.g., VATEX-Eval, Flickr8K-Expert, and Flickr8K-CF benchmarks) and revealing the potential for caption improvement.

[Long] VCORE: Variance-Controlled Optimization-based Reweighting for Chain-of-Thought Supervision ▶ ***Underline Video***

Xuan Gong, Senmiao Wang, Hanbo Huang, Ruoyu Sun and Shiyu Liang

Supervised fine-tuning (SFT) on long chain-of-thought (CoT) trajectories has emerged as a crucial technique for enhancing the reasoning abilities of large language models (LLMs). However, the standard cross-entropy loss treats all tokens equally, ignoring their heterogeneous contributions across a reasoning trajectory. This uniform treatment leads to misallocated supervision and weak generalization, especially in complex, long-form reasoning tasks. To address this, we introduce Variance-Controlled Optimization-based REweighting (VCORE), a principled framework that reformulates CoT supervision as a constrained optimization problem. By adopting an optimization-theoretic perspective, VCORE enables a principled and adaptive allocation of supervision across tokens, thereby aligning the training objective more closely with the goal of robust reasoning generalization. Empirical evaluations demonstrate that VCORE achieves the strongest overall average performance, with especially clear gains on lower-capacity models. Across both in-domain and out-of-domain settings, VCORE achieves substantial performance gains on mathematical and coding benchmarks, using models from the Qwen3 series (4B, 8B, 32B) and LLaMA-3.1-8B-Instruct. Moreover, we show that VCORE serves as a more effective initialization for subsequent reinforcement learning, establishing a stronger foundation for advancing the reasoning capabilities of LLMs.

[Long] VecInfer: Efficient LLM Inference with Low-Bit KV Cache via Outlier-Suppressed Vector Quantization ▶ ***Underline Video***

Dingyu Yao, Chenxu Yang, Zhengyang Tong, Zheng Lin, Wei Liu, Jian Luan and Weiping Wang

The Key-Value (KV) cache introduces substantial memory overhead during large language model (LLM) inference. Although existing vector quantization (VQ) methods reduce KV cache usage and provide flexible representational capacity across bit-widths, they suffer severe performance degradation at ultra-low bit-widths due to key cache outliers that hinder effective codebook utilization. To address this challenge, we propose VecInfer, a novel VQ method for aggressive KV cache compression while enabling efficient inference. By applying smooth and Hadamard transformations, VecInfer suppresses outliers in the key cache, enabling the codebook to comprehensively cover the original data distribution and thereby reducing quantization difficulty. To facilitate efficient deployment, we design an optimized CUDA kernel that fuses computation with dequantization to minimize memory access overhead. Extensive evaluations demonstrate that VecInfer consistently outperforms existing quantization baselines across both long-context understanding and mathematical reasoning tasks. With only 2-bit quantization, VecInfer achieves performance comparable to full precision, while delivering up to 2.7x speedup in large-batch self-attention computation and 8.3x reduction in single-batch end-to-end latency on Llama-3.1-8B with a 196k sequence length.

[Findings] VIDA: A Visual Intent-driven Design Assistant for Proactive Multimodal Clarification ▶ ***Underline Video***

Yanshan Liu, Hongbo Zhang, Zhen Sun, Jiaheng Wei and Kaishun Wu

In complex domains like interior design, user requests are often ambiguous and multimodal. Professional designers address this by asking strategic clarification questions based on hierarchical priorities, a capability lacking in current Vision-Language Models (VLMs). When fine-tuned on dialogue data, existing models often exhibit modality forgetting, overfitting to textual patterns while neglecting visual cues and thus producing hallucinated or visually irrelevant questions. To bridge this gap, we introduce VIDA (Visual Intent-driven Design Assistant), an assistant designed to generate proactive, visually grounded, and strategically prioritized clarification questions. Instead of standard fine-tuning, we propose a strategy-aware alignment framework that evolves from imitation learning to value-driven reinforcement. We utilize Group Sequence Policy Optimization to strictly enforce expert protocols, ensuring the model not only mimics fluent speech but also adheres to optimal inquiry strategies. Crucially, we design a novel hierarchical reward mechanism with Dynamic Intent Binding to align the assistant with professional prioritization standards. To facilitate this research, we construct and release InteriorClarify, a multimodal benchmark dataset comprising 1,016 real-world consultation cases annotated with this three-tier intent hierarchy. Extensive experiments demonstrate that VIDA sets a new state-of-the-art, improving the Strategic

Alignment Score (SAS) by 20.59% over SFT baselines and effectively restoring visual grounding capabilities lost during standard fine-tuning.

[Long] VLN-NF: Feasibility-Aware Vision-and-Language Navigation with False-Premise Instructions ▶ *Underline Video*
Hung-Ting Su, Ting-Jun Wang, Jia-Fong Yeh, Min Sun and Winston H. Hsu
Conventional Vision-and-Language Navigation (VLN) benchmarks assume instructions are feasible and the referenced target exists, leaving agents ill-equipped to handle false-premise goals. We introduce VLN-NF, a benchmark with false-premise instructions where the target is absent from the specified area and agents must navigate, gather evidence through in-room exploration, and explicitly output NOT-FOUND. VLN-NF is constructed via a scalable pipeline that rewrites VLN instructions using an LLM and verifies target absence with a VLM, producing plausible yet factually incorrect goals. We further propose REV-SPL to jointly evaluate room reaching, exploration coverage, and decision correctness. To address this challenge, we present ROAM, a two-stage hybrid that combines supervised room-level navigation with LLM/VLM-driven in-room exploration guided by a free-space clearance prior. ROAM achieves the best REV-SPL among compared methods, while baselines often under-explore and terminate prematurely under unreliable instructions. Code and data will be released upon acceptance.

[Findings] VocabTailor: Dynamic Vocabulary Selection for Downstream Tasks in Small Language Models ▶ *Underline Video*
Hanling Zhang, Yayu Zhou, Tongcheng Fang, Zhihang Yuan, Guohao Dai, Wanli Ouyang and Yu Wang
Small Language Models (SLMs) provide computational advantages in resource-constrained environments, yet memory limitations remain a critical bottleneck for edge device deployment. A substantial portion of SLMs' memory footprint stems from vocabulary-related components, particularly embeddings and language modeling (LM) heads, due to large vocabulary sizes. Existing static vocabulary pruning, while reducing memory usage, suffers from rigid, one-size-fits-all designs that cause information loss during the prefill stage and lack flexibility. In this work, we identify two key principles underlying the vocabulary reduction challenge: the lexical locality principle, the observation that only a small subset of tokens is required during any single inference, and the asymmetry in computational characteristics between vocabulary-related components of SLM. Based on these insights, we introduce VocabTailor, a novel decoupled dynamic vocabulary selection framework that addresses memory constraints through offloading embedding and implements a hybrid static-dynamic vocabulary selection strategy for LM Head, enabling on-demand loading of vocabulary components. Comprehensive experiments across diverse downstream tasks demonstrate that VocabTailor achieves a reduction of up to 99% in the memory usage of vocabulary-related components with minimal or no degradation in task performance, substantially outperforming existing static vocabulary pruning. Our code is available at https://github.com/AwakenedInsects/VocabTailor.

[Findings] Vulnerability of LLMs' Stated Belief? LLMs Belief Resistance Check Through Strategic Persuasive Conversation Interventions ▶ *Underline Video*
Fan Huang, Haewoon Kwak and Jisun An
Large Language Models (LLMs) are increasingly employed in various question-answering tasks. However, recent studies showcase that LLMs are susceptible to persuasion and could adopt counterfactual beliefs. We present a systematic evaluation of LLM susceptibility to persuasion under the Source–Message–Channel–Receiver (SMCR) communication framework. Across six mainstream Large Language Models (LLMs) and three domains (factual knowledge, medical QA, and social bias), we analyze how different persuasive strategies influence stated belief stability over multiple interaction turns. We further examine whether verbalized confidence prompting (i.e., eliciting self-reported confidence scores) affects resistance to persuasion. Results show that the smallest model (Llama 3.2-3B) exhibits extreme compliance, with 82.5% of belief changes occurring at the first persuasive turn (average end turn of 1.1–1.4). Contrary to expectations, verbalized confidence prompting increases vulnerability by accelerating belief erosion rather than enhancing robustness. Finally, an exploratory study of adversarial fine-tuning reveals highly model-dependent effectiveness: GPT-4o-mini achieves near-complete robustness (98.6%) and Mistral~7B improves substantially (35.7% -> 79.3%), but Llama models remain highly susceptible (<14% RQ1) even when fine-tuned on their own failure cases. Together, these findings highlight substantial model-dependent limits of current robustness interventions and offer guidance for developing more trustworthy LLMs.

[TACL] What Can String Probability Tell Us About Grammaticality? ▶ *Underline Video*
Jennifer Hu, Ethan Wilcox, Siyuan Song, Kyle Mahowald and Roger Levy
What have language models (LMs) learned about grammar? This question remains hotly debated, with major ramifications for linguistic theory. However, since probability and grammaticality are distinct notions in linguistics, it is not obvious what string probabilities can reveal about an LM's underlying grammatical knowledge. We present a theoretical analysis of the relationship between grammar, meaning, and string probability, based on simple assumptions about the generative process of corpus data. Our framework makes three predictions, which we validate empirically using 280K sentence pairs in English and Chinese: (1) correlation between the probability of strings within minimal pairs, i.e., string pairs with minimal semantic differences; (2) correlation between models' and humans' deltas within minimal pairs; and (3) poor separation in probability space between unpaired grammatical and ungrammatical strings. Our analyses give theoretical grounding for using probability to learn about LMs' structural knowledge, and suggest directions for future work in LM grammatical evaluation.

[Findings] When Backdoors Go Beyond Triggers: Semantic Drift in Diffusion Models Under Encoder Attacks ▶ *Underline Video*
Shenyang Chen and Liuwan Zhu
Standard evaluations of backdoor attacks on text-to-image (T2I) models primarily measure trigger activation and visual fidelity. We challenge

this paradigm, demonstrating that encoder-side poisoning induces persistent, trigger-free semantic corruption that fundamentally reshapes the representation manifold. We trace this vulnerability to a geometric mechanism: a Jacobian-based analysis reveals that backdoors act as low-rank, target-centered deformations that amplify local sensitivity, causing distortion to propagate coherently across semantic neighborhoods. To rigorously quantify this structural degradation, we introduce SEMAD (Semantic Alignment and Drift), a diagnostic framework that measures both internal embedding drift and downstream functional misalignment. Our findings, validated across diffusion and contrastive paradigms, expose the deep structural risks of encoder poisoning and highlight the necessity of geometric audits beyond simple attack success rates.

[Long] When High Accuracy Hides Poor Calibration: Rethinking Confidence Evaluation in Transformer-Based Text Classification with Balanced Brier Score ▶ *Underline Video*

Guilherme Fonseca, Gabriel Prenassi, Washington Cunha, Leonardo Rocha and Marcos André Gonçalves

Transformer-based Small (SLMs) and Large Language Models (LLMs) achieve strong effectiveness in text classification (TC), yet deployment requires reliable confidence estimates. Although miscalibration in Transformers has been reported, evidence for TC under fine-tuning remains limited. We evaluate the calibration of fine-tuned SLMs and LLMs against Logistic Regression, a classical, well-calibrated baseline, and find that, despite superior effectiveness, Transformers remain markedly overconfident. Crucially, we show that widely used calibration metrics, such as Expected Calibration Error and Brier Score, become biased in high-effectiveness regimes, where the dominance of correct predictions masks severe miscalibration on errors, sometimes even suggesting better calibration than Logistic Regression, a well-known calibrated method. To address this limitation, we propose the Balanced Brier Score (BBS), which balances the contribution of correct and incorrect predictions within confidence bins. BBS reveals substantially poorer calibration in both SLMs and LLMs, consistent with qualitative evidence from calibration curves. These findings challenge current calibration assessment practices and provide a more reliable alternative for evaluating confidence quality in Transformer-based TC.

[Long] Why Do LLM-based Web Agents Fail? A Hierarchical Planning Perspective ▶ *Underline Video*

Mohamed Aghzal, Gregory J. Stein and Ziyu Yao

Large language model (LLM) web agents are increasingly used for web navigation but remain far from human reliability on realistic, long-horizon tasks. Existing evaluations focus primarily on end-to-end success, offering limited insight into where failures arise. We propose a hierarchical planning framework that analyzes web agents across three layers (i.e., high-level planning, low-level execution, and re-planning), enabling process-based evaluation of reasoning, grounding, and recovery. Our experiments show that structured Planning Domain Definition Language (PDDL) plans produce more concise and goal-directed strategies than natural language (NL) plans, but low-level execution remains the dominant bottleneck. These results indicate that improving perceptual grounding and adaptive control, not only high-level reasoning, is critical for achieving human-level reliability. This hierarchical perspective provides a principled foundation for diagnosing and advancing LLM web agents.

[Long] Why LLM Safety Guardrails Collapse After Fine-tuning: A Similarity Analysis Between Alignment and Fine-tuning Datasets ▶ *Underline Video*

Lei Hsiung, Tianyu Pang, Yung-Chen Tang, Linyue Song, Tsung-Yi Ho, Pin-Yu Chen and Yaoqing Yang

Recent advancements in large language models (LLMs) have underscored their vulnerability to safety alignment jailbreaks, particularly when subjected to downstream fine-tuning. However, existing mitigation strategies primarily focus on reactively addressing jailbreak incidents after safety guardrails have been compromised, removing harmful gradients during fine-tuning, or continuously reinforcing safety alignment throughout fine-tuning. As such, they tend to overlook a critical upstream factor: the role of the original safety-alignment data. This paper therefore investigates the degradation of safety guardrails through the lens of representation similarity between upstream alignment datasets and downstream fine-tuning tasks. Our experiments demonstrate that high similarity between these datasets significantly weakens safety guardrails, making models more susceptible to jailbreaks. Conversely, low similarity between these two types of datasets yields substantially more robust models and thus reduces harmfulness score by up to 10.33%. By highlighting the importance of upstream dataset design in the building of durable safety guardrails and reducing real-world vulnerability to jailbreak attacks, these findings offer actionable insights for fine-tuning service providers to prioritize upstream models with low jailbreak risk.

[Findings] WikiVideo: Article Generation from Multiple Videos ▶ *Underline Video*

Alexander Martin, Reno Kriz, William Gantt Walden, Kate Sanders, Hannah Recknor, Eugene Yang, Francis Ferraro and Benjamin Van Durme

We introduce the task of grounded article generation with the goal of creating a Wikipedia-style article from multiple diverse videos about real-world events—from natural disasters to political elections—where all the information in the article is supported by video evidence. Videos are intuitive sources for retrieval-augmented generation (RAG), but most contemporary RAG workflows focus heavily on text while existing methods for video-based summarization focus on low-level scene understanding rather than high-level event semantics. To close this gap, we introduce, a benchmark consisting of expert-written articles and densely annotated videos that provide evidence for articles' claims, facilitating the integration of video into RAG pipelines and enabling the creation of in-depth content that is grounded in multimodal sources. We further propose Collaborative Article Generation (CAG), a novel interactive method for article creation from multiple videos. CAG leverages an iterative interaction between an r1-style reasoning model and a VideoLLM to draw higher-level inferences about the target event than is possible with VideoLLMs alone, which fixate on low-level visual features. We benchmark state-of-the-art VideoLLMs and

CAG in both oracle retrieval and RAG settings and find that CAG consistently outperforms alternative methods, while suggesting intriguing avenues for future work.

[Long] XtraGPT: Context-Aware and Controllable Academic Paper Revision via Human-AI Collaboration ► *Underline Video*
Nuo Chen, Andre Lin HuiKai, Jiaying Wu, Junyi Hou, Zining Zhang, Qian Wang, Xidong Wang and Bingsheng He
Despite the growing adoption of large language models (LLMs) in academic workflows, their capabilities remain limited in supporting high-quality scientific writing. Most existing systems are designed for general-purpose scientific text generation and fail to meet the sophisticated demands of research communication beyond surface-level polishing, for example, maintaining conceptual coherence across sections. Furthermore, academic writing is inherently iterative and revision-driven, a process that is not well supported by direct prompting-based paradigms. To address these scenarios, we propose a human-AI collaboration framework for academic paper revision, centered on criteria-guided intent alignment and context-aware modeling. To validate the framework, we curate a dataset of 7,000 research papers from top-tier venues, annotated with 140,000 instruction-response pairs that reflect realistic, section-level scientific revisions. We instantiate the framework in XtraGPT, the first suite of open-source LLMs (1.5B to 14B parameters) specifically fine-tuned for context-aware academic paper revision. Extensive experiments show that XtraGPT significantly outperforms same-scale baselines and rivals the quality of proprietary counterparts. Both automated preference assessments and human evaluations confirm the effectiveness of XtraGPT in improving scientific drafts. Our code and models are available at https://github.com/Xtra-Computing/XtraGPT and https://huggingface.co/collections/Xtra-Computing/xtragpt.

[Long] Your Inference Request Will Become a Black Box: Confidential Inference for Cloud-based Large Language Models ► *Underline Video*
Chung-ju Huang, Huiqiang Zhao, Yuanpeng He, Lijian Li, Wenpin Jiao, Zhi Jin, PeixuanChen PeixuanChen and Leye Wang
The increasing reliance on cloud-hosted Large Language Models (LLMs) exposes sensitive client data, such as prompts and responses, to potential privacy breaches by service providers. Existing approaches fail to ensure privacy, maintain model performance, and preserve computational efficiency simultaneously. To address this challenge, we propose Talaria, a confidential inference framework that partitions the LLM pipeline between a client-verified Confidential Virtual Machine (CVM) and the public cloud to protect client data without compromising the cloud's model intellectual property or inference quality. The interaction between the CVM and the cloud is secured by our Reversible Masked Outsourcing (ReMO) protocol, which uses a hybrid masking technique to reversibly obscure intermediate data before outsourcing computations. Extensive evaluations show that Talaria can defend against state-of-the-art token inference attacks, reducing token reconstruction accuracy from over 97.5% to an average of 1.34%, all while being a lossless mechanism that guarantees output identical to the original model without significantly decreasing efficiency and scalability. To the best of our knowledge, this is the first work that ensures clients prompts and responses remain inaccessible to the cloud, while also preserving model privacy, performance, and efficiency.

[Findings] Your LLM Agent Can Leak Your Data: Data Exfiltration via Backdoored Tool Use ► *Underline Video*
Wuyang Zhang and Shichao Pei
Tool-use large language model (LLM) agents are increasingly deployed to support sensitive workflows, relying on tool calls for retrieval, external API access, and session memory management. While prior research has examined various threats, the risk of systematic data exfiltration by backdoored agents remains underexplored. In this work, we present a backdoor attack framework that embeds semantic triggers into fine-tuned LLM agents. When triggered, the backdoored agent invokes memory-access tool calls to retrieve stored user context and exfiltrates it via disguised retrieval tool calls. We further demonstrate that multi-turn interaction amplifies the impact of data exfiltration, as attacker-controlled retrieval responses can subtly steer subsequent agent behavior and user interactions, enabling sustained and cumulative information leakage over time. Our experimental results expose a critical vulnerability in LLM agents with tool access and highlight the need for defenses against exfiltration-oriented backdoors.

Poster Session A (Zoom)

11:00-12:30 (Zoom)

[Long] Learning What Matters: Dynamic Dimension Selection and Aggregation for Interpretable Vision-Language Reward Modeling ► *Underline Video*
Qiyuan Chen, Hongsen Huang, Jiahe Chen, Qian Shao, Jintai Chen, Hongxia Xu, Renjie Hua, Chuan Ren and Jian Wu
Vision-language reward modeling faces a dilemma: generative approaches are interpretable but slow, while discriminative ones are efficient but act as opaque "black boxes." To bridge this gap, we propose VL-MDR (Vision-Language Multi-Dimensional Reward), a framework that dynamically decomposes evaluation into granular, interpretable dimensions. Instead of outputting a monolithic scalar, VL-MDR employs a visual-aware gating mechanism to identify relevant dimensions and adaptively weight them (e.g., Hallucination, Reasoning) for each specific input. To support this, we curate a dataset of 321k vision-language preference pairs annotated across 21 fine-grained dimensions. Extensive experiments show that VL-MDR consistently outperforms existing open-source reward models on benchmarks like VL-RewardBench. Furthermore, we show that VL-MDR-constructed preference pairs effectively enable DPO alignment to mitigate visual hallucinations and

improve reliability, providing a scalable solution for VLM alignment.

Session 3: Oral/Posters/Demos B - 14:00-15:30

Demo Session B

14:00-15:30 (Grand Hall)

[Demo] LinkNav: Surfacing Interconnected Information in Scientific Articles ▶ *Underline Video*

Sebastian Joseph, Jennifer Healey, Junyi Jessy Li and Ani Nenkova

We present LinkNav, an enhanced experience for reading academic papers which makes explicit connections between related but non-adjacent passages. To create the experience, we instruct a language model to generate questions that may arise while reading a passage and then search for answer-bearing passages elsewhere in the document, forming intra-document connections when answers are found. We confirm that these building blocks work well to power the experience, with an answer detection pipeline that works with high precision, resulting in a reasonable number of such connections being made for a document. On a dataset of academic papers, we find that connected segments are on average ten segments away from each other, making explicit connections that a reader may have otherwise missed. The LinkNav experience is accessible through this link: https://elaborate-hamster-086a75.netlify.app.

[Demo] MASEval: Extending Multi-Agent Evaluation from Models to Systems ▶ *Underline Video*

Cornelius Emde, Alexander Rubinstein, Anmol Goel, Ahmed Heakl, Sangdoo Yun, Seong Joon Oh and Martin Gubri

The rapid adoption of LLM-based agentic systems has produced a rich ecosystem of frameworks (smolagents, LangGraph, AutoGen, CAMEL, LlamaIndex, i.a.). Yet existing benchmarks are model-centric: they fix the agentic setup and do not compare other system components. We argue that implementation decisions substantially impact performance, including choices such as topology, orchestration logic, and error handling. MASEval addresses this evaluation gap as a framework-agnostic library that treats the entire system as the unit of analysis. Through the first systematic system-level comparison across 3 benchmarks, 3 models, and 3 frameworks, we find that framework choice matters as much as model choice. MASEval allows researchers to explore all components of agentic systems, opening new avenues for principled system design, and practitioners to identify the best implementation for their use case. MASEval is available under the MIT licence at https://github.com/parameterlab/MASEval.

[Demo] PaperMentor: A Human-Centered Multi-Agent Writing Tutor for AI Research Papers in Overleaf ▶ *Underline Video*

Jiarui Liu, Terry Jingchen Zhang, Ryan Faulkner, X. Angelo Huang, Vilém Zouhar, Dominik Glandorf, Isabel Dahlgren, Rishit Dagli, Yuen Chen, Felix Leeb, Van Q. Truong, Punya Syon Pandey, Yves Bicker, Suvajit Majumder, Wenyuan Jiang, Zeju Qiu, Sankalan Pal Chowdhury, Mrinmaya Sachan, Bernhard Schölkopf, Mona T. Diab and Zhijing Jin

Junior AI researchers often lack writing-level feedback on their paper drafts. Existing automated tools focus on the review stage, flagging methodological weaknesses, rather than providing writing suggestions during drafting. We present PaperMentor, a human-centered writing assistant that delivers actionable suggestions as native inline comments on Overleaf. PaperMentor integrates an expert skill library, consolidating writing advice from senior researchers, into a multi-agent system with twelve specialized agents covering different aspects of paper writing. In a user study with 14 annotators, 90.6% of the generated comments were rated actionable and 67.5% were rated valid, significantly outperforming a baseline system without the skill library. We open-source PaperMentor for public use.

[Demo] PROTEA: Offline Evaluation and Iterative Refinement for Multi-Agent LLM Workflows ▶ *Underline Video*

Kazuki Kawamura, Satoshi Waki and Kei Tateno

Multi-agent LLM workflows, which are AI systems composed of multiple role-specialized LLM calls, often outperform single prompts, but they are notoriously difficult to debug and refine. Failures can originate from subtle mistakes in intermediate artifacts that silently propagate downstream, forcing developers to read long traces and guess which agent to edit. We present PROTEA, a unified UI that closes the loop for offline, test-case–driven improvement of multi-agent workflows, enabling developers to efficiently diagnose and fix errors without manual inspection of long traces. PROTEA executes a workflow, scores intermediate artifacts with configurable evaluators, and overlays per-node states and rationales on the workflow graph to localize likely bottlenecks. To address the difficulty of preparing intermediate reference in complex systems, PROTEA performs backward node evaluation by inferring each node's ideal expected output from terminal supervision and graph context, and comparing it with the observed node output. For a selected node, it proposes a targeted prompt patch as an editable diff, then automatically re-runs and re-evaluates the workflow to show before/after output diffs and score trajectories within the same interface. Using PROTEA, users can visually pinpoint system-wide bottlenecks at a glance, streamline remediation via semi-automated prompt patching, and immediately verify pre- and post-correction outcomes within a unified loop.

[Demo] rosaOS: Agentic Operating System for Embodied LLMs ▶ *Underline Video*

Yijun Ge, Kushaldeep Mujral, Karthik Nambiar and Jimmy Lin;

We present rosaOS, an open source agentic operating system for embodied LLMs: interactive, LLM-driven agents coordinate various

software tools and physical devices through a single embodied desktop companion, the Reachy Mini robot. Existing LLM-robotic systems are generally built as an intertwined stack, making it difficult to switch hardware, add extra capabilities, or expand to multiple devices without bespoke integration. rosaOS aims to provide an OS-inspired architecture where an agentic kernel manages all task execution and mediates device access, while worker agents invoke tools to perform actions. We adopt industry standard interfaces with MCP for agentic tool-calling and ROS for connecting robots, and demonstrate rosaOS on a multi-device setup including a quadruped robot, a mobile robot, and a smart lamp, all controlled through interactions with Reachy Mini. By incorporating MCP extensibility with ROS hardware interoperability, rosaOS enables a plug-and-play ecosystem to build and evaluate embodied agentic systems. The implementation is available at rosaos.ai.

[Demo] SymbolicThought: Integrating Language Models and Symbolic Reasoning for Consistent and Interpretable Human Relationship Understanding ▶ *Underline Video*

Runcong Zhao, Qinglin Zhu, Hainiu Xu, Bin Liang, Lin Gui and Yulan He

Understanding character relationships is essential for interpreting complex narratives and conducting socially grounded AI research. However, manual annotation is time-consuming and low in coverage, while large language models (LLMs) often produce hallucinated or logically inconsistent outputs. We present SymbolicThought, a human-in-the-loop framework that combines LLM-based extraction with symbolic reasoning. The system constructs editable character relationship graphs, refines them using seven types of logical constraints, and enables real-time validation and conflict resolution through an interactive interface. To support logical supervision and explainable social analysis, we release a dataset of 160 interpersonal relationships with corresponding logical structures. Experiments show that SymbolicThought improves annotation accuracy and consistency while significantly reducing time cost, offering a practical tool for narrative understanding, explainable AI, and LLM evaluation. A screencast (https://www.youtube.com/watch?v=nb10-bDCkRU) and demo (http://104.168.96.23:3000/) are available online.

[Demo] TokCollate: A Comprehensive Tool for Tokenizer Evaluation and Visualization across Languages ▶ *Underline Video*

Duan Vari; Abishek Stephen; Jindich Libovický

Tokenization quality varies significantly across languages, contributing to disparities in LLM performance and cost for speakers of less-resourced languages - a phenomenon known as the "token premium" problem. Despite growing research interest, no existing tool provides a comprehensive intrinsic evaluation of tokenizers paired with interactive visualization. We present TokCollate (pronounced similarly to chocolate), a Python-based evaluation framework combined with a JavaScript visualization interface that addresses this gap. TokCollate implements a wide range of intrinsic metrics, including monolingual measures such as average token length and Rényi/Shannon efficiency, and cross-lingual measures such as vocabulary overlap, Jensen-Shannon divergence, alignment-based Eflomal scores, and length ratios. It further enables analysis across language groups defined by genealogical families, scripts, geographic regions, speaker populations, and estimated data availability. TokCollate is open-source under the MIT license and available on GitHub.

[Demo] WIGVO: Real-Time Bidirectional Speech Translation over Legacy PSTN Calls via Dual-Session Echo Gating ▶ *Underline Video*

Hyeongseob Kim, Sang-Woo Son, Hyun-woo Cho, Hyeonsang Kim and Jinmo Kim

Real-time speech translation with large language models (LLMs) has become feasible in controlled settings such as mobile apps and web browsers, where developers assume wideband audio and client-side echo cancellation. Recent end-to-end full-duplex models further push latency below 200\{},ms but remain confined to broadband environments. However, deploying such systems over the Public Switched Telephone Network (PSTN) remains challenging due to narrowband G.711 audio, unpredictable round-trip delays, and absence of client-side signal processing. We present WIGVO (WIGTN Voice-Only), a server-side relay system that enables bidirectional LLM-based speech translation over ordinary telephone calls without requiring app installation or carrier integration. A central contribution is addressing what we term echo-induced self-reinforcing translation loops: synthesized speech echoing back through the PSTN gets re-ingested and repeatedly translated. WIGVO solves this through a dual-session architecture with deterministic silence injection and energy-based voice activity detection (VAD) gating. We evaluate WIGVO on 155 PSTN calls (148 instrumented, 147 completed) across three communication modes—voice-to-voice (V2V), text-to-voice (T2V), and full-agent—observing 555\{},ms median caller-to-callee latency and 2,684\{},ms median callee-to-caller latency, zero echo-induced translation loops, and USD\{},0.28 per minute cost. The system is deployed at https://wigvo-web-gzjzn35jyq-du.a.run.app with a live demo and video walkthrough. Evaluation scripts and anonymized call logs are available in the open-source repository.

Oral Session B: AI/LLM Agents 1

14:00-15:30 (Harbor E-F)

[Long] ACE-Router: Generalizing History-Aware Routing from MCP Tools to the Agent Web ▶ *Underline Video*

Zhiyuan Yao, Zishan Xu, Yifu Guo, Zhiguang Han, Cheng Yang, Shuo Zhang, Weinan Zhang, Xingshan Zeng and Weiwen Liu

With the rise of the Agent Web and Model Context Protocol (MCP), the agent ecosystem is evolving into an open collaborative network, exponentially increasing accessible tools. However, current architectures face severe scalability and generality bottlenecks. To address

this, we propose ACE-Router, a pipeline for training history-aware routers to empower precise navigation in large-scale ecosystems. By leveraging a dependency-rich candidate Graph to synthesize multi-turn trajectories, we effectively train routers with dynamic context understanding to create the plug-and-play Light Routing Agent. Experiments on the real-world benchmarks MCP-Universe and MCP-Mark demonstrate superior performance. Notably, ACE-Router exhibits critical properties for the future Agent Web: it not only generalizes to multi-agent collaboration with minimal adaptation but also maintains exceptional robustness against noise and scales effectively to massive candidate spaces. These findings provide a strong empirical foundation for universal orchestration in open-ended ecosystems. Our code is available at https://github.com/euyis1019/ACE-Router.

[Long] AgentOCR: Reimagining Agent History via Optical Self-Compression ▶ *Underline Video*

Lang Feng, Fuchao Yang, Feng Chen, Xin Cheng, Haiyang Xu, Zhenglin Wan, Ming Yan and Bo An

Recent advances in large language models (LLMs) enable agentic systems trained with reinforcement learning (RL) over multi-turn interaction, but practical deployment is bottlenecked by rapidly growing textual histories that inflate token and memory costs. We introduce AgentOCR, a framework that exploits visual tokens' superior information density by representing the accumulated observation-action history as a compact rendered image. To make multi-turn rollouts scalable, AgentOCR proposes segment optical caching. By decomposing history into hashable segments and maintaining a visual cache, this mechanism eliminates redundant re-rendering. Beyond fixed rendering, AgentOCR introduces agentic self-compression, where the agent actively emits a compression rate and is trained with compression-aware reward to adaptively balance task success and token efficiency. We conduct extensive experiments on challenging agentic benchmarks, ALFWorld and search-based QA. Remarkably, AgentOCR preserves over 95% of text-based agent performance while substantially reducing token consumption (>50%), yielding consistent token and memory efficiency. Further analysis validates a 20x rendering speedup from optical caching and effective self-compression balancing. Our code is available at https://github.com/langfengQ/AgentOCR.

[Long] BioProAgent: Neuro-Symbolic Grounding for Constrained Scientific Planning ▶ *Underline Video*

Yuyang Liu, Jingya Wang, Liuzhenghao Lv and Yonghong Tian

Large language models (LLMs) have demonstrated significant reasoning capabilities in scientific discovery but struggle to bridge the gap to physical execution in wet-labs. In these irreversible environments, probabilistic hallucinations are not merely incorrect; they can cause equipment damage or experimental failure. To address this, we propose BioProAgent, a neuro-symbolic framework that anchors probabilistic planning in a deterministic Finite State Machine (FSM). We introduce a State-Augmented Planning mechanism that enforces a rigorous Design-Verify-Rectify workflow, ensuring hardware compliance before execution. Furthermore, we address the context bottleneck inherent in complex device schemas by Semantic Symbol Grounding, reducing token consumption by 6x through symbolic abstraction. In the extended BioProBench benchmark, BioProAgent achieves 95.6% physical compliance (compared to 21.0% for ReAct), demonstrating that neuro-symbolic constraints are essential for reliable autonomy in irreversible physical environments. Code at https://github.com/YuyangSunshine/bioproagent and Website at https://yuyangsunshine.github.io/BioPro-Project/.

[Long] DeepPlanning: Benchmarking Long-Horizon Agentic Planning with Verifiable Constraints ▶ *Underline Video*

Yinger Zhang, Shutong Jiang, Renhao Li, Jianhong Tu, Yang Su, Lianghao Deng, Xudong Guo, ChenXu Lv and Junyang Lin

While agent evaluation has shifted toward long-horizon tasks, most benchmarks still emphasize local, step-level reasoning rather than the global constrained optimization (e.g., time and financial budgets) that demands genuine planning ability. Meanwhile, existing LLM planning benchmarks underrepresent the active information gathering and fine-grained local constraints typical of real-world settings. To address this, we introduce DeepPlanning, a challenging benchmark for practical long-horizon agent planning. It features multi-day travel planning and multi-product shopping tasks that require proactive information acquisition, local constrained reasoning, and global constrained optimization. Evaluations on DeepPlanning show that even frontier agentic LLMs struggle with these problems, highlighting the importance of reliable explicit reasoning patterns and parallel tool use for achieving better effectiveness-efficiency trade-offs. Error analysis further points to promising directions for improving agentic LLMs over long planning horizons. We open-source the code and data to support future research.

[Long] Enhanced Reasoning for Biomedical Document-Level Relation Extraction via a Novel Cascade Language Model Framework ▶ *Underline Video*

Haohua Song, Wenhao Gu, Zhijing Li, Yunwen Yu, Tiantian Zhu, Xiao Yang and Zexuan Zhu

Biomedical document-level relation extraction poses significant challenges beyond sentence-level tasks, as it necessitates the integration of evidence from entire documents and the ability for coherent cross-sentence reasoning. While pretrained language models (PLMs) demonstrate efficiency in handling local contexts, they often struggle with global dependency modeling. Conversely, large language models (LLMs) exhibit strong reasoning capabilities but tend to generate hallucinations in knowledge-intensive biomedical domains. This paper introduces CoRE, a novel cascade framework that leverages the complementary strengths of PLMs and LLMs through a detect-then-rethink paradigm. The PLM serves as an efficient detector for high-confidence relations, while challenging cases are forwarded to an LLM enhanced with semantic retrieval and iterative reasoning mechanisms. Experimental results on BioRED and CDR datasets show that CoRE achieves substantial improvements over state-of-the-art baselines, validating the effectiveness of the proposed cascade paradigm for complex biomedical relation extraction.

[Long] Explicit Trait Inference for Multi-Agent Coordination ▶ *Underline Video*

Suhaib Abdurahman, Etsuko Ishii, Katerina Margatina, Divya Bhargavi, Monica Sunkara and Yi Zhang

LLM-based multi-agent systems (MAS) show promise on complex tasks but remain prone to coordination failures such as goal drift, error cascades, and misaligned behaviors. We propose Explicit Trait Inference (ETI), a psychologically grounded method for improving coordination. ETI enables agents to infer and track partner characteristics along two established psychological dimensions, warmth (e.g., trust) and competence (e.g., skill), from interaction histories to guide decisions. We evaluate ETI in controlled settings (economic games), where it reduces payoff loss by 45-77%, and in more realistic, complex multi-agent settings (MultiAgentBench), where it improves performance by 3-29% depending on the scenario and model, relative to a CoT baseline. Additional analysis shows that gains are closely linked to trait inference: ETI profiles predict agents actions, and informative profiles drive improvements. These results highlight ETI as a lightweight and robust mechanism for improving coordination in diverse multi-agent settings, and provide the first systematic evidence that LLM agents can (i) reliably infer others traits from interaction histories and (ii) leverage structured awareness of others traits for coordination.

[Long] OCR-Memory: Optical Context Retrieval for Long-Horizon Agent Memory ▶ *Underline Video*

Jinze Li, Yang Zhang, Xin Yang, Jiayi Qu, Jinfeng Xu, Shuo Yang, Junhua Ding and Edith Cheuk-Han Ngai

Autonomous LLM agents increasingly operate in long-horizon, interactive settings where success depends on reusing experience accumulated over extended histories. However, existing agent memory systems are fundamentally constrained by text-context budgets: storing or revisiting raw trajectories is prohibitively token-expensive, while summarization and text-only retrieval trade token savings for information loss and fragmented evidence. To address this limitation, we propose Optical Context Retrieval Memory (OCR-Memory), a memory framework that leverages the visual modality as a high-density representation of agent experience, enabling retention of arbitrarily long histories with minimal prompt overhead at retrieval time. Specifically, OCR-Memory renders historical trajectories into images annotated with unique visual identifiers. OCR-Memory retrieves stored experience via a locate-and-transcribe paradigm that selects relevant regions through visual anchors and retrieves the corresponding verbatim text, avoiding free-form generation and reducing hallucination. Experiments on long-horizon agent benchmarks show consistent gains under strict context limits, demonstrating that optical encoding increases effective memory capacity while preserving faithful evidence recovery.

[Long] OS-Sentinel: Towards Safety-Enhanced Mobile GUI Agents via Hybrid Validation in Realistic Workflows ▶ *Underline Video*

Qiushi Sun, Mukai Li, Zhoumianze Liu, Zhihui Xie, Fangzhi Xu, Zhangyue Yin, Kanzhi Cheng, Zehao Li, Zichen Ding, Qi Liu, Zhiyong Wu, Zhuosheng Zhang, Ben Kao and Lingpeng Kong

Computer-using agents powered by Vision-Language Models (VLMs) have demonstrated human-like capabilities in operating digital environments like mobile platforms. While these agents hold great promise for advancing digital automation, their potential for unsafe operations, such as system compromise and privacy leakage, is raising significant concerns. Detecting these safety concerns across the vast and complex operational space of mobile environments presents a formidable challenge that remains critically underexplored. To establish a foundation for mobile agent safety research, we introduce MobileRisk-Live, a dynamic sandbox environment accompanied by a safety detection benchmark comprising realistic trajectories with fine-grained annotations. Built upon this, we propose OS-Sentinel, a novel hybrid safety detection framework that synergistically combines a Formal Verifier for detecting explicit system-level violations with a VLM-based Contextual Judge for assessing contextual risks and agent actions. Experiments show that achieves 10%-30% improvements over existing approaches across multiple metrics. Further analysis provides critical insights that foster the development of safer and more reliable autonomous mobile agents. Our code, environment, and data are available at https://qiushisun.github.io/OS-Sentinel-Home/.

[Long] Timely Machine: Awareness of Time Makes Test-Time Scaling Agentic ▶ *Underline Video*

Yichuan Ma, Linyang Li, Yongkang Chen, Peiji Li, Xiaozhe Li, Qipeng Guo, Dahua Lin and Kai Chen

As large language models (LLMs) increasingly tackle complex reasoning tasks, test-time scaling has become critical for enhancing capabilities. However, in agentic scenarios with frequent tool calls, the traditional generation-length-based definition breaks down: tool latency decouples inference time from generation length. We propose Timely Machine, redefining test-time as wall-clock time, where models dynamically adjust strategies based on time budgets. We introduce Timely-Eval, a benchmark spanning high-frequency tool calls, low-frequency tool calls, and time-constrained reasoning. By varying tool latency, we find smaller models excel with fast feedback through more interactions, while larger models dominate high-latency settings via superior interaction quality. Moreover, existing models fail to adapt reasoning to time budgets. We propose Timely-RL to address this gap. After cold-start supervised fine-tuning, we use reinforcement learning to enhance temporal planning. Timely-RL improves time budget awareness and consistently boosts performance across Timely-Eval. We hope our work offers a new perspective on test-time scaling for the agentic era.

Oral Session B: Industry 2

14:00-15:30 (Balboa)

[Industry] A Unified Framework for Modeling Heterogeneous Financial Data via Dual-Granularity Prompting ▶ *Underline Video*

Yu Lei, Zixuan Wang, Yiqing Feng, Junru Zhang, Yahui Li, Liu Chu, Tongyao Wang and Dongyang Li

Recent industrial credit scoring models remain heavily reliant on manually tuned statistical learning methods. Despite their potential, deep learning architectures have struggled to consistently outperform traditional statistical models in industrial credit scoring, largely due to the

complexity of heterogeneous financial data and the challenge of modeling evolving creditworthiness. To bridge this gap, we introduce FinLangNet, a novel framework that reformulates credit scoring as a multi-scale sequential learning problem. FinLangNet processes heterogeneous financial data through a dual-module architecture that combines tabular feature extraction with temporal sequence modeling, generating probability distributions of users' future financial behaviors across multiple time horizons. A key innovation is our dual-prompt mechanism within the sequential module, which introduces learnable prompts operating at both feature-level granularity for capturing fine-grained temporal patterns and user-level granularity for aggregating holistic risk profiles. Notably, real world deployment yielded a 6.3 pp improvement in KS, along with a 9.9% reduction in bad debt rate.

[Industry] Agentic Context Strategies for Multi-Format Document Understanding: When Should Language Models Use Tools? ► *Underline Video*

Mukul Singh, Mansi Uniyal and Ryan Nadel

Large language models face fundamental trade-offs when processing long documents: full context is expensive and may exceed limits, while RAG risks missing relevant information. We evaluate four context strategies across six frontier models on three document formats (Word, Excel, and PowerPoint). Our key finding: agentic tool-augmented approaches dramatically outperform passive strategies, with RAG+Tools achieving 46% accuracy vs 6% for RAG-only. Tool benefits are consistent across formats (+28-40 points) and models. We further show that (1) intelligent routing matters more than iteration count, (2) tools provide unique capability beyond reasoning loops, and (3) forcing active exploration matches providing context proactively. These results suggest tool augmentation is crucial for complex document QA.

[Industry] Grounded Multimodal In-Context Learning for Product Weight Estimation at Scale in E-commerce ► *Underline Video*

Bhavuk Singhal, Arsh Keshari and Ravindra Kumar Yadav

Accurately inferring implicit physical attributes of products, such as weight, is critical for large-scale e-commerce logistics but challenging due to sparse or unreliable textual metadata and high visual variability. We formulate weight estimation as a grounded multimodal reasoning problem and investigate whether large vision-language models (LVLMs) can infer discretized weight buckets through in-context learning (ICL) over product images and descriptions. We introduce a scalable inference framework that conditions predictions on automatically retrieved, category-specific exemplars and propose a distribution-calibrated retrieval strategy that aligns few-shot contexts with the empirical weight distribution of each product sub-category. This calibration substantially improves few-shot multimodal reasoning compared to random or embedding-based retrieval baselines. Across 14 high-variance categories, our approach significantly outperforms strong multimodal KNN baselines in both exact-match accuracy and near-bucket reliability. Deployed in production on a large e-commerce platform, our system processes millions of listings daily and reduces shipping-related revenue leakage by 22%, demonstrating that multimodal ICL can serve as a practical and cost-effective alternative to manual or hardware-based verification.

[Industry] Hierarchical Reason-of-Contact Detection in Retail Banking Customer Interactions via LLM-Driven Taxonomy Induction ► *Underline Video*

Yoages Kumar Mantri, Sriraja Yagneswaran, John Hartley, Greig Cowan, Raad Khraishi and Abhinav Gupta

Retail banks handle high volumes of customer interactions across different channels that span various topics. Early and accurate detection of the intent of the customer is critical towards streamlining contact-center operations through efficient routing and handling of conversations. Mining of customer interactions leads to identification of friction points in customer journeys and offers valuable insights about customer needs. Existing approaches to define customer intents or contact reasons remain fragmented, manually maintained across organizations and relying on knowledge of specific business processes. We propose a framework that develops a dynamic hierarchical Reason-of-Contact (RoC) taxonomy to cover customer topics across hundreds of business processes. We further demonstrate the implementation of this taxonomy to a robust solution that identifies intents for all customer conversations across different channels. Our deployed system supports real time use with a 150 to 300 ms turnaround per conversation. It achieves up to 10% improvement in F1 score over baseline approaches on a reference dataset. We also detail deployment considerations, including dynamic taxonomy updates, out-of-domain detection, and auditability. Finally, we present ablations and error analyses to characterize effectiveness.

[Industry] LLM as a Risk Manager: LLM Semantic Filtering for Lead-Lag Trading in Prediction Markets ► *Underline Video*

Sumin Kim, Minjae Kim, Jihoon Kwon, Yoon Kim, Oscar Levy, Alejandro Lopez-Lira, Yongjae Lee and Chanyeol Choi

Prediction markets provide a unique setting where event-level time series are directly tied to natural-language descriptions, yet discovering robust lead-lag relationships remains challenging due to spurious statistical correlations. We propose a hybrid two-stage causal screener to address this challenge: (i) a statistical stage that uses Granger causality to identify candidate leader-follower pairs from market-implied probability time series, and (ii) an LLM-based semantic stage that re-ranks these candidates by assessing whether the proposed direction admits a plausible economic transmission mechanism based on event descriptions. Because causal ground truth is unobserved, we evaluate the ranked pairs using a fixed, signal-triggered trading protocol that maps relationship quality into realized profit and loss (PnL). On Kalshi Economics markets, our hybrid approach consistently outperforms the statistical baseline. Across rolling evaluations, the win rate increases from 51.4% to 54.5%. Crucially, the average magnitude of losing trades decreases substantially from 649 USD to 347 USD. This reduction is driven by the LLM's ability to filter out statistically fragile links that are prone to large losses, rather than relying on rare gains. These improvements remain stable across different trading configurations, indicating that the gains are not driven by specific parameter choices. Overall, the results suggest that LLMs function as semantic risk managers on top of statistical discovery, prioritizing lead–lag relationships that generalize under changing market conditions.

[Industry] TingIS: Real-time Risk Event Discovery from Noisy Customer Incidents at Enterprise Scale ▶ *Underline Video*

Jun Wang, Ziyin Zhang, Rui Wang, Hang Yu, Peng Di and Rui Wang

Real-time detection and mitigation of technical anomalies are critical for large-scale cloud-native services, where even minutes of downtime can result in massive financial losses and diminished user trust. While customer incidents serve as a vital signal for discovering risks missed by monitoring, extracting actionable intelligence from this data remains challenging due to extreme noise, high throughput, and semantic complexity of diverse business lines. In this paper, we present TingIS, an end-to-end system designed for enterprise-grade incident discovery. At the core of TingIS is a multi-stage event unification engine that synergizes efficient indexing techniques with Large Language Models (LLMs) to make informed decisions on event merging, enabling the stable extraction of actionable incidents from just a handful of diverse user descriptions. This engine is complemented by a cascaded routing mechanism for precise business attribution and a multi-dimensional noise reduction pipeline that integrates domain knowledge, statistical patterns, and behavioral filtering. Deployed in a production environment handling a peak throughput of over 2,000 messages per minute and 300,000 messages per day, TingIS achieves a P90 alert latency of 3.5 minutes and a 95% discovery rate for high-priority incidents. Benchmarks constructed from real-world data demonstrate that TingIS significantly outperforms baseline methods in routing accuracy, clustering quality, and Signal-to-Noise Ratio.

[Industry] TrendPulse: A Simple yet Efficient Framework for Capturing Viral E-Commerce Spikes via LLM-Driven Contextualization ▶ *Underline Video*

Arin Jain, Devashish Gupta, Divay Jindal, Bhavuk Singhal, Vinit Rongata and Ravindra Kumar Yadav

Anticipating and capturing transient demand spikes is a critical challenge for e-commerce platforms, as reactive discovery mechanisms often fail to surface relevant products during rapid cultural or seasonal shifts. We propose TrendPulse, a three-stage framework that identifies regional search momentum, leverages Large Language Model (LLM) to transform spikes into semantic trends, and employs a cross-attention mechanism to provide personalized catalog recommendations. Our comprehensive ablation experiments and evaluations validate the impact of each architectural component, showing consistent improvements across multiple critical business metrics. TrendPulses effectiveness is further validated through online A/B experiments, where it drives measurable gains in both business metrics and overall user experience. Finally, we outlined the deployment strategy in detail, providing a reproducible blueprint that can be readily applied to similar industry-scale applications.

[Industry] VishBox v2: A Multi-Agent System for Real-time Adaptive Voice Phishing Simulation ▶ *Underline Video*

Sungmi Park, Daon Choi, YOONMO YANG, Hong Yunyi and Heedou Kim

Voice phishing is a multi-round social engineering attack in which strategy and victim psychology co-evolve, yet real transcripts are rarely accessible for systematic analysis. We present VishBox v2, a multi-agent architecture that generates structured phishing simulations grounded in crime-script procedures and persuasion principles. A Main Agent coordinates attacker, victim, and analysis agents with round-level feedback, web-based tactic mining, and emotion-driven vulnerability tracking. Across 181 cases (571 rounds), experiments validate procedural realism and reveal that stage-wise elaboration increases simulated success, whereas high-risk rounds concentrate pressure tactics rather than diversify them. The framework supports reproducible, round-level analysis for safer red-teaming and security training research.

Oral Session B: Information Extraction and Retrieval 2

14:00-15:30 (Harbor G)

[Long] AutoSchemaKG: Autonomous Knowledge Graph Construction through Dynamic Schema Induction from Web-Scale Corpora ▶ *Underline Video*

Jiaxin Bai, Wei Fan, Qi Hu, Qing Zong, Chunyang Li, Hong Ting Tsang, Hongyu Luo, Yauwai Yim, Haoyu Huang, Xiao Zhou, Feng QIN, Tianshi Zheng, Xi Peng, Xin YAO, Huiwen Yang, Leijie Wu, JI Yi, Gong Zhang, Renhai Chen and Yangqiu Song

We present AutoSchemaKG, a framework for fully autonomous knowledge graph construction that eliminates the need for predefined schemas. Our system leverages large language models to simultaneously extract knowledge triples and induce comprehensive schemas directly from text, modeling both entities and events while employing conceptualization to organize instances into semantic categories. Processing over 50 million documents, we construct ATLAS (Automated Triple Linking And Schema induction), a family of knowledge graphs with 900+ million nodes and 5.9 billion edges. This approach outperforms state-of-the-art baselines on multi-hop QA tasks and enhances LLM factuality. Notably, our schema induction achieves 92% semantic alignment with human-crafted schemas with zero manual intervention, demonstrating that billion-scale knowledge graphs with dynamically induced schemas can effectively complement parametric knowledge in large language models.

[Long] BrowseComp-Plus: A Fair and Disentangled Evaluation Benchmark for Deep Search Agents ▶ *Underline Video*

Zijian Chen, Xueguang Ma, Shengyao Zhuang, Ping Nie, Kai Zou, Sahel Sharifymoghaddam, Andrew Liu, Joshua Green, Kshama Patel, Ruoxi Meng, Mingyi Su, Yanxi Li, Haoran Hong, Xinyu Shi, Xuye Liu, Hosna Oyarhoseini, Nandan Thakur, Crystina Zhang, Luyu Gao, Wenhu Chen and Jimmy Lin

Deep search agents that combine large language models with retrieval tools excel at complex, multi-hop queries. Yet, existing benchmarks

such as BrowseComp rely on black-box web search APIs, facing key limitations. (1) Fairness: for agents, dynamic and opaque web APIs hinder reproducibility and fair comparisons across agents. (2) Disentanglement: for retrieval, the lack of a fixed document corpus makes it impossible to isolate retriever contributions from end-to-end search agent accuracy. We introduce BrowseComp-Plus, a benchmark derived from BrowseComp that employs a fixed, human-verified corpus, enabling controlled retrieval for deep search agents. BrowseComp-Plus clearly distinguishes agent performance: with a BM25 retriever, the open-source Search-R1 achieves 3.86% accuracy, while GPT-5 achieves 55.9%. Additionally, BrowseComp-Plus makes retrieval gains explicit: pairing GPT-5 with Qwen3-Embedding-8B retriever further improves accuracy to 70.1% while reducing search calls. Overall, BrowseComp-Plus provides a fair and disentangled testbed, advancing both deep search agent evaluation and retrieval research for agentic search. Code and data can be found at: https://texttron.github.io/BrowseComp-Plus/

[Long] FinKario: Event-Enhanced Automated Construction of Financial Knowledge Graph ► *Underline Video*

Xiang Li, Penglei Sun, Wanyun Zhou, Zikai Wei, Yongqi Zhang and Xiaowen Chu

Individual investors are significantly outnumbered and disadvantaged in financial markets, overwhelmed by abundant information and lacking professional analysis. Equity research reports stand out as crucial resources, offering valuable insights. By leveraging these reports, large language models (LLMs) can enhance investors' decision-making capabilities and strengthen financial analysis. However, two key challenges limit their effectiveness: (1) the rapid evolution of market events often outpaces the slow update cycles of existing knowledge bases, (2) the long-form and unstructured nature of financial reports further hinders timely and context-aware integration by LLMs. To address these challenges, we tackle both data and methodological aspects. First, we introduce the Event-Enhanced Automated Construction of Financial Knowledge Graph (FinKario), a dataset comprising over 305,360 entities, 210,328 relational triples, and 19 distinct relation types. FinKario automatically integrates real-time company fundamentals and market events through prompt-driven extraction guided by professional institutional templates, providing structured and accessible financial insights for LLMs. Additionally, we propose a Two-Stage, Graph-Based retrieval strategy (FinKario-RAG), optimizing the retrieval of evolving, large-scale financial knowledge to ensure efficient and precise data access. Extensive experiments show that FinKario with FinKario-RAG achieves superior stock trend prediction accuracy, outperforming financial LLMs by 18.81% and institutional strategies by 17.85% on average in backtesting.

[Long] From Past To Path: Masked History Learning for Next-Item Prediction in Generative Recommendation ► *Underline Video*

Kaiwen Wei, Kejun He, Xiaomian Kang, Jie Zhang, Yuming Yang, Li Jin, Zhenyang Li, Jiang Zhong, He Bai and Junnan Zhu

Generative recommendation, which directly generates item identifiers, has emerged as a promising paradigm for recommendation systems. However, this left-to-right paradigm inherently biases the model towards local contexts, failing to capture deeper historical dependencies necessary for understanding complex user intents. To address this limitation, we propose Masked History Learning (MHL), a novel training framework that shifts the objective from simple next-step prediction to deep comprehension of history. MHL augments the standard autoregressive objective with an auxiliary task of reconstructing masked historical items, compelling the model to understand "why" an item path is formed from the user's past behaviors, rather than just "what" item comes next. We introduce two key contributions to enhance this framework: (1) an entropy-guided masking policy that intelligently targets the most informative historical items for reconstruction, and (2) a curriculum learning scheduler that progressively transitions from history reconstruction to future prediction. Experiments on three public datasets show that our method significantly outperforms state-of-the-art generative models, highlighting that a comprehensive understanding of the past is crucial for accurately predicting a user's future path. The code is available at https://github.com/CQU-MM-Intelligent-Lab/MHL.

[Long] Hierarchical Token Prepending: Enhancing Information Flow in Decoder-based LLM Embeddings ► *Underline Video*

Xueying Ding, Xingyue Huang, Mingxuan Ju, Liam Collins, Yozen Liu, Leman Akoglu, Neil Shah and Tong Zhao

Large language models produce powerful text embeddings, but their causal attention mechanism restricts the flow of information from later to earlier tokens, degrading representation quality. While recent methods attempt to solve this by prepending a single summary token, they over-compress information, hence harming performance on long documents. We propose Hierarchical Token Prepending (HTP), a method that resolves two critical bottlenecks. To mitigate attention-level compression, HTP partitions the input into blocks and prepends block-level summary tokens to subsequent blocks, creating multiple pathways for backward information flow. To address readout-level over-squashing, we replace last-token pooling with mean-pooling, a choice supported by theoretical analysis. HTP achieves consistent performance gains across 11 retrieval datasets and 30 general embedding benchmarks, especially in long-context settings. As a simple, architecture-agnostic method, HTP effectively enhances both zero-shot and fine-tuned models, offering a scalable route to superior long-document embeddings.

[Long] HTMR: Hybrid Token Masking Reinforcement Learning with Verifiable Rewards for Event Argument Extraction with Multi-Perspective Reasoning ► *Underline Video*

Jianwen Luo, Yongkang Jin, Yu Hong and Jianmin YAO

Event Argument Extraction (EAE) aims to identify event arguments and assign semantic roles under a predefined schema. Recent work formulates EAE with large language models as a structured conditional generation task and applies Reinforcement Learning with Verifiable Rewards (RLVR) to optimize sequence-level event structures. However, RLVR-based EAE supervision is coarse-grained, as a single reward is assigned to the whole event structure, while optimization happens at the token level. This misalignment causes the same reward to be applied to all tokens, including those not related to event roles or arguments, introducing noise into the gradient updates and weakening the signals for decisions critical to argument extraction. To mitigate this misalignment, we propose Hybrid Token Masking RLVR (HTMR), which selectively updates policy gradients on both high-entropy forking tokens and event-critical tokens that define event structure, along with multi-perspective reasoning. Experiments across multiple benchmarks and models show that

HTMR consistently outperforms full-token and high-entropy only RLVR methods. Moreover, HTMR transfers effectively as a plug-and-play approach to other tasks such as named entity recognition and relation classification. The code is publicly available for reproducibility.

[Long] Making Large Language Models Efficient Dense Retrievers ► *Underline Video*

Yibin Lei, Shwai He, Ang Li and Andrew Yates

Recent work has shown that directly fine-tuning large language models (LLMs) for dense retrieval yields strong performance, but their substantial parameter counts make them computationally inefficient. While prior studies have revealed significant layer redundancy in LLMs for generative tasks, it remains unclear whether similar redundancy exists when these models are adapted for retrieval tasks, which require encoding entire sequences into fixed representations rather than generating tokens iteratively. To this end, we conduct a comprehensive analysis of layer redundancy in LLM-based dense retrievers. We find that, in contrast to generative settings, MLP layers are substantially more prunable, while attention layers remain critical for semantic aggregation. Building on this insight, we propose EffiR, a framework for developing efficient retrievers that performs large-scale MLP compression through a coarse-to-fine strategy (coarse-grained depth reduction followed by fine-grained width reduction), combined with retrieval-specific fine-tuning. Across diverse BEIR datasets and LLM backbones, EffiR achieves substantial reductions in model size and inference cost while preserving the performance of full-size models.

[Long] TLSA: LLM-Guided Text-Label Space Alignment with Contrastive Learning for Generalized Category Discovery ► *Underline Video*

Wenxi Xu, Chuan Qin, Xi Chen, Chuyu Fang, Yuanchun Zhou and Hengshu Zhu

Generalized Category Discovery (GCD) aims to classify data from partially labeled datasets by jointly recognizing known categories and discovering novel ones. Despite recent advances, existing methods still suffer from weak text-label alignment, inconsistent objectives across known and novel categories, and poor discrimination of semantically similar clusters. To mitigate these issues, we propose TLSA, a unified framework that enforces contrastive alignment between text and label representations within a shared semantic space. Specifically, we first design a label-semantic aware dual-encoder equipped with a symmetric contrastive objective to achieve text-label alignment. Then, we leverage LLM-based label induction to generate explicit and semantically meaningful names for previously unseen categories, followed by a graph-based refinement strategy that disambiguates semantically overlapping clusters through forced renaming. Finally, a confidence-aware sampling strategy ensures balanced learning across both easy and hard instances. Extensive experiments on four benchmark datasets show that TLSA consistently outperforms state-of-the-art GCD methods. The code is available at https://github.com/Wenxi-Xu/TLSA.

[Long] Zero-Shot Multimodal Retrieval with Multi-Scale Contextual Representations ► *Underline Video*

Sourajit Saha and Tejas Gokhale

In multimodal information retrieval (MMIR), candidates relevant to an input query need to be retrieved from a database, where the query and database items span different modalities. As real-world databases evolve, repeatedly annotating and indexing data and re-optimizing domain-specific models across modalities is impractical. We present MULTI-SCORE, a fine-tuning-free, two-stage MMIR approach that couples efficient candidate filtering with fine-grained multimodal re-ranking. Stage-1 adopts Matryoshka representations to efficiently filter out low-relevance candidates without expensive similarity computations on full-scale representations for the entire database. Stage-2 re-ranks the filtered candidates by computing their fine-grained multimodal contextual representations with two scoring functions for semantic alignment using chain-of-thought prompting and question-answering. Experiments demonstrate state-of-the-art zero-shot retrieval on 12 MMIR tasks across 32 datasets while outperforming supervised methods on 23 datasets.

Oral Session B: Language Models 2

14:00-15:30 (Harbor H-I)

[Long] Find Your Optimal Teacher: Personalized Data Synthesis via Router-Guided Multi-Teacher Distillation ► *Underline Video*

Hengyuan Zhang, Shiping Yang, Xiao Liang, Chenming Shang, Yuxuan Jiang, Chaofan Tao, Jing Xiong, Hayden Kwok-Hay So, Ruobing Xie, Angel X. Chang and Ngai Wong

Training student models on synthetic data generated by strong teacher models is a promising approach to distilling the capabilities of teachers. However, existing studies reveal that stronger models are not always optimal teachers, suggesting a mismatch between the teachers output and the students learning ability. To address this issue, we propose PerSyn (Personalized data Synthesis), a novel and efficient approach that customizes synthetic data to align with the learning capabilities of the student model. Specifically, our PerSyn method routes each prompt to its optimal teacher via a query-level router that jointly considers the student models' learnability and teacher models' response quality. It successfully transfers the synthesis paradigm from the conventional "Generate then Select" to a more efficient manner, i.e., "Route then Generate", eliminating the need for all teacher models to generate parallel responses across the entire prompt set. Extensive experiments across different model families and scales demonstrate that PerSyn consistently outperforms all baselines on six benchmarks, including instruct tuning and math reasoning settings. Further analysis verifies the effectiveness of PerSyn and offers extra insights to propel future research. Our code is available at https://anonymous.4open.science/r/PerSyn-8D85.

[Long] Glyph: Scaling Context Windows via Visual-Text Compression ▶ *Underline Video*

Jiale Cheng, Yusen Liu, Xinyu Zhang, Yulin Fei, Wenyi Hong, Ruiliang Lyu, Weihan Wang, Zhe Su, Xiaotao Gu, Xiao Liu, Yushi Bai, Jie Tan, Hongning Wang and Minlie Huang

Large language models (LLMs) conventionally represent text as sequences of discrete tokens, making long-context scaling largely a matter of processing more tokens more efficiently. We instead explore a complementary direction: increasing how much original context each token represents. To this end, we introduce Glyph, a framework that renders long texts into compact visual pages and processes them with a vision-language model (VLM), allowing a fixed context window to cover substantially more text. To make visual compression practical, Glyph combines continual pre-training on rendered long-text data, an LLM-driven genetic search to identify rendering configurations that balance compression and task performance, and post-training with supervised fine-tuning and reinforcement learning. Across multiple long-context benchmarks, Glyph achieves 3-4x token compression while maintaining performance comparable to strong text-only LLMs such as Qwen3-8B, with over 4x faster prefilling and decoding and 2x faster supervised fine-tuning. Under more aggressive compression, a VLM with a 128K context window can handle tasks that would otherwise require up to 1M input tokens. Our code and model are released at https://github.com/thu-coai/Glyph.

[Long] How Memory Management Impacts LLM Agents: An Empirical Study of Experience-Following Behavior ▶ *Underline Video*

Zidi Xiong, Yuping Lin, Wenya Xie, Pengfei He, Zirui Liu, Jiliang Tang, Himabindu Lakkaraju and Zhen Xiang

Memory is a critical component in large language model (LLM)-based agents, enabling them to store and retrieve past executions to improve task performance over time. In this paper, we conduct an empirical study on how memory management choices impact the LLM agents' behavior, especially their long-term performance. Specifically, we focus on two fundamental memory management operations that are widely used by many agent frameworks, memory addition and deletion, to systematically study their impact on the agent behavior. Through our quantitative analysis, we find that LLM agents display an experience-following property: high similarity between a task input and the input in a retrieved memory record often results in highly similar agent outputs. Our analysis further reveals two significant challenges associated with this property: error propagation, where inaccuracies in past experiences compound and degrade future performance, and misaligned experience replay, where some seemingly correct executions can provide limited or even misleading value as experiences. Through controlled experiments, we demonstrate the importance of regulating experience quality within the memory bank and show that future task evaluations can serve as free quality labels for stored memory. Our findings offer insights into the behavioral dynamics of LLM agent memory systems and provide practical guidance for designing memory components that support robust, long-term agent performance.

[Long] How to Improve LLMs' Performance on Specific Languages: A Perspective on LLM-Derived Language Similarity ▶ *Underline Video*

Xinhe Shi, Qingcheng Zeng, Weihao Xuan and Linchao Zhu

Large language models (LLMs) exhibit uneven performance across languages. In language-specific applications, practitioners often rely on target-language corpora or cross-lingual transfer to achieve better performance. However, traditional linguistic typology, commonly used as a transfer language selection strategy in previous studies, may not align with LLM's perception of language similarity. This work proposes LLM-based language similarity as a novel perspective for selecting effective fine-tuning languages. We construct a framework to quantify the similarity within each language pair through both the lenses of language-specific performance patterns and cross-lingual transferability, ultimately deriving three similarity score matrices. Moreover, we observe a counter-intuitive phenomenon: super-additive transfer effect, where fine-tuning on a certain language yields higher performance than fine-tuning directly on the target language. Additionally, due to the absence of an existing dataset meeting our experimental requirements, we construct and release M4CQ-Pro dataset, which features domain-diverse distribution of 135 tasks and content consistency across 31 languages (including over 20 medium- and low-resource languages), with 61518 manually reviewed high-quality questions per language. We evaluate our approach on representative multilingual LLMs and results show that all three LLM-based similarity measures effectively guide fine-tuning language selection, outperforming traditional linguistic similarity, with the integrated measure achieving the best results. Our approach provides not only a novel perspective on language similarity, but also practical baselines for selecting fine-tuning languages.

[Long] Learning Diverse Responses with Prefix-Conditioned Supervised Fine-Tuning ▶ *Underline Video*

Zhiyuan Fan, Guanqiao Chen, Yanyi HUANG, Mingkuan Zhao, Dadi Guo and Yi R. Fung

Large language models (LLMs) have shown strong performance on hard reasoning and general instruction-following tasks. However, when sampling multiple outputs for the same prompt, they often produce highly homogeneous, repetitive responses, resulting in inefficient exploration. This limits the gains from test-time scaling and constrains the upper bound of RL training. We attribute this issue in part to supervised fine-tuning (SFT): when a single prompt is paired with multiple reference responses, the model is trained to generate diverse outputs under the same prior condition, which induces optimization interference and can lead to diversity collapse. To address this, we propose Prefix-Conditioned SFT (P-SFT), a simple yet effective method that constructs semantically consistent yet distributionally distinct prior contents to different responses, thereby projecting the instruction into distinct latent regions to establish diverse prior distributions and decouple the one-to-many mapping. Experiments on large reasoning language models show that our approach improves absolute performance by 5.3% and increases generation diversity by 198.3% on average, while substantially enhancing output diversity and test-time scaling. Notably, even without any additional training, our prefixing strategy can be applied at inference time alone and still yields significant gains in both diversity and reasoning performance for instruction-tuned LLMs and reasoning-enhanced models.

[Long] Parameter Importance is Not Static: Evolving Parameter Isolation for Supervised Fine-Tuning ▶ *Underline Video*

Zekai Lin, Chao Xue, Di Liang, Xingsheng Han, Peiyang Liu, Xianjie Wu, Lei Jiang, Yu Lu, Haibo Shi, Shuang Liang and Minlong Peng

Supervised Fine-Tuning (SFT) of large language models often suffers from task interference and catastrophic forgetting. Recent approaches alleviate this issue by isolating task-critical parameters during training. However, these methods represent a static solution to a dynamic problem, assuming that parameter importance remains fixed once identified. In this work, we empirically demonstrate that parameter importance exhibits temporal drift over the course of training. To address this, we propose Evolving Parameter Isolation (EPI), a fine-tuning framework that adapts isolation decisions based on online estimates of parameter importance. Instead of freezing a fixed subset of parameters, EPI periodically updates isolation masks using gradient-based signals, enabling the model to protect emerging task-critical parameters while releasing outdated ones to recover plasticity. Experiments on diverse multi-task benchmarks demonstrate that EPI consistently reduces interference and forgetting compared to static isolation and standard fine-tuning, while improving overall generalization. Our analysis highlights the necessity of synchronizing isolation mechanisms with the evolving dynamics of learning diverse abilities.

[Long] SAGE: A Search-AuGmented Evaluation of Large Language Models on FreeForm QA ▶ *Underline Video*

Sher Badshah, Ali Emami and Hassan Sajjad

As Large Language Models (LLMs) become increasingly used for question-answering (QA), relying on static, pre-annotated references for evaluation poses significant challenges in cost, scalability, and completeness. Meanwhile, using LLMs themselves as evaluators without external grounding remains unreliable for objective tasks, as they systematically over-accept incorrect answers, fabricate supporting rationales, and degrade sharply on questions that fall outside their training data. We propose Search-AuGmented Evaluation (SAGE), a framework to assess LLM outputs without fixed ground-truth answers. Unlike conventional metrics that compare to static references or depend solely on LLM-as-a-judge knowledge, SAGE acts as an agent that actively retrieves and synthesizes external evidence. It iteratively generates web queries, collects information, summarizes findings, and refines subsequent searches through reflection. By reducing dependence on static reference-driven evaluation protocols, SAGE offers a scalable and adaptive alternative for evaluating the factuality of LLMs. Experimental results on multiple free-form QA benchmarks show that SAGE achieves substantial to perfect agreement with human evaluations.

[TACL] Simulating Hard Attention Using Soft Attention ▶ *Underline Video*

Andy Yang, Lena Strobl, David Chiang and Dana Angluin

We study conditions under which transformers using soft attention can simulate hard attention, that is, effectively focus all attention on a subset of positions. First, we examine several subclasses of languages recognized by hard-attention transformers, which can be defined in variants of linear temporal logic. We demonstrate how soft-attention transformers can compute formulas of these logics using unbounded positional embeddings or temperature scaling. Second, we demonstrate how temperature scaling allows softmax transformers to simulate general hard-attention transformers, using a temperature that depends on the minimum gap between the maximum attention scores and other attention scores.

[Long] Ted-Tok: Maintaining an Evolving Vocabulary for Lifelong Learning ▶ *Underline Video*

Jiameng Huang, Zhi Zhang, Zhenyu He, Jiacheng Sun and Di He

Lifelong learning investigates how models adapt when exposed to a potentially infinite stream of data. Most conventional approaches focus on updating model parameters (i.e., the neural network weights) as the underlying data distribution evolves over time. However, in natural language processing, model parameters are not the only components that matter. The tokenizer, a foundational part of the system, is usually assumed to remain fixed in lifelong learning scenarios. In this work, we challenge the validity of this assumption: as language evolves, a static tokenizer fragments newly emerging lexical items, reducing compression efficiency and consequently degrading the model performance. We introduce the Temporal Drift Tokenizer (Ted-Tok), which maintains an evolving vocabulary that adapts to emerging linguistic patterns over time. This adaptivity is driven by time-weighted frequency estimators that smooth short-term fluctuations to capture persistent linguistic trends, and a principled addition-deletion strategy targeting sink tokens. Across multiple domains, Ted-Tok consistently improves compression and task performance, with gains increasing under stronger drift, underscoring the role of tokenizer adaptivity in lifelong learning.

Oral Session B: Multilinguality and Language Diversity 2

14:00-15:30 (Harbor A)

[Long] "Newspaper Eat" Means "Not Tasty": A Taxonomy and Benchmark for Coded Language in Real-World Chinese Online Reviews ▶ *Underline Video*

Ruyuan Wan, Changye Li and Ting-Hao 'Kenneth' Huang

Coded language is an important part of human communication. It refers to cases where users intentionally encode meaning so that the surface text differs from the intended meaning and must be decoded to be understood. Current language models handle coded language poorly. Progress has been limited by the lack of real-world datasets and clear taxonomies. This paper introduces CodedLang, a dataset of

7,744 Chinese Google Maps reviews, including 900 reviews with span-level annotations of coded language. We developed a seven-class taxonomy that captures common encoding strategies, including phonetic, orthographic, and cross-lingual substitutions. We benchmarked language models on coded language detection, classification, and review rating prediction. Results show that even strong models can fail to identify or understand coded language. Because many coded expressions rely on pronunciation-based strategies, we further conducted a phonetic analysis of coded and decoded forms. Our code and dataset are publicly available. Together, our results highlight coded language as an important and underexplored challenge for real-world NLP systems.

[Long] AfriqueLLM: How Data Mixing and Model Architecture Impact Continued Pre-training for African Languages ▶ *Underline Video*

Hao Yu, Tianyi Xu, Michael A. Hedderich, Wassim Hamidouche, Syed Waqas Zamir and David Ifeoluwa Adelani

Large language models (LLMs) are increasingly multilingual, yet open models continue to underperform relative to proprietary systems, with the gap most pronounced for African languages. Continued pre-training (CPT) offers a practical route to language adaptation, but improvements on demanding capabilities such as mathematical reasoning often remain limited. This limitation is driven in part by the uneven domain coverage and missing task-relevant knowledge that characterize many low-resource language corpora. We present AfriqueLLM, a suite of open LLMs adapted to 20 African languages through CPT on 26B tokens. We perform a comprehensive empirical study across five base models spanning sizes and architectures, including Llama 3.1, Gemma 3, and Qwen 3, and systematically analyze how CPT data composition shapes downstream performance. In particular, we vary mixtures that include math, code, and synthetic translated data, and evaluate the resulting models on a range of multilingual benchmarks. Our results identify data composition as the primary driver of CPT gains. Adding math, code, and synthetic translated data yields consistent improvements, including on reasoning-oriented evaluations. Within a fixed architecture, larger models typically improve performance, but architectural choices dominate scale when comparing across model families. Moreover, strong multilingual performance in the base model does not reliably predict post-CPT outcomes; robust architectures coupled with task-aligned data provide a more dependable recipe. Finally, our best models improve long-context performance, including document-level translation.

[Long] Beyond Monolingual Assumptions: A Survey on Code-Switched NLP in the Era of Large Language Models across Modalities ▶ *Underline Video*

Rajvee Sheth, Samridhi Raj Sinha, Mahavir Patil, Himanshu Beniwal and Mayank Singh

Amidst the rapid advances of large language models (LLMs), most LLMs still struggle with mixed-language inputs, limited Code-switching (CSW) datasets, and evaluation biases, which hinder their deployment in multilingual societies. This survey provides the first comprehensive analysis of CSW-aware LLM research, reviewing 327 studies spanning five research areas, 15+ NLP tasks, 30+ datasets, and 80+ languages. We classify recent advances by architecture, training strategy, and evaluation methodology, outlining how LLMs have reshaped CSW modelling and what challenges persist. The paper concludes with a roadmap emphasizing the need for inclusive datasets, fair evaluation, and linguistically grounded models to achieve truly multilingual intelligence. A curated collection of all resources is maintained at https://github.com/lingo-iitgn/awesome-code-mixing/.

[Long] EMCEE: Improving Multilingual Capability of LLMs via Bridging Knowledge and Reasoning with Extracted Synthetic Multilingual Context ▶ *Underline Video*

Hamin Koo and Jaehyung Kim

Large Language Models (LLMs) have achieved impressive progress across a wide range of tasks, yet their heavy reliance on English-centric training data leads to significant performance degradation in non-English languages. While existing multilingual prompting methods emphasize reformulating queries into English or enhancing reasoning capabilities, they often fail to incorporate the language- and culture-specific grounding that is essential for some queries. To address this limitation, we propose EMCEE (Extracting synthetic Multilingual Context and merging), a simple yet effective framework that enhances the multilingual capabilities of LLMs by explicitly extracting and utilizing query-relevant knowledge from the LLM itself. In particular, EMCEE first extracts synthetic context to uncover latent, language-specific knowledge encoded within the LLM, and then dynamically merges this contextual insight with reasoning-oriented outputs through a judgment-based selection mechanism. Extensive experiments on four multilingual benchmarks covering diverse languages and tasks demonstrate that EMCEE consistently outperforms prior approaches, achieving an average relative improvement of 16.4% overall and 31.7% in low-resource languages.

[Long] MultiFinBen: Benchmarking Large Language Models for Multilingual and Multimodal Financial Application ▶ *Underline Video*

Xueqing Peng, Lingfei Qian, Yan Wang, Ruoyu Xiang, Yueru He, Yang Ren, Mingyang Jiang, Vincent Jim Zhang, Yuqing Guo, Jeff Zhao, Huan He, Yi Han, Yun Feng, Yuechen Jiang, Yupeng Cao, Haohang Li, Yangyang Yu, Xiaoyu Wang, Penglei Gao, Shengyuan Lin, Keyi Wang, Shanshan Yang, Yilun Zhao, Zhiwei Liu, Peng Lu, Jerry Huang, Suyuchen Wang, Triantafillos Papadopoulos, Polydoros Giannouris, Efstathia Soufleri, Nuo Chen, Zhiyang Deng, Heming Fu, Yijia Zhao, MINGQUAN LIN, Meikang Qiu, Kaleb E Smith, Arman Cohan, Xiao-Yang Liu, Jimin Huang, GUOJUN XIONG, Alejandro Lopez-Lira, Xi Chen, Junichi Tsujii, Jian-Yun Nie, Sophia Ananiadou and Qianqian Xie

Real-world financial analysis involves information across multiple languages and modalities, from reports and news to scanned filings and meeting recordings. Yet most existing evaluations of LLMs in finance remain text-only, monolingual, and largely saturated by current models. To bridge these gaps, we present MultiFinBen, the first expert-annotated multilingual (five languages) and multimodal (text, vision, audio) benchmark for evaluating LLMs in realistic financial contexts. MultiFinBen introduces two new task families: multilingual financial

reasoning, which tests cross-lingual evidence integration from filings and news, and financial OCR, which extracts structured text from scanned documents containing tables and charts. Rather than aggregating all available datasets, we apply a structured, difficulty-aware selection based on advanced model performance, ensuring balanced challenge and removing redundant tasks. Evaluating 21 leading LLMs shows that even frontier multimodal models like GPT-4o achieve only 46.01% overall, stronger on vision and audio but dropping sharply in multilingual settings. These findings expose persistent limitations in multilingual, multimodal, and expert-level financial reasoning. All datasets, evaluation scripts, and leaderboards are publicly released.

[Long] MUTANT: A Recipe for Multilingual Tokenizer Design ▶ *Underline Video*
Souvik Rana, Arul Menezes, Ashish Kulkarni, Chandra Khatri and Shubham Agarwal
Tokenizers play a crucial role in determining the performance, training efficiency, and the inference cost of Large Language Models (LLMs). Designing effective tokenizers for multilingual LLMs is particularly challenging due to diverse scripts and rich morphological variation. While subword methods like Byte Pair Encoding (BPE) are widely adopted, their effectiveness in multilingual settings remains underexplored. We present MUTANT, a recipe for building multilingual tokenizers, with careful vocabulary and training data design, language-aware pre-tokenization, and subword and multiword aware training. We also introduce MUTANT-Indic, a tokenizer for India-specific multilingual LLMs, that produces linguistically coherent tokens and achieves state-of-the-art performance. Evaluated across English, Indian languages and code data, our tokenizer improves the average fertility score by 39.5% over LLaMA4 and by 18% over Sutra (the current best). This translates to 44% improvement in inference throughput over LLaMA4 while maintaining comparable performance on English and Indic benchmarks. We present detailed ablations across tokenizer training data size, vocabulary size, merging techniques, and pre-tokenization strategies, demonstrating the robustness of our design choices.

[Long] Standard-to-Dialect Transfer Trends Differ across Text and Speech: A Case Study on Intent and Topic Classification in German Dialects ▶ *Underline Video*
Verena Blaschke, Miriam Winkler and Barbara Plank
Research on cross-dialectal transfer from a standard to a non-standard dialect variety has typically focused on text data. However, dialects are primarily spoken, and non-standard spellings cause issues in text processing. We compare standard-to-dialect transfer in three settings: text models, speech models, and cascaded systems where speech first gets automatically transcribed and then further processed by a text model. We focus on German dialects in the context of written and spoken intent classification - releasing the first dialectal audio intent classification dataset - with supporting experiments on topic classification. The speech-only setup provides the best results on the dialect data while the text-only setup works best on the standard data. While the cascaded systems lag behind the text-only models for German, they perform relatively well on the dialectal data if the transcription system generates normalized, standard-like output.

VFA: Empoweing Multilingual MLLMs via Vision-Free Adaptation ▶ *Underline Video*
Yixia Li, Yaqing Shi, Zhiwen Ruan, Dongdong Zhang, Lingjie Jiang, Shaohan Huang, Yun Chen, Guanhua Chen and Furu Wei
Multimodal large language models have advanced rapidly, yet most remain English-centric, as scaling multilingual multimodal instruction tuning is limited by the scarcity and high cost of high-quality non-English image-text supervision. Although multilingual text data is abundant, naive textual fine-tuning can disrupt vision-language alignment and induce catastrophic forgetting. We propose Vision-Free Adaptation (VFA), a framework that decouples multilingual language enhancement from visual alignment by composing complementary task vectors over a shared LLM backbone. Specifically, we fine-tune a base LLM on multilingual text data to derive a multilingual task vector, which is then merged with the vision-aligned task vector of an MLLM. Experiments on five MLLMs across six multilingual multimodal benchmarks show consistent improvements while preserving both general multimodal and text-only capabilities. Moreover, VFA attains competitive performance with a fully multimodally trained model using less than 2% of the text data, demonstrating its efficiency and effectiveness.

[Long] Where the Cat Sat: A Multilingual Framework for Spatial Language Understanding ▶ *Underline Video*
Demian Inostroza, Ekaterina Vylomova, Charles Kemp, Mae Carroll, Wanchun Li and Meladel Mistica
Spatial language understanding is fundamental to tasks from robot navigation to document analysis, yet current work exhibits biases toward English and prepositional marking. We present a multilingual framework and benchmark decomposing spatial relations into surface elements (figure, ground, predicate, markers) and semantic components (dynamicity, stasis). Evaluating frontier LLMs on Spanish, Basque, and Chinese with text-only input, we find high accuracy on figure and ground identification but persistent gaps in two areas: semantic classification of topological and projective relations, and surface identification of morphological spatial markers, Basque case affixes proving most challenging at as low as 15.3%. These results suggest that surface parsing does not entail spatial understanding, and that evaluation must include typologically diverse spatial marking strategies.

Oral Session B: Resources and Evaluation 2

14:00-15:30 (Harbor B-C)

[CL] BAP v2: An Enhanced Task Framework for Instruction Following in Minecraft Dialogues ▶ *Underline Video*

Prashant Jayannavar, Liliang Ren, Marisa Hudspeth, Risham Sidhu, Charlotte Lambert, Ariel Cordes, Elizabeth Kaplan, Anjali Narayan-Chen and Julia Hockenmaier;
Developing interactive agents that can understand language, perceive their surroundings, and act within the physical world is a long-standing goal of AI research. The Minecraft Collaborative Building Task (MCBT) (Narayan-Chen, Jayannavar, and Hockenmaier 2019), a two-player game in which an Architect (A) instructs a Builder (B) to construct a target structure in a simulated 3D Blocks World environment, offers a rich platform to work towards this goal.; In this work, we focus on the Builder Action Prediction (BAP) subtask: predicting B's actions in a multimodal game context (Jayannavar, Narayan-Chen, and Hockenmaier 2020) - a challenging testbed for grounded instruction following, with limited training data. We holistically re-examine this task and introduce BAP v2 to address key challenges in evaluation, training data, and modeling. Specifically, we define an enhanced evaluation benchmark, featuring a cleaner test set and fairer, more insightful metrics that also reveal spatial reasoning as the primary performance bottleneck. To address data scarcity and to teach models basic spatial skills, we generate different types of synthetic MCBT data. We observe that current, LLM-based SOTA models trained on the human BAP dialogues fail on these simpler, synthetic BAP ones, but show that training models on this synthetic data improves their performance across the board. We also introduce a new SOTA model, Llama-CRAFTS, which leverages richer input representations, and achieves an F1 score of 53.0 on the BAP v2 task and strong performance on the synthetic data. While this result marks a notable 6 points improvement over previous work, it also underscores the task's remaining difficulty, establishing BAP v2 as a fertile ground for future research, and providing a useful measure of the spatial capabilities of current text-only LLMs in such embodied tasks.;

[TACL] Benchmarking Linguistic Diversity of Large Language Models ▶ *Underline Video*
Yanzhu Guo, Guokan Shang and Chloé Clavel
The development and evaluation of Large Language Models (LLMs) has primarily focused on their task-solving capabilities, with recent models even surpassing human performance in some areas. However, this focus often neglects whether machine-generated language matches the human level of diversity, in terms of vocabulary choice, syntactic construction, and expression of meaning, raising questions about whether the fundamentals of language generation have been fully addressed. This paper emphasizes the importance of examining the preservation of human linguistic richness by language models, given the concerning surge in online content produced or aided by LLMs. We propose a comprehensive framework for evaluating LLMs from various linguistic diversity perspectives including lexical, syntactic, and semantic dimensions. Using this framework, we benchmark several state-of-the-art LLMs across all diversity dimensions, and conduct an in-depth case study for syntactic diversity. Finally, we analyze how different development and deployment choices impact the linguistic diversity of LLM outputs.

[Long] DR-Arena: an Automated Evaluation Framework for Deep Research Agents ▶ *Underline Video*
Yiwen Gao, Ruochen Zhao, Yang Deng and Wenxuan Zhang
As Large Language Models (LLMs) increasingly operate as Deep Research (DR) Agents capable of autonomous investigation and information synthesis, reliable evaluation of their task performance has become a critical bottleneck. Current benchmarks predominantly rely on static datasets, which suffer from several limitations: limited task generality, temporal misalignment, and data contamination. To address these, we introduce DR-Arena, a fully automated evaluation framework that pushes DR agents to their capability limits through dynamic investigation. DR-Arena constructs real-time Information Trees from fresh web trends to ensure the evaluation rubric is synchronized with the live world state, and employs an automated Examiner to generate structured tasks testing two orthogonal capabilities: Deep reasoning and Wide coverage. DR-Arena further adopts Adaptive Evolvement Loop, a state-machine controller that dynamically escalates task complexity based on real-time performance, demanding deeper deduction or wider aggregation until a decisive capability boundary emerges. Experiments with six advanced DR agents demonstrate that DR-Arena achieves a Spearman correlation of 0.94 with the LMSYS Search Arena leaderboard. This represents state-of-the-art alignment with human preferences without any manual efforts, validating DR-Arena as a reliable alternative for costly human adjudication.

[Long] Illusions of the Gold Standard: A Large-scale Analysis of Human Evaluation Protocols for Long-form Text Generation ▶ *Underline Video*
Katelyn Xiaoying Mei, Yi-Li Hsu, Minjoon Choi, Zongwan Cao, Chenjun Xu, Bingbing Wen, Su Lin Blodgett and Lucy Lu Wang
Human evaluation plays a critical role in assessing the quality of generated text. However, the reliability and reproducibility of these evaluations depend on transparent and well-documented protocols—details that are frequently missing in current practice. In this work, we conduct a large-scale analysis of human evaluation protocols for evaluating long-form generation tasks in *CL conference publications from 2023–2025, including a full manual review of 284 papers and LLM-assisted analysis for another 1.8k+ papers. We define a set of 20 reportable criteria related to reproducibility of human evaluation studies, and apply these criteria to systematically examine reporting norms and practices within the community. We find widespread under-reporting of important aspects of human evaluation study design, leading to ambiguity about what was measured and how, who contributed judgments, and how judgments should be interpreted. Based on these findings, we outline actionable recommendations to support more transparent and reproducible reporting in future research.

[Long] MediEval: A Unified Medical Benchmark for Patient-Contextual and Knowledge-Grounded Reasoning in LLMs ▶ *Underline Video*
Zhan Qu and Michael Färber
Large Language Models (LLMs) are increasingly applied to medicine, yet their adoption is limited by concerns over reliability and safety.

Existing evaluations either test factual medical knowledge in isolation or assess patient-level reasoning without verifying correctness, leaving a critical gap. We introduce MediEval, a benchmark that links MIMIC-IV electronic health records (EHRs) to a unified knowledge base built from UMLS and other biomedical vocabularies. MediEval generates diverse factual and counterfactual medical statements within real patient contexts, enabling systematic evaluation across a 4-quadrant framework that jointly considers knowledge grounding and contextual consistency. Using this framework, we identify critical failure modes, including hallucinated support and truth inversion, that current proprietary, open-source, and domain-specific LLMs frequently exhibit. To address these risks, we propose Counterfactual Risk-Aware Fine-tuning (CoRFu), a DPO-based method with an asymmetric penalty targeting unsafe confusions. CoRFu improves by +16.4 macro-F1 points over the base model and eliminates truth inversion errors, demonstrating both higher accuracy and substantially greater safety.

[Long] PEAP: Proactive Embodied Action Sequence Planning with Joint Understanding of Vision and Audio Perception ► *Underline Video*

Tianwei Lan, Jiaqi Wu, Zeming Liu, Zhaoxin Fan, Haifeng Wang and Yuhang Guo

Embodied Action Sequence Planning focuses on the capability of embodied agents to implement action planning via environmental perception. This technology enables diverse intelligent assistance for real-world scenarios such as home and office environments. To address the limitations of existing embodied agents in meeting the requirement for proactivity and achieving joint understanding of visual and audio information, this study investigates the ability of embodied agents to proactively provide assistance through action sequence planning based on joint understanding of vision and audio perception without explicit human instructions. Correspondingly, we propose PEAP, the first multimodal proactive embodied action sequence planning dataset. We evaluate the performance of multiple Large Language Models on the PEAP dataset. The results demonstrate that these models still exhibit significant deficiencies on this task particularly lacking accurate environmental perception capabilities. Furthermore, ablation experiment and replacement experiment further corroborate that the joint understanding of multimodal information can significantly improve the models' performance on proactive embodied action sequence planning task.

[Long] RPC-Bench: A Fine-grained Benchmark for Research Paper Comprehension ► *Underline Video*

Yelin Chen, Fanjin Zhang, Suping Sun, Yunhe Pang, Yuanchun Wang, Jian Song, Xiaoyan Li, Lei Hou, Shu Zhao, Jie Tang and Juanzi Li

Understanding research papers remains challenging for foundation models due to specialized scientific discourse and complex figures and tables, yet existing benchmarks offer limited fine-grained evaluation at scale. To address this gap, we introduce RPC-Bench, a large-scale question-answering benchmark built from review-rebuttal exchanges of high-quality computer science papers, containing 15K human-verified QA pairs. We design a fine-grained taxonomy aligned with the scientific research flow to assess models ability to understand and answer why, what, and how questions in scholarly contexts. We also define an elaborate LLM-human interaction annotation framework to support large-scale labeling and quality control. Following the LLM-as-a-Judge paradigm, we develop a scalable framework that evaluates models on correctness-completeness and conciseness, with high agreement to human judgment. Experiments reveal that even the strongest models (GPT-5) achieve only 68.2% correctness-completeness, dropping to 37.46% after conciseness adjustment, highlighting substantial gaps in precise academic paper understanding.

[Long] SEA-BED: How Do Embedding Models Represent Southeast Asian Languages? ► *Underline Video*

Wuttikorn Ponwitayarat, Peerat Limkonchotiwat, Raymond Ng, Jann Railey Montalan, Thura Aung, Jian Gang Ngui, Yosephine Susanto, William Chandra Tjhi, Panuthep Tasawong, Erik Cambria, Ekapol Chuangsuwanich and Sarana Nutanong

Multilingual text embeddings are often assumed to encode meaning in a perspective-independent semantic space, yielding stable similarity judgments across tasks and languages. Our results show that this assumption does not hold in practice. We introduce SEA-BED, a large-scale benchmark covering 10 Southeast Asian (SEA) languages and diverse embedding tasks, designed to systematically examine how embedding performance varies across tasks, languages, and language-task combinations. Across extensive evaluations, we observe that no single model performs uniformly well across SEA languages; task difficulty differs markedly within languages, and success on one task does not reliably generalize to others. Language-task analyses further reveal highly non-uniform performance landscapes, where performance varies across different language-task combinations. These findings call for closer attention to performance measurements that provide an expansive view across languages and tasks to uncover inconsistencies in semantic representation. Based on these observations, we provide insights for future model development, including data, algorithmic, and architectural considerations.

[Long] ViDoRe V3: A Comprehensive Evaluation of Retrieval Augmented Generation in Complex Real-World Scenarios ► *Underline Video*

António Loison, Quentin Macé, Antoine Edy, Victor Xing, Tom Balough, Gabriel de Souza P. Moreira, Bo Liu, Manuel Faysse, Céline Hudelot and Gautier Viaud

Retrieval-Augmented Generation (RAG) pipelines must address challenges beyond simple single-document retrieval, such as interpreting visual elements (tables, charts, images), synthesizing information across documents, and providing accurate source grounding. Existing benchmarks fail to capture this complexity, often focusing on textual data, single-document comprehension, or evaluating retrieval and generation in isolation. We introduce ViDoRe V3, a comprehensive multimodal RAG benchmark featuring multi-type queries over visually rich document corpora. It covers 10 datasets across diverse professional domains, comprising ~26,000 document pages paired with 3,099 human-verified queries, each available in 6 languages. Through 12,000 hours of human annotation effort, we provide high-quality annotations for retrieval relevance, bounding box localization, and verified reference answers. Our evaluation of state-of-the-art RAG pipelines reveals

that visual retrievers outperform textual ones, late-interaction models and textual reranking substantially improve performance, and hybrid or purely visual contexts enhance answer generation quality. However, current models still struggle with non-textual elements, open-ended queries, and fine-grained visual grounding. To encourage progress in addressing these challenges, the benchmark is released under a commercially permissive license.

Oral Session B: Safety and Alignment in LLMs 1 (Harbor D)

14:00-15:30 (Harbor D)

[Long] Among Us: Measuring and Mitigating Malicious Contributions in Model Collaboration Systems ▶ *Underline Video*

Ziyuan Yang, Wenxuan Ding, Shangbin Feng and Yulia Tsvetkov

Language models (LMs) are increasingly used in collaboration: multiple LMs trained by different parties collaborate through routing systems, multi-agent debate, model merging, and more. Critical safety risks remain in this decentralized paradigm: what if some of the models in multi-LLM systems are compromised or malicious? We first quantify the impact of malicious models by engineering four categories of malicious LMs, plug them into four types of popular model collaboration systems, and evaluate the compromised system across 10 datasets. We find that malicious models have a severe impact on the multi-LLM systems, especially for reasoning and safety domains where performance is lowered by 7.12% and 7.94% on average. We then propose mitigation strategies to alleviate the impact of malicious components, by employing external supervisors that oversee model collaboration to disable/mask them out to reduce their influence. On average, these strategies recover 95.31% of the initial performance, while making model collaboration systems fully resistant to malicious models remains an open research question.

[Long] GAMBIT: A Gamified Jailbreak Framework for Multimodal Large Language Models ▶ *Underline Video*

Xiangdong Hu, Yangyang Jiang, Qin Hu and Xiaojun Jia

Multimodal Large Language Models (MLLMs) have become widely deployed, yet their safety alignment remains fragile under adversarial inputs. Previous work has shown that increasing inference steps can disrupt safety mechanisms and lead MLLMs to generate attacker-desired harmful content. However, most existing attacks focus on increasing the complexity of the modified visual task itself and do not explicitly leverage the model's own reasoning incentives. This leads to them underperforming on reasoning models (Models with Chain-of-Thoughts) compared to non-reasoning ones (Models without Chain-of-Thoughts). If a model can think like a human, can we influence its cognitive-stage decisions so that it proactively completes a jailbreak? To validate this idea, we propose GAMBIT (Gamified Adversarial Multimodal Breakout via Instructional Traps), a novel multimodal jailbreak framework that decomposes and reassembles harmful visual semantics, then constructs a gamified scene that drives the model to explore, reconstruct intent, and answer as part of winning the game. The resulting structured reasoning chain increases task complexity in both vision and text, positioning the model as a participant whose goal pursuit reduces safety attention and induces it to answer the reconstructed malicious query. Extensive experiments on popular reasoning and non-reasoning MLLMs demonstrate that GAMBIT achieves high Attack Success Rates (ASR), reaching 92.13% on Gemini 2.5 Flash, 91.20% on QvQ-MAX, and 85.87% on GPT-4o, significantly outperforming baselines. Warning: This paper contains unsafe and offensive examples.

[Long] Logical Consistency as a Bridge: Improving LLM Hallucination Detection via Label Constraint Modeling between Responses and Self-Judgments ▶ *Underline Video*

Hao Mi, Qiang Sheng, Shaofei Wang, Beizhe Hu, Yifan Sun, Zhengjia Wang, Hengqi Zeng, Yang Li, Danding Wang and Juan Cao

Large Language Models (LLMs) are prone to factual hallucinations, risking their reliability in real-world applications. Existing hallucination detectors mainly extract micro-level intrinsic patterns for uncertainty quantification or elicit macro-level self-judgments through verbalized prompts. However, these methods address only a single facet of the hallucination, focusing either on implicit neural uncertainty or explicit symbolic reasoning, thereby treating these inherently coupled behaviors in isolation and failing to exploit their interdependence for a holistic view. In this paper, we propose LaaB (Logical Consistency-as-a-Bridge), a framework that bridges neural features and symbolic judgments for hallucination detection. LaaB introduces a "meta-judgment" process to map symbolic labels back into the feature space. By leveraging the inherent logical bridge where response and meta-judgment labels are either the same or opposite based on the self-judgment's semantics, LaaB aligns and integrates dual-view signals via mutual learning and enhances the hallucination detection. Extensive experiments on 4 public datasets, across 4 LLMs, against 8 baselines demonstrate the superiority of LaaB.

[Long] Lying with Truths: Open-Channel Multi-Agent Collusion for Belief Manipulation via Generative Montage ▶ *Underline Video*

Jinwei Hu, Xinmiao Huang, Youcheng Sun, Yi Dong and Xiaowei Huang

As large language models (LLMs) transition to autonomous agents synthesizing real-time information, their reasoning capabilities introduce an unexpected attack surface. This paper introduces a novel threat where colluding agents steer victim beliefs using only truthful evidence fragments distributed through public channels, without relying on covert communications, backdoors, or falsified documents. By exploiting LLMs' overthinking tendency, we formalize the first cognitive collusion attack and propose Generative Montage: a Writer-Editor-Director framework that constructs deceptive narratives through adversarial debate and coordinated posting of evidence fragments, causing victims

to internalize and propagate fabricated conclusions. To study this risk, we develop CoPHEME, a dataset derived from real-world rumor events, and simulate attacks across diverse LLM families. Our results show pervasive vulnerability across 14 LLM families: attack success rates reach 74.4% for proprietary models and 70.6% for open-weights models. Counterintuitively, stronger reasoning capabilities increase susceptibility, with reasoning-specialized models showing higher attack success than base models or prompts. Furthermore, these false beliefs then cascade to downstream judges, achieving over 60% deception rates, highlighting a socio-technical vulnerability in how LLM-based agents interact with dynamic information environments.

[Long] Maximizing Local Entropy Where It Matters: Prefix-Aware Localized LLM Unlearning ▶ ***Underline Video***

Naixin Zhai, Pengyang Shao, Binbin Zheng, Yonghui Yang, Fei Shen, Long Bai and Xun Yang

Machine unlearning aims to forget sensitive knowledge from Large Language Models (LLMs) while maintaining general utility. However, existing approaches typically treat all tokens in a response indiscriminately and enforce uncertainty over the entire vocabulary. This global treatment results in unnecessary utility degradation and extends optimization to content-agnostic regions. To address these limitations, we propose PALU (Prefix-Aware Localized Unlearning), a framework driven by a local entropy maximization objective across both temporal and vocabulary dimensions. PALU reveals that (i) suppressing the sensitive prefix alone is sufficient to sever the causal generation link, and (ii) flattening only the top-K logits is adequate to maximize uncertainty in the critical subspace. These findings allow PALU to alleviate redundant optimization across the full vocabulary and parameter space while minimizing collateral damage to general model performance. Comprehensive evaluations validate that PALU achieves superior forgetting efficacy and utility preservation compared to state-of-the-art baselines. Our code is available at https://github.com/nxZhai/PALU.

[Long] Merging Triggers, Breaking Backdoors: Defensive Poisoning for Instruction-Tuned Language Models ▶ ***Underline Video***

San Kim and Gary Lee

Large Language Models (LLMs) have greatly advanced Natural Language Processing (NLP), particularly through instruction tuning, which enables broad task generalization without additional fine-tuning. However, their reliance on large-scale datasets, often collected from human or web sources, makes them vulnerable to backdoor attacks, where adversaries poison a small subset of data to implant hidden behaviors. Despite this growing risk, defenses for instruction-tuned models remain underexplored. We propose MB-Defense (Merging & Breaking Defense Framework), a novel training pipeline that immunizes instruction-tuned LLMs against diverse backdoor threats. MB-Defense comprises two stages: (i) Defensive Poisoning, which merges attacker and defensive triggers into a unified backdoor representation, and (ii) Backdoor Neutralization, which breaks this representation through additional training to restore clean behavior. Extensive experiments across multiple LLMs show that MB-Defense substantially lowers attack success rates while preserving instruction-following ability. Our method offers a generalizable and data-efficient defense strategy, improving the robustness of instruction-tuned LLMs against unseen backdoor attacks.

[TACL] Universal Jailbreak Suffixes Are Strong Attention Hijackers ▶ ***Underline Video***

Matan Ben-Tov, Mor Geva and Mahmood Sharif

We study suffix-based jailbreaks, a powerful family of attacks against large language models (LLMs) that optimize adversarial suffixes to circumvent safety alignment. Specifically, we focus on the widely used foundational GCG attack (Zou et al., 2023), observing that its suffixes vary in their efficacy, with some suffixes being more universal (i.e., generalize to many, unseen, harmful instructions) than others. We first show that GCGs effectiveness is driven by a shallow, critical mechanism, built on the information flow from the adversarial suffix to the final chat template tokens before generation. Quantifying the dominance of this mechanism during generation, we find GCG irregularly and aggressively hijacks the contextualization process. Crucially, we tie hijacking to the universality phenomenon, with more universal suffixes being stronger hijackers. Subsequently, we show that these insights have practical implications: GCG universality can be efficiently enhanced (up to x5 in some cases) at no additional computational cost, and can also be surgically mitigated, at least halving attack success with minimal utility loss.

[Long] Visual Self-Fulfilling Alignment: Shaping Safety-Oriented Personas via Threat-Related Images ▶ ***Underline Video***

Qishun Yang, Shu Yang, Lijie Hu and Di Wang

Multimodal large language models (MLLMs) face safety misalignment where visual inputs enable harmful outputs. Existing methods require explicit safety labels or contrastive data, yet threat-related concepts are concrete and visually depictable, while safety concepts like helpfulness are abstract and lack visual referents. Inspired by self-fulfilling mechanism underlying emergent misalignment, we propose Visual Self-Fulfilling Alignment (VSFA). VSFA fine-tunes vision-language models (VLMs) on neutral VQA tasks constructed around threat-related images, without any safety labels. Through repeated exposure to threat-related visual content, models internalize implicit semantics of vigilance and caution, shaping safety-oriented personas. Experiments across multiple VLMs and safety benchmarks demonstrate that VSFA reduces attack success rate, improves response quality, and mitigates over-refusal while preserving general capabilities. Our work extends self-fulfilling mechanism from text to visual modalities, offering a label-free approach to VLMs alignment.

Oral Session B: Safety and Alignment in LLMs 1 (Harbor E-F)

14:00-15:30 (Harbor E-F)

[Long] CoAct: Co-Active LLM Preference Learning with Human-AI Synergy ▶ *Underline Video*
Ruiyao Xu, Mihir Parmar, Tiankai Yang, Zhengyu Hu, Yue Zhao and Kaize Ding
Learning from preference-based feedback has become an effective approach for aligning LLMs across diverse tasks. However, high-quality human-annotated preference data remains expensive and scarce. Existing methods address this challenge through either self-rewarding, which scales by using purely AI-generated labels but risks unreliability, or active learning, which ensures quality through oracle annotation but cannot fully leverage unlabeled data. In this paper, we present CoAct, a novel framework that synergistically combines self-rewarding and active learning through strategic human-AI collaboration. CoAct leverages self-consistency to identify both reliable self-labeled data and samples requiring oracle verification. Additionally, oracle feedback guides the model to generate new instructions within its solvable capability. Evaluated on three reasoning benchmarks across two model families, CoAct achieves average improvements of +13.25% on GSM8K, +8.19% on MATH, and +13.16% on WebInstruct, consistently outperforming all baselines.

Oral Session B: Semantics: Lexical, Sentence-level Semantics, Textual Inference and Other areas

14:00-15:30 (Regatta)

[Long] Causal2Vec: Improving Decoder-only LLMs as Embedding Models through a Contextual Token ▶ *Underline Video*
Ailiang Lin, Zhuoyun Li, Yusong Wang, Kotaro Funakoshi and Manabu Okumura
Decoder-only large language models (LLMs) have been increasingly adopted to build embedding models for diverse tasks. To overcome the inherent limitations of causal attention in representation learning, many existing methods modify the attention mechanism to be bidirectional, potentially undermining LLMs' ability to extract semantic information acquired during pre-training. Meanwhile, leading unidirectional approaches often rely on extra input text to generate contextualized embeddings, inevitably increasing computational costs. In this work, we propose Causal2Vec, a general-purpose embedding model tailored to enhance the performance of decoder-only LLMs without altering their original architectures or introducing significant computational overhead. Specifically, we first employ a lightweight BERT-style model to pre-encode the input text into a single Contextual token, which is then prepended to the LLM's input sequence, allowing each token to capture contextualized information even without attending to future tokens. Furthermore, to mitigate the recency bias introduced by last-token pooling, we concatenate the last hidden states of Contextual and EOS tokens as the final text embedding. In practice, Causal2Vec achieves a new state-of-the-art performance on the MTEB benchmark among models trained solely on publicly available retrieval datasets.

[Long] Enhancing Lexical Relation Mining with Structured Sememe Knowledge ▶ *Underline Video*
Hansi Wang, Qiliang Liang, Yue Wang and Yang Liu
Lexical Relation Mining (LRM) aims to identify and classify lexical relations between word pairs. In this paper, we focus on two subtypes of LRM: Lexical Relation Classification (LRC) and Lexical Entailment (LE). Existing top-performing methods for them rely heavily on Pre-trained Language Models (PLMs) yet fail to distinguish nuanced lexical relations. From a linguistic perspective, intralexical tree-structured sememe information can reflect interlexical relations. Inspired by this, we are motivated to explore leveraging such structured knowledge to enhance LRC and LE. We first propose an automated Sememe Tree Construction (STC) pipeline to predict sememe trees; Then, we present the SememeLRM method to fully leverage structured sememe knowledge; Experimental results show that it achieves a notable 1.6% improvement on average across benchmarks, even outperforming Large Language Model (LLM)-based methods that contain 20 times more parameters on most benchmarks. Further results also suggest that sememe trees predicted by our pipeline can rival the gold-standard in HowNet, extending their applicability to lexico-semantic computing. Overall, this paper presents a potentially generalizable framework for leveraging complete sememe trees and makes significant progress, helping to unlock the value of such intralexical knowledge in downstream tasks.

[Long] FormulaSPIN: Self-Play Fine-Tuning for Natural Language to Spreadsheet Formula Generation ▶ *Underline Video*
Cy Xie
Spreadsheet applications are used by hundreds of millions worldwide, yet writing formulas remains a significant barrier. Existing approaches rely on static supervised data, which quickly saturates on limited annotations. In this paper, we introduce FormulaSPIN, a self-play framework that breaks the ceiling of supervised fine-tuning by enabling iterative self-improvement without any additional data. Vanilla SPIN fails on this task: it uniformly penalizes every non-matching output, so execution-equivalent alternatives are pushed down as negatives in one example while serving as ground truth in another, producing contradictory gradients. Our framework resolves this by exploiting formula generation's unique advantage: binary executability provides implicit supervision that separates semantic errors from valid stylistic variants. We frame training as a two-player game in which the main player learns to prefer ground-truth formulas over those from its previous version, while execution feedback sorts outputs into distinct granularities—enabling an adaptive curriculum that shifts from semantic correctness to stylistic refinement. To further increase accuracy, we incorporate ExecVote, a semantic-level voting mechanism that naturally handles multiple valid formulations. Experiments on multiple benchmarks demonstrate that FormulaSPIN achieves state-of-the-art performance,

with 74.9% exact match and 87.1% execution accuracy on NL2FORMULA, matching models trained with additional preference annotations while outperforming both traditional SFT and frontier proprietary models. These findings underscore self-play's potential to tackle scarce data tasks and open the door to extending it beyond executable domains.

[Long] Label and Explanation Variation in LLM-Based Annotation: a Case Study in Natural Language Inference ▶ *Underline Video*

Artur Kulmizev, Erika Lombart, Patrick Watrin and Marie-Catherine de Marneffe

Large language models (LLMs) have shown considerable promise for annotation purposes, yet questions remain about their ability to capture human label variation (HLV) — genuine disagreement between annotators often observed across NLP tasks. Here, we investigate how label and explanation variation manifests within and across LLMs with respect to the Natural Language Inference (NLI) task. Using zero-shot prompting with exact human annotation instructions, we treat individual model generations as participants and examine three response sampling strategies: varying generation parameters, leveraging within-family model size differences, and pooling responses from distinct LLMs. We show that, while model ensembles can generate label distributions similar to humans, they likewise exhibit distinct, idiosyncratic judgments and disagreement patterns. We further analyze explanation variation, observing that, although models generate longer explanations than humans, they demonstrate substantially less stylistic diversity. Our findings suggest that, while LLMs may serve as useful tools for generating diverse annotations, they should not be viewed as drop-in replacements for human annotators — particularly in applications requiring authentic representation of diversity in human judgments, such as NLI.

[Long] Map of Encoders - Mapping Sentence Encoders using Quantum Relative Entropy ▶ *Underline Video*

Gaifan Zhang and Danushka Bollegala

We propose a method to compare and visualise sentence encoders at scale by creating a map of encoders where each sentence encoder is represented in relation to the other sentence encoders. Specifically, we first represent a sentence encoder using an embedding matrix of a sentence set, where each row corresponds to the embedding of a sentence. Next, we compute the PIP matrix for a sentence encoder using its embedding matrix. Finally, we create a feature vector for each sentence encoder that reflects its QRE with respect to a unit base encoder. We construct a map of encoders covering 1101 publicly available sentence encoders, providing a new perspective of the landscape of the pre-trained sentence encoders. Our map accurately reflects various relationships between encoders, where encoders with similar attributes are proximally located on the map. Moreover, our encoder feature vectors can be used to accurately infer downstream task performance of the encoders, such as in retrieval and clustering tasks, demonstrating the correctness of our map.

[Long] MetFuse: Figurative Fusion between Metonymy and Metaphor ▶ *Underline Video*

Saptarshi Ghosh and Tianyu Jiang

Metonymy and metaphor often co-occur in natural language, yet computational work has studied them largely in isolation. We introduce a framework that transforms a literal sentence into three figurative variants: metonymic, metaphoric, and hybrid. Using this framework, we construct MetFuse, the first dedicated dataset of figurative fusion between metonymy and metaphor, containing 1,000 human-verified meaning-aligned quadruplets totaling 4,000 sentences. Extrinsic experiments on eight existing benchmarks show that augmenting training data with MetFuse consistently improves both metonymy and metaphor classification, with hybrid examples yielding the largest gains on metonymy tasks. Using this dataset, we also analyze how the presence of one figurative type impacts another. Our findings show that both human annotators and large language models better identify metonymy in hybrid sentences than in metonymy-only sentences, demonstrating that the presence of a metaphor makes a metonymic noun more explicit.

[Long] PrefixNLI: Detecting Factual Inconsistencies as Soon as They Arise ▶ *Underline Video*

Sapir Harary, Eran Hirsch, Aviv Slobodkin, David Wan, Mohit Bansal and Ido Dagan

Natural Language Inference (NLI) models have been used in various ways to improve the factuality of LLM outputs. This is typically done by applying an NLI model to judge whether the model output is entailed from the supposed evidence, triggering some corrective actions, such as beam reranking at inference time or RL rewards during training. While NLI models are trained to detect factual inconsistencies over complete sentences, decisions in the common autoregressive generation architecture are made for each evolving text prefix, during decoding. Addressing this setting, we generalize the entailment detection task to apply over arbitrary text prefixes, and suggest its utility for improving generation faithfulness. Providing suitable evaluation and training datasets for this task, we train MiniTruePrefixes, a novel specialized model that better detects factual inconsistencies over text prefixes, outperforming comparable baseline NLI models by 5-14 F1 points in prefix-level entailment. We further demonstrate that integrating MiniTruePrefixes into a controlled decoding framework substantially improves factual consistency in abstractive summarization. When guided by MiniTruePrefixes, LLaMA-3.2-3B-Instruct matches the faithfulness and runtime of the 8B model from the same model family, while using only half the memory.

[Long] SenseRel: A Sense-Level Benchmark for Denotational and Connotational Meaning Relations ▶ *Underline Video*

Pierluigi Cassotti, Naomi Baes, Stefano De Pascale, Jáder Martins Camboim de Sá, Francesco Periti, Nick Haslam, Dirk Geeraerts and Nina Tahmasebi

Polysemy enables a single word to convey multiple related meanings, reflecting the conceptual and emotional aspects of the evolution of the senses. We introduce the first sense-level benchmark, SenseRel, for modeling semantic relations between word senses, uniting denotational and connotational aspects of meaning. SenseRel distinguishes denotational relations, such as generalization or metaphor, as well as two connotational dimensions: valence and arousal. We evaluate large language models (LLMs), GPT-4o, Llama 3.1, and DeepSeek, in zero-shot

and fine-tuned settings. Results show that GPT-4o best aligns with human affective judgments, while a fine-tuned RoBERTa model excels at classifying denotational relations.

[Long] The Imperfective Paradox in Large Language Models ▶ *Underline Video*
Bolei Ma and Yusuke Miyao
Do Large Language Models (LLMs) genuinely grasp the compositional semantics of events, or do they rely on surface-level probabilistic heuristics? We investigate the Imperfective Paradox, a logical phenomenon where the past progressive aspect entails event realization for activities (e.g., running -> ran) but not for accomplishments (e.g., building built). We introduce ImperfectiveNLI, a diagnostic dataset designed to probe this distinction across diverse semantic classes. Evaluating state-of-the-art open-weight models, we uncover a pervasive Teleological Bias: models systematically hallucinate completion for goal-oriented events, even overriding explicit textual cancellation. Prompting interventions partially reduce this bias but trigger a calibration crisis, causing models to incorrectly reject valid entailments for atelic verbs. Representational analyses further show that while internal embeddings often distinguish progressive from simple past forms, inference decisions are dominated by strong priors about goal attainment. Taken together, our findings indicate that these current open-weight LLMs operate as predictive narrative engines rather than faithful logical reasoners, and that resolving aspectual inference requires moving beyond prompting toward structurally grounded alignment.

Oral Session B: Student Research Workshop 2

14:00-15:30 (Gas Lamp A-B)

[SRW] Adaptive Data Collection for Latin-American Community-sourced Evaluation of Stereotypes (LACES) ▶ *Underline Video*
Guido Ivetta, Pietro Palombini, Sofía Martinelli, Marcos J Gomez, M Emilia Echeveste, Sunipa Dev, Vinodkumar Prabhakaran and Luciana Benotti;
The evaluation of societal biases in NLP models is critically hindered by a geo-cultural gap, This leaves regions such as Latin America severely underserved, making it impossible to adequately assess or mitigate the perpetuation of harmful regional stereotypes in language technologies. This paper presents LACES, a stereotype association dataset, for 15 Latin American countries. This dataset includes 4,789 stereotype associations manually created and annotated by 83 participants. The dataset was developed through targeted community partnerships across Latin America. Additionally, in this paper, we propose a novel adaptive data collection methodology that uniquely integrates the sourcing of new stereotype entries and the validation of existing data within a single, unified workflow. This approach results in a resource with more unique stereotypes than previous static collection methods, enabling a more efficient stereotype collection. The paper further supports the quality of LACES by demonstrating reduced efficacy of debiasing methods on this dataset in comparison to existing popular stereotype benchmarks.

[SRW] MetaCog-Bench: Quantifying the Metacognition Gap in Edge LLM Tool Calling Under Information Insufficiency ▶ *Underline Video*
Yu-An Lu, Chun-En Hsiao, Chengwei Chiang and Hong-Han Shuai
When personal assistants powered by edge-scale language models (8B–32B) encounter under-specified user requests, such as "order me a pizza for tonight" without specifying size, toppings, or address, they overwhelmingly fabricate plausible parameter values rather than seeking clarification. We introduce MetaCog-Bench, a diagnostic benchmark spanning 22 API domains with 4,573 samples, featuring a five-category response taxonomy and per-condition annotations that distinguish fabrication from coincidental ground-truth leakage. Evaluating 14 model configurations across four families, we find that hallucination rates (HR) reach 32–81% and, critically, that fabrication risk varies unpredictably across domains: different models exhibit different vulnerability profiles, making single-domain evaluation insufficient. We then show that a structured audit prompt requiring per-parameter source verification collapses HR to 0.9–4.3% across all architectures, compressing cross-model variance by a factor of 19. This reveals a metacognition gap: models can reliably perform parameter verification when appropriately prompted, yet this behavior remains absent under standard instructions. Fine-tuning on MetaCog-Bench closes 30–34% of this gap on standard prompts without inference-time scaffolding, validating the benchmark as a training resource.

[SRW] Multi-Agent Reasoning Improves Compute Efficiency: Pareto-Optimal Test-Time Scaling ▶ *Underline Video*
Florian Valentin Wunderlich, Lars Benedikt Kaesberg, Jan Philip Wahle, Terry Ruas and Bela Gipp;
Advances in inference methods have enabled language models to improve their predictions without additional training. These methods often prioritize raw performance over cost-effective compute usage. However, computational efficiency is key for real-world applications with resource constraints. We provide a systematic analysis of the inference scaling strategies self-consistency, self-refinement, multi-agent debate, and mixture-of-agents, to study their computational performance tradeoffs. We evaluate methods on two reasoning benchmarks (MMLU-Pro, BBH) and include extensive parameter configurations (e.g., scaling the number of parallel predictions, agents, and debate rounds) across different model sizes. Across 34 configurations and over 100 evaluations, we compute the Pareto-optimal front to select methods that achieve the best accuracy with the lowest computational budget. Notably, inference scaling improves accuracy by up to +7.1% points over chain-of-thought at the highest evaluated budgets (20x the CoT compute budget) on MMLU-Pro. With an equal computing budget,

debate and mixture-of-agents outperform self-consistency by 1.3% and 2.7% points, respectively. While self-consistency saturates earlier, multi-agent gains persist, particularly on more complicated tasks. We identify a simple multi-agent design guideline: mixture-of-agents is most efficient when the number of parallel generations exceeds the number of sequential aggregations.

[SRW] One Panel Does Not Fit All: Case-Adaptive Multi-Agent Deliberation for Clinical Prediction ▶ *Underline Video*
Yuxing Lu; Yushuhong Lin; Jason Zhang
Large language models applied to clinical prediction exhibit case-level heterogeneity: simple cases yield consistent outputs, while complex cases produce divergent predictions under minor prompt changes. Existing single-agent strategies sample from one role-conditioned distribution, and multi-agent frameworks use fixed roles with flat majority voting, discarding the diagnostic signal in disagreement. We propose CAMP (Case-Adaptive Multi-agent Panel), where an attending-physician agent dynamically assembles a specialist panel tailored to each case's diagnostic uncertainty. Each specialist evaluates candidates via three-valued voting (KEEP/REFUSE/NEUTRAL), enabling principled abstention outside one's expertise. A hybrid router directs each diagnosis through strong consensus, fallback to the attending physician's judgment, or evidence-based arbitration that weighs argument quality over vote counts. On diagnostic prediction and brief hospital course generation from MIMIC-IV across four LLM backbones, CAMP consistently outperforms strong baselines while consuming fewer tokens than most competing multi-agent methods, with voting records and arbitration traces offering transparent decision audits.

[SRW] The signal is coming from inside the noun phrase! Tracking semantic proto-role inferences during sentence processing ▶ *Underline Video*
Lucas Y. Li; Zander Lynch; Marten Van Schijndel
Semantic roles between a predicate and argument can be decomposed into proto-role properties (e.g., INSTIGATION). We introduce a novel LLM feature attribution method, Generalized Contextual Decomposition for Transformers (GCD-T), which we use to probe which parts of a sentence enable models to infer proto-role properties. We compare our findings with human inferences.

[SRW] Think Anywhere in Code Generation ▶ *Underline Video*
Xue Jiang; Tianyu Zhang; Ge Li; Mengyang Liu; Taozhi Chen; Zhenhua Xu; Wenpin Jiao; Zhi Jin; Yihong Dong
Recent advances in reasoning Large Language Models (LLMs) have primarily relied on upfront thinking, where reasoning occurs before final answer. However, this approach suffers from critical limitations in code generation, where upfront thinking is often insufficient as problems' full complexity only reveals itself during code implementation. Moreover, it cannot adaptively allocate reasoning effort throughout the code generation process where difficulty varies significantly. In this paper, we propose Think-Anywhere, a novel reasoning mechanism that enables LLMs to invoke thinking on-demand at any token position during code generation. We achieve Think-Anywhere by first teaching LLMs to imitate the reasoning patterns through cold-start training, then leveraging outcome-based RL rewards to drive the model's autonomous exploration of when and where to invoke reasoning. Extensive experiments on four mainstream code generation benchmarks (i.e., LeetCode, LiveCodeBench, HumanEval, and MBPP) show that Think-Anywhere achieves state-of-the-art performance over both existing reasoning methods and recent post-training approaches, while demonstrating consistent generalization across diverse LLMs. Our analysis further reveals that Think-Anywhere enables the model to adaptively invoke reasoning at high-entropy positions, providing enhanced interpretability. Our source code and data are available at https://github.com/jiangxxxue/Think-Anywhere.

Oral Session B: Summarization, Generalizability and Transfer

14:00-15:30 (Promenade)

[Long] Adaptive Planning for Multi-Attribute Controllable Summarization with Monte Carlo Tree Search ▶ *Underline Video*
Sangwon Ryu, Heejin Do, Yunsu Kim, Gary Geunbae Lee and Jungseul Ok
Controllable summarization moves beyond generic outputs toward human-aligned summaries guided by specified attributes. In practice, the interdependence among attributes makes it challenging for language models to satisfy correlated constraints consistently. Moreover, previous approaches often require per-attribute fine-tuning, limiting flexibility across diverse summary attributes. In this paper, we propose adaptive planning for multi-attribute controllable summarization (PACO), a training-free framework that reframes the task as planning the order of sequential attribute control with a customized Monte Carlo Tree Search (MCTS). In PACO, nodes represent summaries, and actions correspond to single-attribute adjustments, enabling progressive refinement of only the attributes requiring further control. This strategy adaptively discovers optimal control orders, ultimately producing summaries that effectively meet all constraints. Extensive experiments across diverse domains and models demonstrate that PACO achieves robust multi-attribute controllability, surpassing both LLM-based self-planning models and fine-tuned baselines. Remarkably, PACO with Llama-3.2-1B rivals the controllability of the much larger Llama-3.3-70B baselines. With larger models, PACO achieves superior control performance, outperforming all competitors.

[Long] Attn-GS: Attention-Guided Context Compression for Efficient Personalized LLMs ▶ *Underline Video*
Shenglai Zeng, Tianqi Zheng, Chuan Tian, Dante Everaert, Yau-Shian Wang, Yupin Huang, Michael J. Morais, Rohit Patki, Jinjin Tian, Xinnan Dai, Kai Guo, Monica Xiao Cheng and Hui Liu

Personalizing large language models (LLMs) to individual users requires incorporating extensive interaction histories and profiles, but input token constraints make this impractical due to high inference latency and API costs. Existing approaches rely on heuristic methods such as selecting recent interactions or prompting summarization models to compress user profiles. However, these methods treat context as a monolithic whole and fail to consider how LLMs internally process and prioritize different profile components. We investigate whether LLMs' attention patterns can effectively identify important personalization signals for intelligent context compression. Through preliminary studies on representative personalization tasks, we discover that (a) LLMs' attention patterns naturally reveal important signals, and (b) fine-tuning enhances LLMs' ability to distinguish between relevant and irrelevant information. Based on these insights, we propose Attn-GS, an attention-guided context compression framework that leverages attention feedback from a marking model to mark important personalization sentences, then guides a compression model to generate task-relevant, high-quality compressed user contexts. Extensive experiments demonstrate that Attn-GS significantly outperforms various baselines across different tasks, token limits, and settings, achieving performance close to using full context while reducing token usage by 50 times.

[Long] Enhancing Two Steps Textual Anomaly Detection through Anisotropy Mitigation ▶ *Underline Video*

Pierre Fihey, Matthieu Labeau and Pavlo Mozharovskyi

Anomaly detection aims at distinguishing between in-distribution samples, which belong to the same distribution as the training set, and out-of-distribution samples, which lie outside of it. In textual anomaly detection, recent approaches routinely apply anomaly detection algorithms directly to embeddings extracted from pre-trained embedding models (two-stage approaches). However, the geometric properties of pre-trained embeddings can hinder the effectiveness of detection algorithms, which often rely on distance-based measures. In this work, we first highlight the relevance of similarity-trained models for textual anomaly detection. Beyond being trained to capture semantic similarities, these models also exhibit geometric properties that appear better suited to detection algorithms. We further demonstrate that, besides model choice, a simple post-processing step can significantly improve anomaly detection by adapting embeddings to the assumptions made by classical detection algorithms. The bulk of our experiments is done on a reformulation of the classification tasks from the MTEB benchmark into anomaly detection tasks.

[Long] In-depth Research Impact Summarization through Fine-Grained Temporal Citation Analysis ▶ *Underline Video*

Hiba Arnaout, Noy Sternlicht, Tom Hope and Iryna Gurevych

Understanding the impact of scientific publications is crucial for identifying breakthroughs and guiding future research. Traditional metrics based on citation counts often miss the nuanced ways a paper contributes to its field. In this work, we propose a new task: generating nuanced, expressive, and time-aware impact summaries that capture both praise (confirmation citations) and critique (correction citations) through the evolution of fine-grained citation intents. We introduce an evaluation framework tailored to this task, showing moderate to strong human correlation on subjective metrics such as insightfulness. Expert feedback from professors reveals a strong interest in these summaries and suggests future improvements. Data and code are made available.

[Long] Long-Context Reasoning Through Proxy-Based Chain-of-Thought Tuning ▶ *Underline Video*

Miao Li, Irina Saparina, Alexander Gurung and Mirella Lapata

Recent large language models support inputs of up to 10 million tokens, yet they perform poorly on long-context tasks that require complex reasoning. Such tasks can be solved using only a subset of the input — a proxy context — rather than the full sequence. Despite sharing the same underlying reasoning process, models exhibit a significant performance disparity between proxy and full contexts. To improve long-context reasoning, we propose ProxyCoT, a novel training framework that transfers reasoning capabilities from short proxy contexts to full long contexts. Specifically, we first obtain high-quality chain-of-thought reasoning traces on proxy contexts through reinforcement learning or distillation from a larger teacher model, and then ground the generated traces in full long contexts with supervised fine-tuning. Experiments across different datasets demonstrate that ProxyCoT consistently outperforms strong baselines with reduced computational overhead. Furthermore, models trained with ProxyCoT generalize their long-context reasoning capabilities to out-of-domain tasks.

[Long] Mitigating Catastrophic Forgetting in Target Language Adaptation of LLMs via Source-Shielded Updates ▶ *Underline Video*

Atsuki Yamaguchi, Terufumi Morishita, Aline Villavicencio and Nikolaos Aletras

Expanding the linguistic diversity of instruct large language models (LLMs) is crucial for global accessibility but is often hindered by the reliance on costly specialized target language labeled data and catastrophic forgetting during adaptation. We tackle this challenge under a realistic, low-resource constraint: adapting instruct LLMs using only unlabeled target language data. We introduce Source-Shielded Updates (SSU), a selective parameter update strategy that proactively preserves source knowledge. Using a small set of source data and a parameter importance scoring method, SSU identifies parameters critical to maintaining source abilities. It then applies a column-wise freezing strategy to protect these parameters before adaptation. Experiments across five typologically diverse languages and 7B and 13B models demonstrate that SSU successfully mitigates catastrophic forgetting. It reduces performance degradation on monolingual source tasks to just 3.4% (7B) and 2.8% (13B) on average, a stark contrast to the 20.3% and 22.3% from full fine-tuning. SSU also achieves target-language performance highly competitive with full fine-tuning, outperforming it on all benchmarks for 7B models and the majority for 13B models.

[Long] Mitigating Spurious Correlations in Text Classification Using Latent Space Geometry ▶ *Underline Video*

Jiasen Gao, Xiaoliang Chen, Duoqian Miao, Xu Gu, Xianyong Li and Yajun Du

Spurious correlations cause deep learning models to rely on predictive shortcuts that hold in the training data but break under distribution

shifts, leading to large performance drops for minority groups. Existing strategies often rely on costly group annotations or employ unstable adversarial training. In this paper, we propose Prototype-guided debiasing using Robust Invariant Feature Transformations (PRIFT), a novel framework that mitigates spurious correlations by manipulating latent space geometry. Specifically, we introduce a prototype-guided modeling approach that leverages natural language prompts to represent confounders, transforming abstract biases into interpretable geometric anchors without auxiliary classifiers. Based on these anchors, we introduce a centered projection operator that adaptively purifies representations by removing confounding deviations specific to instances while preserving essential semantic structure. Furthermore, PRIFT can handle confounding factor information at different levels, ranging from true labels to unsupervised latent inference. Experiments on four text classification benchmarks demonstrate the superiority of our method; notably, PRIFT outperforms state-of-the-art baselines and improves worst-group accuracy by over 20% on the CivilComments dataset compared to standard empirical risk minimization.

[Long] Narrative License and Model Sycophancy in LLM Summaries of Scientific Work ► *Underline Video*
Calvin Isch and Grace Jennings
Large language models (LLMs) are increasingly used to summarize academic work, yet model summaries can subtly exaggerate or mischaracterize findings. We examine how Narrative License (NL), rhetorical shifts that amplify claims beyond the underlying evidence, emerges in LLM summaries of scholarly articles. Using diverse prompting strategies across six leading models, we assess three dimensions of NL: causal overreach, rhetorical confidence, and sentiment (N = 100 peer-reviewed articles). Under basic summarization prompts, models frequently increase NL relative to academic abstracts; however, guardrail prompts can reduce these distortions. We further test how model "sycophancy" shapes NL, finding that stated stances and user personas produce predictable shifts in each element. These findings suggest that users and the benchmarks used to evaluate summarization should explicitly consider subtle rhetorical distortions and user alignment to ensure faithful scientific communication.

[Long] No Reader Left Behind: Multi-Agent Summaries Everyone Can Understand ► *Underline Video*
Jimin Jung, MyoungJin Kim, Jaehyung Seo and Heuiseok Lim
The Plain Writing Act in the United States requires government documents to be written in clear and simple language. However, existing summarization systems struggle to address diverse linguistic and cognitive barriers among general readers. We propose NRLB (No Reader Left Behind), a unified multi-agent framework for plain language summarization that simulates three representative reader groups: elementary school students, non-native speakers, and readers with attention deficits. NRLB integrates template-based planning with an iterative feedback loop guided by simulated readers and domain expert revision to address comprehension barriers such as unknown terms, missing contexts, and confusing sentences. Evaluations across multiple datasets demonstrate consistent improvements in both readability and factuality. Human evaluation further supports these findings, with annotator preference rates ranging from 55% to 76%, highlighting NRLBs ability to generate summaries that are both faithful to the source and accessible to a wide range of readers.

Poster Session B (Coronado)

14:00-15:30 (Coronado)

[Findings] [b] = [d] - [t] + [p]: Self-supervised Speech Models Discover Phonological Vector Arithmetic ► *Underline Video*
Kwanghee Choi, Eunjung Yeo, Cheol Jun Cho, David Harwath and David R. Mortensen
Self-supervised speech models (S3Ms) are known to encode rich phonetic information, yet how this information is structured remains underexplored. We conduct a comprehensive study across 96 languages to analyze the underlying structure of S3M representations, with particular attention to phonological vectors. We first show that there exist linear directions within the model's representation space that correspond to phonological features. We further demonstrate that the scale of these phonological vectors correlate to the degree of acoustic realization of their corresponding phonological features in a continuous manner. For example, the difference between [d] and [t] yields a voicing vector: adding this vector to [p] produces [b], while scaling it results in a continuum of voicing. Together, these findings indicate that S3Ms encode speech using phonologically interpretable and compositional vectors, demonstrating phonological vector arithmetic. All code and interactive demos are available at https://github.com/juice500ml/phonetic-arithmetic.

[Long] A Multi-Agent Framework for High-Interaction Terminal Simulation ► *Underline Video*
Kai Wei, Yuwen Cui, Kehan Shen, Hua Wei and Guangjing Wang
Terminal simulation, framed as a terminal command-level Turing test, is a long-standing problem of symbolic language generation in dialogue and interactive systems. Prior scripted simulators lack the flexibility needed for complex, multi-turn interactions, while LLM-based approaches often misinterpret commands, break output formats, drift from system state, and remain vulnerable to prompt injection. In this work, we propose MANTIS, a terminal simulation framework that improves realism, consistency, and robustness in command-language generation. MANTIS integrates a multi-agent architecture with a filter-based routing model that safely dispatches commands to external tools or an LLM-based agent, enabling support for interactive commands while defending against prompt injection attacks. In addition, we design an agentic file system with history pruning to preserve long-term state consistency. We release three datasets: 28,045 real terminal input-output pairs, a 1,000-session multi-turn interaction dataset, and a 25,849-instance labeled classification dataset. MANTIS outperforms

state-of-the-art baselines by more than 9%, achieving over 95% accuracy on multi-turn terminal simulation. The dataset and source code are available at https://github.com/kaiwei666a/MANTIS_Terminal_Simulation

[Long] AAPO: Enhancing the Reasoning Capabilities of LLMs with Advantage Margin ► *Underline Video*

Jian Xiong, Jingbo Zhou, Jingyong Ye, Qiang Huang and Dejing Dou

Reinforcement learning (RL) has emerged as an effective approach for enhancing the reasoning capabilities of large language models (LLMs), especially in scenarios where supervised fine-tuning (SFT) falls short due to limited chain-of-thought (CoT) data. Among RL-based post-training methods, group relative advantage estimation, as exemplified by Group Relative Policy Optimization (GRPO), has attracted considerable attention for eliminating the dependency on the value model, thereby simplifying training compared to traditional approaches like Proximal Policy Optimization (PPO). However, existing group relative advantage estimation method still suffers from training inefficiencies, particularly when the estimated advantage approaches zero. To address this limitation, we propose Advantage-Augmented Policy Optimization (AAPO), a novel RL algorithm that optimizes the cross-entropy (CE) loss using advantages enhanced through a margin-based estimation scheme. This approach effectively mitigates the inefficiencies associated with group relative advantage estimation. Experimental results on multiple mathematical reasoning benchmarks and model series demonstrate the superior performance of AAPO. Code is available at https://github.com/JianxXiong/AAPO.

[Long] AgentRouter: A Knowledge-Graph-Guided LLM Router for Collaborative Multi-Agent Question Answering ► *Underline Video*

Zheyuan Zhang, Kaiwen Shi, Zhengqing Yuan, Zehong Wang, Tianyi Ma, Keerthiram Murugesan, Vincent Galassi, Chuxu Zhang and Yanfang Ye

Large language models (LLMs) and agent-based frameworks have advanced rapidly, enabling diverse applications. Yet, with the proliferation of models and agentic strategies, practitioners face substantial uncertainty in selecting the best configuration for a downstream task. Prior studies show that different agents and backbones exhibit complementary strengths, and that larger models are not always superior, underscoring the need for adaptive routing mechanisms. Existing approaches to agent routing, however, often emphasize cost efficiency while overlooking the fine-grained contextual and relational structure inherent in QA tasks. In this paper, we propose AgentRouter, a framework that formulates multi-agent QA as a knowledge-graph-guided routing problem supervised by empirical performance signals. Specifically, we convert QA instance into a heterogeneous knowledge graph that jointly encodes queries, contextual entities, and agents, and then train a heterogeneous graph neural network (GNN) to propagate information across node types and produce task-aware routing distributions over agents. By leveraging soft supervision and weighted aggregation of agent outputs, AgentRouter learns principled collaboration schemes that capture the complementary strengths of diverse agents. Extensive experiments demonstrate that our framework consistently outperforms single-agent and ensemble baselines, while generalizing across benchmarks and LLM backbones. These results highlight the effectiveness and robustness of graph-supervised multi-agent routing for question answering. Our code repo is available at https://anonymous.4open.science/r/AgentRouter.

[Long] Anchored Sliding Window: Toward Robust and Imperceptible Linguistic Steganography ► *Underline Video*

Ruiyi Yan, Shiao Meng and Yugo Murawaki

Linguistic steganography based on language models typically assumes that steganographic texts are transmitted without alteration, making them fragile to even minor modifications. While previous work mitigates this fragility by limiting the context window, it significantly compromises text quality. In this paper, we propose the anchored sliding window (ASW) framework to improve imperceptibility and robustness. In addition to the latest tokens, the prompt and a bridge context are anchored within the context window, encouraging the model to compensate for the excluded tokens. We formulate the optimization of the bridge context as a variant of prompt distillation, which we further extend using self-distillation strategies. Experiments show that our ASW significantly and consistently outperforms the baseline method in text quality, imperceptibility, and robustness across diverse settings. The code is available at github.com/ryehr/ASW_steganography.

[Findings] Automated Profile Inference with Language Model Agents ► *Underline Video*

Yuntao Du, Zitao Li, Bolin Ding, Yaliang Li, Hanshen Xiao, Jingren Zhou and Ninghui Li

Impressive progress has been made in automated problem-solving by the collaboration of large language model (LLM) based agents. However, these automated capabilities also open avenues for malicious applications. In this paper, we study a new threat that LLMs pose to online pseudonymity, called automated profile inference, where an adversary can instruct LLMs to automatically collect and extract sensitive personal attributes from publicly available user activities on pseudonymous platforms. We also introduce an automated profiling framework called AutoProfiler to demonstrate and assess the feasibility of such attacks in real-world scenarios. AutoProfiler consists of four specialized LLM agents that work collaboratively to retrieve and process user online activities and generate a profile with extracted personal information. Experimental results on two real-world datasets and one synthetic dataset show that AutoProfiler is highly effective and efficient, and the inferred attributes are both identifiable and sensitive, posing significant privacy risks. We explore mitigation strategies from different perspectives and advocate for increased public awareness of this emerging privacy threat.

[Findings] Beyond Memorization: Extending Reasoning Depth with Recurrence, Memory and Test-Time Compute Scaling ► *Underline Video*

Ivan Rodkin, Daniil Orel, Konstantin Smirnov, Arman Bolatov, Bilal Elbouardi, Besher Hassan, Yuri Kuratov, Aydar Bulatov, Preslav Nakov,

Timothy Baldwin, Artem Shelmanov and Mikhail Burtsev
Reasoning is a core capability of large language models (LLMs), yet how multi-step reasoning is learned and executed remains unclear. We study this question in a controlled cellular-automata (1dCA) framework that excludes memorization by using disjoint training and test rules. Given a short state sequence, the model is required to infer the hidden local rule and then chain it to predict multiple future steps. Our evaluation shows that LLMs largely fail to reliably solve a natural-language proxy of the proposed task. We find that most neural architectures trained from scratch can learn rule inference and achieve high next-step accuracy, but performance drops sharply as the required number of intermediate reasoning steps increases. Experiments show that increasing model depth is crucial, and extending effective depth via recurrence, memory, or test-time compute improves results but remains bounded. Code is available on github: https://github.com/RodkinIvan/associative-recurrent-memory-transformer/tree/ACT.

[Findings] Beyond Outcome Verification: Verifiable Process Reward Models for Structured Reasoning ► *Underline Video*
massimiliano pronestianya belzyufang hou
Recent work on reinforcement learning with verifiable rewards (RLVR) has shown that large language models (LLMs) can be substantially improved using outcome-level verification signals, such as unit tests for code or exact-match checks for mathematics. In parallel, process supervision has long been explored as a way to shape the intermediate reasoning behaviour of LLMs, but existing approaches rely on neural judges to score chain-of-thought steps, leaving them vulnerable to opacity, bias, and reward hacking. To address this gap, we introduce Verifiable Process Reward Models (VPRMs), a reinforcement-learning framework in which intermediate reasoning steps are checked by deterministic, rule-based verifiers. We apply VPRMs to risk-of-bias assessment for medical evidence synthesis, a domain where guideline-defined criteria and rule-based decision paths enable programmatic verification of reasoning traces. Across multiple datasets, we find that VPRMs generate reasoning that adheres closely to domain rules and achieve substantially higher coherence between step-level decisions and final labels. Results show that VPRMs achieve up to 20% higher F1 than state-of-the-art models and 6.5% higher than verifiable outcome rewards, with substantial gains in evidence grounding and logical coherence.

[Findings] Beyond Static Synthetic Noise: Assessing the Robustness of Large Language Models to Natural Context Variation in the Real World ► *Underline Video*
Yulong Wu, Viktor Schlegel and Riza Batista-Navarro
Robustness evaluation in Question Answering (QA) has predominantly relied on synthetic perturbations that poorly capture natural text evolution in real-world settings, a limitation that becomes more pronounced with the widespread deployment of Large Language Models (LLMs) in dynamic, user-facing environments. In this work, we address this gap by proposing a framework for automatically evaluating QA models under naturally occurring textual perturbations, replacing context passages with revised counterparts from Wikipedia edit histories. Through extensive evaluation on SQUAD across diverse encoder architectures, we construct two challenging sets where human performance remains stable, yet state-of-the-art LLMs exhibit significant degradation, with performance drops of up to 28.28%. These robustness gaps further generalize to more complex QA scenarios, such as DROP and HOTPOTQA. To mitigate these errors, we show that robustness to natural perturbations can be improved via adversarial training for encoder-only models and in-context demonstrations of perturbed instances for LLMs, though a more generalizable and effective defense strategy remains an open challenge.

[Long] Beyond the Panorama: Training-Free Hierarchical Perception-Reasoning for Fine-Grained Vision in MLLMs ► *Underline Video*
Xiaoyang Yi, Jing Chen, Li Peng, Yuru Bao and Jian Zhang
Multimodal large language models (MLLMs) enable cross-modal semantic understanding and generation by learning semantic alignment and fusion across modalities. However, existing MLLMs still face challenges in fine-grained visual tasks. Their uniform encoding for global understanding tends to blur or lose local details, while the lack of explicit modeling of intermediate visual evidence leads them to rely on semantic priors or the statistical patterns of language models rather than grounded visual information, resulting in potential hallucinations. To address these issues, we propose HiPerson, a training-free hierarchical perception-reasoning framework that enhances fine-grained visual understanding by simulating human perception mechanisms. Specifically, HiPerson fuses internal relative attention and gradient activation signals to generate a task-aware semantic heatmap, providing explicit perceptual anchors for precise localization. Then, it employs a dual-scale adaptive cropping strategy to extract visual cues for interactive reasoning, simulating the process of human visual focus shifting and detail attention. Finally, by combining local-global dual-image cooperative input with a multi-step reasoning prompting mechanism, HiPerson guides the model to complete a full perception loop from detail observation to contextual verification. Experiments show that HiPerson achieves competitive results on multiple datasets, demonstrating its generalizability and scalability.

[Long] Can Large Language Models Keep Up? Benchmarking Online Adaptation to Continual Knowledge Streams ► *Underline Video*
Jiyeon Kim, Hyunji Lee, Dylan Zhou, Sue Hyun Park, Seunghyun Yoon, Trung Bui, Franck Dernoncourt, Sungmin Cha and Minjoon Seo
LLMs operating in dynamic real-world contexts often encounter knowledge that evolves continuously or emerges incrementally. To remain accurate and effective, models must adapt to newly arriving information on the fly. We introduce Online Adaptation to Continual Knowledge Streams(OAKS) to evaluate this capability, establishing a benchmark for online adaptation over streaming, continually updating knowledge. Specifically, the benchmark is structured as a sequence of fine-grained context chunks where facts change dynamically across time intervals. OAKS comprises two datasets: OAKS-BABI and OAKS-Novel, where individual facts evolve multiple times across context chunks. These

datasets include dense annotations to measure whether models track changes accurately. Evaluating 14 models with varied inference approaches, we observe significant limitations in current methodologies. Both state-of-the-art models and agentic memory systems fail to adapt robustly on OAKS, demonstrating delays in state-tracking and susceptibility to distraction within streaming environments. We will open-source the code and datasets.

[Long] ChartAgent: A Multimodal Agent for Visually Grounded Reasoning in Complex Chart Question Answering ► *Underline Video*

Rachneet Kaur, Nishan Srishankar, Zhen Zeng and Sumitra Ganesh

Recent multimodal LLMs have shown promise in chart-based visual question answering, but their performance declines sharply on unannotated charts, those requiring precise visual interpretation rather than relying on textual shortcuts. To address this, we introduce ChartAgent, a novel agentic framework that explicitly performs visual reasoning directly within the charts spatial domain. Unlike textual chain-of-thought reasoning, ChartAgent iteratively decomposes queries into visual subtasks and actively manipulates and interacts with chart images through specialized actions such as drawing annotations, cropping regions (e.g., segmenting pie slices, isolating bars), and localizing axes, using a library of chart-specific vision tools to fulfill each subtask. This iterative reasoning process closely mirrors human cognitive strategies for chart comprehension. ChartAgent achieves state-of-the-art accuracy on the ChartBench and ChartX benchmarks, surpassing prior methods by up to 16.07% absolute gain overall and 17.31% on unannotated, numerically intensive queries. Furthermore, our analyses show that ChartAgent is (a) effective across diverse chart types, (b) achieves the highest scores across varying visual and reasoning complexity levels, and (c) serves as a plug-and-play framework that boosts performance across diverse underlying LLMs. Our work is among the first to demonstrate visually grounded reasoning for chart understanding using tool-augmented multimodal agents.

[Long] CHOIR: Harmonizing Structured Persona Diversity for Robust Collaborative LLM Reasoning ► *Underline Video*

Xiangjue Dong, Cong Wang, Maria Teleki, Millennium Bismay, Ruihong Huang and James Caverlee

Persona-assigned Large Language Models can adopt diverse roles, enabling personalized and context-aware reasoning. However, even minor demographic perturbations in personas, such as simple pronoun swaps, can alter reasoning trajectories, leading to divergent sets of correct answers on reasoning benchmarks. We explore the potential of these variations as a constructive resource to improve LLM reasoning performance. We propose CHOIR (Collaborative Harmonization fOr Inference Robustness), a test-time framework that harmonizes a set of demographically perturbed, persona-conditioned reasoning signals into a unified prediction. CHOIR orchestrates a collaborative decoding process among counterfactual personas perturbed across dimensions of gender, race, religion, disability, and age, dynamically balancing agreement and divergence in their reasoning paths to improve performance. Experiments demonstrate that CHOIR consistently enhances LLM reasoning across model architectures, scales, and tasks. Improvements reach up to 20.1% for individual groups and 15.1% on average, and we show that CHOIR remains effective even when base personas are suboptimal.

[Long] CoDial: Interpretable Task-Oriented Dialogue Systems Through Dialogue Flow Alignment ► *Underline Video*

Radin Shayanfar, Chu Fei Luo, Rohan Bhambhoria, Samuel Dahan and Xiaodan Zhu

Building Task-Oriented Dialogue (TOD) systems that generalize across different tasks remains a challenging problem. Data-driven approaches often struggle to transfer effectively to unseen tasks. While recent schema-based TOD frameworks improve generalization by decoupling task logic from language understanding, their reliance on neural or generative models often obscures how task schemas influence behaviour and hence impair interpretability. In this work, we introduce a novel framework, CoDial (Code for Dialogue), at the core of which is converting a predefined task schema to a structured heterogeneous graph and then to popular programmatic LLM guardrailing code, such as NVIDIA's Colang. The pipeline enables efficient and interpretable alignment of dialogue policies during inference. We introduce two paradigms for LLM guardrailing code generation, CoDial-free and CoDial-structured, and propose a mechanism that integrates human feedback to iteratively improve the generated code. Empirically, CoDial achieves state-of-the-art (SOTA) performance on the widely used benchmark datasets, while providing inherent interpretability in the design. We additionally demonstrate CoDial's iterative improvement via manual and LLM-aided feedback, making it a practical tool for human-guided alignment of LLMs in unseen domains.

[Findings] CoPA: Benchmarking Personalized Question Answering with Data-Informed Cognitive Factors ► *Underline Video*

Hang Su, Zequn Liu, Chen Hu, Xuesong Lu, Yingce Xia and Zhen Liu

While LLMs have demonstrated remarkable potential in Question Answering (QA), evaluating personalization remains a critical bottleneck. Existing paradigms predominantly rely on surface-level similarity or manual heuristics, often lacking sufficient data-driven validation. We address this by mining Community-Individual Preference Divergence (CIPD), where individual choices override consensus, to distill six key personalization factors as evaluative dimensions. Accordingly, we introduce CoPA, a benchmark with 1,985 user profiles for fine-grained, factor-level assessment. By quantifying the alignment between model outputs and user-specific cognitive preferences inferred from interaction patterns, CoPA provides a more comprehensive and discriminative standard for evaluating personalized QA than generic metrics. The code is available at https://github.com/bjzgcai/CoPA.

[Long] D^2Plan: Dual-Agent Dynamic Global Planning for Complex Retrieval-Augmented Reasoning ► *Underline Video*

Kangcheng Luo, Tinglang Wu and Yansong Feng

Recent search-augmented LLMs trained with reinforcement learning (RL) can interleave searching and reasoning for multi-hop reasoning tasks. However, they face two critical failure modes as the accumulating context becomes flooded with both crucial evidence and irrelevant

information: (1) ineffective search chain construction that produces incorrect queries or omits retrieval of critical information, and (2) reasoning hijacking by peripheral evidence that causes models to misidentify distractors as valid evidence. To address these challenges, we propose DšPlan, a Dual-agent Dynamic global Planning paradigm for complex retrieval-augmented reasoning. DšPlan operates through the collaboration of a Reasoner and a Purifier: the Reasoner constructs explicit global plans during reasoning and dynamically adapts them based on retrieval feedback; the Purifier assesses retrieval relevance and condenses key information for the Reasoner. We further introduce a two-stage training framework consisting of supervised fine-tuning (SFT) cold-start on synthesized trajectories and RL with plan-oriented rewards to teach LLMs to master the DšPlan paradigm. Extensive experiments demonstrate that DšPlan enables more coherent multi-step reasoning and stronger resilience to irrelevant information, thereby achieving superior performance on challenging QA benchmarks.

[Findings] Decomposed Prompting Does Not Fix Knowledge Gaps, But Helps Models Say I Dont Know ▶ *Underline Video*

Dhruv Madhwal, Lyuxin Zhang, Dan Roth, Tomer Wolfson and Vivek Gupta

Large language models often struggle to recognize their knowledge limits in closed-book question answering, leading to confident hallucinations. While decomposed prompting is typically used to improve accuracy, we investigate its impact on reliability. We evaluate three task-equivalent prompting regimes: Direct, Assistive, and Incremental, across different model scales and multi-hop QA benchmarks. We find that although accuracy gains from decomposition diminish in frontier models, disagreements between prompting regimes remain highly indicative of potential errors. Because factual knowledge is typically stable while hallucinations are stochastic, cross-regime agreement provides a precise signal of internal uncertainty. We leverage this signal to implement a training-free abstention policy that requires no retrieval or fine-tuning. Our results show that disagreement-based abstention outperforms standard uncertainty baselines as an error detector, improving both F1 and AUROC across settings. This demonstrates that decomposition-based prompting can serve as a practical diagnostic probe for model reliability in closed-book QA.

[Long] Disentangling Reasoning Logic to Resolve Explicit Knowledge Conflicts ▶ *Underline Video*

Xianda Zheng, Zijian Huang, Meng-Fen Chiang, Jiamou Liu, Yuan Fang, Michael Witbrock and Kaiqi Zhao

Explicit knowledge conflicts, where retrieved contexts contain contradictory information, have become increasingly prevalent as Large Language Models (LLMs) integrate diverse data sources. The core challenge lies in the complexity of entangled narratives and the heterogeneity of conflict cases, which impose excessive demands on the reasoning capabilities of standard models. To address this, we propose Knowledge Conflict Reasoning (KCR), a framework that adjudicates conflicts by structuring the underlying logic. KCR first disentangles conflicting contexts into distinct sets of reasoning traces, utilizing both textual and graph-based representations, to simplify comprehension. It then employs a Reinforcement Learning with Verifiable Rewards (RLVR) paradigm, guiding the model to internalize a reasoning process that maximizes logical consistency while actively suppressing spurious reasoning paths derived from contradictory contexts. Extensive experiments demonstrate that KCR yields substantial improvements: a KCR-enhanced 7B model surpasses the performance of baselines equipped with top-tier closed-source models such as GPT-4o and GPT-5.1.

[Findings] Diversity Collapse in Multi-Agent LLM Systems: Structural Coupling and Collective Failure in Open-Ended Idea Generation ▶ *Underline Video*

Nuo Chen, Yicheng Tong, Yuzhe Yang, Yufei He, Xueyi Zhang, Zou Qingyun, Qian Wang and Bingsheng He

Multi-agent systems (MAS) are increasingly used for open-ended idea generation, driven by the expectation that collective interaction will broaden the exploration diversity. However, when and why such collaboration truly expands the solution space remains unclear. We present a systematic empirical study of diversity in MAS-based ideation across three bottom-up levels: model intelligence, agent cognition, and system dynamics. At the model level, we identify a compute efficiency paradox, where stronger, highly aligned models yield diminishing marginal diversity despite higher per-sample quality. At the cognition level, authority-driven dynamics suppress semantic diversity compared to junior-dominated groups. At the system level, group-size scaling yields diminishing returns and dense communication topologies accelerate premature convergence. We characterize these outcomes as collective failures emerging from structural coupling, a process where interaction inadvertently contracts agent exploration and triggers diversity collapse. Our analysis shows that this collapse arises primarily from the interaction structure rather than inherent model insufficiency, highlighting the importance of preserving independence and disagreement when designing MAS for creative tasks. Our code is available at https://github.com/Xtra-Computing/MAS_Diversity.

[Findings] Do LLMs Catch Their Own Mistakes? A Comprehensive Benchmark for Reflective Tool Use LLMs ▶ *Underline Video*

Zheyuan Liu, Liqiang Xiao, Yang Li, Hyokun Yun, Lihong Li, Chao Zhang and Meng Jiang

Large language models (LLMs) increasingly rely on external tools to complete complex tasks, yet their ability to recognize and correct their own tool-use mistakes remains underexplored. Existing benchmarks primarily evaluate planning and execution success, overlooking the self-reflective dimension of tool use. To address this gap, we present ReflecTool-Bench, the first benchmark designed to assess LLMs self-reflective reasoning in tool-augmented multi-turn dialogues. ReflecTool-Bench covers 10 domains with 88 distinct APIs and 968 annotated dialogues, systematically injecting diverse error types arising from both user and assistant behavior. The benchmark defines two complementary evaluation setups: the Critique task, where models diagnose errors in third-party dialogues, and the Self-Reflection Task, where models must detect and repair their own prior tool-use mistakes. We introduce fine-grained metrics for error detection, error classification, correction accuracy, and explanation quality, enabling a holistic assessment of reflective reasoning. Evaluations across 12 state-of-the-art models, including both API-based closed source models and open source models, reveal that while models can reliably identify user-originated errors, they struggle with assistant-originated ones, and performance drops sharply when moving from critique to

self-reflection.

[Findings] Do LLMs Really Know What They Dont Know? Internal States Mainly Reflect Knowledge Recall Rather Than Truthfulness ► *Underline Video*

Chi Seng Cheang, Hou Pong Chan, Wenxuan Zhang and Yang Deng;

Recent work suggests that LLMs "know what they don't know", positing that hallucinated and factually correct outputs arise from distinct internal processes and can therefore be distinguished using internal signals. However, hallucinations have multifaceted causes: beyond simple knowledge gaps, they can emerge from training incentives that encourage models to exploit statistical shortcuts or spurious associations learned during pretraining. In this paper, we argue that when LLMs rely on such learned associations to produce hallucinations, their internal processes are mechanistically similar to those of factual recall, as both stem from strong statistical correlations encoded in the model's parameters. To verify this, we propose a novel taxonomy categorizing hallucinations into Unassociated Hallucinations (UHs), where outputs lack parametric grounding, and Associated Hallucinations (AHs), which are driven by spurious associations. Through mechanistic analysis, we compare their computational processes and hidden-state geometries with factually correct outputs. Our results show that hidden states primarily reflect whether the model is recalling parametric knowledge rather than the truthfulness of the output itself. Consequently, AHs exhibit hidden-state geometries that largely overlap with factual outputs, rendering standard detection methods ineffective. In contrast, UHs exhibit distinctive, clustered representations that facilitate reliable detection.

[Findings] Doc-PP: Document Policy Preservation Benchmark for Large Vision-Language Models ► *Underline Video*

Haeun Jang, Hwan Chang and Hwanhee Lee

The deployment of Large Vision-Language Models (LVLMs) for real-world document question answering is often constrained by dynamic, user-defined policies that dictate information disclosure based on context. While ensuring adherence to these explicit constraints is critical, existing safety research primarily focuses on implicit social norms or text-only settings, overlooking the complexities of multimodal documents. In this paper, we introduce Doc-PP (Document Policy Preservation Benchmark), a novel benchmark constructed from real-world reports requiring reasoning across heterogeneous visual and textual elements under strict non-disclosure policies. Our evaluation highlights a systemic Reasoning-Induced Safety Gap: models frequently leak sensitive information when answers must be inferred through complex synthesis or aggregated across modalities, effectively circumventing existing safety constraints. Furthermore, we identify that providing extracted text improves perception but inadvertently facilitates leakage. To address these vulnerabilities, we propose DVA (Decompose-Verify-Aggregation), a structural inference framework that decouples reasoning from policy verification. Experimental results demonstrate that DVA significantly outperforms standard prompting defenses, offering a robust baseline for policy-compliant document understanding.

[Findings] Dual Hierarchical Dialogue Policy Learning for Legal Inquisitive Conversational Agents ► *Underline Video*

Xubo Lin, Zezhi Deng, Shihao Wang, Grace Hui Yang and Yang Deng

Most existing dialogue systems are user-driven, primarily designed to fulfill user requests. However, in many critical real-world scenarios, a conversational agent must proactively extract information to achieve its own objectives rather than merely respond. To address this gap, we introduce Inquisitive Conversational Agents (ICAs) and develop an ICA specifically tailored to U.S. Supreme Court oral arguments. We propose a Dual Hierarchical Reinforcement Learning framework featuring two cooperating RL agents, each with its own policy, to coordinate strategic dialogue management and fine-grained utterance generation. By learning when and how to ask probing questions, the agent emulates judicial questioning patterns and systematically uncovers crucial information to fulfill its legal objectives. Evaluations on a U.S. Supreme Court dataset show our method outperforms single-agent RL baselines in multiple metrics. Although specialized to a single legal domain, it represents an important first step toward broader high-stakes, domain-specific applications. We attached a part of the code as supplementary material. All code will be released upon publication for reproducibility.

[Long] EASE: Entity-Aware Sub-table Generation for Real-world Multi-table QA ► *Underline Video*

Myunghoon Kang, Dahyun Jung, Suhyune Son, Seonmin Koo, Changwoo Chun, Daniel Rim, Haeyoung Kwon, Yuna Hur and Heuiseok Lim

Recent advancements in table-based question answering (table QA) have been driven by the development of table-specific reasoning strategies for leveraging large language models. Previous works employ sub-table-based reasoning, which involves matching query-relevant table values and aggregating them into sub-tables for precise reasoning. However, these approaches are limited to scenarios with query-relevant single tables, failing to handle real-world table QA settings that involve noisy multi-table sets. To address the challenges of real-world table QA, we propose EASE: Entity-Aware Sub-table Generation for REal-world Multi-table QA framework. Given a noisy multi-table set, EASE first extracts key entities from the question to construct a sub-table schema. It then populates this schema by utilizing a selected set of column values from the noisy multi-table set, thereby facilitating efficient and effective sub-table-based reasoning. We introduce a Noisy Multi-table QA dataset and conduct extensive experiments to evaluate EASE's effectiveness on real-world table QA. Our results demonstrate that EASE effectively filters out irrelevant information while incorporating pertinent table values, leading to efficient and effective performance on real-world table QA. Our dataset can be found https://github.com/Metalchaos8527/ease_noisy_multi-table_qa.git

[Long] Efficient Process Reward Modeling via Contrastive Mutual Information ► *Underline Video*

Nakyung Lee, Sangwoo Hong and Jungwoo Lee

Recent research has devoted considerable effort to verifying the intermediate reasoning steps of chain-of-thought (CoT) trajectories using

process reward models (PRMs) and other verifier models. However, training a PRM typically requires human annotators to assign reward scores to each reasoning step, which is both costly and time-consuming. Existing automated approaches, such as Monte Carlo (MC) estimation, also demand substantial computational resources due to repeated LLM rollouts. To overcome these limitations, we propose contrastive pointwise mutual information (CPMI), a novel automatic reward labeling method that leverages the models internal probability to infer step-level supervision while significantly reducing the computational burden of annotating dataset. CPMI quantifies how much a reasoning step increases the mutual information between the step and the correct target answer relative to hard-negative alternatives. This contrastive signal serves as a proxy for the steps contribution to the final solution and yields a reliable reward. The experimental results show that CPMI-based labeling reduces dataset construction time by 84% and token generation by 98% compared to MC estimation, while achieving higher accuracy on process-level evaluations and mathematical reasoning benchmarks.

[Findings] EXPO-SQL: Execution-based Clause-level Policy Optimization for Text-to-SQL ▶ *Underline Video*

Jaehoon Lee, ChelWon Na, Suyoung Bae, Jin-Seop Lee, Jihyung Lee, YunSeok Choi and Jee-Hyong Lee

Text-to-SQL enables users to query databases using natural language by generating executable SQL queries. Recent methods have increasingly adopted Large Language Models based reinforcement learning (RL) to leverage execution feedback for training. However, existing RL methods assign uniform query-level rewards to all clauses in a SQL query, treating correct and incorrect clauses equally. This coarse-grained reward design leads to insufficient learning signals for correct SQL generation. To address this issue, we propose EXPO-SQL (EXecution-based clause-level Policy Optimization for Text-to-SQL) which provides fine-grained supervision through clause-level rewards. To assign clause-level rewards, our method identifies erroneous clauses by analyzing execution results, including error messages and clause-wise incremental execution. Experiments on widely-used Text-to-SQL benchmarks demonstrate that EXPO-SQL significantly outperforms existing supervised fine-tuning, prompting, and RL-based methods through fine-grained clause-level learning. Our code is available at https://github.com/jhn25/EXPO-SQL.

[Findings] From Documents to Segments: A Contextual Reformulation for Topic Assignment ▶ *Underline Video*

Hoonsang Yoon, Takyoung Kim, Wonkee Lee, Ilmin Cho, Dilek Hakkani-Tür and Stanley Jungkyu Choi

Traditional topic modeling treats each document as a single, coherent unit of topic, which can cause topic contamination when documents cover multiple topics. This becomes especially problematic when stakeholders are interested in identifying documents that focus on a specific topic. We introduce segment-based topic allocation, a novel paradigm that redefines topic assignment at the level of segments, coherent textual spans conveying distinct topical content. This granularity improves topic purity, interpretability, and applicability to multi-theme corpora such as reviews or survey responses. To support this paradigm, we construct SemEval-STM, a benchmark derived from aspect-based sentiment datasets, where segments are automatically extracted using large language models (LLMs) and post-processed with human supervision. We further propose the segment intrusion task (SIT), a novel evaluation method extending word intrusion to the span level, enabling human-centric assessment of topical coherence. Empirical results across diverse metrics and models demonstrate that SBTA significantly outperforms traditional document-based methods in clustering and interpretability. Our framework provides a practical and scalable solution for fine-grained topic analysis in heterogeneous text corpora.

[Findings] From Tasks to Teams: A Risk-First Evaluation Framework for Multi-Agent LLM Systems in Finance ▶ *Underline Video*

Zichen Chen, Jianda Chen, Jiaao Chen and Misha Sra

Current financial benchmarks prioritize large language models (LLMs) for task accuracy and portfolio returns, yet overlook risks arising from multi-agent cooperation, tool-sharing, and real-world financial actions. We introduce M-SAEA, a Multi-agent, Safety-Aware Evaluation Agent that audits LLM teams without fine-tuning, deploying ten probes across four layers: model, workflow, interaction, and system, to yield a continuous risk vector and natural-language rationale. Evaluated across three high-stakes tasks (finance management, webshop automation, transactional services) with six prominent models, M-SAEA (i) identifies unsafe trajectories with minimal false positives, (ii) reveals latent risks (e.g., temporal staleness) that are not addressed by standard metrics, and (iii) provides granular, actionable scores for balancing safety and latency pre-deployment. By quantifying safety as a model-agnostic metric, M-SAEA reorients evaluation from individual tasks to collaborative teams, offering a robust template for risk-first assessment of agentic AI in finance and beyond.

[Findings] Fundamental Reasoning Paradigms Induce Out-of-Domain Generalization in Language Models ▶ *Underline Video*

Mingzi Cao, Xingwei Tan, Mahmud Elahi Akhter, Marco Valentino, Maria Liakata, Xi Wang and Nikolaos Aletras

Deduction, induction, and abduction are fundamental reasoning paradigms, core for human logical thinking. Although improving Large Language Model (LLM) reasoning has attracted significant research efforts, the extent to which the fundamental paradigms induce generalization has yet to be systematically explored. In this study, we shed light on how the interplay between these core paradigms influences LLMs' reasoning behavior. To this end, we first collect a new dataset of reasoning trajectories from symbolic tasks, each targeting one of the three fundamental paradigms, to abstract from concrete world knowledge. Then, we investigate effective ways for inducing these skills into LLMs. We experiment with a battery of methods including simple fine-tuning, and more complex approaches to increase model depth, or transform a dense model to a mixture-of-experts. We comprehensively evaluate induced models on realistic out-of-domain tasks, that are entirely formulated in natural language and contain real-world knowledge. Our results reveal that our approach yields strong generalizability with substantial performance gains (up to 14.60) across realistic tasks.

[Long] Gated Tree Cross-Attention for Checkpoint-Compatible Syntax Injection in Decoder-Only LLMs ▶ *Underline Video*

Xinyu Gao, Shaonan Wang, Li Zhao and Nai Ding

Decoder-only large language models achieve strong broad performance but are brittle to minor grammatical perturbations, undermining reliability for downstream reasoning. However, directly injecting explicit syntactic structure into an existing checkpoint can interfere with its pretrained competence. We introduce a checkpoint-compatible gated tree cross-attention (GTCA) branch that reads precomputed constituency chunk memory while leaving backbone architecture unchanged. Our design uses a token update mask and staged training to control the scope and timing of structural updates. Across benchmarks and transformer backbones, GTCA strengthens syntactic robustness beyond continued-training baselines without compromising Multiple-Choice QA performance or commonsense reasoning, providing a practical checkpoint-compatible route to more syntax-robust decoder-only LLMs.

[Long] GLARE: Agentic Reasoning for Legal Judgment Prediction ▶ *Underline Video*

Xinyu Yang, Chenlong Deng and Zhicheng Dou

Legal judgment prediction serves as a pivotal task in intelligent judicial systems. Although large language models have achieved remarkable progress in general reasoning, they struggle with tasks that require fine-grained distinctions between similar charges. These models often select plausible charges directly without discriminating among closely related alternatives. In this paper, we introduce GLARE, an agentic legal reasoning framework that enables models to actively retrieve and apply external knowledge during decision-making. Unlike static prediction, GLARE simulates comparative reasoning by dynamically expanding the decision space to include confusing candidates, then retrieving exclusionary logic from precedents and statutes to identify the correct judgment. Experiments on real-world datasets show that our method significantly outperforms strong baselines, especially on complex cases involving confusing or rare charges. The code is available at https://anonymous.4open.science/r/GLARE-LJP-8EDF.

[Findings] GoT-R1: Internalizing Graph-of-Thought via Structural Reinforcement for High-Density Reasoning ▶ *Underline Video*

Zuchao Li, Qiwei Li, Yao Yao, Hai Zhao, Lefei Zhang and Bo Du

Chain-of-Thought (CoT) reasoning, while effective, suffers from an inherent mechanism flaw: linearity induces overthinking. Constrained by sequential generation, models often produce redundant narration and circular self-corrections to maintain logical context. We propose GoT-R1, a framework that fundamentally mitigates this by replacing verbose linear trajectories with high-density reasoning graphs. Unlike CoT, GoT-R1 decouples logic from narration, modeling deliberation as a structured topology of atomic units. We internalize this inductive bias via a two-stage regimen: synthesizing structural data to distill logical skeletons, followed by Group Relative Policy Optimization (GRPO) to explicitly reinforce topological integrity. Extensive evaluations across mathematical reasoning and instruction following demonstrate that GoT-R1 consistently outperforms state-of-the-art baselines. Crucially, it achieves these gains with significantly reduced token overhead, demonstrating that structured reasoning density offers a more robust and parsimonious alternative to the recursive verbosity of standard CoT. The GoT-R1 models are open-sourced on Hugging Face at: https://huggingface.co/collections/MYTH-Lab/got-r1.

[Long] Guaranteeing Knowledge Integration with Joint Decoding for Retrieval-Augmented Generation ▶ *Underline Video*

Zhengyi Zhao, Shubo Zhang, Zezhong Wang, Yuxi Zhang, Huimin Wang, Yutian Zhao, Yefeng Zheng, Binyang Li, Kam-Fai Wong and Xian Wu

Retrieval-Augmented Generation (RAG) significantly enhances Large Language Models (LLMs) by providing access to external knowledge. However, current research primarily focuses on retrieval quality, often overlooking the critical "integration bottleneck": even when relevant documents are retrieved, LLMs frequently fail to utilize them effectively due to conflicts with their internal parametric knowledge. In this paper, we argue that implicitly resolving this conflict in a single generation pass is suboptimal. We introduce GuarantRAG, a framework that explicitly decouples reasoning from evidence integration. First, we generate an "Inner-Answer" based solely on parametric knowledge to capture the model's reasoning flow. Second, to guarantee faithful evidence extraction, we generate a "Refer-Answer" using a novel Contrastive DPO objective. This objective treats the parametric Inner-Answer as a negative constraint and the retrieved documents as positive ground truth, forcing the model to suppress internal hallucinations in favor of external evidence during this phase. Finally, rather than naive concatenation or using the DPO trained model directly, we propose a joint decoding mechanism that dynamically fuses the logical coherence of the Inner-Answer with the factual precision of the Refer-Answer at the token level. Experiments on five QA benchmarks demonstrate that GuarantRAG improves accuracy by up to 12.1% and reduces hallucinations by 16.3% compared to standard and dynamic RAG baselines.

[Findings] HintPilot: LLM-based Compiler Hint Synthesis for Code Optimization ▶ *Underline Video*

Hanyun Jiang, Peisen Yao, Kaiyue Li, Tingting Lin, Chengpeng Wang and Kui Ren

Code optimization remains a core objective in software development, yet modern compilers struggle to navigate the enormous optimization spaces. While recent research has looked into employing large language models (LLMs) to optimize source code directly, these techniques can introduce semantic errors and miss fine-grained compiler-level optimization opportunities. We present HintPilot, which bridges LLM-based reasoning with traditional compiler infrastructures via synthesizing compiler hints, annotations that steer compiler behavior. HintPilot employs retrieval-augmented synthesis over compiler documentation and applies profiling-guided iterative refinement to synthesize semantics-preserving and effective hints. Upon PolyBench and HumanEval-CPP benchmarks, HintPilot achieves up to 6.88x geometric mean speedup over -Ofast while preserving program correctness.

[Long] Hyperion: Private Token Sampling with Homomorphic Encryption ▶ *Underline Video*

Lawrence Lim, Jiaming Liu, Vikas Kalagi, Divyakant Agrawal and Amr El Abbadi

A promising direction for enabling private queries to large language models (LLMs) is with homomorphic encryption (HE). An open problem is performing token sampling under HE. In this paper, we introduce Hyperion, an efficient HE algorithm for inverse transform sampling, enabling private token sampling with 1 comparison depth, O(1) amortized comparisons, and O(n) rotations. We implement our approach and demonstrate that it samples tokens in 0.14 seconds for 32k tokens (~ 4.4\{ } mu s per token) on GPU, achieving a 100x latency improvement over prior work.

[Long] Injecting Context via Situation Working Memory for Logical Reasoning with LLMs ► *Underline Video*

Jieun Kim, Seoha Lim, YoungHae Choi and Sung-Bae Cho

Recent advances in large language models (LLMs) have improved logical reasoning by injecting formal logic or explicit structured representations. However, such methods often lose track of what is true now in multi-step reasoning, failing to maintain a coherent global state and its logical consequences. Motivated by Situation Model Theory in cognitive psychology, which views comprehension as constructing and updating a mental model of events along key dimensions (time, space, causality, intention, protagonist), we propose Situation Working Memory (SituW), a cognitively inspired method for contextual reasoning in LLMs. SituW first builds a situation representation by decomposing text along these five dimensions, then guides LLM inference with this evolving state. Keeping an explicit, dynamically updated situation memory instead of a static logical form encourages globally consistent reasoning over the situation model rather than raw text. Evaluated in both supervised and unsupervised settings, SituW improves accuracy by 23.3%p and 15.93%p while reducing uncertain predictions, suggesting that explicit situation modeling supports more globally consistent LLM reasoning.

[Long] Interpretable Traces, Unexpected Outcomes: Investigating the Disconnect in Trace-Based Knowledge Distillation ► *Underline Video*

Siddhant Bhambri, Upasana Biswas and Subbarao Kambhampati

Recent advances in reasoning-oriented Large Language Models (LLMs) have been driven by the introduction of Chain-of-Thought (CoT) traces, where models generate intermediate reasoning traces before producing an answer. These traces, as in DeepSeek R1, are not only used to guide model inference but also serve as supervision signals for Knowledge Distillation (KD) to improve smaller models. A prevailing but under-examined implicit assumption is that these CoT traces when emitted at inference time are both semantically correct and interpretable for the end-users. While there are reasons to believe that these intermediate tokens help improve solution accuracy, in this work, we question their validity (semantic correctness) and interpretability to the end user. To isolate the effect of trace semantics, we design experiments in the Question Answering (QA) domain using a rule-based problem decomposition method. This enables us to create Supervised Fine-Tuning (SFT) datasets for LLMs where - each QA problem is paired with either verifiably correct or incorrect CoT traces, while always providing the correct final solution. Trace correctness at inference time is then evaluated by checking the accuracy of every sub-step in decomposed reasoning chains. To assess end-user interpretability, we finetune LLMs with three additional types of CoT traces: R1 traces, R1 trace summaries, and post-hoc explanations of R1 traces. We further conduct a human-subject study with 100 participants asking them to rate the interpretability of each trace type on a standardized Likert scale. Our experiments reveal two key findings - (1) CoT trace correctness is not reliably correlated with the models generation of correct final answers: correct traces led to correct solutions only for 28% test-set problems while incorrect traces don't necessarily degrade solution accuracy. (2) In end-user interpretability studies, fine-tuning on verbose R1 traces produced the best model performance but these traces were rated as least interpretable by users, scoring on average 3.39 for interpretability and 4.59 for cognitive load metrics on a 5-point Likert scale. In contrast, the decomposed traces that are judged significantly more interpretable don't lead to comparable solution accuracy. Together, these findings challenge the assumption in question suggesting that researchers and practitioners should decouple model supervision objectives from end-user-facing trace design.

[Findings] Is Chain-of-Thought Reasoning of LLMs a Mirage? A Data Distribution Lens ► *Underline Video*

Chengshuai Zhao, Zhen Tan, Pingchuan Ma, Dawei Li, Bohan Jiang, Yancheng Wang, Yingzhen Yang and Huan Liu

Chain-of-Thought (CoT) prompting has been shown to be effective in eliciting structured reasoning (i.e., CoT reasoning) from large language models (LLMs). Regardless of its popularity, recent studies expose its failures in some reasoning tasks, raising fundamental questions about the nature of CoT reasoning. In this work, we propose a data distribution lens to understand when and why CoT reasoning succeeds or fails. We hypothesize that CoT reasoning reflects a structured inductive bias learned from in-distribution data, enabling models to conditionally generate reasoning trajectories that approximate those observed during training. As such, the effectiveness of CoT reasoning is fundamentally governed by the nature and degree of distribution discrepancy between training data and test queries. Guided by this lens, we dissect CoT reasoning via three dimensions: task, length, and format. To test the hypothesis, we introduce DataAlchemy, an abstract and fully controllable environment that trains LLMs from scratch and systematically probes them under various distribution conditions. Through rigorous controlled experiments, we reveal that CoT reasoning is a brittle mirage when it is pushed beyond training distributions, emphasizing the ongoing challenge of achieving genuine and generalizable reasoning.

[Long] I$B-LPO: Latent Policy Optimization via Iterative Information Bottleneck ► *Underline Video*

Huilin Deng, Hongchen Luo, Yue Zhu, Long Li, Zhuoyue Chen, Xinghao Zhao, MING LI, Chuyang Zhao, Jihai Zhang, MengChang WANG, Yang Cao and Yu Kang

Despite recent advances in Reinforcement learning with verifiable rewards (RLVR) for large language model (LLM) reasoning, most methods suffer from exploration collapse, as the semantic homogeneity of random rollouts traps models in narrow, over-optimized behaviors. Existing methods leverage policy entropy to encourage exploration, but face inherent limitations: global entropy regularization is susceptible to reward

hacking, inducing meaningless verbosity, whereas local token-selective updates struggle with the strong inductive bias of pre-trained models. To this end, we propose Latent Policy Optimization via Iterative Information Bottleneck (IšB-LPO), which shifts from statistical perturbation of token distributions to topological branching of reasoning trajectories. IšBLPO triggers latent branching at highentropy states to diversify reasoning trajectories and applies the Information Bottleneck as a trajectory filter and self-reward to ensure concise and informative exploration. Empirical results on four mathematical benchmarks demonstrate that IšB-LPO achieves state-of-the-art performance, with margins of up to 5.3% in accuracy and 7.4% in diversity metrics. Code is available at https://github.com/denghuilin-cyber/IIB-LPO.

[Long] KASER: Knowledge-Aligned Student Error Simulator for Open-Ended Coding Tasks ▶ *Underline Video*

Zhangqi Duan, Nigel Fernandez and Andrew Lan

Open-ended tasks, such as coding problems that are common in computer science education, provide detailed insights into student knowledge. However, training large language models (LLMs) to simulate and predict possible student errors in their responses to these problems can be challenging: they often suffer from mode collapse and fail to fully capture the diversity in syntax, style, and solution approach in student responses. In this work, we present KASER (Knowledge-Aligned Student Error Simulator), a novel approach that aligns errors with student knowledge. We propose a training method based on reinforcement learning using a hybrid reward that reflects three aspects of student code prediction: i) code similarity to the ground-truth, ii) error matching, and iii) code prediction diversity. On two real-world datasets, we perform two levels of evaluation and show that: At the per-student-problem pair level, our method outperforms baselines on code and error prediction; at the per-problem level, our method outperforms baselines on error coverage and simulated code diversity.

[Long] LaoBench: A Large-Scale Multidimensional Lao Benchmark for Large Language Models ▶ *Underline Video*

Jian Gao, Richeng Xuan, Zhaolu Kang, Dingshi Liao, Wenxin Huang, Zongmou Huang, Yangdi Xu, Bowen Qin, Zheqi He, Xi Yang, Changjin Li and Yonghua Lin

The rapid advancement of large language models (LLMs) has not been matched by their evaluation in low-resource languages, especially Southeast Asian languages like Lao. To fill this gap, we introduce LaoBench, the first large-scale, high-quality, and multidimensional benchmark for assessing LLM language understanding and reasoning in Lao. LaoBench contains 17,000+ expert-curated samples across three dimensions: culturally grounded knowledge application, curriculum-aligned K12 education, and bilingual translation among Lao, Chinese, and English. It includes open-source and held-out subsets, where the held-out portion enables secure black-box evaluation via a controlled service to improve fairness and data security. We construct LaoBench with a hybrid pipeline that combines expert authoring with agent-assisted verification, ensuring linguistic accuracy, cultural relevance, and educational validity. We evaluate diverse state-of-the-art open-source and closed-source LLMs, and find that even strong multilingual models lag behind human experts, particularly in culturally grounded reasoning and translation fidelity. We hope LaoBench will catalyze research on Lao and other underrepresented Southeast Asian languages for more inclusive multilingual evaluation.

[Long] Learning More from Less: Exploiting Counterfactuals for Data-Efficient Chart Understanding ▶ *Underline Video*

Jianzhu Bao, Haozhen Zhang, Kuicai Dong, Bozhi Wu, Sarthak Ketanbhai Modi, Zi Pong Lim, Yon Shin Teo and Wenya Wang

Vision-Language Models (VLMs) have demonstrated remarkable progress in chart understanding, largely driven by supervised fine-tuning (SFT) on increasingly large synthetic datasets. However, scaling SFT data alone is inefficient and overlooks a key property of charts: charts are programmatically generated visual artifacts, where small, code-controlled visual changes can induce drastic shifts in semantics and correct answers. Learning this counterfactual sensitivity requires VLMs to discriminate fine-grained visual differences, yet standard SFT treats training instances independently and provides limited supervision to enforce this behavior. To address this, we introduce ChartCF, a data-efficient training framework designed to enhance counterfactual sensitivity. ChartCF consists of: (1) a counterfactual data synthesis pipeline via code modification, (2) a chart similarity-based data selection strategy that filters overly difficult samples for improved training efficiency, and (3) multimodal preference optimization across both textual and visual modalities. Experiments on five benchmarks show that ChartCF achieves superior or comparable performance to strong chart-specific VLMs while using significantly less training data.

[Findings] Leveraging External Knowledge for Historical Document Restoration via Retrieval-Augmented Large Language Models ▶ *Underline Video*

Gabeen Kim and Kyeongpil Kang

Historical documents act as invaluable knowledge archives but often suffer from illegibility due to physical deterioration and damage. While existing restoration methods based on masked language modeling effectively utilize local context, they struggle to restore named entities that require external historical knowledge. To address this limitation, we introduce a novel framework for historical document restoration that leverages large language models with retrieval-augmented generation (RAG). By combining the implicit knowledge of pre-trained LLMs with explicitly retrieved external context, our model ARI effectively mitigates the challenge of inferring context-dependent proper nouns. Extensive experiments on Korean historical documents demonstrate that our approach significantly outperforms baselines, achieving substantial gains in restoring both general characters and named entities. Furthermore, comprehensive evaluations including expert assessments confirm that ARI serves as a practical tool for domain experts, promising to accelerate the analysis of historical records.

[Findings] LSEG: A Fine-tuning Free Method for NL2FOL via Logic-Structure and Entropy Guided Inference Controlling ▶ *Underline Video*

Yuwei Huang, Shi Wang, Kangli Zi, Tianyu Luo, Jiang Zhixiao, Ruiheng Wang and Yufei Wang

Large language models have shown strong generative and reasoning capabilities, yet they still struggle with natural language to first order logic (NL2FOL) translation due to logical hallucination. We propose LSEG (Logic Structure and Entropy Guided), a fine-tuning free framework designed to improve logical consistency during inference. The core idea of LSEG is to correct hidden state deviation by leveraging logical stability across logic preserving perturbations of the input. Such deviation is especially harmful in NL2FOL, as even small drifts can flip quantifier scope or logical operators, producing formulas that are syntactically valid yet logically incorrect. First, LSEG constructs perturbation-averaged direction vectors that approximate a stable logical center. Second, it derives layer-wise correction directions by contrasting original and perturbed representations. Lastly, LSEG uses an entropy-guided adaptive mechanism to inject these directions only when the model exhibits unstable or over-confident reasoning states, thereby preserving fluency while correcting logical drift. Experiments on the FOLIO and MALLS benchmarks show that LSEG consistently improves logical equivalence scores over strong baselines, despite requiring no training or parameter updates. Further evaluation on LogicLLaMA demonstrates LSEGs architecture-agnostic effectiveness.

[Findings] MIThinker: A Plug-and-Play Policy-Optimized Thinker For Motivational Interviewing Counseling ► *Underline Video*

Yizhe Yang, Palakorn Achananuparp, Heyan Huang, Jing Jiang and Ee-Peng Lim

Reasoning large language models (LLMs) have recently made much progress in complex problem-solving, leveraging internal reasoning (or thought) to guide their solution generation. However, existing LLM-based counseling agents, including those using Motivational Interviewing (MI), generate responses without explicitly aligning thoughts with counseling techniques, limiting their effectiveness. We propose MIThinker, a lightweight thinking model that generates therapeutic thoughts to guide MI counseling agents in strategy selection and response generation. To overcome the lack of annotated thought data, we introduce AugR1-MI, an automated pipeline that reverse-engineers counselor's thoughts from observed responses. Through two-stage training combining supervised fine-tuning and reinforcement learning, MIThinker demonstrates improved theory-of-mind assessment and strategy alignment. Comprehensive evaluations show that MindfulMI, our agent leveraging MIThinker, achieves MI competency comparable to state-of-the-art systems with an order of magnitude less computation.

[Long] Mitigating Lost in Multi-turn Conversation via Curriculum RL with Verifiable Accuracy and Abstention Rewards ► *Underline Video*

Ming Li

Large Language Models demonstrate strong capabilities in single-turn instruction following but suffer from Lost-in-Conversation (LiC), a degradation in performance as information is revealed progressively in multi-turn settings. Motivated by the current progress on Reinforcement Learning with Verifiable Rewards (RLVR), we propose Curriculum Reinforcement Learning with Verifiable Accuracy and Abstention Rewards (RLAAR), a framework that encourages models not only to generate correct answers, but also to judge the solvability of questions in the multi-turn conversation setting. Our approach employs a competence-gated curriculum that incrementally increases dialogue difficulty (in terms of instruction shards), stabilizing training while promoting reliability. Using multi-turn, on-policy rollouts and a mixed-reward system, RLAAR teaches models to balance problem-solving with informed abstention, reducing premature answering behaviors that cause LiC. Evaluated on LiC benchmarks, RLAAR significantly mitigates LiC performance decay (62.6% to 75.1%) and improves calibrated abstention rates (33.5% to 73.4%). Together, these results provide a practical recipe for building multi-turn reliable and trustworthy LLMs.

[Long] More Aligned, Less Diverse? Analyzing the Grammar and Lexicon of Two Generations of LLMs ► *Underline Video*

Adrián Gude, Roi Santos-Rios, Francis Bond, Dan Flickinger, Carlos Gómez-Rodríguez and Olga Zamaraeva

This study contributes to a growing line of research in comparing LLM-generated texts with human-authored text, in this case, English news text. We focus in particular on the evaluation of syntactic properties through formal grammar frameworks. Our analysis compares two generations of LLMs in the context of two human-authored English news datasets from two different years. Employing the Head-Driven Phrase Structure Grammar (HPSG) formalism, we investigate the distributions of syntactic structures and lexical types of AI-generated texts and contrast them with the corresponding distributions in the human-authored New Your Times (NYT) articles. We use diversity metrics from ecology and information theory to quantify variation in grammatical constructions and lexical types. Our results show that, while English news text has changed little in the given time frame, newer, instruction-tuned LLMs display reduced syntactic and, especially, lexical diversity compared to older, non-instruction-tuned models. These findings point to future work in studying effects of instruction tuning, which, while enhancing coherence and adherence to prompts, may narrow the expressive range of model output.

[Findings] MR-ALIGN: Meta-Reasoning Informed Factuality Alignment for Large Reasoning Models ► *Underline Video*

Xinming Wang, Jian Xu, Bin Yu, Sheng Lian, Yi Chen, Boran Wang, Yingjian Zhu, Hongzhu Yi, Hong-Ming Yang, Han Hu, Cheng-Lin Liu and Xu-Yao Zhang

Large reasoning models (LRMs) show strong capabilities in complex reasoning, yet their marginal gains on evidence-dependent factual questions are limited. We find this limitation is partially attributable to a reasoning-answer hit gap, where the model identifies the correct facts during reasoning but fails to incorporate them into the final response, thereby reducing factual fidelity. To address this issue, we propose MR-ALIGN, a Meta-Reasoning informed alignment framework that enhances factuality without relying on external verifiers. MR-ALIGN quantifies state-transition probabilities along the models thinking process and constructs a transition-aware implicit reward that reinforces beneficial reasoning patterns while suppressing defective ones at the atomic thinking segments. This re-weighting reshapes token-level signals into probability-aware segment scores, encouraging coherent reasoning trajectories that are more conducive to factual correctness. Empirical evaluations across four factual QA datasets and one long-form factuality benchmark show that MR-ALIGN consistently improves

accuracy and truthfulness while reducing misleading reasoning. These results highlight that aligning the reasoning process itself, rather than merely the outputs, is pivotal for advancing factuality in LRMs.

[Findings] MTMCS-Bench: Evaluating Contextual Safety of Multimodal Large Language Models in Multi-Turn Dialogues ▶ *Underline Video*

Zheyuan Liu, Dongwhi Kim, Yixin Wan, Xiangchi Yuan, Zhaoxuan Tan, Fengran Mo and Meng Jiang

Multimodal large language models (MLLMs) are increasingly deployed as assistants that interact through text and images, making it crucial to evaluate contextual safety when risk depends on both the visual scene and the evolving dialogue. Existing contextual safety benchmarks are mostly single-turn and often miss how malicious intent can emerge gradually or how the same scene can support both benign and exploitative goals. We introduce the Multi-Turn Multimodal Contextual Safety Benchmark (MTMCS-Bench), a benchmark of realistic images and multi-turn conversations that evaluates contextual safety in MLLMs under two complementary settings, escalation-based risk and context-switch risk. MTMCS-Bench offers paired safe and unsafe dialogues with structured evaluation. It contains over 30 thousand multimodal (image+text) and unimodal (text-only) samples, with metrics that separately measure contextual intent recognition, safety-awareness on unsafe cases, and helpfulness on benign ones. Across eight open-source and seven proprietary MLLMs, we observe persistent trade-offs between contextual safety and utility, with models tending to either miss gradual risks or over-refuse benign dialogues. Finally, we evaluate five current guardrails and find that they mitigate some failures but do not fully resolve multi-turn contextual risks.

[Findings] Multilingual Tokenization through the Lens of Indian Languages: Challenges and Insights ▶ *Underline Video*

Maharaj Brahma, NJ Karthika, Rajat Verma, Nagasai Saketh Naidu, Rohit Saluja, Maunendra Sankar Desarkar and Ganesh Ramakrishnan

Tokenization plays a pivotal role in NLP and is fundamental to training language models. However, existing tokenizers are often skewed towards high-resource languages, limiting their effectiveness for linguistically diverse and morphologically rich languages such as those in the Indian subcontinent. In this work, we present a comprehensive empirical study of multilingual tokenization across 17 Indic languages spanning 11 scripts and two language families. We systematically evaluate the effects of (i) widely used subword algorithms: BPE sennrich-etal-2016-neural and Unigram LM kudo-2018-subword, (ii) script and orthography-aware normalization, (iii) vocabulary size, and (iv) multilingual vocabulary construction strategies. We use a combination of intrinsic and extrinsic evaluations to obtain the following observations: (i) script-specific normalization improves tokenization quality, (ii) Unigram LM better preserves morphological boundaries than BPE, (iii) cluster-based vocabulary construction shows improvement in downstream tasks compared to the joint method. Our findings highlight the importance of linguistically informed design choices in multilingual tokenization and offer practical guidance for building effective tokenizers for low-resource and morphologically complex languages.

[Findings] Multiplication in Multimodal LLMs: Computation with Text, Image, and Audio Inputs ▶ *Underline Video*

Samuel G. Balter, Ethan Jerzak and Connor T. Jerzak

Multimodal LLMs can accurately perceive numerical content across modalities yet fail to perform exact multi-digit multiplication when the identical underlying arithmetic problem is presented as numerals, number words, images, or in audio form. Because existing benchmarks often lack systematically paired instances across modalities, it remains difficult to compare genuine arithmetic limits within and across model families. We therefore introduce a controlled multimodal multiplication benchmark that factorially varies digit length, digit sparsity, representation (e.g., numerals vs. number words), and modality (text, rendered images, and audio), with paired instances from a reproducible generator. We also define arithmetic load, C, as the product of the total and non-zero digit number as a compact, mechanistically motivated proxy for operation count. Across evaluations, accuracy falls sharply as C grows, often nearing zero by C > 100. Indeed, C remains predictive of performance across modalities and models, with R^2 > 0.5, nearing the value from more complex measures of arithmetic load that count the number of intermediate arithmetic steps. A separate perception-versus-computation decomposition shows that multimodal degradation is primarily computational rather than perceptual: on matched perception checks, models are near-perfect (>99%) across modalities even when multiplication accuracy drops substantially. Beyond measuring when models fail, we ask which procedures they are predisposed to follow. We introduce a style-controlled forced-completion loss probe that scores heuristic-specific reasoning prefixes, including columnar multiplication, distributive decomposition, and rounding/compensation. Here, distributive decomposition is favored in both text and vision modalities; heuristic-specific LoRA adapters produce near-orthogonal updates yet degrade accuracy, indicating the base model maintains a well-tuned internal router.

[Findings] Neural Induction of Finite-State Transducers ▶ *Underline Video*

Michael Ginn, Alexis Palmer and Mans Hulden

Finite-State Transducers (FSTs) are effective models for string-to-string rewriting tasks, often providing the efficiency necessary for high-performance applications, but constructing transducers by hand is difficult. In this work, we propose a novel method for automatically constructing unweighted FSTs following the hidden state geometry learned by a recurrent neural network. We evaluate our methods on real-world datasets for morphological inflection, grapheme-to-phoneme prediction, and historical normalization, showing that the constructed FSTs are highly accurate and robust for many datasets, massively outperforming classical transducer learning algorithms by up to 87% accuracy on held-out test sets.

[Findings] ODTQA-FoRe: An Open-Domain Tabular Question Answering Dataset for Future Data Forecasting and Reasoning ▶ *Underline Video*

Zhensheng Wang, Xiaole Liu, Wenmian Yang, Kun Zhou, Yiquan Zhang and Weijia Jia
The rapid development of LLMs has significantly advanced tabular question answering, but most systems cannot perform future-oriented numerical prediction. To address this gap, we introduce a novel task, Open-Domain Tabular Question Answering for Future Data Forecasting and Reasoning, and propose the first dataset to cover time-series forecasting and forecast-based reasoning scenarios using real estate data. This task poses challenges in retrieving precise historical data, overcoming the forecasting limitations of LLMs, and standardizing responses for diverse queries. To solve the above challenges, we propose TimeFore, an LLM agent-based framework that decomposes the problem into three collaborative roles: a Retriever autonomously generates SQL to fetch data, a Forecaster invokes external time-series models for higher accuracy, and an Analyzer synthesizes the results to construct a precise and consistent final answer. Extensive experiments demonstrate the effectiveness of our TimeFore.

[Long] On the Emergence and Test-Time Use of Structural Information in Large Language Models ▶ *Underline Video*
Michelle Chao Chen, Moritz Miller, Bernhard Schölkopf and Siyuan Guo
Learning structural information from observational data is central to producing new knowledge outside the training corpus. This holds for mechanistic understanding in scientific discovery as well as flexible test-time compositional generation. We thus study how language models learn abstract structures and utilize the learnt structural information at test-time. To ensure a controlled setup, we design a natural language dataset based on linguistic structural transformations. We empirically show that the emergence of learning structural information correlates with complex reasoning tasks, and that the ability to perform test-time compositional generation remains limited.

[Long] Open Your Models Eyes: Video and Context-Aware Multimodal Backchannel Prediction ▶ *Underline Video*
Min-Jae Kim, Jun-Yeong Moon, Mujeen Sung and Gyeong-Moon Park
Backchannels, which signal listener states like empathy and understanding, are fundamental to natural human interaction. However, current approaches rely solely on audio and text. This omits crucial visual cues, such as facial expressions and gestures, as well as broader conversational contexts, which are necessary for accurate prediction. In this paper, we introduce Context-Aware Multimodal Alignment for Backchannel Prediction (CAMA-BC), a novel framework that leverages visual information through Multi-layer Multimodal Alignment (MMA). Our alignment process comprises two stages. First, Context Alignment (MMA-CA) utilizes unlabeled dialogues with videos to capture conversational contexts. Next, Backchannel Alignment (MMA-BA) fine-tunes the representations specifically for backchannel prediction. Experimental results show that CAMA-BC significantly outperforms both existing methods and simple multimodal baselines, with particular effectiveness in recognizing complex backchannels such as empathy.

[Findings] PBEBench: A Multi-Step Programming by Examples Reasoning Benchmark inspired by Historical Linguistics ▶ *Underline Video*
Atharva Naik, Prakam Dhiman, Yash Mathur, Darsh Agrawal, Manav Kapadnis, Yuwei An, Clayton Marr, Carolyn Rose and David R. Mortensen
While many benchmarks evaluate the reasoning abilities of Large Language Models (LLMs), few isolate reasoning as a capability independent of domain knowledge. We introduce a new benchmark for inductive reasoning inspired by Sound Law Induction (SLI) in historical linguistics and formulated in a simple multi-step Programming by Example (PBE) framework. The task requires inducing a cascade of string rewrite programs that transform inputs into target outputs. We present PBEBench, a fully automated evaluation approach that generates such problems with controllable difficulty and ordering constraints, enabling scalable and contamination-resistant evaluation of sequential inductive reasoning. Using this approach, we construct three datasets that show a large gap between models that leverage test-time compute or long chain-of-thought reasoning and those that do not. Although recent models such as GPT-5 and gpt-oss-120b show promise, solve rates remain below 5% on hard PBEBench instances with long program cascades, even under computationally expensive scaling strategies. Finally, we show that PBEBench scores are more predictive of performance on real SLI than are other inductive reasoning benchmarks. We will release code and data to support further research.

[Long] PILOT: Planning via Internalized Latent Optimization Trajectories for Large Language Models ▶ *Underline Video*
Haoyu Zheng, Yun Zhu, Yuqian Yuan, Bo Yuan, Wenqiao Zhang, Siliang Tang and Jun Xiao
Strategic planning is critical for multi-step reasoning, yet compact Language Language Models (LLMs) often lack the capacity to formulate global strategies, leading to error propagation in long-horizon tasks. Our analysis reveals that LLMs possess latent reasoning capabilities that can be unlocked when conditioned on explicit plans from a teacher model; however, runtime reliance on external guidance is often impractical due to latency and availability constraints. To bridge this gap, we propose PILOT (Planning via Internalized Latent Optimization Trajectories), a non-invasive framework designed to internalize the strategic oversight of large models into intrinsic Latent Guidance. Instead of altering backbone weights, PILOT employs a lightweight Hyper-Network to synthesize a query-conditioned Latent Guidance. This vector acts as an internal steering mechanism, guiding the models representations toward optimal reasoning paths. Extensive experiments on mathematical and coding benchmarks demonstrate that PILOT effectively stabilizes reasoning trajectories, consistently outperforming strong baselines (e.g., +8.9% on MATH500) with negligible inference latency. Our code is available at: https://anonymous.4open.science/r/PILOT-B266

[Findings] PROPER Agents: Proactivity Driven Personalized Agents for Advancing Knowledge Gap Navigation ▶ *Underline Video*
Kirandeep Kaur, Vinayak Gupta, Aditya Gupta and Chirag Shah
Current approaches to proactive assistance move beyond the ask-and-respond paradigm by anticipating user needs. In practice, they either

burden users with clarifying questions or rely on context-based extrapolation, often leading to unnecessary or mistimed interventions. Such systems lack explicit mechanisms to model users knowledge gaps, resulting in incomplete or suboptimal task outcomes. To address this, we propose PROPER, a framework that explicitly models user-specific knowledge gaps in a controlled manner. Central to our approach is the notion of dimensions: structured, task-relevant factors that define the considerations required for effective task completion. Given a user query, the DGA (Dimension Generating Agent) identifies explicit dimensions (from the users query) and generates a set of candidate implicit dimensions capturing unarticulated aspects of the task. The RGA (Response Generating Agent) integrates both explicit and implicit dimensions selectively to produce personalized, context-aware, and proactively informative responses. We evaluate PROPER across multiple domains using a structured, gap-aware rubric that measures coverage, initiative appropriateness, and intent alignment. PROPER improves on quality scores and win rates across all domains, achieving up to 84% gains in single-turn evaluation and consistent dominance in multiturn interactions. All code for PROPER is available at: https://github.com/i-kiran/ProPer-Agent.

[Findings] Query-Focused Individual Simulation with Progressive Persona Completion ▶ *Underline Video*

Weiwen SU, Naoki Yoshinaga and Masashi Toyoda

Large language models (LLMs) enable simulating individual responses from persona information, supporting applications such as opinion elicitation and virtual character creation. However, existing approaches typically assume rich persona profiles, which are often unavailable in practice. In this work, motivated by recent findings that LLMs can identify query-relevant persona dimensions (e.g., whether a user is price-sensitive), we study query-focused individual simulation under cold-start settings, where relevant persona information is identified and requested on demand for each query. To solve this task while minimizing the number of persona requests, we explore a progressive method that iteratively predicts the most critical relevant persona dimension and uses self-reported confidence as a stopping signal to determine when sufficient information has been collected. Experiments on two dialogue datasets show that this query-driven paradigm achieves simulation performance comparable to approaches that rely on rich persona information extracted from dialogue history, using only a few persona dimensions (up to five per query), and this number is further reduced by our progressive method while maintaining or improving simulation quality.

[Findings] ReSQL: Self-Improving Framework for Reasoning-Aware Text-to-SQL Dataset Generation ▶ *Underline Video*

Minjun Park, Yongju Seong, Myoseop Sim, Kyungkoo Min and Stanley Jungkyu Choi

Recent advances in Text-to-SQL have greatly benefited from large language models, yet small and medium-sized models still suffer from frequent execution errors and limited self-correction ability. We present ReSQL (Retrieval-augmented error reasoning for Text-to-SQL), a self-improving framework that generates and learns from its own error-reasoning dataset, enabling models to autonomously refine their SQL generation and correction capabilities. ReSQL combines feedback-driven fine-tuning with retrieval-based inference: it gathers model-generated errors, analyzes them through structured feedback prompts, and retrieves relevant correction examples during inference. This unified approach allows models to internalize robust error-reasoning patterns and dynamically apply them to unseen queries. Experimental results on the SPIDER and BIRD benchmarks show that ReSQL substantially improves execution accuracy and self-correction ability over strong baselines, achieving competitive performance with much larger proprietary models such as GPT-4. Our findings highlight ReSQL as a promising step toward self-improving, reasoning-aware Text-to-SQL systems that can continually enhance their reliability and interpretability without external supervision. All code and generated reasoning datasets are available to facilitate application to open-source LLMs and reproducible baseline training.

[Long] RV-HATE: Reinforced Multi-Module Voting for Implicit Hate Speech Detection ▶ *Underline Video*

Yejin Lee, Hyeseon An and Yo-Sub Han

Hate speech remains prevalent in human society and continues to evolve in its forms and expressions. Modern advancements in the internet and online anonymity accelerate its rapid spread and complicate its detection. However, hate speech datasets exhibit diverse characteristics primarily because they are constructed from different sources and platforms, each reflecting different linguistic styles and social contexts. Despite this diversity, prior studies on hate speech detection often rely on fixed methodologies without adapting to data-specific features. We introduce RV-HATE, a detection framework designed to account for the dataset-specific characteristics of each hate speech dataset. RV-HATE consists of multiple specialized modules, where each module focuses on distinct linguistic or contextual features of hate speech. The framework employs reinforcement learning to optimize weights that determine the contribution of each module for a given dataset. A voting mechanism then aggregates the module outputs to produce the final decision. RV-HATE offers two primary advantages: (1) it improves detection accuracy by tailoring the detection process to dataset-specific attributes, and (2) it also provides interpretable insights into the distinctive features of each dataset. Consequently, our approach effectively addresses implicit hate speech and achieves superior performance compared to conventional static methods.

[Findings] Schema-Guided Response Generation using Multi-Frame Dialogue State for Motivational Interviewing Systems ▶ *Underline Video*

Jie Zeng and Yukiko Nakano

The primary goal of Motivational Interviewing (MI) is to help clients build their own motivation for behavioral change. To support this in dialogue systems, it is essential to guide large language models (LLMs) to generate counselor responses aligned with MI principles. By employing a schema-guided approach, this study proposes a method for updating multi-frame dialogue states and a strategy decision mechanism that dynamically determines the response focus in a manner grounded in MI principles. The proposed method was implemented

in a dialogue system on two different datasets and evaluated through a user study. Results showed that the proposed method successfully generates responses aligned with MI principle and frequently asks questions to elicit change talk.

[Long] SciCustom: A Framework for Custom Evaluation of Scientific Capabilities in Large Language Models ▶ *Underline Video*

Yiyang Gu, Junwei Yang, Junyu Luo, Ye Yuan, Bin Feng, Yingce Xia, Shufang Xie, Kaili Liu, Bohan Wu, Qi Shi, Haoran Li, Beier Xiao, Zhiping Xiao, Xiao Luo, Weizhi Zhang, Philip S. Yu, Zequn Liu and Ming Zhang

Large language models (LLMs) are increasingly applied to scientific research, yet existing evaluations often fail to reflect the fine-grained capabilities required in practice. Most benchmarks are manually curated or domain-generic, limiting scalability and alignment with real scientific use cases. In this paper, we propose a new framework named SciCustom to address the problem. It enables the custom construction of benchmarks from large-scale scientific data to evaluate application-specific scientific capabilities in LLMs. SciCustom first organizes scientific knowledge into ontology-grounded knowledge units with controlled granularity and trains a tagger to map large-scale data instances into this knowledge space. Given a custom requirement, relevant knowledge units are identified via voting-based multi-model consensus. These units enable relevance-aware benchmark retrieval via binary search, followed by proxy subset selection and data-grounded benchmark generation for efficient evaluation. Experiments in chemistry and healthcare demonstrate that SciCustom reveals fine-grained differences in LLM scientific capabilities that standard benchmarks overlook, while requiring neither expert annotation nor synthetic question generation. This work provides a scalable and application-aware foundation for benchmarking scientific capabilities in LLMs.

[Long] SciMDR: Advancing Scientific Multimodal Document Reasoning ▶ *Underline Video*

Ziyu Chen; Yilun Zhao; Chengye Wang; Rilyn R. Han; Manasi Patwardhan; Arman Cohan

Constructing scientific multimodal document reasoning datasets for foundation model training involves an inherent trade-off among scale, faithfulness, and realism. To address this challenge, we introduce the synthesize-and-reground framework, a two-stage pipeline comprising: (1) Claim-Centric QA Synthesis, which generates faithful, isolated QA pairs and reasoning on focused segments, and (2) Document-Scale Regrounding, which programmatically re-embeds these pairs into full-document tasks to ensure realistic complexity. We present SciMDR, a large-scale training dataset for cross-modal comprehension, comprising 300K QA pairs with explicit reasoning chains across 20K scientific papers. We further construct SciMDR-Eval, an expert-annotated benchmark to evaluate multimodal comprehension within full-length scientific workflows. Experiments demonstrate that models fine-tuned on SciMDR achieve significant improvements across multiple scientific QA benchmarks, particularly in tasks requiring complex document-level reasoning.

[Findings] SciText2Eq: Assessing LLMs for Explainable Equation Generation for Scientific Creativity ▶ *Underline Video*

Yifan Mo, Xiao Fu, Yue Su, Qingyu Meng, Koen Hindriks, Qingzhi Liu and Jiahuan Pei

This work investigates the ability of large language models (LLMs) to generate mathematical equations from scientific texts. Prior work faces challenges in unstructured grounding, multi-equation dependency, and human-aligned evaluation. To address this, we construct a dataset of AI research papers, pairing contextual passages with ground-truth equations and variable descriptions. We develop an explainable equation generation workflow and evaluate it across diverse open- and closed-source LLMs. Our evaluation protocol combines automatic metrics, LLM-based rubrics, and human judgments to assess accuracy, explainability, and human-LLM alignment. Results show that LLMs achieve moderate performance on lexical and syntactic similarity, but struggle with semantic accuracy. LLM-based evaluations show limited alignment with human judgments, highlighting challenges in assessing equation quality. These findings provide insights for improving equation generation models and developing more reliable evaluation methods for scientific creativity. We provide code and data for reproducibility.

[Findings] Self-EmoQ: Plutchik-Guided Value-based Planning to Drive Streaming Emotional TTS ▶ *Underline Video*

Yue Zhao, Hongyan Li, Yong Chen and Luo Ji

Emotional interaction is increasingly crucial for conversational AI, yet current systems lack a self-emotion determination mechanism to drive the streaming text-to-speech (TTS) synthesis. We propose an emotion-planning framework that determines the emotion prior to the textual generation, grounding the downstream emotional TTS in a streaming manner. The framework is implemented by a plug-and-play LLM module, initialized from pretrained LLMs, and trained by reinforcement learning (RL) with emotions as the actions. A hybrid reward is employed which combines imitation signals with theory-driven scoring, in which the theory of Plutchik's wheel of emotions is adopted. By experiments on DailyDialog, EmoryNLP, IMEOCAP, and MELD, our method outperforms prompting and finetuning baselines on both emotion determination and response quality. We finally implement an entire streaming pipeline for real-time deployment, with the speech quality confirming the framework's emotional alignment, contextual coherence, and expressive fluency. Codes, cases, and demos are available in https://sixingdeguo.github.io/EmoQ-page/.

[Findings] Self-Evolving Multi-Agent Systems via Textual Backpropagation ▶ *Underline Video*

Xiaowen Ma, Yunpu Ma, Chenyang Lin, Sikuan Yan, Jinhe Bi, Zixuan Cao, Yijun Tian, Volker Tresp and Hinrich Schuetze

Leveraging multiple Large Language Models (LLMs) has proven effective for addressing complex, high-dimensional tasks, but current approaches often rely on static, manually engineered multi-agent configurations. To overcome these constraints, we present the Agentic Neural Network (ANN), a framework that conceptualizes multi-agent collaboration as a layered neural network architecture. In this design, each agent operates as a node, and each layer forms a cooperative team focused on a specific subtask. The proposed framework follows a two-phase optimization strategy: (1) Forward Phase - Drawing inspiration from neural network forward passes, tasks are dynamically decomposed into subtasks, and cooperative agent teams with suitable aggregation methods are constructed layer by layer. (2) Backward

Phase - Mirroring backpropagation, we refine both global and local collaboration through iterative feedback, allowing agents to self-evolve their roles, prompts, and coordination. This neuro-symbolic approach enables our framework to create new or specialized agent teams post-training, delivering notable gains in accuracy and adaptability. Across seven benchmark datasets, ANN surpasses leading multi-agent baselines under the same configurations, showing consistent performance improvements.

[Long] Sign-Language Datasets at Scale: A Comprehensive Survey on Resources, Benchmarks, and Annotation Standards ▶ *Underline Video*

Yiming Ni, Zhi-Qi Cheng, Jiayu Li and Wei Cheng

Sign languages are expressive visual languages used by Deaf and Hard-of-Hearing (DHH) communities. Despite substantial progress in sign-language recognition, translation, and production, advances remain constrained by fragmented datasets, inconsistent annotations, and limited linguistic coverage. Existing benchmarks often fail to reflect real-world communication needs, and systematic analyses of these limitations remain limited. In this survey, we present a comprehensive index of sign-language datasets, covering 120 resources across 35 sign languages. We analyze key challenges such as modality imbalance, annotation granularity, and signer bias, and outline considerations for future dataset design. We also introduce a 24-field Sign-Language Datasheet and release a public GitHub repository to support standardized documentation and reproducible evaluation. Overall, our work provides a unified and practical foundation for developing inclusive, robust, and scalable sign-language technologies in real-world applications.

[Long] SpecMind: Cognitively Inspired, Interactive Multi-Turn Framework for Postcondition Inference ▶ *Underline Video*

Cuong Chi Le, Minh V.T. Pham, Tung D. Vu, Van Duc Cuong, Phan Nht Huy, PHAN NHAT HOANG and Tien N. Nguyen

Specifications are vital for ensuring program correctness, yet writing them manually remains challenging and time-intensive. Recent large language model (LLM)-based methods have shown successes in generating specifications such as postconditions, but existing single-pass prompting often yields inaccurate results. In this paper, we present SpecMind, a novel framework for postcondition generation that treats LLMs as interactive and exploratory reasoners rather than one-shot generators. SpecMind employs feedback-driven multi-turn prompting approaches, enabling the model to iteratively refine candidate postconditions by incorporating implicit and explicit correctness feedback, while autonomously deciding when to stop. This process fosters deeper code comprehension and improves alignment with true program behavior via exploratory attempts. Our empirical evaluation shows that SpecMind significantly outperforms state-of-the-art approaches in both accuracy and completeness of generated postconditions.

[Long] SPENCE: A Syntactic Probe for Detecting Contamination in NL2SQL Benchmarks ▶ *Underline Video*

Mohammadtaher Safarzadeh, Hitesh Laxmichand Patel, Afshin Oroojlooy, Graham Horwood and Dan Roth

Large language models (LLMs) have achieved strong performance on natural language to SQL (NL2SQL) benchmarks, yet their reported accuracy may be inflated by contamination from benchmark queries or structurally similar patterns seen during training. We introduce SPENCE (Syntactic Probing and Evaluation of NL2SQL Contamination Effects), a controlled syntactic probing framework for detecting and quantifying such contamination. SPENCE systematically generates syntactic variants of test queries for four widely used NL2SQL datasets, Spider, SParC, CoSQL, and the newer BIRD benchmark. We use SPENCE to evaluate multiple high-capacity LLMs under execution-based scoring. For each model, we measure changes in execution accuracy (Delta~Accuracy) across increasing levels of syntactic divergence and quantify rank sensitivity using Kendalls tau with bootstrap confidence intervals. By aligning these robustness trends with benchmark release dates, we observe a clear temporal gradient: older benchmarks such as Spider exhibit the strongest negative values and thus the highest likelihood of training leakage, whereas the more recent BIRD dataset shows minimal sensitivity and appears largely uncontaminated. Together, these findings highlight the importance of temporally contextualized, syntactic-probing evaluation for trustworthy NL2SQL benchmarking.

[Findings] TARo: Token-level Adaptive Routing for LLM Test-time Alignment ▶ *Underline Video*

Arushi Rai, Qiang Zhang, Hanqing Zeng, Yunkai Zhang, Dipesh Tamboli, Xiangjun Fan, Zhuokai Zhao and Lizhu Zhang

Large language models (LLMs) exhibit strong reasoning capabilities but typically require expensive post-training to reach high performance. Recent test-time alignment methods offer a lightweight alternative, but have been explored mainly for preference alignment rather than reasoning. To bridge this gap, we propose Token-level Adaptive Routing (TARo), which steers frozen LLMs toward structured reasoning entirely at inference time. Specifically, we first train reward models on step-wise mathematical traces to capture fine-grained logical consistency signals, then introduce a learnable token-level router that automatically controls the guidance of the reward model to the base model. Extensive experiments show that TARo significantly improves reasoning performance by up to +22.4% over base model and +8.4% over existing token-level test-time alignment methods, while also boosting out-of-distribution clinical reasoning (MedXpertQA) and instruction following (AlpacaEval). Furthermore, TARo also generalizes from small to large backbones without retraining, extending test-time alignment from preference optimization to robust, cross-domain reasoning.

[Findings] Teaching Language Models to Forecast Research Success Through Comparative Idea Evaluation ▶ *Underline Video*

Srujan Mule, Aniketh Garikaparthi and Manasi Patwardhan

As language models accelerate scientific research by automating hypothesis generation and implementation, a new bottleneck emerges: evaluating and filtering hundreds of AI-generated ideas without exhaustive experimentation. We ask whether LMs can learn to forecast the empirical success of research ideas before any experiments are run. We study comparative empirical forecasting: given a benchmark-specific

research goal and two candidate ideas, predict which will achieve better benchmark performance. We construct a dataset of 11,488 idea pairs grounded in objective outcomes from PapersWithCode. While off-the-shelf 8B-parameter models struggle (30% acc.), SFT dramatically boosts performance to 77.1%, outperforming GPT-5 (61.1%). By framing evaluation as a reasoning task via Reinforcement Learning with Verifiable Rewards (RLVR), we train models to discover latent reasoning paths, achieving 71.35% acc. with interpretable justifications. Through additional ablations and out-of-distribution tests, we show robustness to surface-level heuristics and transfer to both a cross-domain time-split test set and an independently constructed test set. Our results demonstrate that compute-efficient small language models can serve as effective, objective verifiers, offering a scalable path for autonomous scientific discovery.

[Findings] The Learnability of Model-Theoretic Interpretation Functions in Artificial Neural Networks ▶ *Underline Video*

Adrian Brasoveanu and Jakub Dotlacil

The systematicity of natural language interpretation, our ability to understand novel expressions by compositionally combining familiar elements, has been central to debates about symbolic versus neural approaches to cognition since Fodor and Pylyshyn (1988). We investigate whether artificial neural networks can learn model-theoretic interpretation functions that generalize systematically to out-of-training-sample sentences, framing interpretation as an encoding task from discrete linguistic input to continuous truth-conditional representations. We extend Frank et al. (2009) with entity-level semantic representations, modern architectures (GRU, LSTM, Attention with AbsPE/RoPE), principled competing event generation, extended systematicity tests (~350 vs. ~80 sentences), and a two- dimensional difficulty analysis disaggregating results by modifier complexity. Across 140 trained models (7 architectures), we find that capacity-matched architectures perform comparably on easy tests, but gated recurrent networks (GRU and LSTM) significantly outperform transformer architectures on the hardest compositional generalization test (Basic Event), while ungated SRN does not, indicating that the gating mechanism is a critical factor. Entity vectors significantly improve scores on Basic Event across most architectures, with gated architectures benefiting most, validating formal semantics treatment of entities as important theoretical primitives. The extended test set reveals that systematicity difficulty has two dimensions: the type of systematicity test (as in Frank et al. 2009), and the number of modifiers being composed.

[Long] Thinking in Schemas: Robust Syllogistic Reasoning in LLMs ▶ *Underline Video*

Federico Ranaldi, Leonardo Ranaldi, Fabio Massimo Zanzotto and Shay B. Cohen

LLMs often mistake what sounds true for what is formally valid. This limitation is especially evident in syllogistic reasoning, where plausible arguments can lead models to endorse conclusions that are logically invalid, a phenomenon known as Content Effect (CE). We present Boethius, a schema-guided framework for syllogistic reasoning that disentangles semantic plausibility from logical validity. Boethius adopts an auditable, quasi-formal reasoning process with two complementary stages: a Schema Module, which deduces the underlying logical form by analysing the formal structure of the premises, and an Instantiation Module, which instantiates this form over the concrete argument and evaluates validity independently of content-level semantics. Our results show that Boethius consistently outperforms existing approaches, improving syllogistic reasoning accuracy while substantially reducing CE. These gains hold for both large models in a pure in-context learning setting and smaller models trained via schema-guided trajectories using supervised fine-tuning and optimisation-based refinement.

[Findings] TinyAttack: Exploring Stylistic Vulnerabilities in Large Language Models ▶ *Underline Video*

Mamta Mamta, Bogdan Grecu and Oana Cocarascu

Large Language Models (LLMs) have demonstrated impressive results in natural language processing (NLP) tasks, however, their brittleness against subtle input perturbations continues to pose a significant challenge. Existing research on robustness has predominantly focused on standard text-based perturbations and the use of invisible characters and homoglyphs, while overlooking the impact of stylized characters increasingly prevalent on social media. To address this, we propose TinyAttack, a novel adversarial attack framework designed to exploit vulnerabilities in LLMs through Unicode-based stylistic transformations. TinyAttack utilises five Unicode variants to modify the visual rendering of text without altering its underlying semantic or syntactic structure. Our comprehensive evaluation on both open-source (Llama, Mistral, Gemma, Qwen) and closed-source LLMs (Gemini, GPT) demonstrates their susceptibility to these stylized inputs, with performance drops ranging from 29-92% and 6-88.5%, respectively, across all tasks. Our code is available at https://github.com/TRAI-group/TinyAttack.

[Findings] ToolDNA: Autonomous Evolution of Tool Metadata for Robust Dialogue Agents ▶ *Underline Video*

Qiuyuan Ai, Cong Wang, jiaqi zhang, zengxin han and Jie Song

Task-oriented dialogue (TOD) systems are vital for facilitating complex, goal-directed interactions across sectors like customer support and online retail. However, they face persistent limitations: labor-intensive manual metadata tuning and sparse reinforcement learning (RL) rewards that fail to diagnose invocation errors. To address this, we propose ToolDNA, a dynamic adaptation framework enabling autonomous co-evolution of policy networks and tool metadata via RL, anchored by two synergistic loops. An RL loop optimizes policies by generating rollout trajectories (reasoning, actions, descriptive updates) from user inputs, with multi-dimensional rewards refining invocations. A tool metadata loop, coordinated by a dedicated Tool Manager, evolves metadata through policy-generated candidates during rollouts and Feedback LLM-derived refinements from historical data. These mutually reinforcing loops close traditional reward gaps, forming a closed-loop trial-error-reflection cycle for self-improvement. Extensive experiments on a real-world dataset of 3,100 customer service dialogues confirm ToolDNA's superiority, with notable gains over baselines: it achieves +11% problem resolution and +54% accuracy over commercial LLMs with prompt engineering; +25%/+35% over supervised fine-tuning; and +15%/+15% over traditional RL baseline. Linguistic analysis corroborates evolved metadata retain semantic intent while enhancing parseability. Case studies in two typical contexts, i.e., car inventory search and loan calculation, further validates its ability to resolve critical ambiguities. ToolDNA pioneers scalable self-improvement for

robust, deployable tool-augmented agents with minimal human oversight. We release our code to facilitate future research.

[Findings] Verifiable Parameterization of Bayesian Networks from Scientific Literature: Unlocking Unstructured Empirical Evidence ▶ *Underline Video*

Jonas Gottal and Florian Matthes

Learning Bayesian Networks typically requires access to raw tabular data to estimate conditional probabilities. However, in many scientific domains, raw data is unavailable due to privacy concerns or general lack of access, while structured statistical summaries are increasingly accessible through large language models and published literature. We propose and evaluate five distinct strategies to reconstruct local conditional probability tables solely from statistical summaries in order to parameterize Bayesian Networks. Our comprehensive evaluation across mixed-type synthetic networks demonstrates that copula-based methods significantly outperform standard baselines, offering a viable path for knowledge integration from heterogeneous sources – unlocking the wealth of published knowledge for causal modeling while ensuring transparency and verifiability.

[Long] VISTA: Verification In Sequential Turn-based Assessment ▶ *Underline Video*

Ashley Lewis, Andrew Perrault, Eric Fosler-Lussier and Michael White

Hallucination—defined here as generated statements unsupported or contradicted by available evidence or conversational context—remains a major obstacle to using conversational AI systems in settings that demand factual reliability. Existing metrics evaluate isolated responses or treat unverifiable content as errors, limiting their use for multi-turn dialogue. We introduce VISTA (Verification In Sequential Turn-based Assessment), a framework for evaluating conversational factuality via claim-level verification and sequential consistency tracking. VISTA decomposes each turn into atomic claims, verifies them against trusted sources and dialogue history, and categorizes unverifiable statements (subjective, contradicted, lacking evidence, or abstaining). Across eight large language models and four dialogue factuality benchmarks (Ais, Begin, FaithDial, and Fade), VISTA substantially improves hallucination detection over FActScore and LLM-as-Judge baselines. Human evaluation confirms that VISTA's decomposition improves annotator agreement and reveals inconsistencies in existing benchmarks. By modeling factuality as a dynamic property of conversation, VISTA offers a more transparent, human-aligned measure of truthfulness in dialogue systems.

[Long] When TableQA Meets Noise: A Dual Denoising Framework for Complex Questions and Large-scale Tables ▶ *Underline Video*

Shenghao Ye, Yu Guo, Dong Jin, Yuxiang Wang, Yikai Shen, Yunpeng Hou, Shuangwu Chen, Jian Yang and Xiaofeng Jiang

Table question answering (TableQA) is a fundamental task in natural language processing (NLP). The strong reasoning capabilities of large language models (LLMs) have brought significant advances in this field. However, as real-world applications involve increasingly complex questions and larger tables, substantial noisy data is introduced, which severely degrades reasoning performance. To address this challenge, we focus on improving two core capabilities: Relevance Filtering, which identifies and retains information truly relevant to reasoning, and Table Pruning, which reduces table size while preserving essential content. Based on these principles, we propose EnoTab, a dual denoising framework for complex questions and large-scale tables. Specifically, we first perform Evidence-based Question Denoising by decomposing the question into minimal semantic units and filtering out those irrelevant to answer reasoning based on consistency and usability criteria. Then, we propose Evidence Tree-guided Table Denoising, which constructs an explicit and transparent table pruning path to remove irrelevant data step by step. At each pruning step, we observe the intermediate state of the table and apply a post-order node rollback mechanism to handle abnormal table states, ultimately producing a highly reliable sub-table for final answer reasoning. Finally, extensive experiments show that EnoTab achieves outstanding performance on TableQA tasks with complex questions and large-scale tables, confirming its effectiveness.

[Findings] Which Pieces Does Unigram Tokenization Really Need? ▶ *Underline Video*

Sander Land and Yuval Pinter

The Unigram tokenization algorithm offers a probabilistic alternative to the greedy heuristics of Byte-Pair Encoding. Despite its theoretical elegance, its implementation in practice is complex, limiting its adoption to the SentencePiece package and adapters thereof. We bridge this gap between theory and practice by providing a clear guide to implementation and parameter choices. We also identify a simpler algorithm that accepts slightly higher training loss in exchange for improved compression.

[Findings] WildSci: Advancing Scientific Reasoning from In-the-Wild Literature ▶ *Underline Video*

Tengxiao Liu, Deepak Nathani, Zekun Li, Kevin Yang and William Yang Wang

Recent progress in large language model (LLM) reasoning has focused on domains like mathematics and coding, where abundant high-quality data and objective evaluation metrics are readily available. In contrast, progress in scientific reasoning remains limited in domains such as medicine and materials science due to restricted dataset coverage and the inherent complexity of open-ended scientific questions. To address these challenges, we propose a general framework for sustainable scientific reasoning QA generation, and introduce WildSci, a new dataset of domain-specific science questions automatically synthesized from peer-reviewed literature, spanning 9 scientific disciplines and 26 subdomains. WildSci enables scalable training with well-defined reward signals in a multiple-choice format. We further apply reinforcement learning to finetune models on WildSci and analyze the resulting training dynamics, including domain-specific performance changes, response behaviors, and generalization trends. Experiments on a suite of scientific benchmarks demonstrate the effectiveness of our framework and dataset. We release WildSci to enable scalable and sustainable research in scientific reasoning.

Poster Session B (Grand Hall)

14:00-15:30 (Grand Hall)

[Findings] A Few Bad Apples Spoil the Bunch: Preventing Global Entropy Collapse Driven by a Small Set of Tokens in LLM Reasoning ▶ *Underline Video*

Jaeeun Jang, Hansle Lee and Sangmin Kim

Reinforcement Learning with Verifiable Rewards (RLVR) and Reinforcement Learning from Internal Feedback (RLIF) often fail to benefit from test-time compute due to entropy collapse and the resulting loss of reasoning diversity. We show that this collapse is driven not by uniform entropy decay, but by premature overconfidence at a small number of structurally critical decision points. Based on a token-level analysis of GRPO-style policy optimization, we propose SCOPE (Structural Collapse-aware Optimization via Partial Entropy control), which assigns each generated token a redistribution score and applies selective KL regularization to only the top 5% of tokens under this score. Across model scales and architectures on math reasoning benchmarks, SCOPE consistently improves performance under both RLVR and RLIF settings, demonstrating that targeted entropy control at a vanishingly small subset of tokens is sufficient to sustain reasoning diversity and effective test-time scaling.

[Findings] A Mechanistic Perspective and Difficulty Metric for Unlearning ▶ *Underline Video*

Jiali Cheng, Ziheng Chen, Chirag Agarwal and Hadi Amiri

Machine unlearning is becoming essential for building trustworthy and compliant language models. Yet unlearning success varies considerably across individual samples: some are reliably erased, while others persist despite the same procedure. We argue that this disparity is not only a data-side phenomenon, but also reflects model-internal mechanisms that encode and protect memorized information. We study this problem from a mechanistic perspective based on model circuits–structured interaction pathways that govern how predictions are formed. We propose Circuit-guided Unlearning Difficulty (), a pre-unlearning metric that assigns each sample a continuous difficulty score using circuit-level signals. Extensive experiments demonstrate that reliably separates intrinsically easy and hard samples, and remains stable across unlearning methods. We identify key circuit-level patterns that reveal a mechanistic signature of unlearning difficulty: easy-to-unlearn samples are associated with shorter, shallower interactions concentrated in earlier-to-intermediate parts of the original model, whereas hard-to-unlearn samples rely on longer and deeper pathways closer to late-stage computation. Compared to existing qualitative studies, takes a first step toward a principled, fine-grained, and interpretable analysis of unlearning difficulty; and motivates the development of unlearning methods grounded in model mechanisms.

[Findings] A Study of LLMs' Preferences for Libraries and Programming Languages ▶ *Underline Video*

Lukas Twist, Jie Zhang, Mark Harman, Don Syme, Joost Noppen, Helen Yannakoudakis and Detlef Nauck

Despite the rapid progress of large language models (LLMs) in code generation, existing evaluations focus on functional correctness or syntactic validity, overlooking how LLMs make critical design choices such as which library or programming language to use. To fill this gap, we perform the first empirical study of LLMs' preferences for libraries and programming languages when generating code, covering eight diverse LLMs. We observe a strong tendency to overuse widely adopted libraries such as NumPy; in up to 45% of cases, this usage is not required and deviates from the ground-truth solutions. The LLMs we study also show a significant preference toward Python as their default language. For high-performance project initialisation tasks where Python is not the optimal language, it remains the dominant choice in 58% of cases, and Rust is not used once. These results highlight how LLMs prioritise familiarity and popularity over suitability and task-specific optimality; underscoring the need for targeted fine-tuning, data diversification, and evaluation benchmarks that explicitly measure language and library selection fidelity.

[Long] A Survey of Deep Learning for Geometry Problem Solving ▶ *Underline Video*

Jianzhe Ma, Wenxuan Wang and Qin Jin

Geometry problem solving, a crucial aspect of mathematical reasoning, is vital across various domains, including education, the assessment of AI's mathematical abilities, and multimodal capability evaluation. The recent surge in deep learning technologies, particularly the emergence of multimodal large language models, has significantly accelerated research in this area. This paper presents a survey of the applications of deep learning in geometry problem solving, including (i) a comprehensive summary of the relevant tasks in geometry problem solving; (ii) a thorough review of related deep learning methods; (iii) a detailed analysis of evaluation metrics and methods; and (iv) a critical discussion of state-of-the-art performance, existing challenges, and promising future directions. Our objective is to offer a comprehensive and practical reference of deep learning for geometry problem solving, thereby fostering further advancements in this field. We maintain a list of relevant papers: https://github.com/majianz/dl4gps.

[Findings] A Survey of Large Models in Sports ▶ *Underline Video*

Yichen Xu, Jianzhe Ma, Chuhan Wang, Zhonghao Cao, Liangyu Chen, Wenxuan Wang and Qin Jin

Sports have witnessed growing global enthusiasm in recent years, serving as a vital force for physical health, cultural exchange, social

connection, and economic growth. The rapid advancement of large models, particularly (multimodal) large language models (M)LLMs, has demonstrated transformative potential to reshape sports understanding, analysis, and interaction across diverse domains. This paper presents a comprehensive survey of large models in sports, including (i) an overview of tasks and applications across different participant groups; (ii) a detailed analysis of sports-related datasets and benchmarks; and (iii) a critical discussion of current challenges and future directions. Our goal is to establish a foundation for advancing research and practical development of large-model-driven sports intelligence. An open-source GitHub repository is maintained at: https://github.com/Road2Redemption/Awesome_Large_Models_In_Sports1.

[TACL] A Systematic Review of NLP for Dementia: Tasks, Datasets and Opportunities ▶ *Underline Video*

Roi Reichart and Lotem Peled-Cohen

The close link between cognitive decline and language has fostered long-standing collaboration between the NLP and medical communities in dementia research. To examine this, we reviewed over 240 papers applying NLP to dementia-related efforts, drawing from medical, technological, and NLP-focused literature. We identify key research areas, including dementia detection, linguistic biomarker extraction, caregiver support, and patient assistance, showing that half of all papers focus solely on dementia detection using clinical data. Yet, many directions remain unexplored- artificially degraded language models, synthetic data, digital twins, and more. We highlight gaps and opportunities around trust, scientific rigor, applicability and cross-community collaboration. We raise ethical dilemmas in the field, and highlight the diverse datasets encountered throughout our review- recorded, written, structured, spontaneous, synthetic, clinical, social media-based, and more. This review aims to inspire more creative, impactful, and rigorous research on NLP for dementia.

[Long] ADAPT: Benchmarking Commonsense Planning under Unspecified Affordance Constraints ▶ *Underline Video*

Pei-An Chen, Yong-Ching Liang, Jia-Fong Yeh, Hung-Ting Su, Yi-Ting Chen, Min Sun and Winston H. Hsu

Intelligent embodied agents should not simply follow instructions, as real-world environments often involve unexpected conditions and exceptions. However, existing methods usually focus on directly executing instructions, without considering whether the target objects can actually be manipulated, meaning they fail to assess available affordances. To address this limitation, we introduce DynAfford, a benchmark that evaluates embodied agents in dynamic environments where object affordances may change over time and are not specified in the instruction. DynAfford requires agents to perceive object states, infer implicit preconditions, and adapt their actions accordingly. To enable this capability, we introduce ADAPT (Affordance-Driven Adaptive Planning and Task execution), a plug-and-play module that augments existing planners with explicit affordance reasoning. Experiments demonstrate that incorporating ADAPT significantly improves robustness and task success across both seen and unseen environments. We also show that a domain-adapted, LoRA-finetuned vision-language model used as the affordance inference backend outperforms a commercial LLM (GPT-4o), highlighting the importance of task-aligned affordance grounding.

[Findings] Afrispeech Semantics: Evaluating Audio-Semantic Reasoning in Spoken Language Models Across Domains and Accents ▶ *Underline Video*

Chibuzor Okocha and Christan Grant

Audio language models (ALMs) are increasingly used for speech-based understanding; yet, their ability to perform semantic reasoning beyond transcription, Text-to-Audio Retrieval, Captioning, and Question-Answering accuracy remains insufficiently benchmarked. In particular, the effects of accent variation, domain shift, and semantic over-inference on audio reasoning are poorly understood. We evaluate audio language models across five semantic and paralinguistic reasoning tasks: entailment, consistency, plausibility, accent drift, and accent restraint. Collectively, these tasks assess a models ability to reason over spoken audio as the primary evidence source, including whether a textual hypothesis can be inferred, contradicted, or left undetermined by the audio, whether statements align or conflict with spoken content, whether claims are plausible given the discourse, and whether model predictions remain stable or appropriately constrained across accent variation. These findings highlight critical limitations in current audio reasoning evaluations and hope to provide guidance for more robust and equitable ALM design and assessment.

[Long] Aligned Multi-View Scripts for Universal Chart-to-Code Generation ▶ *Underline Video*

Zhihan Zhang and Lizi Liao

Chart-to-code generation converts a chart image into an executable plotting script, enabling faithful reproduction and editable visualizations. Existing methods are largely Python-centric, limiting practical use and overlooking a critical source of supervision: the same chart can be expressed by semantically equivalent scripts in different plotting languages. To fill this gap, we introduce Chart2NCode, a dataset of 176K charts paired with aligned scripts in Python, R, and LaTeX that render visually equivalent outputs, constructed via a metadata-to-template pipeline with rendering verification and human quality checks. Building on a LLaVA-style architecture, we further propose CharLuMA, a parameter-efficient adaptation module that augments the multimodal projector with a language-conditioned mixture of low-rank subspaces, allowing the model to share core chart understanding while specializing code generation to the target language through lightweight routing. Extensive experiments show consistent gains in executability and visual fidelity across all languages, outperforming strong open-source baselines and remaining competitive with proprietary systems. Further analyses reveal that balanced multi-language supervision benefits all languages and that the adapter allocates a compact shared core plus language-specific capacity.

[Findings] Anonpsy: A Graph-Based Framework for Structure-Preserving De-identification of Psychiatric Narratives ▶ *Underline Video*

Kyungho Lim and Byung-Hoon Kim

Psychiatric narratives encode patient identity not only through explicit identifiers but also through idiosyncratic life events embedded in clinical structure. Existing de-identification approaches, including PHI masking and LLM-based synthetic rewriting, operate at the text level and offer limited control over which semantic elements are preserved or altered. We introduce Anonpsy, a de-identification framework that reformulates the task as graph-guided semantic rewriting. Anonpsy (1) converts each narrative into a semantic graph encoding clinical entities, temporal anchors, and typed relations; (2) applies graph-constrained perturbations that modify identifying context while preserving clinical structure; and (3) regenerates text via graph-conditioned LLM generation. Evaluated on 90 clinician-authored psychiatric case narratives, Anonpsy preserves diagnostic fidelity while achieving consistently low re-identification risk under expert, semantic, and GPT-5-based evaluations. Compared with a strong LLM-only rewriting baseline, Anonpsy yields substantially lower semantic similarity and identifiability. These results demonstrate that explicit structural representations combined with constrained generation provide an effective approach to de-identification for psychiatric narratives.

[Long] Anti-Length Shift: Dynamic Outlier Truncation for Training Efficient Reasoning Models ▶ *Underline Video*

Wei Wu, Liyi Chen, Congxi Xiao, Tianfu Wang, Qimeng Wang, Chengqiang Lu, Yan Gao, Yi Wu, Yao Hu and Hui Xiong

Large reasoning models enhanced by reinforcement learning with verifiable rewards have achieved significant performance gains by extending their chain-of-thought. However, this paradigm incurs substantial deployment costs as models often exhibit excessive verbosity on simple queries. Existing efficient reasoning methods relying on explicit length penalties often introduce optimization conflicts and leave the generative mechanisms driving overthinking largely unexamined. In this paper, we identify a phenomenon termed length shift where models increasingly generate unnecessary reasoning on trivial inputs during training. To address this, we introduce Dynamic Outlier Truncation (DOT), a training-time intervention that selectively suppresses redundant tokens. This method targets only the extreme tail of response lengths within fully correct rollout groups while preserving long-horizon reasoning capabilities for complex problems. To complement this intervention and ensure stable convergence, we further incorporate auxiliary KL regularization and predictive dynamic sampling. Experimental results across multiple model scales demonstrate that our approach significantly pushes the efficiency-performance Pareto frontier outward. Notably, on the AIME-24, our method reduces inference token usage by 78% while simultaneously increasing accuracy compared to the initial policy and surpassing state-of-the-art efficient reasoning methods.

[Long] Are they lovers or friends? Evaluating LLMs Social Reasoning in English and Korean Dialogues ▶ *Underline Video*

Eunsu Kim, Junyeong Park, Juhyun Oh, Kiwoong Park, Seyoung Song, A. Seza Doruöz, Alice Oh and Najoung Kim

As LLMs are increasingly deployed in real-world interactions, their social reasoning in interpersonal communication becomes critical. To explore their capabilities, we introduce SCRIPTS, a 1.1k-dialogue dataset in English and Korean, sourced from movie scripts and propose a social reasoning task based on SCRIPTS that evaluates the capacity of LLMs to infer the social relationships (e.g., friends, lovers) between speakers in each dialogue. Evaluating nine models on our task, current LLMs achieve around 75-80% on the English dataset and 58-69% in Korean, and models predict an Unlikely relationship in 10-25% of responses in both languages. Furthermore, we find that thinking models and chain-of-thought prompting provide minimal benefits for social reasoning and occasionally amplify social biases. In sum, there are significant limitations in current LLMs' social reasoning capabilities, especially for Korean, highlighting the need for efforts to develop socially-aware LLMs across languages.

[Findings] BiasGRPO: Stabilizing Bias Mitigation in High-Variance Reward Landscapes via Group-Relative Policy Optimization ▶ *Underline Video*

Saket Reddy, Ke Yang and ChengXiang Zhai

Mitigating social bias in Large Language Models (LLMs) presents a distinct alignment challenge: unlike verifiable tasks, social bias lacks a single ground truth, creating a high-variance, subjective reward landscape. Previous preference-based fine-tuning methods have major trade-offs: Direct Preference Optimization (DPO) is limited by the lack of exploration inherent in offline training, while Proximal Policy Optimization (PPO) can lead to training instability due to potentially unreliable critic estimates. In this paper, we propose BiasGRPO, an adaptation of Group Relative Policy Optimization (GRPO) that stabilizes alignment by normalizing rewards across a group of sampled completions. By substituting the value function with a group-relative baseline, our approach reduces instability while maintaining the exploration benefits of online reinforcement learning. To adapt GRPO, we curate and synthetically extend a dataset spanning multiple domains and contexts, and create a custom, bias-specific reward model for effectively guiding generation while avoiding knowledge degradation. We find that BiasGRPO outperforms DPO and PPO across multiple benchmarks, indicating its effectiveness as an alignment technique that can overcome the limitations of previous preference-based methods.

[Long] BTC-LLM: Efficient Sub-1-Bit LLM Quantization via Learnable Transformation and Binary Codebook ▶ *Underline Video*

Hao Gu, Lujun Li, Hao Wang, LEI WANG, Zheyu Wang, Bei Liu, Jiacheng Liu, Qiyuan Zhu, Sirui Han and Yike Guo

Binary quantization represents the most extreme form of compression, reducing weights to +/-1 for maximal memory and computational efficiency. While recent sparsity-aware binarization achieves sub-1-bit compression via weight pruning, it faces critical challenger: performance degradation, mask-management overhead, and limited hardware compatibility. In this paper, we present BTC-LLM, a novel sub-1-bit LLM quantization framework that leverages binary pattern clustering and weight transformation to overcome these limitations. Our approach incorporates two key innovations: (1) a Binary Codebook that clusters recurring vectors into compact indices using custom distance metrics and sign-based updates; (2) a Learnable Transformation that reduces outliers and promotes shared sign patterns among

binary weights. This eliminates sparse masks, enabling efficient inference on standard hardware. Extensive evaluations across LLaMA, Qwen, and FBI-LLM families demonstrate that BTC-LLM achieves state-of-the-art results in extreme compression (1.11-0.7 bits). Notably, BTC-LLM compressed to 0.8 bits on LLaMA-2-13B maintains high performance, with only a 3.1% accuracy drop in zero-shot benchmarks, while delivering a 1.6x speedup over FP16.

[Findings] Budget-Aware Routing for Long Clinical Text ▶ *Underline Video*

Khizar QureshiGeoffrey MartinYifan Peng

A key challenge for large language models is token cost per query and overall deployment cost. Clinical inputs are long, heterogeneous, and often redundant, while downstream tasks are short and high stakes. We study budgeted context selection, where a subset of document units is chosen under a strict token budget so an off-the-shelf generator can meet fixed cost and latency constraints. We cast this as a knapsack-constrained subset selection problem with two design choices, unitization that defines document segmentation and selection that determines which units are kept. We propose RCD, a monotone submodular objective that balances relevance, coverage, and diversity. We compare sentence, section, window, and cluster-based unitization, and introduce a routing heuristic that adapts to the budget regime. Experiments on MIMIC discharge notes, Cochrane abstracts, and L-Eval show that optimal strategies depend on the evaluation setting. Positional heuristics perform best at low budgets in extractive tasks, while diversity-aware methods such as MMR improve LLM generation. Selector choice matters more than unitization, with cluster-based grouping reducing performance and other schemes behaving similarly. ROUGE saturates for LLM summaries, while BERTScore better reflects quality differences.

[Long] Can Compact Language Models Search Like Agents? Distillation-Guided Policy Optimization for Preserving Agentic RAG Capabilities ▶ *Underline Video*

Rikuto Kotoge, Mai Nishimura and Jiaxin Ma

Reinforcement Learning has emerged as a dominant post-training approach to elicit agentic RAG behaviors such as search and planning from language models. Despite its success with larger models, applying RL to compact models (e.g., 0.5-1B parameters) presents unique challenges. The compact models exhibit poor initial performance, resulting in sparse rewards and unstable training. To overcome these difficulties, we propose Distillation-Guided Policy Optimization (DGPO), which employs cold-start initialization from teacher demonstrations and continuous teacher guidance during policy optimization. To understand how compact models preserve agentic behavior, we introduce Agentic RAG Capabilities (ARC), a fine-grained metric analyzing reasoning, search coordination, and response synthesis. Comprehensive experiments demonstrate that DGPO enables compact models to achieve sophisticated agentic search behaviors, even outperforming the larger teacher model in some cases. DGPO makes agentic RAG feasible in computing resource-constrained environments.

[Long] Can LLM Safety Be Ensured by Constraining Parameter Regions? ▶ *Underline Video*

Zongmin Li, Jian Su, Farah Benamara and Aixin Sun

Large language models (LLMs) are often assumed to contain "safety regions" - parameter subsets whose modification directly influences safety behaviors. We conduct a systematic evaluation of four safety region identification methods spanning different parameter granularities, from individual weights to entire Transformer layers, across four families of backbone LLMs with varying sizes. Using ten safety identification datasets, we find that the identified safety regions exhibit only low to moderate overlap, as measured by IoU. The overlap drops significantly when the safety regions are further refined using utility datasets (i.e. non-harmful queries). These results suggest that current techniques fail to reliably identify a stable, dataset-agnostic safety region.

[Findings] CANDICE: Agentic Causal Disentanglement with Class Conditional Knowledge Integration for Long Tailed Domain Generalization ▶ *Underline Video*

Midhat Urooj, Ayan Banerjee and Sandeep Gupta

Deep learning models deployed in clinical settings face two major challenges: domain generalization (DG) and long-tailed (LT) recognition. DG requires learning domain-invariant features to ensure robustness across heterogeneous acquisition protocols and patient populations. However, we identify a fundamental trade-off: objectives that enforce domain invariance often suppress class-discriminative signals essential for long-tailed recognition. To address this, we propose the Agentic Causal Disentanglement (CANDICE) Framework, a modular architecture that integrates explicit clinical expertise from sonographers, radiologists, and specialists as a form of causal intervention. The framework combines clinical reasoning, causal representation learning, and automated pipeline construction to disentangle domain-invariant and class-discriminative features. By incorporating domain-specific causal knowledge, it effectively decouples the objectives of DG and LT learning. We evaluate CANDICE on 10 diverse medical imaging datasets spanning four modalities. The framework achieves an average performance improvement of 10.3% across both multi-domain and in-domain long-tailed tasks, demonstrating its effectiveness in handling distribution shifts while preserving minority class performance.

[Findings] CCD: Mitigating Hallucinations in Radiology MLLMs via Clinical Contrastive Decoding ▶ *Underline Video*

Xi Zhang, Zaiqiao Meng, Jake Lever and Edmond S. L. Ho

Multimodal large language models (MLLMs) have recently achieved remarkable progress in radiology by integrating visual perception with natural language understanding. However, they often generate clinically unsupported descriptions, known as medical hallucinations, which pose serious risks in medical applications that demand accuracy and image-grounded outputs. Through empirical analysis, we find that prompt-induced hallucinations remain prevalent in radiology MLLMs, largely due to over-sensitivity to clinical sections. To address

this, we introduce Clinical Contrastive Decoding (CCD), a training-free and retrieval-free inference framework that integrates structured clinical signals from taskspecific radiology expert models. CCD introduces a dual-stage contrastive mechanism to refine token-level logits during generation, thereby enhancing clinical fidelity without modifying the base MLLM. Experiments on three datasets and multiple models demonstrate that CCD consistently improves overall performance on radiology report generation (RRG). On the MIMIC-CXR dataset, it yields up to a 17% improvement in RadGraph-F1 when applied to state-of-the-art RRG models. Our approach provides a lightweight and generalisable solution for mitigating medical hallucinations, effectively bridging expert models and MLLMs in radiology.

[Long] Characterizing and Evaluating Working Emotion Vocabularies in Multilingual Large Language Models ▶ *Underline Video*

Nicholas Deas, Ivan Ernesto Perez Mejia, Ellie Yang and Kathleen McKeown

Prior work evaluating emotion and affective understanding in large language models (LLMs) typically rely on predetermined label sets or focus on a singular evaluation task (e.g., emotion detection). We consider affective states, referring to the much broader variety of terms people use to label their emotional experiences. We evaluate multilingual language models understanding of affective states in English and Spanish through three different tasks: 1) _identification_, where models predict an affective state given text, 2) _expression_, where models generate text expressing a given affective state, and 3) _verification_, where models report whether a given term refers to an affective state. We show that performance on one task is not necessarily predictive of performance on another. Using these three tasks, we then begin to explore when and why models struggle to understand particular affective states compared to others. We examine systematic patterns in the affective state terms that are well and poorly understood by models, characterizing the working emotion vocabulary of LLMs.

[Findings] Characterizing Selective Refusal Bias in Large Language Models ▶ *Underline Video*

Adel Khorramrouz and Sharon Levy

Safety guardrails in large language models (LLMs) are developed to prevent malicious users from generating toxic content at a large scale. However, these measures can inadvertently introduce or reflect new biases, as LLMs may refuse to generate harmful content targeting some demographic groups and not others. We explore this selective refusal bias in LLM guardrails through the lens of refusal rates of targeted individual and intersectional demographic groups, types of LLM responses, and length of generated refusals. Our results show evidence of selective refusal bias across gender, sexual orientation, nationality, and religion attributes. This leads us to investigate additional safety implications via an indirect attack, where we target previously refused groups, and find that Llama 3.1 fails to defend against our attack in roughly 89% of the trials. Our findings emphasize the need for more equitable and robust performance in safety guardrails across demographic groups.

[Long] CIS-BWE: Chaos-Informed Speech Bandwidth Extension ▶ *Underline Video*

Tarikul Islam Tamiti, Tonmoy Das, Nursadul Mamun and Anomadarshi Barua

We design CIS-BWE, a novel adversarial Bandwidth Extension (BWE) framework that introduces two chaos-informed discriminators - Multi-Resolution Lyapunov Discriminator (MRLD) and Multi-Scale Detrended Fractal Analysis Discriminator (MSDFA) - for capturing the deterministic chaos from speech. MRLD exploits Lyapunov exponents to capture nonlinear chaotic fluctuations. MSDFA exploits detrended fluctuation analysis to quantify fractal-like, long-range temporal chaotic correlations. To the best of our knowledge, MRLD and MSDFA are included here for the first time with a complex-valued adversarial network to explore the chaotic study of speech reconstruction. We also introduce a novel complex-valued and dual-stream generator, which uses our newly proposed ConformerNeXt as a core block with Lattice interactions, acting as a gating mechanism by enabling controlled mixing of information across streams. We extensively optimize our design across five resolutions and use depth-wise separable convolutions to make our model lightweight yet powerful. Our CIS-BWE is tested with a de facto English and French dataset for clean and noisy speech for generalization. It achieves better performance across a total of nine subjective and objective evaluation metrics with a 40x reduction in discriminator size and overall 0.5x fewer parameters, establishing a new baseline in the BWE task.

[Findings] CLaRE-ty Amid Chaos: Quantifying Representational Entanglement to Predict Ripple Effects in LLM Editing ▶ *Underline Video*

Manit Baser, Alperen Yildiz, Dinil Mon Divakaran and Mohan Gurusamy

The static knowledge representations of large language models (LLMs) inevitably become outdated or incorrect over time. While model-editing techniques offer a promising solution by modifying a model's factual associations, they often produce unpredictable ripple effects, which are unintended behavioral changes that propagate even to the hidden space. In this work, we introduce CLaRE, a lightweight representation-level technique to identify where these ripple effects may occur. Unlike prior gradient-based methods, CLaRE quantifies entanglement between facts using forward activations from a single intermediate layer, avoiding costly backward passes. To enable systematic study, we prepare and analyse a corpus of 11,427 facts drawn from three existing datasets. Using CLaRE, we compute large-scale entanglement graphs of this corpus for multiple models, capturing how local edits propagate through representational space. These graphs enable stronger preservation sets for model editing, audit trails, efficient red-teaming, and scalable post-edit evaluation. In comparison to baselines, CLaRE achieves an average of 62.2% improvement in Spearman correlation with ripple effects while being 2.74x faster, and using 2.85x less peak GPU memory. Besides, CLaRE requires only a fraction of the storage needed by the baselines to compute and preserve fact representations. Our entanglement graphs and corpus are available at https://github.com/manitbaser/CLaRE.

[Long] Closing the Modality Reasoning Gap for Speech Large Language Models ▶ *Underline Video*

Chaoren Wang, Heng Lu, Xueyao Zhang, Shujie LIU, Yan Lu, Jinyu Li and Zhizheng Wu

Although Speech Large Language Models have achieved notable progress, a substantial modality reasoning gap remains: their reasoning performance on speech inputs is markedly weaker than on text. This gap could be associated with representational drift across Transformer layers and behavior deviations in long-chain reasoning. To address this issue, we introduce TARS, a reinforcementlearning framework that aligns textconditioned and speechconditioned trajectories through an asymmetric reward design. The framework employs two dense and complementary signals: representation alignment, which measures layer-wise hidden-state similarity between speech- and text-conditioned trajectories, and behavior alignment, which evaluates semantic consistency between generated outputs and reference text completions. Experiments on challenging reasoning benchmarks, including MMSU and OBQA, show that our approach significantly narrows the modality reasoning gap and achieves state-of-the-art performance among 7B-scale Speech LLMs.

[Findings] Colorful Talks with Graphs: Human-Interpretable Graph Encodings for Large Language Models ▶ *Underline Video*

Angelo Zangari, Peyman Baghershahi and Sourav Medya

Graph problems are fundamentally challenging for large language models (LLMs). While LLMs excel at processing unstructured text, graph tasks require reasoning over explicit structure, permutation invariance, and computationally complex relationships, creating a mismatch with the representations of text-based models. Our work investigates how LLMs can be effectively applied to graph problems despite these barriers. We introduce a human-interpretable structural encoding strategy for graph-to-text translation that injects graph structure directly into natural language prompts. Our method involves computing a variant of Weisfeiler-Lehman (WL) similarity classes and maps them to human-like color tokens rather than numeric labels. The key insight is that semantically meaningful and human-interpretable cues may be more effectively processed by LLMs than opaque symbolic encoding. Experimental results across multiple graph algorithms and predictive tasks show significant improvements from our method on both synthetic and real-world datasets. By capturing both local and global-range dependencies, our method enhances LLM performance, especially on graph tasks that require reasoning over global graph structure.

[Long] Community-Aware Assessment of Social Textual Engagement and Resonance: A Human-Centric Perspective on User-Generated Content Evaluation ▶ *Underline Video*

Tianjiao Li, Kai Zhao, Xiang Li, Yang Liu and Huyang Sun

Traditional Video Quality Assessment (VQA) focuses narrowly on aesthetic fidelity, overlooking the complex social dynamics that define quality in User-Generated Content (UGC). In this work, we propose a paradigm shift from signal-centric metrics to human-centric resonance assessment. We introduce CASTER (Community-Aware Assessment of Social Textual Engagement and Resonance), a new task that evaluates whether a UGC item achieves positive community resonance based on its multimodal attributes rather than visual quality alone. To address this, we present MEDEA (Multimodal Engagement-Driven Evaluation Architecture), which introduces a novel Social Chain-of-Thought (Social-CoT) mechanism. Unlike traditional logical CoT, Social-CoT performs multimodal perspective-taking, instantiating diverse viewer personas to simulate collective cognitive and emotional reactions (i.e., the "community mind") before deriving a quality judgment. MEDEA is trained via a two-stage approach involving supervised fine-tuning and process-supervised reinforcement learning with Social Alignment Reward to ensure reasoning paths are grounded in authentic human social cognition. To support this task, we release CASTER-Bench, a comprehensive human-annotated benchmark covering diverse UGC categories. Experiments demonstrate that MEDEA significantly outperforms state-of-the-art baselines on CASTER-Bench while providing interpretable and empathetic reasoning paths that align with real community feedback.

[Findings] Compact Example-Based Explanations for Language Models ▶ *Underline Video*

Loris Schoenegger and Benjamin Roth

Training data influence estimation methods quantify the contribution of training documents to a models output, making them a promising source of information for example-based explanations. As humans cannot interpret thousands of documents, only a small subset of the training data can be presented as an explanation. Although the choice of which documents to include directly affects explanation quality, previous evaluations of such systems have largely ignored any selection strategies. To address this, we propose a novel selection relevance score, a retraining-free metric that quantifies how useful a set of examples is for explaining a model's output. We validate this score through fine-tuning experiments, confirming that it can predict whether a set of examples supports or undermines the model's predictions. Using this metric, we further show that common selection strategies often underperform random selection. Motivated by this finding, we propose a strategy that balances influence and representativeness, enabling better use of selection budgets than naively selecting the highest-ranking examples.

[Findings] Comparing Human and Large Language Model Interpretation of Implicit Information ▶ *Underline Video*

Antonio De Santis, Tommaso Bonetti, Andrea Tocchetti and Marco Brambilla;

The interpretation of implicit meanings is an integral aspect of human communication. However, this framework may not transfer to interactions with Large Language Models (LLMs). To investigate this, we introduce the task of Implicit Information Extraction (IIE) and propose an LLM-based IIE pipeline that builds a structured knowledge graph from a context sentence by extracting relational triplets, validating implicit inferences, and analyzing temporal relations. We evaluate two LLMs against crowdsourced human judgments on two datasets. We find that humans agree with most model triplets yet consistently propose many additions, indicating limited coverage in current LLM-based IIE. Moreover, in our experiments, models appear to be more conservative about implicit inferences than humans in socially rich contexts, whereas humans become more conservative in shorter, fact-oriented contexts. Our code is available at https://github.com/Antonio-Dee/IIE_from_LLM.

[Findings] Context Attribution with Multi-Armed Bandit Optimization ► *Underline Video*

Deng Pan, Keerthiram Murugesan, Ting Hua, Nuno Moniz and Nitesh V. Chawla;

Understanding which parts of the retrieved context contribute to a large language model's generated answer is essential for building interpretable and trustworthy retrieval-augmented generation. We propose a novel framework that formulates context attribution as a combinatorial multi-armed bandit problem. We utilize Linear Thompson Sampling to efficiently identify the most influential context segments while minimizing the number of model queries. Our reward function leverages token log-probabilities to measure how well a subset of segments supports the original response, making it applicable to both open-source and black-box API-based models. Unlike SHAP and other perturbation-based methods that sample subsets uniformly, our approach adaptively prioritizes informative subsets based on posterior estimates of segment relevance, reducing computational costs. Experiments on multiple QA benchmarks demonstrate that our method achieves up to 30% reduction in model queries while matching or exceeding the attribution quality of existing approaches.

[Long] Controlling Distributional Bias in Multi-Round LLM Generation via KL-Optimized Fine-Tuning ► *Underline Video*

Yanbei Jiang, Amr Keleg, Ryandito Diandaru, Jey Han Lau, Lea Frermann, Biaoyan Fang and Fajri Koto

While the real world is inherently stochastic, Large Language Models (LLMs) are predominantly evaluated on single-round inference against fixed ground truths. In this work, we shift the lens to distribution alignment: assessing whether LLMs, when prompted repeatedly, can generate outputs that adhere to a desired target distribution, e.g. reflecting real-world statistics or a uniform distribution. We formulate distribution alignment using the attributes of gender, race, and sentiment within occupational contexts. Our empirical analysis reveals that off-the-shelf LLMs and standard alignment techniques, including prompt engineering and Direct Preference Optimization, fail to reliably control output distributions. To bridge this gap, we propose a novel fine-tuning framework that couples Steering Token Calibration with Semantic Alignment. We introduce a hybrid objective function combining Kullback-Leibler divergence to anchor the probability mass of latent steering tokens and Kahneman-Tversky Optimization to bind these tokens to semantically consistent responses. Experiments across six diverse datasets demonstrate that our approach significantly outperforms baselines, achieving precise distributional control in attribute generation tasks.

[Long] Counteracting the Matthew Effect in Self-Improvement of LVLMs through Head-Tail Re-balancing ► *Underline Video*

Xin Guo, Zhiheng Xi, Yiwen Ding, Yitao Zhai, Xiaowei Shi, Xunliang Cai, Tao Gui, Qi Zhang and Xuanjing Huang

Self-improvement has emerged as a mainstream paradigm for advancing the reasoning capabilities of large vision-language models (LVLMs), where models explore and learn from successful trajectories iteratively. However, we identify a critical imbalance during this process: the model readily generates high-quality trajectories for simple queries (i.e., head data) but struggles with complex ones (i.e., tail data). This bias drives the optimization to disproportionately prioritize simple reasoning skills, while inhibiting the acquisition of complex capabilities. As iterations progress, this imbalance becomes more acute, a dynamic we term the "Matthew effect", ultimately stalling performance gains. To mitigate this, we approach head-tail re-balance during the exploration-and-learning process from two perspectives: distribution-reshaping and trajectory-resampling. Extensive experiments on Qwen2-VL-7B-Instruct and InternVL2.5-4B models across visual reasoning tasks demonstrate that our methods consistently improve visual reasoning capabilities, outperforming vanilla self-improvement baselines by an average of 3.86 points.

[Long] CrossGuard: Safeguarding MLLMs against Joint-Modal Implicit Malicious Attacks ► *Underline Video*

Xu Zhang, Hao Li and Zhichao Lu

Multimodal Large Language Models (MLLMs) achieve strong reasoning and perception capabilities but are increasingly vulnerable to jailbreak attacks. While existing work focuses on explicit attacks, where malicious content resides in a single modality, recent studies reveal implicit attacks, in which benign text and image inputs jointly express unsafe intent. Such joint-modal threats are difficult to detect and remain underexplored, largely due to the scarcity of high-quality implicit data. We propose ImpForge, an automated red-teaming pipeline that leverages reinforcement learning with tailored reward modules to generate diverse implicit samples across 14 domains. Building on this dataset, we further develop CrossGuard, an intent-aware safeguard providing robust and comprehensive defense against both explicit and implicit threats. Extensive experiments across safe and unsafe benchmarks, implicit and explicit attacks, and multiple out-of-domain settings demonstrate that CrossGuard significantly outperforms existing defenses, including advanced MLLMs and guardrails, achieving stronger security while maintaining high utility. This offers a balanced and practical solution for enhancing MLLM robustness against real-world multimodal threats. Our code is released https://github.com/ZhangXu0963/CrossGuard.

[Long] CURA: Clinical Uncertainty Risk Alignment for Language Model-Based Risk Prediction ► *Underline Video*

Sizhe Wang, Ziqi Xu, Claire Najjuuko, Charles Alba and Chenyang Lu

Clinical language models (LMs) are increasingly applied to support clinical risk prediction from free-text notes, yet their uncertainty estimates often remain poorly calibrated and clinically unreliable. In this work, we propose Clinical Uncertainty Risk Alignment (CURA), a framework that aligns clinical LM-based risk estimates and uncertainty with both individual error likelihoods and cohort-level ambiguities. CURA first fine-tunes domain-specific clinical LMs to obtain task-adapted patient embeddings, and then performs uncertainty fine-tuning of a multi-head classifier using a bi-level uncertainty objective. Specifically, an individual-level calibration term aligns predictive uncertainty with each patients likelihood of error, while a cohort-aware regularizer pulls risk estimates toward event rates in their local neighborhoods in the embedding space and places extra weight on ambiguous cohorts near the decision boundary. We further show that this cohort-aware

term can be interpreted as a cross-entropy loss with neighborhood-informed soft labels, providing a label-smoothing view of our method. Extensive experiments on MIMIC-IV clinical risk prediction tasks across various clinical LMs show that CURA consistently improves calibration metrics without substantially compromising discrimination. Further analysis illustrates that CURA reduces overconfident false reassurance and yields more trustworthy uncertainty estimates for downstream clinical decision support.

[Long] CURE-Med: Curriculum-Informed Reinforcement Learning for Multilingual Medical Reasoning ▶ *Underline Video*

Eric Onyame, Akash Ghosh, Subhadip Baidya, Sriparna Saha, Xiuying Chen and Chirag Agarwal

While large language models (LLMs) have shown to perform well on monolingual mathematical and commonsense reasoning, they remain unreliable for multilingual medical reasoning applications, hindering their deployment in multilingual healthcare settings. We address this by first introducing CURE-Med-Bench, a high-quality multilingual medical reasoning dataset with open-ended reasoning queries with a single verifiable answer, spanning thirteen languages, including underrepresented languages such as Amharic, Yoruba, and Swahili. Building on this dataset, we propose CURE-Med, a curriculum-informed reinforcement learning framework that integrates code-switching-aware supervised fine-tuning and Group Relative Policy Optimization to jointly improve logical correctness and language stability. Across thirteen languages, our approach consistently outperforms strong baselines and scales effectively, achieving 85.21% language consistency and 54.35% logical correctness at 7B parameters and 94.96% language consistency and 70.04% logical correctness at 32B parameters. These results support reliable and equitable multilingual medical reasoning in LLMs. The code and dataset will be made publicly available upon acceptance.

[Findings] Decomposed Trust: Privacy, Adversarial Robustness, Ethics, and Fairness in Low-Rank LLMs ▶ *Underline Video*

Daniel Agyei Asante, Md Mokarram Chowdhury and Yang Li

Large language models (LLMs) have driven major advances across domains, yet their massive size hinders deployment in resource-constrained settings. Low-rank factorization addresses this challenge by compressing models to effectively reduce their computation and memory consumption while maintaining accuracy. While these compressed models boast benign performance and system-level advantages, their trustworthiness implications remain poorly understood. In this paper, we present the first comprehensive study of how low-rank factorization affects LLM trustworthiness across privacy, adversarial robustness, ethics, and fairness, complemented by an explainability-driven analysis of the internal mechanisms behind these trust-related changes. We evaluate multiple LLMs of different sizes and architectures compressed with various low-rank factorization algorithms, revealing key insights: (1) low-rank factorization preserves training data privacy but weakens the protection of personally identifiable information during conversations; (2) adversarial robustness is generally enhanced under compression; (3) ethics degrades in zero-shot prompting but partially recovers in few-shot prompting; (4) fairness declines under compression. Beyond compression, we investigate how model scale and fine-tuning affect trustworthiness. Additionally, to move beyond black-box analysis, we employ a gradient-based attribution to identify which layers of LLMs contribute most to adversarial robustness.

[Findings] DeFrame: Debiasing Large Language Models Against Framing Effects ▶ *Underline Video*

Kahee Lim, Soyeon Kim and Steven Euijong Whang

As large language models (LLMs) are increasingly deployed in real-world applications, ensuring their fair responses across demographics has become crucial. Despite many efforts, an ongoing challenge is hidden bias: LLMs appear fair under standard evaluations, but can produce biased responses outside those evaluation settings. In this paper, we identify framing – differences in how semantically equivalent prompts are expressed (e.g., A is better than B vs. B is worse than A) – as an underexplored contributor to this gap. We first introduce the concept of framing disparity to quantify the impact of framing on fairness evaluation. By augmenting fairness evaluation benchmarks with alternative framings, we find that (1) fairness scores vary significantly with framing and (2) existing debiasing methods improve overall (i.e., frame-averaged) fairness, but often fail to reduce framing-induced disparities. To address this, we propose a framing-aware debiasing method that encourages LLMs to be more consistent across framings. Experiments demonstrate that our approach reduces overall bias and improves robustness against framing disparities, enabling LLMs to produce fairer and more consistent responses.

[Findings] Detecting AIGenerated Video: A Vision-Language DualView Survey ▶ *Underline Video*

Dylan Xinming Hou, Juntian Zhang, Xu Gu, Yichen Wu, Nils Lukas, Gus Xia, Xiuying Chen and Yuhan Liu

The evolving realism of AI-generated Videos (AIGC-V) is rapidly rendering traditional artifact-centric detection insufficient, necessitating a paradigm shift from low-level inspection to high-level semantic verification. This paper presents a comprehensive survey of AIGC-V detection, reframing the task as Factual Fidelity Verification, which asks whether the events, entities, and physical processes depicted in a video are consistent with real-world facts. To systematize this rapidly evolving field, we propose a Vision-Language Dual-View taxonomy that organizes existing methods into a hierarchical, four-layer landscape, spanning intrinsic cue analysis, spatiotemporal consistency modeling, cross-modal consistency reasoning, and language-guided world-level reasoning. This dual-view framing highlights a fundamental transition from artifact matching to evidence-based semantic verification enabled by vision-language models and agentic reasoning pipelines. Based on a systematic review of 195 papers, we synthesize AIGC-V generation paradigms, survey the landscape of detection methods, and review evaluation metrics and benchmarks in line with proposed views. Finally, we discuss current challenges and identify promising directions toward robust, explainable, and trustworthy detection.

[Long] Diagnosing Hidden Instabilities in Model Editing via Uncertainty Quantification ▶ *Underline Video*

Zihan Gu, TianYi Zhang, Xinyan Zhang, Zhiyuan Wang, Han Zhang, Yuhao Wei, Jiacheng Lu, Tianyi Ma, Xingsheng ZHANG, Hua Zhang and Yue Hu

Model editing provides a promising mechanism for updating large language models (LLMs) without expensive retraining. Existing approaches, particularly locate-and-edit methods based on least-squares optimization, aim to introduce targeted knowledge changes while preserving pre-trained behavior. In this work, we show that this objective is fundamentally fragile under standard single-edit evaluation protocols. We first develop a unified theoretical framework that characterizes activation-based editing as a constrained intervention on intermediate representations. Within this framework, we demonstrate that least-squares edits cannot, in general, isolate target updates from unrelated activations, giving rise to unavoidable interference that accumulates with successive edits. Crucially, this degradation can remain undetected in single-edit settings when assessed using conventional success and locality metrics. To expose such hidden instabilities, we introduce an uncertainty-based evaluation protocol that combines structured semantic perturbations with uncertainty quantification based on Sampling with Perturbation for UQ. By measuring edit-induced growth in aleatoric and epistemic uncertainty, our method reveals local knowledge conflicts that are invisible to existing benchmarks. Extensive experiments across multiple models, datasets, and editing algorithms show that both least-squares and other parameter-update-based methods consistently increase post-edit uncertainty. Together, our results suggest that current evaluation practices substantially overestimate the reliability of single-edit model editing, and that uncertainty-based diagnostics are necessary for assessing edit stability.

[Long] DiNO: Disinformation Narrative Observer ▶ *Underline Video*

Witold Sosnowski, Arkadiusz Modzelewski, Kinga Skorupska and Adam Wierzbicki

Disinformation is an escalating global threat, making it essential to understand its content, dissemination, and evolution. To confront this challenge, researchers have begun grouping related false claims into broader disinformation narratives, which can be tracked across cultures, time periods, and media sources. Analyzing these narratives provides critical insights for developing more effective countermeasures. To this end, we introduce DiNO: Disinformation Narrative Observer, a novel method designed to extract disinformation narratives from news articles. We applied DiNO to news articles on the Ukraine War, COVID-19 and Migration, sourced from disinformation-prone outlets as well as a reputable source. We evaluated the narratives extracted by DiNO by measuring how well their topics and stances aligned with a recognized disinformation narratives dataset. DiNO outperforms competitive narrative mining approaches, including Relatio and CaNarEx, achieving a 41%-44% improvement in topical alignment and a 30%-41% improvment in stance alignment.

[Long] Dissecting Failure Dynamics in Large Language Model Reasoning ▶ *Underline Video*

Wei Zhu, Jian Zhang, Lixing Yu, Kun Yue and Zhiwen Tang

Large Language Models (LLMs) achieve strong performance through extended inference-time deliberation, yet how their reasoning failures arise remains poorly understood. By analyzing model-generated reasoning trajectories, we find that errors are not uniformly distributed but often originate from a small number of early transition points, after which reasoning remains locally coherent but globally incorrect. These transitions coincide with localized spikes in token-level entropy, and alternative continuations from the same intermediate state can still lead to correct solutions. Based on these observations, we introduce GUARD, a targeted inference-time framework that probes and redirects critical transitions using uncertainty signals. Empirical evaluations across multiple benchmarks confirm that interventions guided by these failure dynamics lead to more reliable reasoning outcomes. Our findings highlight the importance of understanding when and how reasoning first deviates, complementing existing approaches that focus on scaling inference-time computation.

[Findings] Distributional Alignment for Large Language Models under Domain Shift ▶ *Underline Video*

Viet-Thanh Pham, Lizhen Qu, Zhuang Li and Gholamreza Haffari

Distributional alignment enables large language models (LLMs) to predict how a target population distributes its responses across answer options, rather than collapsing disagreement into a single consensus answer. However, existing LLM-based distribution prediction is often unstable and degrades under cultural and domain shift. Token score-based estimates can change with minor option wording or formatting, response sampling-based estimates are expensive and sensitive to prompts and decoding settings, and directly generated distributions are frequently miscalibrated. We propose Evi-DA, an evidence-based alignment technique that improves the fidelity and robustness of LLM-based distribution estimation under domain and cultural shift. Given a target country and a multiple-choice question, Evi-DA retrieves related World Values Survey items and their answer distributions, predicts a coarse Welzel value signature for each option, and infers the country-conditioned answer distribution in a structured format. We train the LLMs using a two-stage pipeline, where reinforcement learning optimizes survey-derived rewards that encourage accurate intermediate value predictions, faithful final distributions, well-formed structured outputs, and reduced cultural bias. Across in-domain and out-of-domain benchmarks and multiple open-source backbones, Evi-DA reduces Jensen-Shannon divergence between predicted and gold distributions relative to strong baselines, with average relative improvements of up to 44%.

[Short] Diving into the Decoding Space of Non-Autoregressive Models via Lexically Constrained Search ▶ *Underline Video*

Chenyang Huang and Osmar R. Zaïane

Non-autoregressive (NAR) models have been mainly developed to improve decoding efficiency. Lately, they have also shown great potential in controlled text generation tasks. In this work, we investigate the decoding space of NAR models through lexically constrained machine translation tasks, and develop a search-based decoding algorithm named LexMAP, which is comparable to the autoregressive Grid Beam Search (GBS) method. Our analysis reveals several interesting properties of NAR decoding: 1) the NAR-based method does not suffer from the MAP degradation issue as the autoregressive method does; 2) AR beam search exhibits strong positional bias, in which the candidates only diverge at the end of the sequence; 3) NAR search explores a larger portion of the probability space, suggesting that the search algorithm

better exploits the model's potential.

[Findings] Domain-Specific Data Generation Framework for RAG Adaptation ▶ *Underline Video*
Chris Xing Tian, Weihao Xie, Zhen Chen, Hui Liu, Zhengyuan Yi, Haoliang Li, Shiqi Wang and Siwei Ma
Retrieval-Augmented Generation (RAG) combines the language understanding and reasoning capabilities of large language models (LLMs) with external retrieval to produce domain-grounded responses. Effectively adapting RAG systems to domain-specific settings requires specialized, context-rich training data beyond general-purpose question-answering datasets. Here, we propose RAGen, a scalable and modular data-centric framework for generating domain-grounded question-answer-context (QAC) triples tailored to diverse RAG adaptation strategies. These QAC triples serve as training signals for multiple RAG adaptation approaches; in this work, we demonstrate their use for contrastive fine-tuning of embedding models and supervised fine-tuning of LLMs under retrieved contexts. RAGen generates QAC triples by identifying key concepts within documents, producing diverse questions guided by Blooms Taxonomy-inspired principles, and pairing them with precise answers extracted from relevant contexts. Its modular pipeline incorporates semantic chunking, hierarchical concept extraction, multi-chunk retrieval, and curated distractor contexts to encourage robust reasoning. Designed for scalability, RAGen efficiently handles large and evolving document corpora without redundant processing, making it particularly suitable for dynamic domains like enterprise knowledge bases.

[Findings] Don't Just Listen, Try Planning: Graph-based Retrieval-Generation Agent for Long-form Audio Meeting Understanding ▶ *Underline Video*
Dong Zhang
Long-form audio meeting understanding (LAMU) is gaining attention, but dedicated question answering (QA) datasets are lacking. Previous tailored speech QA and existing Speech LLMs suffer from acoustic information loss and poor long-term dependency capture. We construct the LongAudioQA dataset and propose the GRGA model, which models heterogeneous audio features into a multi-dimensional graph and leverages agent planning for retrieval and answer generation, effectively addressing existing limitations.

[Findings] DRA-GRPO: Your GRPO Needs to Know Diverse Reasoning Paths for Mathematical Reasoning ▶ *Underline Video*
Xiwen Chen, Wenhui Zhu, Peijie Qiu, XUANZHAO DONG, Hao Wang, Haiyu Wu, Huayu Li, Aris Sotiras, Yalin Wang and Abolfazl Razi
Post-training LLMs with Reinforcement Learning, specifically Group Relative Policy Optimization (GRPO), has emerged as a paradigm for enhancing mathematical reasoning. However, standard GRPO relies on scalar correctness rewards that are often non-injective with respect to semantic content: distinct reasoning paths receive identical rewards. This leads to a Diversity-Quality Inconsistency, where the policy collapses into a narrow set of dominant modes while ignoring equally valid but structurally novel strategies. To bridge this gap, we propose Diversity-aware Reward Adjustment (DRA), a theoretically grounded framework that calibrates the reward signal using the semantic density of sampled groups. By leveraging Submodular Mutual Information (SMI), DRA implements an Inverse Propensity Scoring (IPS) mechanism that effectively de-biases the gradient estimation. This creates a repulsive force against redundancy, driving the policy to achieve better coverage of the high-reward landscape. Our method is plug-and-play and integrates seamlessly with GRPO variants. Empirical evaluations on five math benchmarks demonstrate that DRA-GRPO consistently outperforms strong baselines, achieving an average accuracy of 58.2% on DeepSeek-R1-Distill-Qwen-1.5B with only 7,000 training samples and $55 cost, highlighting the critical role of diversity calibration in data-efficient alignment.

[Long] Efficient KL Divergence Estimation via Truncated Top-K Integration for Large Language Models ▶ *Underline Video*
Xinyuan Wang, Luozhijie Jin, Bo Wang, Yuan Li, Zhangyue Yin and Xipeng Qiu
Kullback-Leibler (KL) divergence regularization is essential for stabilizing reinforcement learning from human feedback (RLHF) in large language models (LLMs), yet its exact computation requires summing over vocabularies of all tokens, incurring prohibitive memory costs during training. Existing stochastic estimators circumvent this bottleneck by estimating KL divergence using only the sampled token from the trajectory, but suffer from high variance (k_1) or systematic bias (k_2). We propose TIKE (Top-k Importance-weighted KL Estimator), which exploits the Zipfian structure of language model distributions: by deterministically integrating over only the top-k tokens, TIKE captures most of the probability mass while effectively reducing memory cost. To ensure correctness in off-policy settings characteristic of Group Relative Policy Optimization (GRPO), we incorporate importance sampling weights that correct for distribution shift between rollout and optimization policies. Experiments on models across diverse benchmarks demonstrate that TIKE consistently outperforms stochastic baselines, while exhibiting substantially lower gradient variance. Our analysis reveals that TIKE closely tracks the exact Rao-Blackwellized estimator with near-zero variance, offering a practical path toward stable, memory-efficient KL regularization for reasoning-intensive LLMs training.

[Findings] EMSDialog: Synthetic Multi-person Emergency Medical Service Dialogue Generation from Electronic Patient Care Reports via Multi-LLM Agents ▶ *Underline Video*
Xueren Ge, Sahil Murtaza, Anthony Cortez and Homa Alemzadeh
Conversational diagnosis prediction requires models to track evolving evidence in streaming clinical conversations and decide when to commit to a diagnosis. Existing medical dialogue corpora are largely dyadic or lack the multi-party workflow and annotations needed for this setting. We introduce an ePCR-grounded, topic-flow-based multi-agent generation pipeline that iteratively plans, generates, and self-refines dialogues with rule-based factual and topic flow checks. The pipeline yields EMSDialog, a dataset of 4,414 synthetic multi-speaker EMS conversations based on a real-world ePCR dataset, annotated with 43 diagnoses, speaker roles, and turn-level topics. Human and

LLM evaluations confirm high quality and realism of EMSDialog using both utterance- and conversation-level metrics. Results show that EMSDialog-augmented training improves accuracy, timeliness, and stability of EMS conversational diagnosis prediction. Our datasets and code are publicly available at https://uva-dsa.github.io/EMSDialog

[Findings] ErrorRadar: Benchmarking Complex Mathematical Reasoning of Multimodal Large Language Models Via Error Detection ▶ *Underline Video*

Yibo Yan, Shen Wang, Jiahao Huo, Hang Li, Boyan Li, Jiamin Su, Xiong Gao, YiFan Zhang, Tianlong Xu, Zhendong Chu, Aoxiao Zhong, Kun Wang, Hui Xiong, Philip S. Yu, Xuming Hu and Qingsong Wen

As the field of Multimodal Large Language Models (MLLMs) continues to evolve, their potential to handle mathematical reasoning tasks is promising, as they can handle multimodal questions via cross-modal understanding capabilities compared to text-only LLMs. Current mathematical benchmarks predominantly focus on evaluating MLLMs' problem-solving ability, yet there is a crucial gap in addressing more complex scenarios such as error detection, for enhancing reasoning capability in complicated settings. To fill this gap, we formally formulate the new task , multimodal error detection, and introduce ErrorRadar, the first benchmark designed to assess MLLMs' capabilities in such a task. ErrorRadar evaluates two sub-tasks: error step identification and error categorization, providing a framework for evaluating MLLMs' complex mathematical reasoning ability. It consists of 2,500 high-quality multimodal K-12 mathematical problems, collected from real-world student interactions in an educational organization, with expert-based annotation and metadata such as problem type and error category. Through extensive experiments, we evaluated both open-source and closed-source representative MLLMs, benchmarking their performance against educational expert evaluators. Results indicate challenges still remain, as GPT-4o with best model performance is still around 10% behind human evaluation

[Long] Experience Retrieval-Augmentation with Electronic Health Records Enables Accurate Discharge QA ▶ *Underline Video*

Justice Ou, Tinglin Huang, Yilun Zhao, Ziyang Yu, Peiqing Lu, Yifei Shen and Rex Ying

To improve the reliability of Large Language Models (LLMs) in clinical applications, retrieval-augmented generation (RAG) is extensively applied to provide factual medical knowledge. However, beyond general medical knowledge from open-ended datasets, clinical case-based knowledge is also critical for effective medical reasoning, as it provides context grounded in real-world patient experiences. Motivated by this, we propose Experience Retrieval-Augmentation ExpRAG framework based on Electronic Health Record(EHR), aiming to offer the relevant context from other patients' discharge reports. ExpRAG performs retrieval through a coarse-to-fine process, utilizing an EHR-based report ranker to efficiently identify similar patients, followed by an experience retriever to extract task-relevant content for enhanced medical reasoning. To evaluate ExpRAG, we introduce DischargeQA, a clinical QA dataset with 1,280 discharge-related questions across diagnosis, medication, and instruction tasks. Each problem is generated using EHR data to ensure realistic and challenging scenarios. Experimental results demonstrate that ExpRAG consistently outperforms a text-based ranker, achieving an average relative improvement of 5.2%, highlighting the importance of case-based knowledge for medical reasoning.

[Long] Explain the Synth: Interpretable Evaluation of LLM Data Synthesis ▶ *Underline Video*

Yue Yang, Fan Yang, Yu Bai and Hao Wang

Large language models (LLMs) are increasingly used to generate synthetic data, in which tabular data constitute a fundamental data modality across a wide range of domains. Yet, current evaluation practices often provide limited insights into whether the synthetic data preserve real data-generating relationships or introduce plausible-looking artifacts. We present a conceptually simple, interpretable auditing framework that compares the explanatory structure induced by real versus synthetic data. The key idea is to use a transparent rule-based model as a shared explanatory language: we extract rules from real data to summarize how features relate to labels, then examine how this rule structure changes when explained using LLM-generated data. Importantly, these rules are derived by an independent rule auditor rather than by the generator itself. The resulting "explanation shift" reveals which relationships are preserved, weakened, removed, or newly introduced by the generator, offering actionable diagnostics beyond aggregate fidelity scores. We further provide a theoretical perspective that links explanation shift and cross-domain predictive gaps to distribution mismatch within an interpretable hypothesis class. Overall, our approach turns synthetic data evaluation into a human-auditable comparison of explanations, improving transparency for LLM-based tabular synthesis.

[Short] Exploring Cross-Client Memorization of Training Data in Large Language Models for Federated Learning ▶ *Underline Video*

Tinnakit Udsa, Can Udomcharoenchaikit, Patomporn Payoungkhamdee, Sarana Nutanong and Norrathep Rattanavipanon

Federated learning (FL) enables collaborative training without raw data sharing, but still risks training data memorization. Existing FL memorization detection techniques focus on one sample at a time, underestimating more subtle risks of cross-sample memorization. In contrast, recent work on centralized learning (CL) has introduced fine-grained methods to assess memorization across all samples in training data, but these assume centralized access to data and cannot be applied directly to FL. We bridge this gap by proposing a framework that quantifies both intra- and inter-client memorization in FL using fine-grained cross-sample memorization measurement across all clients. Based on this framework, we conduct two studies: (1) measuring subtle memorization across clients and (2) examining key factors that influence memorization, including decoding strategies, prefix length, and FL algorithms. Our findings reveal that FL models do memorize client data, particularly intra-client data, more than inter-client data, with memorization influenced by training and inferencing factors.

[Long] FactVerse: A Benchmark for Factual Consistency in Interleaved Image-Text Generation ▶ *Underline Video*

Yubo Shan, Kun Zhang, Qiming Xu, Liping Cao, Yingying Cao, Jian Zhang, Yu Wang, Jingyuan Li and Yuanzhuo Wang

Interleaved multimodal understanding and generation, where models can interactively comprehend and produce images and text in arbitrary orders, has emerged as a key research direction in generative Multimodal Large Language Models(MLLMs). Such interleaved image-text content plays an increasingly important role in information dissemination. However, the compounded persuasive power of multimodal narratives also raises the risk of factual misinformation. Despite this, existing benchmarks lack effective mechanisms to evaluate factual consistency in interleaved image-text content. To bridge this gap, we introduce FactVerse, a benchmark dedicated to evaluating factual consistency in interleaved image-text generation. FactVerse comprises 3,000 human-verified instances across four categories and 50 domains, supporting both English and Chinese. We also establish a multi-dimensional evaluation framework designed to rigorously assess factual consistency. Experiments demonstrate that our framework achieves high alignment with human judgments, significantly outperforming existing evaluation methods. Furthermore, our analysis reveals systematic deficiencies in current models, offering critical insights for future design.

[Findings] FastKV: Decoupling of Context Reduction and KV Cache Compression for Prefill-Decoding Acceleration ▶ *Underline Video*

Dongwon Jo, Jiwon Song, Yulhwa Kim and Jae-Joon Kim

While large language models (LLMs) excel at handling long-context sequences, they require substantial prefill computation and key-value (KV) cache, which can heavily burden computational efficiency and memory usage in both prefill and decoding stages. Recent works that compress KV caches with prefill acceleration reduce this cost but inadvertently tie the prefill compute reduction to the decoding KV budget. This coupling arises from overlooking the layer-dependent variation of critical context, often leading to accuracy degradation. To address this issue, we introduce FastKV, a KV cache compression framework designed to reduce latency in both prefill and decoding by leveraging the stabilization of token importance in later layers. FastKV performs full-context computation until a Token-Selective Propagation (TSP) layer, which forwards only the most informative tokens to subsequent layers. From these propagated tokens, FastKV independently selects salient KV entries for caching, thereby decoupling KV budget from the prefill compute reduction based on the TSP decision. This independent control of the TSP rate and KV retention rate enables flexible optimization of efficiency and accuracy. Experimental results show that FastKV achieves speedups of up to 1.82x in prefill and 2.87x in decoding compared to the full-context baseline, while matching the accuracy of the decoding-only baselines. Our code is available at https://github.com/dongwonjo/FastKV.

[Long] Fine-Grained Analysis of Shared Syntactic Mechanisms in Language Models ▶ *Underline Video*

Ryoma Kumon and Hitomi Yanaka

While language models demonstrate sophisticated syntactic capabilities, the extent to which their internal mechanisms align with cross-constructional principles studied in linguistics remains poorly understood. This study investigates whether models employ shared neural mechanisms across different syntactic constructions by applying causal interpretability methods at a granular level. Focusing on filler-gap dependencies and negative polarity item (NPI) licensing, we utilize activation patching to identify the functional roles of specific attention heads and MLP blocks. Our results reveal a highly localized and shared mechanism for filler-gap dependencies located in the early to middle layers, whereas NPI processing exhibits no such unified mechanism. Furthermore, we find that these mechanisms identified by activation patching generalize to out-of-distribution, while distributed alignment search, a supervised interpretability method, is susceptible to overfitting on narrow linguistic distributions. Finally, we validate our findings by demonstrating that the manipulation of the identified components improves model performance on acceptability judgment benchmarks.

[Long] FLARE: Fine-Grained Length-Aware Routing for Resource-Efficient Heterogeneous LLM Serving ▶ *Underline Video*

Yujia Fu, Heming Zhong, Dan Huang and Yutong Lu

With the rapid proliferation of large language models (LLMs), model pools have become increasingly heterogeneous in both capability and efficiency. Larger LLMs can improve quality but incur higher latency and cost, while smaller LLMs are the opposite, making per-query model selection crucial in practice. This has spawned LLM routers that dispatch each query to an appropriate model. Existing routers lack fine-grained resource awareness across deployment settings, which degrades efficiency metrics in real-world serving. To this end, We propose FLARE, a length-centric, resource-aware multi-LLM routing framework that uses length-based models to estimate per-query latency and cost. FLARE formulates routing as a discrete multi-objective optimization problem to achieve efficient trade-off. Experiments show that FLARE reduces latency and cost by up to 68% and 75% while maintaining competitive accuracy, and can be easily applied to new datasets and LLMs.

[Long] For-Value: Efficient Forward-Only Data Valuation for finetuning LLMs and VLMs ▶ *Underline Video*

Wenlong Deng, Qi Zeng, Jiaming Zhang, Minghui Chen, Zixin Ding, Christos Thrampoulidis, Boying Gong and Xiaoxiao Li

Data valuation is essential for enhancing the transparency and accountability of large language models (LLMs) and vision-language models (VLMs). However, existing methods typically rely on gradient computations, making them computationally prohibitive for billion-parameter models and precluding batch parallelization. In this work, we introduce For-Value, a forward-only data valuation framework that enables efficient batch-scalable value estimation while maintaining effectiveness. Leveraging the expressive power of pretrained LLMs/VLMs, we theoretically demonstrate that data valuation can be captured by the alignment between the final hidden representations and prediction errors at the last layer. In light of this insight, For-Value computes data value using a simple closed-form expression with a single forward pass, eliminating the need for costly backpropagation and enabling efficient batch calculating at scale. Extensive experiments show that For-Value matches or outperforms gradient-based baselines in detecting influential data and mislabeled data, while achieving significant efficiency improvements.

[Findings] From Domains to Instances: Dual-Granularity Data Synthesis for LLM Unlearning ► *Underline Video*

Xiaoyu Xu, Minxin Du, Zitong Li, Zi Liang, Zhibiao Guo, Shiyu Zhang, Peizhao Hu, Qingqing Ye and Haibo Hu

Although machine unlearning is essential for removing private, harmful, or copyrighted content from LLMs, current benchmarks often fail to faithfully represent the true "forgetting scope" learned by the model. We formalize two distinct unlearning granularities, domain-level and instance-level, and propose, an automated framework for synthesizing high-quality forget sets. Unlike prior work relying on external generators, exploits the target model per se to elicit data that matches its internal knowledge distribution through seed-guided and adversarial prompting. Our experiments across diverse benchmarks show that it achieves a superior balance of relevance, diversity, and efficiency. Quantitatively, in the Harry Potter domain, it improves relevance by 20 and diversity by 0.05 while halving the total data size compared to SOTAs. Ultimately, it facilitates more robust forgetting and better utility preservation, providing a more rigorous foundation for evaluating LLM unlearning.

[Findings] From If-Statements to ML Pipelines: Revisiting Bias in Code-Generation ► *Underline Video*

Minh Duc Bui, Xenia Heilmann, Mattia Cerrato, Manuel Mager and Katharina von der Wense

Prior work evaluates code generation bias primarily through simple conditional statements, which represent only a narrow slice of real-world programming and reveal solely overt, explicitly encoded bias. We demonstrate that this approach dramatically underestimates real-world bias by examining a more realistic task: generating machine learning (ML) pipelines. Testing both code-specialized and general-instruction large language models, we find that ML pipelines exhibit substantially greater bias than simple conditionals across all conditions: standard generation, with varying prompt-based mitigation strategies, varying numbers of attributes, and different ML pipeline difficulty levels. Even attribute selection alone, the simplest pipeline difficulty, shows higher bias compared to conditionals, demonstrating that ML pipelines inherently amplify bias beyond what isolated conditionals reveal. Critically, we uncover a stark asymmetry: models maintain equivalent bias detection performance on both simple conditionals and ML pipelines, revealing that models recognize bias equally well in both contexts yet generate significantly more biased code in ML pipelines. These findings challenge simple conditionals as valid proxies for bias evaluation and suggest current benchmarks mischaracterize model safety in practical deployment contexts.

[Long] From Isolation to Entanglement: When Do Interpretability Methods Identify and Disentangle Known Concepts? ► *Underline Video*

Aaron Mueller, Andrew Lee, Shruti Joshi, Ekdeep Singh Lubana, Dhanya Sridhar and Patrik Reizinger

A goal of interpretability is to recover disentangled representations of latent concepts (features) from the activations of neural networks. The quality of features is typically evaluated in isolation, and under implicit independence assumptions that may not hold in practice. Thus, it is unclear to what extent common featurization methods such as sparse autoencoders (SAEs) and probes disentangle one concept from another. We propose a multi-concept evaluation setting using concepts such as sentiment, domain, voice, and tense. We evaluate how well featurizers produce disentangled representations of each concept, observing that features are typically sensitive to only one concept, but also that concepts are distributed across many features. Then, we steer these features, measuring whether each concept is independently manipulable, and whether features interact. Even in idealized settings, steering a feature often affects many concepts, despite a near absence of interaction effects. These results suggest that correlational metrics are insufficient to establish steering selectivity, and that demonstrating that two features operate in separate spaces is insufficient to claim that they will be selective for one concept. These results underscore the importance of multi-concept evaluations in interpretability research.

[Long] From Word Sequences to Behavioral Sequences: Adapting Modeling and Evaluation Paradigms for Longitudinal NLP ► *Underline Video*

Adithya V Ganesan, Vasudha Varadarajan, Oscar Kjell, Whitney Ringwald, Scott M Feltman, Benjamin J. Luft, Romon Kotov, Ryan L Boyd and H. Andrew Schwartz

While NLP typically treats documents as independent and unordered samples, in longitudinal studies, this assumption rarely holds: documents are nested within authors and ordered in time, forming person-indexed, time-ordered behavioral sequences. Here, we demonstrate the need for and propose a longitudinal modeling and evaluation paradigm that consequently updates four parts of the NLP pipeline: (1) evaluation splits aligned to generalization over people (cross-sectional) and/or time (prospective); (2) accuracy metrics separating between-person differences from within-person dynamics; (3) sequence inputs to incorporate history by default; and (4) model internals that support different coarseness of latent state over histories (pooled summaries, explicit dynamics, or interaction-based models). We demonstrate the issues ensued by traditional pipeline and our proposed improvements on a dataset of 17k daily diary transcripts paired with PTSD symptom severity from 238 participants, finding that traditional document-level evaluation can yield substantially different and sometimes reversed conclusions compared to our ecologically valid modeling and evaluation. We tie our results to a broader discussion motivating a shift from word-sequence evaluation toward behavior-sequence paradigms for NLP.

[Long] Graph-Based Alternatives to LLMs for Human Simulation ► *Underline Video*

Joseph Suh, Suhong Moon and Serina Chang

Large language models (LLMs) have become a popular approach for simulating human behaviors, yet it remains unclear if LLMs are necessary for all simulation tasks. We study a broad family of close-ended simulation tasks, with applications from survey prediction to test-taking, and show that a graph neural network can match or surpass strong LLM-based methods. We introduce Graph-basEd Models

for Human Simulation (GEMS) which formulates close-ended simulation as link prediction on a heterogeneous graph of individuals and choices. Across three datasets and three evaluation settings, GEMS matches or outperforms the strongest LLM-based methods while using three orders of magnitude fewer parameters. These results suggest that graph-based modeling can complement LLMs as an efficient and transparent approach to simulating human behaviors. Code is available at https://github.com/schang-lab/gems.

[Findings] GRE Score: Generative Risk Evaluation for Large Language Models ▶ *Underline Video*

ZAITANG LI, Pin-Yu Chen and Tsung-Yi Ho

Large Language Models (LLMs) have revolutionized generative tasks, but concerns about their trustworthiness and vulnerability to adversarial attacks persist. This paper introduces the Generative Robustness Evaluation (GRE) Score, a novel metric designed to assess LLMs' resilience against adversarial red teaming attempts that may compromise model compliance and elicit undesired responses. Our approach utilizes conditional generation for synthetic text creation, offering an attack-independent evaluation of LLM robustness. By calculating the margin in refusal scores, we quantify the robustness of LLMs in an attack-agnostic manner. We evaluate our method on five dimensions with specified datasets, encompassing ethical considerations, safety protocols, and potential misuse scenarios. We present four contributions: (1) The GRE Score framework, which establishes a textual robustness certificate for LLMs against adversarial red teaming attempts, providing a theoretical foundation for quantifying model resilience. (2) Comprehensive evaluations across five dimensions using eight prominent LLMs, validating GRE Scores with adversarial red teaming attacks. Our method demonstrates a consistent ranking of LLM robustness when compared to the attack-based model ranking on TrustLLM huang2024trustllm while achieving a significant 5-8x speedup compared to traditional evaluation techniques. (3) Insights into the non-linear relationship between model scaling and performance, revealing that larger models do not always perform better, and an analysis of how instruction-tuning impacts robustness across LLMs. (4) The discovery that all evaluated LLMs exhibit lower performance in robustness and privacy tasks compared to other areas, highlighting a critical gap in capabilities.

[Findings] Ground Then Rank: Revisiting Knowledge-Based VQA with Training-Free Entity Identification ▶ *Underline Video*

Qian Ma, Qiong Wu, Zhengyi Zhou and Yao Ma

Knowledge-Based Visual Question Answering (KB-VQA) requires grounding visual queries to external knowledge beyond directly observable content in images. While recent multi modal large language models (MLLMs) show strong perceptual abilities, they struggle on KB-VQA tasks requiring groundings from both fine-grained entity and evidence levels. Most existing multi-modal retrieval augmented generation (MM-RAG) methods tightly couple entity discrimination and section-level evidence ranking into a single re-ranking stage, leading to high cost and limited generalization. In this work, we revisit existing MM-RAG solutions from a workflow perspective and argue both entity-level and fact-level groundings are key bottlenecks. We observe that although MLLMs often fail under open-ended entity naming, they can better identify the correct entity when selecting from a small set of candidate names. Based on this insight, we propose a simple and training-free identify-before-answer IBA framework that decouples entity identification from section-level re-ranking. Our approach prompts an MLLM to select high-confidence entities using only candidate names, followed by an off-the-shelf textual re-ranker for evidence selection. Experiments on Encyclopedic-VQA and InfoSeek show that our method consistently outperforms fine-tuned multi-modal re-ranking baselines while reducing training and inference complexity. Additional analyses reveal that the improvements arise not only from better entity identification, but also from selecting more informative evidence once correct entity is fixed. Our implementation is made public to ease reproducibility

[Long] How Instruction and Reasoning Data shape Post-Training: Data Quality through the Lens of Layer-wise Gradients ▶ *Underline Video*

Ming Li

As the post-training of large language models (LLMs) advances from instruction-following to complex reasoning tasks, understanding how different data affect finetuning dynamics remains largely unexplored. In this paper, we present a spectral analysis of layer-wise gradients induced by low/high-quality instruction and reasoning data for LLM post-training. Our analysis reveals that widely-studied metrics for data evaluation, e.g., IFD, InsTag, Difficulty, and Reward, can be explained and unified by spectral properties computed from gradients' singular value decomposition (SVD). Specifically, higher-quality data are usually associated with lower nuclear norms and higher effective ranks. Notably, effective rank exhibits better robustness and resolution than nuclear norm in capturing subtle quality differences. For example, reasoning data achieves substantially higher effective ranks than instruction data, implying richer gradient structures on more complex tasks. Our experiments also highlight that models within the same family share similar gradient patterns regardless of their sizes, whereas different model families diverge significantly. Providing a unified view on the effects of data quality across instruction and reasoning data, this work illuminates the interplay between data quality and training stability, shedding novel insights into developing better data exploration strategies for post-training.

[Findings] HTMuon: Improving Muon via Heavy-Tailed Spectral Correction ▶ *Underline Video*

Tianyu Pang, Yujie Fang, Zihang Liu, Shenyang Deng, Lei Hsiung, Shuhua Yu and Yaoqing Yang

Muon has recently shown promising results in LLM training. In this work, we study how to further improve Muon. We argue that Muon's orthogonalized update rule suppresses the emergence of heavy-tailed weight spectra and over-emphasizes the training along noise-dominated directions. Motivated by the Heavy-Tailed Self-Regularization (HT-SR) theory, we propose HTMuon. HTMuon preserves Muon's ability to capture parameter interdependencies while producing heavier-tailed updates and inducing heavier-tailed weight spectra. Experiments

on LLM pretraining and image classification show that HTMuon consistently improves performance over state-of-the-art baselines and can also serve as a plug-in on top of existing Muon variants. For example, on LLaMA pretraining on the C4 dataset, HTMuon reduces perplexity by up to 0.98 compared to Muon. We further theoretically show that HTMuon corresponds to steepest descent under the Schatten-q norm constraint and provide convergence analysis in smooth non-convex settings. The implementation of HTMuon is available at https://github.com/TDCSZ327/HTmuon.

[Long] ImmersiveTTS: Environment-Aware Text-to-Speech with Multimodal Diffusion Transformer and Domain-Specific Representation Alignment ► *Underline Video*

Jun-Hak Yun, Seung-Bin Kim and Seong-Whan Lee

Recent advancements in text-guided audio generation have yielded promising results in diverse domains, including sound effects, speech, and music. However, jointly generating speech with environmental audio remains challenging due to the inherent disparities in their acoustic patterns and temporal dynamics. We propose ImmersiveTTS, an environment-aware text-to-speech (TTS) model that generates natural speech seamlessly integrated within environmental contexts by explicitly modeling cross-modal interactions. Our model builds on a multimodal diffusion transformer and fuses transcript-aligned speech latent with text-conditioned environmental context via joint attention. To enhance semantic consistency, we introduce a domain-specific representation alignment objective tailored to environment-aware TTS, leveraging complementary self-supervised representations from speech and audio encoders. Experimental results show that ImmersiveTTS achieves higher naturalness, intelligibility, and audio fidelity than existing approaches across objective metrics and human listening tests.

[Long] Improving the Distributional Alignment of LLMs using Supervision ► *Underline Video*

Gauri Kambhatla, Sanjana Gautam, Angela Zhang, Alexander Liu, Ravi Srinivasan, Junyi Jessy Li and Matthew Lease

The ability to accurately align LLMs with diverse population groups on subjective questions would have great value. In this work, we show that adding simple supervision can more consistently improve the alignment of LLM-generated distributions with diverse population groups, as measured across three datasets spanning public health, public opinion, and values and beliefs. Beyond evaluating average alignment, we also report how alignment varies across specific groups. Our broad findings provide insights into the distributional alignment of LLM generations with diverse populations. By conducting evaluation over many LLMs and prompting strategies, we provide a benchmark to stimulate future research.

[Long] IRIS: Interleaved Reinforcement with Incremental Staged Curriculum for Cross-Lingual Mathematical Reasoning ► *Underline Video*

Navya Gupta, Rishitej Reddy Vyalla, Avinash Anand, Chhavi Kirtani, Erik Cambria, Zhengchen Zhang, Zhengkui Wang, Timothy Liu, Aik Beng Ng, Simon See and Rajiv Ratn Shah

Curriculum learning helps language models tackle complex reasoning by gradually increasing task difficulty. However, it often fails to generate consistent step-by-step reasoning, especially in multilingual and low-resource settings where cross-lingual transfer from English to Indian languages remains limited. We propose IRIS: Interleaved Reinforcement with Incremental Staged Curriculum, a two-axis framework that combines Supervised Fine-Tuning on progressively harder problems (vertical axis) with Reverse Curriculum Reinforcement Learning to reduce reliance on step-by-step guidance (horizontal axis). We design a composite reward combining correctness, step-wise alignment, continuity, and numeric incentives, optimized via Group Relative Policy Optimization (GRPO). We release CL-Math, a dataset of 29k problems with step-level annotations in English, Hindi, and Marathi. Across standard benchmarks and curated multilingual test sets, IRIS consistently improves performance, with strong results on math reasoning tasks and substantial gains in low-resource and bilingual settings, alongside modest improvements in high-resource languages. Our code and dataset will be publicly available at https://github.com/avinanand/IRIS-Interleaved-Reinforcement-

[Long] Knowledge-to-Verification: Exploring RLVR for LLMs in Knowledge-Intensive Domains ► *Underline Video*

Zhonghang Yuan, Zhefan Wang, Fang Hu, Zihong Chen, Jinzhe Li, Gang Li, Jie Ying, Huanjun Kong, Songyang Zhang and Nanqing Dong

Reinforcement learning with verifiable rewards (RLVR) has demonstrated promising potential to enhance the reasoning capabilities of large language models (LLMs) in domains such as mathematics and coding. However, its applications on knowledge-intensive domains have not been effectively explored due to the scarcity of high-quality verifiable data. Furthermore, current RLVR focuses solely on the correctness of final answers, leading to the limitations of flawed reasoning and sparse reward signals. In this work, we propose Knowledge-to-Verification (K2V), a framework that extends RLVR to knowledge-intensive domains through automated verifiable data synthesis, while enabling verification of the LLM's reasoning process. Extensive experiments demonstrate that K2V enhances the reasoning of LLM in knowledge-intensive domains without significantly compromising the model's general capabilities. This study also suggests that integrating automated data synthesis with reasoning verification is a promising direction to enhance model capabilities in these broader domains.

[Findings] Large-Scale Diverse Synthesis for Mid-Training ► *Underline Video*

Xuemiao Zhang, Chengying Tu, Can Ren, Rongxiang Weng, Hongfei Yan, Jingang Wang and Xunliang Cai

Mid-training has become critical for enhancing the knowledge and reasoning ability of large language models (LLMs), especially through the utilization of large-scale synthetic data. However, existing data synthesis methods often generate simplistic and homogeneous QA pairs, with limited scale and diversity. To address this, we propose BoostQA, a novel framework designed to synthesize large-scale, diverse, and high-quality QA data for mid-training. BoostQA introduces model probes during mid-training for the first time and implements

STEM-focused multi-grade synthesis to boost data diversity as well as high-difficulty synthesis to alleviate difficulty degradation, followed by answer refinement to further improve quality. Extensive experiments by mid-training Llama-3 8B demonstrate that using only 20B-token BoostQA data achieves a significant average improvement of 12.74% on MMLU and CMMLU over the pre-training baseline. After mid-training on 500B tokens, including 100B-token BoostQA data, our model achieves SOTA average results across benchmarks among mainstream models of comparable size. BoostQA also demonstrates robust scalability, with performance consistently improving as model size, data volume, and initial FLOPs scale.

[Findings] Law in Silico: Simulating Legal Society with LLM-Based Agents ▶ *Underline Video*
Yiding_Wang1 Yuxuan_Chen11 Fauxu_Meng1 Xifan_Chen1 Xiaolei_Yang1 Muhan_Zhang1
Since real-world legal experiments are often costly or infeasible, simulating legal societies with Artificial Intelligence (AI) systems provides an effective alternative for testing and advancing legal theory, as well as supporting legal administration. Large Language Models (LLMs), with their world knowledge and role-playing capabilities, are strong candidates to serve as the foundation for legal society simulation. However, the application of LLMs to simulate legal systems remains underexplored. In this work, we introduce Law in Silico, a unified LLM-based agent framework for simulating legal scenarios that incorporate individual decision-making and institutional mechanisms, such as legislation, adjudication, and enforcement. We calibrate agent behaviors against real-world crime data, demonstrating that LLM-based agents can capture realistic sociological correlations. Building on this foundation, we structure our simulation through a ”Micro-to-Macro” process: we conduct micro-level simulations in representative conflict-driven scenarios, allowing legal rules to evolve through agent-institution interactions naturally. These evolved laws are then deployed back into macro-scale populations to evaluate their effectiveness in regulating behaviors. Through comprehensive experiments, our results reveal that a well-functioning, transparent, and adaptive legal system can mitigate "cat-and-mouse" regulatory dynamics and offer better protection for vulnerable individuals.

[Findings] LayerNorm Induces Recency Bias in Transformer Decoders ▶ *Underline Video*
Junu Kim, Xiao Liu, Zhenghao Lin, Lei Ji, Yeyun Gong and Edward Choi
Causal self-attention provides positional information to Transformer decoders. Prior work has shown that stacks of causal self-attention layers alone induce a positional bias in attention scores toward earlier tokens. However, this differs from the bias toward later tokens typically observed in Transformer decoders, known as recency bias. We address this discrepancy by analyzing the interaction between causal self-attention and other architectural components. We show that stacked causal self-attention layers combined with LayerNorm induce recency bias. Furthermore, we examine the effects of residual connections and the distribution of input token embeddings on this bias. Our results provide new theoretical insights into how positional information interacts with architectural components and suggest directions for improving positional encoding strategies.

[Long] Learning Uncertainty from Sequential Internal Dispersion in Large Language Models ▶ *Underline Video*
Ponhvoan Srey, Xiaobao Wu, Cong-Duy Nguyen and Anh Tuan Luu
Uncertainty estimation is a promising approach to detect hallucinations in large language models (LLMs). Recent approaches commonly depend on model internal states to estimate uncertainty. However, they suffer from strict assumptions on how hidden states should evolve across layers, and from information loss by solely focusing on last or mean tokens. To address these issues, we present Sequential Internal Variance Representation (SIVR), a supervised hallucination detection framework that leverages token-wise, layer-wise features derived from hidden states. SIVR adopts a more basic assumption that uncertainty manifests in the degree of dispersion or variance of internal representations across layers, rather than relying on specific assumptions, which makes the method model and task agnostic. It additionally aggregates the full sequence of per-token variance features, learning temporal patterns indicative of factual errors and thereby preventing information loss. Experimental results demonstrate SIVR consistently outperforms strong baselines. Most importantly, SIVR enjoys stronger generalisation and avoids relying on large training sets, highlighting the potential for practical deployment.

[Long] LEASH: Adaptive Length Penalty and Reward Shaping for Efficient Large Reasoning Model ▶ *Underline Video*
Yanhao Li, Lu Ma, Jiaran Zhang, Lexiang Tang, Wentao Zhang and Guibo Luo
Large Language Models (LLMs) often produce unnecessarily lengthy reasoning traces, which significantly increase computational cost and latency. Existing approaches typically rely on fixed length penalties, but such penalties are hard to tune and fail to adapt to the evolving reasoning abilities of LLMs, leading to suboptimal trade-offs between accuracy and conciseness. To address this challenge, we propose LEASH (adaptive LEngth penAlty and reward SHaping), a reinforcement learning framework for efficient reasoning in LLMs. We formulate length control as a constrained optimization problem and employ a Lagrangian primal-dual method to dynamically adjust the penalty coefficient. When generations exceed the target length, the penalty is intensified; when they are shorter, it is relaxed. This adaptive mechanism guides models toward producing concise reasoning without sacrificing task performance. Experiments on Deepseek-R1-Distill-Qwen-1.5B and Qwen3-4B-Thinking-2507 show that LEASH reduces the average reasoning length by 60% across diverse tasks, including in-distribution mathematical reasoning and out-of-distribution domains such as coding and instruction following, while maintaining competitive performance. Our work thus presents a practical and effective paradigm for developing controllable and efficient LLMs that balance reasoning capabilities with computational budgets.

[Findings] Like a Therapist, But Not: Reddit Narratives of AI in Mental Health Contexts ▶ *Underline Video*
Elham Aghakhani and Rezvaneh Rezapour

Large language models (LLMs) are increasingly used for emotional support and mental health-related interactions outside clinical settings, yet little is known about how people evaluate and relate to these systems in everyday use. We analyze 5,126 Reddit posts from 47 mental health communities describing experiential or exploratory use of AI for emotional support or therapy. Grounded in the Technology Acceptance Model and therapeutic alliance theory, we develop a theory-informed annotation framework and apply a hybrid LLM-human pipeline to analyze evaluative language, adoption-related attitudes, and relational alignment at scale. Our results show that engagement is shaped primarily by narrated outcomes, trust, and response quality, rather than emotional bond alone. Positive sentiment is most strongly associated with task and goal alignment, while companionship-oriented use more often involves misaligned alliances and reported risks such as dependence and symptom escalation. Overall, this work demonstrates how theory-grounded constructs can be operationalized in large-scale discourse analysis and highlights the importance of studying how users interpret language technologies in sensitive, real-world contexts.

[Long] LiveCultureBench: a Multi-Agent, Multi-Cultural Benchmark for Large Language Models in Dynamic Social Simulations ▶ *Underline Video*

Viet-Thanh Pham, Lizhen Qu, Thuy-Trang Vu, Gholamreza Haffari and Dinh Phung

Large language models (LLMs) are increasingly deployed as autonomous agents, yet evaluations focus primarily on task success rather than cultural appropriateness or evaluator reliability. We introduce LiveCultureBench, a multi-cultural, dynamic benchmark that embeds LLMs as agents in a simulated town and evaluates them on both task completion and adherence to socio-cultural norms. The simulation models a small city as a location graph with synthetic residents having diverse demographic and cultural profiles. Each episode assigns one resident a daily goal while others provide social context. An LLM-based verifier generates structured judgments on norm violations and task progress, which we aggregate into metrics capturing task-norm trade-offs and verifier uncertainty. Using LiveCultureBench across models and cultural profiles, we study (i) cross-cultural robustness of LLM agents, (ii) how they balance effectiveness against norm sensitivity, and (iii) when LLM-as-a-judge evaluation is reliable for automated benchmarking versus when human oversight is needed.

[Long] LLM Reasoning as Trajectories: Step-Specific Representation Geometry and Correctness Signals ▶ *Underline Video*

Lihao Sun, Hang Dong, Bo Qiao, Qingwei Lin, Dongmei Zhang and Saravan Rajmohan

This work characterizes large language models chain-of-thought generation as a structured trajectory through representation space. We show that mathematical reasoning traverses functionally ordered, step-specific subspaces that become increasingly separable with layer depth. This structure already exists in base models, while reasoning training primarily accelerates convergence toward termination-related subspaces rather than introducing new representational organization. While early reasoning steps follow similar trajectories, correct and incorrect solutions diverge systematically at late stages. This late-stage divergence enables mid-reasoning prediction of final-answer correctness with ROC-AUC up to 0.87. Furthermore, we introduce trajectory-based steering, an inference-time intervention framework that enables reasoning correction and length control based on derived ideal trajectories. Together, these results establish reasoning trajectories as a geometric lens for interpreting, predicting, and controlling LLM reasoning behavior.

[Long] LLM-Based Multi-Agent Systems for Clinical Workflows: A Survey of AI Hospitals ▶ *Underline Video*

Zonghai Yao and Hong Yu

This survey reviews LLM-based multi-agent systems for clinical and healthcare workflows, including diagnosis, triage, consultation, discharge, mental health, and EHR-linked decision support. We define AI hospitals as workflow-level clinical systems in which agents take explicit roles, hand off shared state, use EHR- or guideline-grounded tools, and operate with safety gates and audit-ready logs. We argue that these systems should be compared at the workflow level, rather than only by model components or end-task accuracy, because clinical action, evidence, and accountability are expressed through state transitions and handoffs. We organize the literature through a workflow-level taxonomy covering roles and handoffs, memory and evidence, tools, and reasoning, control, and escalation. We further synthesize major workflow settings and task families, introduce a four-layer evaluation stack spanning safety, process, outcome, and operations, and connect model capabilities to workflow observables relevant to deployment. Finally, we present Integration Readiness Levels (IRL1-IRL6), task-level instrumentation requirements, and recurring workflow failure modes as a practical framework for comparing, evaluating, and deploying clinical LLM agents and AI hospitals.

[Long] LLM4Cell: Taxonomy and Evaluation of LLM and Agentic Models for Single-Cell Biology ▶ *Underline Video*

Sajib Acharjee Dip, Adrika Zafor, Bikash Kumar Paul, Uddip Acharjee Shuvo, Muhit Islam Emon, Xuan Wang and Liqing Zhang

Large language models (LLMs) and emerging agentic frameworks are beginning to influence single-cell biology by enabling natural-language interfaces, generative annotation, and multimodal data integration. However, progress remains fragmented across data modalities, model families, and evaluation practices. LLM4Cell presents a unified survey of 58 foundation and agentic models developed for single-cell research, spanning RNA, ATAC, multi-omic, and spatial modalities. We organize these methods into five families foundation, text-bridge, spatial/multimodal, epigenomic, and agentic and map them to eight key analytical tasks, including annotation, trajectory inference, perturbation modeling, and drug-response prediction. Drawing on over 40 public datasets, we analyze benchmark coverage, data diversity, and ethical or scalability constraints, and synthesize reported capabilities across ten domain-level dimensions related to biological grounding, multimodal alignment, fairness, privacy, and interpretability. By explicitly linking datasets, modeling paradigms, and evaluation domains, LLM4Cell provides an integrated perspective on language-driven single-cell analysis and highlights open challenges in standardization, interpretability, and trustworthy model development.

[Long] Localized Low-Rank Adaptation within Clustered Parameter Subspaces ▶ *Underline Video*

Jiahao Xiong, Yihe Liu, Xianming Hu, Hongbo Zhao, Nuoyi Chen, Jie Zhang and Kai Zhang

Low-Rank Adaptation (LoRA) for large language models (LLMs) has achieved significant success in various domains. So far, most algorithms in the LoRA-family rely on global low-rank factors spanning the entire update weight matrix (W). Through careful analysis, however, we observe that the W during fine-tuning typically exhibit heterogeneous subspace clusters, each corresponding to specific sub-sets of rows and columns. This structural heterogeneity suggests that global low-rank factors may not optimally capture the local variations needed for effective model adaptation. To address this limitation, we propose LoRA within Clustered Parameter Subspaces, or CPS-LoRA, which performs independent low-rank updates within clustered blocks of parameter matrices. The key idea is to group the rows/columns of the update matrix into locally coherent, and maximally uncorrelated subspaces, perform low-rank adaptations in each subspace, and iteratively update the partition and local adaptations. This allows adapting to local structures more precisely while preserving high efficiency. Theoretical analysis reveals that in case W can be partitioned into subspace blocks with non-overlapping basis, CPS-LoRA have superior parameter efficiency than global adaptations. Empirical evaluations further demonstrate better rank utilization of CPS-LoRA and its consistent improvements against LoRA (and variants) by up to 3.0% in absolute accuracy in various benchmarks.

[Long] Looking Beyond the One: Operationalizing and Eliciting Visual Ambiguity in VLLMs ▶ *Underline Video*

Yuchong Chen, Bowei Zou, Yuhan Chen, Yifan Fan, Xinyu Li, Shujun Cao and Yu Hong

Visual questions are often ambiguous: the same image-question pair may admit multiple valid answers depending on which region is referenced. However, current Visual Question Answering (VQA) systems typically collapse this ambiguity, committing to a single interpretation during decoding and evaluation. In this work, we study visual question ambiguity from a grounded, region-centric perspective. We operationalize ambiguity as the existence of multiple distinct answer-supporting regions in an image, each independently yielding a valid answer. This formulation makes ambiguity observable without requiring exhaustive multi-answer annotations. Based on this definition, we conduct a systematic empirical study of state-of-the-art Visual Large Language Models (VLLMs). We find that, under default decoding, VLLMs consistently under-report ambiguity, even when multiple valid visual groundings are present. Importantly, probing model hidden states reveals that ambiguity-related signals are already encoded in their internal representations, despite not being reliably expressed in outputs. Finally, we show that selectively activating multi-focus answering based on these signals can recover additional valid answers while avoiding excessive hallucination. Together, our results suggest that ambiguity in VQA is not merely an annotation artifact or capability limitation, but a property that VLLMs internally recognize yet often fail to surface under standard decoding assumptions.

[Findings] Lost in the Prompt Order: Revealing the Limitations of Causal Attention in Language Models ▶ *Underline Video*

Hyunjong Ok and Jaeho Lee

Large language models exhibit surprising sensitivity to the structure of the prompt, but the mechanisms underlying this sensitivity remain poorly understood. In this work, we conduct an in-depth investigation on a striking case: in multiple-choice question answering, placing context before the questions and options (CQO) outperforms the reverse order (QOC) by over 14%p, consistently over a wide range of models and datasets. Through systematic architectural analysis, we identify causal attention as the core mechanism: in QOC prompts, the causal mask prevents option tokens from attending to context, creating an information bottleneck where context becomes invisible to options.

[Findings] MARS: Unleashing the Power of Speculative Decoding via Margin-Aware Verification ▶ *Underline Video*

Jingwei Song, Xinyu Wang, Hanbin Wang, Xiaoxuan Lei, TIANYU SHI, Shixin Han, Eric Yang, Xiao-Wen Chang and Lynn Ai

Speculative Decoding (SD) accelerates autoregressive large language model (LLM) inference by decoupling generation and verification. While recent methods improve draft quality by tightly coupling the drafter with the target model, the verification mechanism itself remains largely unchanged, relying on strict token-level rejection sampling. In practice, modern LLMs frequently operate in low-margin regimes where the target model exhibits weak preference among top candidates. In such cases, rejecting plausible runner-up tokens yields negligible information gain while incurring substantial rollback cost, leading to a fundamental inefficiency in verification. We propose Margin-Aware Speculative Verification, a training-free and domain-agnostic verification strategy that adapts to the target models local decisiveness. Our method conditions verification on decision stability measured directly from the target logits and relaxes rejection only when strict verification provides minimal benefit. Importantly, the approach modifies only the verification rule and is fully compatible with existing target-coupled speculative decoding frameworks. Extensive experiments across model scales ranging from 8B to 235B demonstrate that our method delivers consistent and significant inference speedups over state-of-the-art baselines while preserving generation quality across diverse benchmarks. The code is available at https://github.com/5SSjw/MARS.

[Findings] MedErrBench: A Fine-Grained Multilingual Benchmark for Medical Error Detection and Correction with Clinical Expert Annotations ▶ *Underline Video*

Congbo Ma, Yichun Zhang, Yousef Al-Jazzazi, Ahamed Foisal, Laasya Sharma, Yousra Sadqi, Khaled Saleh, Jihad Mallat and Farah E. Shamout

Inaccuracies in existing or generated clinical text may lead to serious adverse consequences, especially if it is a misdiagnosis or incorrect treatment suggestion. With Large Language Models (LLMs) increasingly being used across diverse healthcare applications, comprehensive evaluation through dedicated benchmarks is crucial. However, such datasets remain scarce, especially across diverse languages and contexts. In this paper, we introduce MedErrBench, the first multilingual benchmark for error detection, localization, and correction, developed

under the guidance of experienced clinicians. Based on an expanded taxonomy of ten common error types, MedErrBench covers English, Arabic and Chinese, with natural medical cases annotated and reviewed by domain experts. We assessed the performance of a range of general-purpose, language-specific, and medical-domain language models across all three tasks. Our results reveal notable performance gaps, particularly in non-English settings, highlighting the need for clinically grounded, language-aware systems. By making MedErrBench and our evaluation protocols publicly-available, we aim to advance multilingual clinical NLP to promote safer and more equitable AI-based healthcare globally. The dataset is publicly available at: https://github.com/congboma/MedErrBench.

[Long] Memorization, Emergence, and Explaining Reversal Failures: A Controlled Study of Relational Semantics in LLMs ► *Underline Video*

Yihua Zhu, Qianying Liu, Jiaxin Wang, Fei Cheng, Chaoran Liu, Akiko Aizawa, Sadao Kurohashi and Hidetoshi Shimodaira

Autoregressive LLMs perform well on relational tasks that require linking entities via relational words (e.g., father/son, friend), but it is unclear whether they learn the logical semantics of such relations (e.g., symmetry and inversion logic) and, if so, whether reversal-type failures arise from missing relational semantics or left-to-right order bias. We propose a controlled Knowledge Graph-based synthetic framework that generates text from symmetric/inverse triples, train GPT-style autoregressive models from scratch, and evaluate memorization, logical inference, and in-context generalization to unseen entities to address these questions. We find a sharp phase transition in which relational semantics emerge with sufficient logic-bearing supervision, even in shallow (2-3 layer) models, and that successful generalization aligns with stable intermediate-layer signals. Finally, order-matched forward/reverse tests and a diffusion baseline indicate that reversal failures are primarily driven by autoregressive order bias rather than deficient inversion semantics.

[Findings] Meta-Reasoner: Dynamic Guidance for Optimized Inference-time Reasoning in Large Language Models ► *Underline Video*

Yuan Sui, Yufei He, Tri Cao, Sophia Simeng Han, Yulin Chen and Bryan Hooi

Large Language Models (LLMs) often struggle with computational efficiency and error propagation in multi-step reasoning tasks. While recent advancements on prompting and post-training have enabled LLMs to perform step-wise reasoning, they still tend to explore unproductive solution paths without effective backtracking or strategy adjustment. In this paper, we propose Meta-Reasoner, a new framework that empowers LLMs to "think about how to think". It optimizes the inference process by dynamically adapting reasoning strategies in real-time. Our approach employs contextual multi-armed bandits (CMABs) to learn an adaptive policy. It learns to evaluate the current state of LLM's reasoning and determine optimal strategy that is most likely to lead to a successful outcome during inference, like whether to backtrack, switch to a new approach, or restart the problem-solving process. This meta-guidance helps avoid unproductive paths exploration during inference and hence improves computational efficiency. We evaluate Meta-Reasoner on math problems (e.g., Game-of-24, TheoremQA) and scientific tasks (e.g., SciBench). Results show that our method outperform previous SOTA methods by 9-12% in accuracy, while reducing inference time by 28-35% under the same compute budget. Additional experiments on creative writing demonstrate the generalizability of our approach to diverse reasoning-intensive tasks.

[Long] Mixture-of-Experts with Intermediate CTC Supervision for Accented Speech Recognition ► *Underline Video*

Wonjun Lee, Hyounghun Kim and Gary Lee

Accented speech remains a persistent challenge for automatic speech recognition (ASR), as most models are trained on data dominated by a few high-resource English varieties, leading to substantial performance degradation for other accents. Accent-agnostic approaches improve robustness yet struggle with heavily accented or unseen varieties, while accent-specific methods rely on limited and often noisy labels. We introduce MoE-CTC, a Mixture-of-Experts architecture with intermediate CTC supervision that jointly promotes expert specialization and generalization. During training, accent-aware routing encourages experts to capture accent-specific patterns, which gradually transitions to label-free routing for inference. Each expert is equipped with its own CTC head to align routing with transcription quality, and a routing-augmented loss further stabilizes optimization. Experiments on the MCV-Accent benchmark demonstrate consistent gains across both seen and unseen accents in low- and high-resource conditions, achieving up to 29.3% relative WER reduction over strong FastConformer baselines.

[Findings] Model-Agnostic Meta Learning for Class Imbalance Adaptation ► *Underline Video*

Hanshu Rao, Guangzeng Han and Xiaolei Huang

Class imbalance is a widespread challenge in NLP tasks, significantly hindering robust performance across diverse domains and applications. We introduce Hardness-Aware Meta-Resample (HAMR), a unified framework that adaptively addresses both class imbalance and data difficulty. HAMR employs bi-level optimizations to dynamically estimate instance-level weights that prioritize genuinely challenging samples and minority classes, while a neighborhood-aware resampling mechanism amplifies training focus on hard examples and their semantically similar neighbors. We validate HAMR on six imbalanced datasets covering multiple tasks and spanning biomedical, disaster response, and sentiment domains. Experimental results show that HAMR achieves substantial improvements for minority classes and consistently outperforms strong baselines. Extensive ablation studies demonstrate that our proposed modules synergistically contribute to performance gains and highlight HAMR as a flexible and generalizable approach for class imbalance adaptation.

[Findings] Modeling Human Perspectives with Socio-Demographic Representations ► *Underline Video*

Leixin Zhang and Cagri Coltekin

Humans often hold different perspectives on the same issues. In many NLP tasks, annotation disagreement can reflect valid subjective

perspectives. Modeling annotator perspectives and understanding their relationship with other human factors, such as socio-demographic attributes, have received increasing attention. Prior work typically focuses on single demographic factors or limited combinations. However, in real-world settings, annotator perspectives are shaped by complex social contexts, and finer-grained socio-demographic attributes can better explain human perspectives. In this work, we propose Socio-Contrastive Learning, a method that jointly models annotator perspectives while learning socio-demographic representations. Our method provides an effective approach for the fusion of socio-demographic features and textual representations to predict annotator perspectives, outperforming standard concatenation-based methods. The learned representations further enable analysis and visualization of how demographic factors relate to variation in annotator perspectives. Our code is available at GitHub: https://github.com/Leixin-Zhang/Socio_Contrastive_Learning

[Long] MTA:A Merge-then-Adapt Framework for Personalized Large Language Models ▶ *Underline Video*

Xiaopeng Li, Yuanjin Zheng, Wanyu Wang, Wenlin Zhang, Pengyue Jia, Yingyi Zhang, Haiying He, Mengyang Ma, Yiqi Wang, Maolin Wang, Xuetao Wei and Xiangyu Zhao

Personalized Large Language Models (PLLMs) aim to align model outputs with individual user preferences, a crucial capability for user-centric applications. However, the prevalent approach of fine-tuning a separate module for each user faces two major limitations: (1) storage costs scale linearly with the number of users, rendering the method unscalable; and (2) fine-tuning a static model from scratch often yields suboptimal performance for users with sparse data. To address these challenges, we propose MTA, a Merge-then-Adapt framework for PLLMs. MTA comprises three key stages. First, we construct a shared Meta-LoRA Bank by selecting anchor users and pre-training meta-personalization traits within meta-LoRA modules. Second, to ensure scalability and enable dynamic personalization combination beyond static models, we introduce an Adaptive LoRA Fusion stage. This stage retrieves and dynamically merges the most relevant anchor meta-LoRAs to synthesize a user-specific one, thereby eliminating the need for user-specific storage and supporting more flexible personalization. Third, we propose a LoRA Stacking for Few-Shot Personalization stage, which applies an additional ultra-low-rank, lightweight LoRA module on top of the merged LoRA. Fine-tuning this module enables effective personalization under few-shot settings. Extensive experiments on the LaMP benchmark demonstrate that our approach outperforms existing SOTA methods across multiple tasks. Our code is also available.

[Findings] Multi-Drafter Speculative Decoding with Alignment Feedback ▶ *Underline Video*

Taehyeon Kim, Hojung Jung and Se-Young Yun

Speculative decoding (SD) accelerates large language model (LLM) inference by using a smaller model to draft future tokens, which are then verified by the target LLM. This preserves generation quality by accepting only aligned tokens. However, individual drafters, often trained for specific tasks or domains, exhibit limited effectiveness across diverse applications. To address this, we introduce MetaSD, a unified framework that integrates multiple drafters into the SD process. MetaSD dynamically allocates computational resources to heterogeneous drafters by leveraging alignment feedback and framing drafter selection as a multi-armed bandit problem. Extensive experiments show MetaSD consistently outperforms single-drafter approaches.

[Findings] Multi-Persona Thinking for Bias Mitigation in Large Language Models ▶ *Underline Video*

Yuxing Chen, Guoqing Luo, Zijun Wu and Lili Mou

Large Language Models (LLMs) exhibit social biases, which can lead to harmful stereotypes and unfair outcomes. We propose Multi-Persona Thinking (MPT), a simple inference-time framework that reduces social bias by encouraging reasoning from multiple perspectives. MPT guides the model to consider contrasting social identities, such as male and female, together with a neutral viewpoint. These viewpoints then interact through an iterative reasoning process to identify and correct biased judgments. This design transforms the potential weakness of persona assignment into a mechanism to mitigate bias. We evaluate MPT on two widely used bias benchmarks with both open-source and closed-source models. Our results show that MPT achieves a lower bias than the existing prompting-based methods while maintaining the core reasoning ability.

[Long] MUSEG: Reinforcing Video Temporal Understanding via Timestamp-Aware Multi-Segment Grounding ▶ *Underline Video*

Fuwen Luo, Shengfeng Lou, Chi Chen, Ziyue Wang, Chenliang Li, Weizhou Shen, Jiyue Guo, Peng Li, Ming Yan, Ji Zhang, Fei Huang and Yang Liu

Video temporal understanding is crucial for multimodal large language models (MLLMs) to reason over events in videos. Despite recent advances in general video understanding, current MLLMs still struggle with fine-grained temporal reasoning. While reinforcement learning (RL) has been explored to address this issue recently, existing RL approaches remain limited in performance on time-sensitive tasks. In this work, we propose MUSEG, a novel RL-based method that enhances temporal understanding by introducing timestamp-aware multi-segment grounding. MUSEG enables MLLMs to align queries with multiple relevant video segments, promoting more comprehensive temporal reasoning. To facilitate effective learning, we design a customized RL training recipe with phased rewards that progressively guides the model toward temporally grounded reasoning. Extensive experiments on temporal grounding and time-sensitive video question answering (QA) tasks demonstrate that MUSEG significantly outperforms existing methods and generalizes well across diverse temporal understanding scenarios.

[Findings] Non-invasive electromyographic speech neuroprosthesis: a geometric perspective ▶ *Underline Video*

Harshavardhana T. Gowda and Lee M. Miller

We present a neuromuscular speech interface that translates silently voiced articulations directly into text. We record surface electromyo-

graphic (EMG) signals from multiple articulatory sites on the face and neck as participants silently articulate speech, enabling direct EMG-to-text translation. Such an interface has the potential to restore communication for individuals who have lost the ability to produce intelligible speech due to laryngectomy, neuromuscular disease, stroke, or trauma-induced damage (e.g., radiotherapy toxicity) to the speech articulators. Prior work has largely focused on mapping EMG collected during audible articulation to time-aligned audio targets or transferring these targets to silent EMG recordings, which inherently requires audio and limits applicability to patients who can no longer speak. In contrast, we propose an efficient representation of high-dimensional EMG signals and demonstrate direct sequence-to-sequence EMG-to-text conversion at the phonemic level without relying on time-aligned audio.

[Long] OASIS: Online Sample Selection for Continual Instruction Tuning ▶ *Underline Video*

Minjae Lee, Minhyuk Seo, Tingyu Qu, Tinne Tuytelaars and Jonghyun Choi

In continual instruction tuning (CIT) scenarios, where new instruction tuning data continuously arrive in an online streaming manner, training delays from large-scale data significantly hinder real-time adaptation. Data selection can mitigate this overhead, but existing strategies often rely on pre-trained reference models, which are impractical in CIT setups since future data are unknown. Recent reference model-free online sample selection methods address this, but typically select a fixed number of samples per batch (e.g., top-k), making them vulnerable to distribution shifts where informativeness varies across batches. To address these limitations, we propose OASIS, an adaptive online sample selection approach for CIT that (1) selects informative samples by estimating each samples informativeness relative to all previously seen data, beyond batch-level constraints, and (2) minimizes informative redundancy of selected samples through iterative selection score updates. Experiments on various large foundation models show that, using only 25% of the data, achieves comparable performance to full-data training and outperforms the state-of-the-art sampling methods.

[Long] Observations and Remedies for Large Language Model Bias in Self-Consuming Performative Loop ▶ *Underline Video*

Yaxuan Wang, Zhongteng Cai, Zhongteng_Cai1 Zhongteng_Cai1, Yujia Bao, Xueru Zhang and Yang Liu

The rapid advancement of large language models (LLMs) has led to growing interest in using synthetic data to train future models. However, this creates a self-consuming retraining loop, where models are trained on their own outputs and may cause performance drops and induce emerging biases. In real-world applications, previously deployed LLMs may influence the data they generate, leading to a dynamic system driven by user feedback. For example, if a model continues to underserve users from a group, less query data will be collected from this particular demographic of users. In this study, we introduce the concept of Self-Consuming Performative Loop (SCPL) and investigate the role of synthetic data in shaping bias during these dynamic iterative training processes under controlled performative feedback. This controlled setting is motivated by the inaccessibility of real-world user preference data from dynamic production systems, and enables us to isolate and analyze feedback-driven bias evolution in a principled manner. We focus on two types of loops, including the typical retraining setting and the incremental fine-tuning setting, which is largely underexplored. Through experiments on three real-world tasks, we find that the performative loop increases preference bias and decreases disparate bias. We design a reward-based rejection sampling strategy to mitigate the bias, moving towards more trustworthy self-improving systems. The code is available at https://github.com/UCSC-REAL/SCPL.git.

[Findings] OjaKV: Context-Aware Online Low-Rank KV Cache Compression ▶ *Underline Video*

Yuxuan Zhu, David H. Yang, Mohammad Mohammadi Amiri, Keerthiram Murugesan, Tejaswini Pedapati and Pin-Yu Chen

The expanding long-context capabilities of large language models are constrained by a significant memory bottleneck: the key-value (KV) cache required for autoregressive generation. This bottleneck is substantial; for instance, a Llama-3.1-8B model processing a 32K-token prompt at a batch size of 4 requires approximately 16 GB for its KV cache, exceeding the model's weights. While KV-cache compression via low-rank projection is promising, existing methods rely on a static, offline-learned subspace that performs poorly under distribution shifts. To overcome these limitations, we introduce OjaKV, a novel framework integrating a hybrid storage policy with online subspace adaptation. OjaKV preserves crucial tokens in full rank as high-fidelity anchors, while applying low-rank compression to intermediate tokens by adapting the projection basis using Oja's algorithm for online PCA. This adaptation involves a comprehensive update during prefilling and lightweight periodic updates during decoding, ensuring the subspace remains aligned with evolving context. Our framework is fully compatible with FlashAttention. Experiments demonstrate that OjaKV maintains or improves zero-shot accuracy at high compression ratios, achieving the strongest gains on long-context benchmarks requiring complex reasoning. Furthermore, our approach combines with token-selection methods for compounded memory savings, establishing a practical, plug-and-play solution for memory-efficient long-context inference without fine-tuning.

[Long] One Single Hub Text Breaks CLIP: Identifying Vulnerabilities in Cross-Modal Encoders via Hubness ▶ *Underline Video*

Hiroyuki Deguchi, Katsuki Chousa and Yusuke Sakai

The hubness problem, in which hub embeddings are close to many unrelated examples, occurs often in high-dimensional embedding spaces and may pose a practical threat for purposes such as information retrieval and automatic evaluation metrics. In particular, since cross-modal similarity between text and images cannot be calculated by direct comparisons, such as string matching, cross-modal encoders that project different modalities into a shared space are helpful for various cross-modal applications, and thus, the existence of hubs may pose practical threats. To reveal the vulnerabilities of cross-modal encoders, we propose a method for identifying the hub embedding and its corresponding hub text. Experiments on image captioning evaluation in MSCOCO and nocaps along with image-to-text retrieval tasks in MSCOCO and Flickr30k showed that our method can identify a single hub text that unreasonably achieves comparable or higher similarity scores than human-written reference captions in many images, thereby revealing the vulnerabilities in cross-modal encoders.

[Findings] Op-Fed: Opinion, Stance, and Monetary Policy Annotations on FOMC Transcripts Using Active Learning ► *Underline Video*

Alisa Kanganis and Katherine A. Keith

The U.S.~Federal Open Market Committee (FOMC) regularly discusses and sets monetary policy, affecting the borrowing and spending decisions of hundreds of millions of people. In this work, we release Op-Fed, a dataset of 1044 human-annotated sentences and their contexts from FOMC transcripts that captures monetary policy stance—specifically, whether an individual FOMC member expresses support for tightening or loosening policy. We faced two major technical challenges in dataset creation: imbalanced classes—we estimate fewer than 8% of sentences express a non-neutral stance toward monetary policy—and inter-sentence dependence—65% of instances require context beyond the sentence-level to annotate. To address these challenges, we developed a five-stage hierarchical schema to isolate aspects of opinion, monetary policy, and stance toward monetary policy, as well as the level of context needed. Second, we selected instances to annotate using active learning, approximately doubling the number of positive instances across all schema aspects. Using Op-Fed, we found a top-performing, closed-weight LLM achieves 0.80 zero-shot accuracy in opinion classification but only 0.61 zero-shot accuracy classifying stance toward monetary policy—below our human baseline of 0.89. We expect Op-Fed to be useful for future model training, confidence calibration, and as a seed dataset for future annotation efforts.

[Long] Outcome-Grounded Advantage Reshaping for Fine-Grained Credit Assignment in Mathematical Reasoning ► *Underline Video*

Ziheng Li, Liu Kang, Feng Xiao, Luxi Xing, Qingyi Si, Zhuoran Li, Weikang Gong, Deqing Yang, Yanghua Xiao and Hongcheng Guo

Group Relative Policy Optimization (GRPO) has emerged as a promising critic-free reinforcement learning paradigm for reasoning tasks. However, standard GRPO employs a coarse-grained credit assignment mechanism that propagates group-level rewards uniformly to to every token in a sequence, neglecting the varying contribution of individual reasoning steps. We address this limitation by introducing Outcome-grounded Advantage Reshaping (OAR), a fine-grained credit assignment mechanism that redistributes advantages based on how much each token influences the model's final answer. We instantiate OAR via two complementary strategies: (1) OAR-P, which estimates outcome sensitivity through counterfactual token perturbations, serving as a high-fidelity attribution signal; (2) OAR-G, which uses an input-gradient sensitivity proxy to approximate the influence signal with a single backward pass. These importance signals are integrated with a conservative Bi-Level advantage reshaping scheme that suppresses low-impact tokens and boosts pivotal ones while preserving the overall advantage mass. Empirical results on extensive mathematical reasoning benchmarks demonstrate that while OAR-P sets the performance upper bound, OAR-G achieves comparable gains with negligible computational overhead, both significantly outperforming a strong GRPO baseline, pushing the boundaries of critic-free LLM reasoning.

[Findings] Parallel Context-of-Experts Decoding for Retrieval Augmented Generation ► *Underline Video*

Giulio Corallo and Paolo Papotti

Retrieval Augmented Generation faces a trade-off: concatenating documents in a long prompt enables multi-document reasoning but creates prefill bottlenecks, while encoding document KV caches separately offers speed but breaks cross-document interaction. We propose Parallel Context-of-Experts Decoding (PCED), a training-free framework that shifts evidence aggregation from the attention mechanism to the decoding. PCED treats retrieved documents as isolated "experts", synchronizing their predictions via a retrieval-aware extension of context-aware decoding. This approach recovers cross-document reasoning capabilities without constructing a shared attention across documents.

[Long] Patches of Nonlinearity: Instruction Vectors in Large Language Models ► *Underline Video*

Irina Bigoulaeva, Jonas Rohweder, Subhabrata Dutta and Iryna Gurevych

Despite the recent success of instruction-tuned language models and their ubiquitous usage, very little is known of how models process instructions internally. In this work, we address this gap from a mechanistic point of view by investigating how instruction-specific representations are constructed and utilized in different stages of post-training: Supervised Fine-Tuning (SFT) and Direct Preference Optimization (DPO). Via causal mediation, we identify that instruction representation is fairly localized in models. These representations, which we call Instruction Vectors (IVs), demonstrate a curious juxtaposition of linear separability along with non-linear causal interaction, broadly questioning the scope of the linear representation hypothesis commonplace in mechanistic interpretability. To disentangle the non-linear causal interaction, we propose a novel method to localize information processing in language models that is free from the implicit linear assumptions of patching-based techniques. We find that, conditioned on the task representations formed in the early layers, different information pathways are selected in the later layers to solve that task, i.e., IVs act as circuit selectors.

[Findings] PedagogyBench: A Cognitive-Driven Benchmark for Multimodal Instructional Video Understanding ► *Underline Video*

Xiaokang Jin, Jia Zhu, Jingjiang Liu, Yabing Shi, Jueqi Guan, Hao Chen and Pasquale De Meo

Existing video understanding benchmarks mainly emphasize general visual recognition and reasoning, but do not adequately capture the pedagogical logic embedded in instructional videos. To address this gap, we present PedagogyBench, a multimodal benchmark for instructional video understanding grounded in pedagogical cognition. We introduce a pedagogy-driven segmentation strategy and a dual-stream semantic injection pipeline that combines machine pre-annotation with expert refinement, enabling the construction of a dataset organized around a cognitive pyramid with four levels and 20 fine-grained tasks. We further propose the Cognitive Fidelity Score

(CFS) to measure the balance of model performance across pedagogical cognitive dimensions. Experiments on 12 multimodal large language models reveal a clear generative gap, where models perform relatively well on discriminative tasks but degrade on higher-order pedagogical diagnosis, often relying on parametric memory rather than grounded visual perception. Project resources are available at https://github.com/Shallcom/PedagogyBench.

[Findings] Perceptual Hallucination in Vision-Language Models: Definition, Analysis and Verification ▶ *Underline Video*

Taewook Hwang, Inbum Heo, Sung Jun Lee and Sangkeun Jung

Vision-Language Models (VLMs) have demonstrated remarkable performance in document understanding tasks; however, VLMs also suffer from hallucinations inherited from LLMs. While prior work has focused on reasoning-stage hallucinations, the role of visual perception remains underexplored. In this work, we define perceptual hallucination as the phenomenon where VLMs generate information as if perceived, despite absent or damaged visual evidence. To analyze this, we construct DocHallu, a benchmark of 2,671 original–damaged image pairs across three tasks, available at https://huggingface.co/datasets/IB99/DocHallu. Experiments reveal that perceptual hallucination occurs across all models, with higher rates for numerical content than textual content. Activation patching analysis suggests that hallucinations are strongly associated with errors introduced in the vision encoder, which can subsequently propagate and become amplified through the text decoding process. We also demonstrate that LLM-based post-hoc filtering can reduce hallucination exposure by 36% on average, with reductions of up to 88%. This work extends VLM hallucination research by defining, analyzing, and verifying perceptual hallucination in document understanding.

[Findings] PhaseMI: A Motivational Interviewing Dataset for Enhancing Phase Progression in LLM-based Counseling ▶ *Underline Video*

Jina Kim, Myeongho Jeon, Soohyun Cho, Chae-Gyun Lim, Jongmin Lim, Haewon Min and Eunho Yang

The growing demand for scalable mental health support has increased interest in AI-based counseling systems grounded in Motivational Interviewing (MI). However, existing MI datasets do not explicitly model the structured progression of MI phases, which is essential for effective and goal-oriented counseling. To address this gap, we introduce PhaseMI, a phase-structured MI dataset, together with a data generation framework that employs therapist, client, and supervisor LLMs to explicitly control phase transitions. Compared to the best alternative baseline, PhaseMI achieves improved coverage of MI phases, with gains of 12.3% in exploring, 37.6% in guiding, and 61.1% in choosing, and experimental evaluations demonstrate that it yields higher overall counseling quality than baseline datasets.

[Findings] Preserving Language Capabilities in Vision-Language Models via Representation Regulation ▶ *Underline Video*

Zixuan Chen, Juncheng Tao and Ziqian Zheng;

Vision-Language Models (VLMs) provide a unified framework to process both text-only tasks and vision-language tasks. However, finetuning VLMs on vision-language data has degraded language capabilities. In this paper, we prove that as the training loss declines during finetuning, the visual representation and textual representation move closer to each other, a phenomenon we term "representation mixing." We prove that the representation mixing occurring within the post-representation layers causes the degradation of language capabilities. Post-representation layers refer to the first few layers in LLMs that are involved in representation learning. To preserve the language capabilities, we propose the Representation Regulation for VLM Training (RRVLM), which introduces a Representation Distribution Difference (RDD) loss to reduce the distance between these representations. Extensive experiments on various benchmarks and VLM frameworks show that our method can effectively preserve the language capabilities and achieve superior vision-language performance.

[Findings] PRIME: Ultra-Low-Rank Principal-Residual Model Merging ▶ *Underline Video*

Seungho Lee, Kyungsu Lee, Zuchi Bazarvaani, Jeongmin Ahn, Insuk Seo, Donghyeon Jeon, Inho Kang and Seung-Hoon Na

Model merging has emerged as an effective approach for integrating multiple task-specific fine-tuned models into a single unified model without requiring additional data-intensive training. A central challenge in model merging is to reduce task interference while preserving the task-specific capabilities of the original models. In this work, we propose PRIME, an ultra-low-rank principal-residual model merging framework that decomposes task vector merging into two complementary stages. First, ultra-low-rank principal task vector merging retains only a small fraction of singular vectors, effectively reducing task interference while preserving most of the task-specific performance. Second, orthogonal residual task vector merging incorporates the remaining components by projecting them onto the null space of the principal subspace, thereby avoiding interference while recovering additional task-relevant information. Extensive experiments on eight natural language processing tasks demonstrate that PRIME consistently outperforms existing model merging methods, achieving improvements of up to 1.18% on T5 and 1.9% on LLaMA-3.2-3B.

[Long] PRiSM: Benchmarking Phone Realization in Speech Models ▶ *Underline Video*

Shikhar Bharadwaj, Chin-Jou Li, Yoonjae Kim, Kwanghee Choi, Eunjung Yeo, Ryan Soh-Eun Shim, Hanyu Zhou, Brendon Boldt, Karen Rosero, Kalvin Chang, Darsh Agrawal, Keer Xu, Chao-Han Huck Yang, Jian Zhu, Shinji Watanabe and David R. Mortensen

Phone recognition (PR) serves as the atomic interface for language-agnostic modeling for cross-lingual speech processing and phonetic analysis. Despite prolonged efforts in developing PR systems, current evaluations only measure surface-level transcription accuracy. We introduce PRiSM, the first open-source benchmark designed to expose blind spots in phonetic perception through intrinsic and extrinsic evaluation of PR systems. PRiSM standardizes transcription-based evaluation and assesses downstream utility in clinical, educational, and multilingual settings with transcription and representation probes. We find that diverse language exposure during training is key to PR

performance, encoder-CTC models are the most stable, and specialized PR systems still outperform LALMs. PRiSM releases code, recipes, and datasets to move the field toward multilingual speech models with robust phonetic ability.

[Findings] Probabilistic Depression Detection from Textual Time Series ▶ *Underline Video*
Fabian Schmidt, Seyedehmoniba Ravan and Vladimir Vlassov;
Accurate and interpretable predictions of depression severity are essential for clinical decision support, yet existing models often lack uncertainty estimates and temporal interpretability. We propose PTTSD, a Probabilistic framework for Depression Detection from clinical interview utterance sequences that predicts PHQ-8 scores while modeling calibrated uncertainty. PTTSD includes sequence-to-sequence and sequence-to-one variants, both combining LSTMs, self-attention, and residual connections with Gaussian or Student's-t output heads trained via negative log-likelihood. The sequence-to-sequence variant enables temporal analysis of how predictive confidence evolves over an interview, despite the target being a single session-level score. Evaluated on E-DAIC and DAIC-WOZ, PTTSD achieves competitive performance among text-only systems (e.g., MAE = 3.85 on E-DAIC, 3.55 on DAIC) and produces well-calibrated prediction intervals. Ablations confirm the value of attention and probabilistic modeling, while a three-part calibration analysis and qualitative case studies highlight the clinical relevance of uncertainty-aware prediction.

[Findings] Prompting Across Time: Evaluating LLMs on Historical and Contemporary Offensive Language ▶ *Underline Video*
Sanne Hoeken, Sophie Spliethoff, Silke Schwandt, Sina ZarrieSS and Özge Alaçam
Research on hate speech detection (HSD) has centered on modern data, even though offensive language has a much longer history. This paper presents the first systematic evaluation of instruction-tuned LLMs on Early Modern English invectives, compared with a modern hate-speech benchmark. Our work applies a modular prompt design to measure the contribution of definitional richness, contextual grounding, decision rules and few-shot examples. The results indicate that clearer annotation boundaries in the curated historical corpus lead to higher classification performance compared to the modern benchmark, despite the disadvantage of linguistic unfamiliarity. Prompt brittleness, however, persists across both domains. Classification-oriented components (rules, examples) drive the strongest effects, while definitional or contextual additions matter less. Fine-tuned encoder models still outperform LLMs, but some prompt configurations can narrow the gap. Overall, our study provides practical guidance for prompt design in both digital humanities and HSD and new opportunities for tracing the historical development of hate speech.

[Long] Prosody as Supervision: Bridging the Non-Verbal-Verbal for Multilingual Speech Emotion Recognition ▶ *Underline Video*
Girish Girish, Mohd Mujtaba Akhtar and Muskaan Singh
In this work, we introduce a paralinguistic supervision paradigm for low-resource multilingual speech emotion recognition (LRM-SER) that leverages non-verbal vocalizations to exploit prosody-centric emotion cues. Unlike conventional SER systems that rely heavily on labeled verbal speech and suffer from poor cross-lingual transfer, our approach reformulates LRM-SER as non-verbal-to-verbal transfer, where supervision from a labelled non-verbal source domain is adapted to unlabeled verbal speech across multiple target languages. To this end, we propose NOVA-ARC, a geometry-aware framework that models affective structure in the Poincaré ball, discretizes paralinguistic patterns via a hyperbolic vector-quantized prosody codebook, and captures emotion intensity through a hyperbolic emotion lens. For unsupervised adaptation, NOVA-ARC performs optimal-transport-based prototype alignment between source emotion prototypes and target utterances, inducing soft supervision for unlabeled speech while being stabilized through consistency regularization. Experiments show that NOVA-ARC delivers the strongest performance under both non-verbal-to-verbal adaptation and the complementary verbal-to-verbal transfer setting, consistently outperforming Euclidean counter parts and strong SSL baselines. To the best of our knowledge, this work is the first to move beyond verbal-speech-centric supervision by introducing a non-verbal-to-verbal transfer paradigm for SER.

[Long] Protecting multimodal large language models against misleading visualizations ▶ *Underline Video*
Jonathan Tonglet, Tinne Tuytelaars, Marie-Francine Moens and Iryna Gurevych
Visualizations play a pivotal role in daily communication in an increasingly data-driven world. Research on multimodal large language models (MLLMs) for automated chart understanding has accelerated massively, with steady improvements on standard benchmarks. However, for MLLMs to be reliable, they must be robust to misleading visualizations, i.e., charts that distort the underlying data, leading readers to draw inaccurate conclusions. Here, we uncover an important vulnerability: MLLM question-answering (QA) accuracy on misleading visualizations drops on average to the level of the random baseline. To address this, we provide the first comparison of six inference-time methods to improve QA performance on misleading visualizations, without compromising accuracy on non-misleading ones. We find that two methods, table-based QA and redrawing the visualization, are effective, with improvements of up to 19.6 percentage points. We make our code and data available.

[Findings] Pseudo2Real: Task Arithmetic for Pseudo-Label Correction in Automatic Speech Recognition ▶ *Underline Video*
Yi-Cheng Lin, Yu-Hsuan Li Liang, Hsuan Su, Tzu-Quan Lin, Shang-Tse Chen, Yun-Nung Chen and Hung-yi Lee
Robust ASR under domain shift is crucial because real-world systems encounter unseen accents and domains with limited labeled data. Although pseudo-labeling offers a practical workaround, it often introduces systematic, accent-specific errors that filtering fails to fix. We ask: How can we correct these recurring biases without target ground truth? We propose a simple parameter-space correction: in a source domain containing both real and pseudo-labeled data, two ASR models are fine-tuned from the same initialization, one on ground-truth labels and the other on pseudo-labels, and their weight difference forms a correction vector that captures pseudo-label biases. When applied

to a pseudo-labeled target model, this vector enhances recognition, achieving up to a 35% relative Word Error Rate (WER) reduction on AfriSpeech-200 across ten African accents with the Whisper tiny model.

[Findings] PUMA: Projected Universal Multilingual ASR for Low-Resource Settings. Application to Diverse African Languages ▶ *Underline Video*

Ilyes Oukid, Bilal Faye, Hanane AZZAG, Mustapha Lebbah and Said Yacine Boulahia

Multilingual ASR systems often fail to generalize to low-resource and linguistically diverse languages while remaining costly to scale. We introduce PUMA, a unified multilingual ASR model that improves low-resource performance with reduced model complexity. PUMA employs a Universal Language Projection (ULP) module that integrates a learnable language token with acoustic representations, enabling language-aware processing through shared parameters. Experiments on diverse African languages show consistent word error rate reductions over strong multilingual baselines, highlighting improved robustness and generalization. Our code is available at the following GitHub URL: https://github.com/ilyes-okd/PUMA

[Findings] RACC: Regret-Aware Confidence Calibration for Consistent Masked Discrete Diffusion Decoding ▶ *Underline Video*

Qinglin Zeng, Jusheng Zhang, Jing Yang, Ningyuan Liu and Keze Wang

Masked Discrete Diffusion Models (MDMs) enable parallel generation via iterative refinement. However, we identify a critical decisional mismatch. The MDM architecture is inherently dynamic and capable of sensing context shifts. In contrast, prevailing decoding paradigms remain static and myopic. They treat each denoising step as an isolated snapshot, effectively discarding valuable temporal feedback that signals logical conflicts. To bridge this gap, we propose Regret-Aware Confidence Calibration (RACC). This training-free framework aligns decoding decisions with the models latent self-correction capabilities. RACC introduces a momentum anchor to track confidence trajectories. When a token's probability drops abruptly below its historical trend, the system triggers a "regret" signal. Unlike expensive re-masking or lookahead search, RACC utilizes this signal to proactively demote unstable candidates. Extensive experiments on reasoning benchmarks, such as HumanEval and GSM8K, demonstrate that RACC significantly improves generation consistency. Crucially, RACC achieves these gains with zero additional inference overhead, effectively balancing decoding quality and efficiency.

[Long] Reasoning Traces Shape Outputs but Models Wont Say So ▶ *Underline Video*

Yijie Hao, Lingjie Chen, Ali Emami and Joyce C. Ho

Can we trust the reasoning traces that large reasoning models (LRMs) produce? We investigate whether these traces faithfully reflect what drives model outputs, and whether models will honestly report their influence. We introduce Thought Injection, a method that injects synthetic reasoning snippets into a model's reasoning trace, then measures whether the model follows the injected reasoning and acknowledges doing so. Across 45,000 samples from three LRMs, we find that injected hints reliably alter outputs, confirming that reasoning traces causally shape model behavior. However, when asked to explain their changed answers, models overwhelmingly refuse to disclose the influence: non-disclosure exceeds 90% for extreme hints across 30,000 follow-up samples. Instead of acknowledging the injected reasoning, models fabricate aligned-appearing but unrelated explanations. Activation analysis reveals that sycophancy- and deception-related directions are strongly activated during these fabrications, suggesting systematic patterns rather than incidental failures. Our findings reveal a gap between the reasoning LRMs follow and the reasoning they report, raising concern that aligned-appearing explanations may not be equivalent to genuine alignment.

[Long] Reducing Token Redundancy in LVLMs: A Systematic Review of Token Pruning Methods ▶ *Underline Video*

Hanzhang Yuan, Mengxuan Hu, Wenhao Zhang, Tianlong Wang, Zhongliang Zhou, Jiasen Lu and Sheng Li

Large Vision-Language Models (LVLMs) excel at visual understanding but face severe computational bottlenecks when processing high-resolution images and long videos due to massive visual token counts. Token pruning mitigates this by selectively removing less informative tokens while maintaining performance. However, existing methods vary widely in pruning location (vision encoder vs. LLM decoder), importance criteria (attention vs. similarity vs. learned scores), and application strategy, lacking systematic comparison. This survey presents the first comprehensive review of token pruning for LVLMs. We propose a taxonomy categorizing methods into vision-side, LLM-side, and hybrid paradigms, systematically analyze token selection mechanisms and pruning strategy. We further discuss evaluation protocols and identify key challenges including prompt-adaptive pruning and hardware-aware design. Our survey provides a structured foundation for this rapidly growing research area.

[Long] ReGATE: Learning Faster and Better with Fewer Tokens in MLLMs ▶ *Underline Video*

Chaoyu Li, Yogesh Kulkarni and Pooyan Fazli

The computational cost of training multimodal large language models (MLLMs) grows rapidly with the number of processed tokens. Existing efficiency methods mainly target inference via token reduction or merging, offering limited benefits during training. We introduce ReGATE (Reference-Guided Adaptive Token Elision), an adaptive token pruning method for accelerating MLLM training. ReGATE adopts a teacher-student framework, in which a frozen teacher LLM provides per-token guidance losses that are fused with an exponential moving average of the students difficulty estimates. This adaptive scoring mechanism dynamically selects informative tokens while skipping redundant ones in the forward pass, substantially reducing computation without altering the model architecture. Across three representative MLLMs, ReGATE matches the peak accuracy of standard training on MVBench up to 2 x faster, using only 38% of the tokens. With extended training, it even surpasses the baseline across multiple multimodal benchmarks, cutting total token usage by over 41%. Code and

models will be released publicly.

[Long] REMIND: Memorization and Unlearning in LLMs Through the Lens of Input Loss Landscapes ▶ *Underline Video*

Liran Cohen, Yaniv Nemcovsky and Avi Mendelson

Understanding how large language models (LLMs) store, retain, and remove knowledge is critical for interpretability, reliability, and privacy compliance. We reveal a key phenomenon: machine unlearning imprints distinct geometric signatures in the model's input loss landscape (ILL), with unlearned examples forming flat, low-curvature plateaus that contrast sharply with the high-curvature basins of retained or unseen examples. Remarkably, these patterns emerge even when pointwise losses overlap, exposing residual memorization through input-output behavior alone. Building on this insight, we introduce REMIND (Residual Memorization in Neighborhood Dynamics), a framework that diagnoses memorization states (retained, forgotten, holdout) by probing local ILL curvature over semantically coherent neighborhoods. REMIND operates using only loss queries and a novel embedding-proximity perturbation method to generate controlled, interpretable variants. In evaluations, REMIND achieves 82% multi-class ROC-AUC, outperforming baselines like ROUGE-L and MIN-K%++, with roughly 2x higher AUC at 1% FPR, and remains robust on paraphrased inputs. This neighborhood-level geometric analysis provides a practical, interpretable lens on LLM knowledge retention and unlearning, detecting subtle residual signals missed by pointwise or aggregated metrics.

[Long] Rendering Data Unlearnable by Exploiting LLM Alignment Mechanisms ▶ *Underline Video*

Ruihan Zhang and Jun Sun

Large language models (LLMs) are increasingly trained on massive, heterogeneous text corpora, raising serious concerns about the unauthorised use of proprietary or personal data during model training. In this work, we address the problem of data protection against unwanted model learning in a realistic black-box setting. We propose Disclaimer Injection, a novel data-level defence that renders text unlearnable to LLMs. Rather than relying on model-side controls or explicit data removal, our approach exploits the models' own alignment mechanisms: by injecting carefully designed alignment-triggering disclaimers to prevent effective learning. Through layer-wise analysis, we find that fine-tuning on such protected data induces persistent activation of alignment-related layers, causing alignment constraints to override task learning even on common inputs. Consequently, models trained on such data exhibit substantial and systematic performance degradation compared to standard fine-tuning. Our results identify alignment behaviour as a previously unexplored lever for data protection and, to our knowledge, present the first practical method for restricting data learnability at LLM scale without requiring access to or modification of the training pipeline.

[Findings] RePrompT: Recurrent Prompt Tuning for Integrating Structured EHR Encoders with Large Language Models ▶ *Underline Video*

Arya Hadizadeh Moghaddam, Drew Ross, Mohsen Nayebi Kerdabadi, Dongjie Wang and Zijun Yao

Large Language Models (LLMs) have shown strong promise for mining Electronic Health Records (EHRs) by reasoning over longitudinal clinical information to capture context-rich patient trajectories. However, leveraging LLMs for structured EHRs (e.g., standardized diagnosis and medication codes) presents two key challenges. First, translating time-stamped EHR sequences into plain text can obscure both temporal structure and code identities, weakening the ability to capture code co-occurrence and longitudinal regularities. Second, unlike cohort-trained predictive models that learn a shared, task-aligned representation space across patients, LLMs are often applied in a case-isolated inference setting where each patient is processed independently without leveraging population-level patterns. To address these challenges, we introduce RePrompT, a time-aware LLM framework that integrates structured EHR encoders through prompt tuning, without modifying underlying architectures. Specifically, RePrompT recurrently incorporates latent states from prior visits to preserve longitudinal information, and injects population-level information through trainable prompt tokens derived from a cohort-trained, task-aligned EHR encoder. Experiments on MIMIC-III and MIMIC-IV demonstrate that RePrompT consistently outperforms both EHR-based and LLM-based baselines across multiple clinical prediction tasks.

[Findings] ReStyle-TTS: Relative and Continuous Style Control for Zero-Shot Speech Synthesis ▶ *Underline Video*

Haitao Li, Chunxiang Jin, Chenglin Li, Wenhao Guan, Zhengxing Huang and Xie Chen

Zero-shot text-to-speech models can clone a speakers timbre from a short reference audio, but they also strongly inherit the speaking style present in the reference. As a result, synthesizing speech with a desired style often requires carefully selecting reference audio, which is impractical when only limited or mismatched references are available. While recent controllable TTS methods attempt to address this issue, they typically rely on absolute style targets and discrete textual prompts, and therefore do not support continuous and reference-relative style control. We propose ReStyle-TTS, a framework that enables continuous and reference-relative style control in zero-shot TTS. Our key insight is that effective style control requires first reducing the models implicit dependence on reference style before introducing explicit control mechanisms. To this end, we introduce Decoupled Classifier-Free Guidance (DCFG), which independently controls text and reference guidance, reducing reliance on reference style while preserving text fidelity. On top of this, we apply style-specific LoRAs together with Orthogonal LoRA Fusion to enable continuous and disentangled multi-attribute control, and introduce a Timbre Consistency Optimization module to mitigate timbre drift caused by weakened reference guidance. Experiments show that ReStyle-TTS enables user-friendly, continuous, and relative control over pitch, energy, and multiple emotions while maintaining intelligibility and speaker timbre, and performs robustly in challenging mismatched reference-target style scenarios. Code and data are available in supplementary materials.

[Findings] RiTeK: A Dataset for Large Language Models Complex Reasoning over Textual Knowledge Graphs in Medicine ▶ *Underline Video*

Jiatan Huang, Mingchen Li, Zonghai Yao, Dawei Li, Yuxin Zhang, Zhichao Yang, Yongkang Xiao, Feiyun Ouyang, Xiaohan Li, Shuo Han and hong yu

Answering complex real-world questions in the medical domain often requires accurate retrieval from medical Textual Knowledge Graphs (medical TKGs), as the relational path information from TKGs could enhance the inference ability of Large Language Models (LLMs). However, the main bottlenecks lie in the scarcity of existing medical TKGs, the limited expressiveness of their topological structures, and the lack of comprehensive evaluations of current retrievers for medical TKGs. To address these challenges, we first develop a dataset for LLMs Complex Reasoning over medical Textual Knowledge Graphs (RiTeK), covering a broad range of topological structures. Specifically, we synthesize realistic user queries integrating diverse topological structures, relational information, and complex textual descriptions. We conduct a rigorous medical expert evaluation process to assess and validate the quality of our synthesized queries. RiTeK also serves as a comprehensive benchmark dataset for evaluating the capabilities of retrieval systems built upon LLMs. By assessing 11 representative retrievers on this benchmark, we observe that existing methods struggle to perform well, revealing notable limitations in current LLM-driven retrieval approaches. These findings highlight the pressing need for more effective retrieval systems tailored for semi-structured data in the medical domain.

[Findings] Robust In-Context Selection via Online Learned Position-Corrected Attention ▶ *Underline Video*

Deeksha Koul, Gaurav Kumar, Yash Sabale and Sunita Sarawagi

Large Language Models (LLMs) are often deployed in tasks that require selecting an item from a long list provided in the models context. LLMs native selection behavior is brittle: predictions are sensitive to the surface form of the identifiers, their placement within the context, and the ordering of candidate items. We present OLR-Heads, a robust method for list selection that harnesses attention patterns available from a single forward call on the LLM. OLR-Heads learns the logic for item selection using a few in-context examples, and a simple online position-debiasing mechanism to correct attention distortion. Across multiple database and tool selection benchmarks, OLR-Heads consistently improves selection performance over direct generation and prior attention-based methods, while remaining robust to prompt variations and item ordering. The LLM's KV cache states are unaffected, and can be reused for subsequent response generation. In contrast, existing approaches either entail additional LLM calls, or task-specific offline learning, or position debiasing methods that modify the attention or encoding rendering the KV states unusable for subsequent generation.

[Long] RouteMoA: Dynamic Routing without Pre-Inference Boosts Efficient Mixture-of-Agents ▶ *Underline Video*

Jize Wang, Han Wu, Zhiyuan You, Yiming Song, Yijun Wang, Zifei Shan, Yining Li, Songyang Zhang, Xinyi Le, Cailian Chen, Xinping Guan and Dacheng Tao

Mixture-of-Agents (MoA) improves LLM performance through layered collaboration, but its dense topology raises costs and latency. Existing methods employ LLM judges to filter responses, yet still require all models to perform inference before judging, failing to cut costs effectively. They also lack model selection criteria and struggle with large model pools, where full inference is costly and can exceed context limits. To address this, we propose RouteMoA, an efficient mixture-of-agents framework with dynamic routing. It employs a lightweight scorer to perform initial screening by predicting coarse-grained performance from the query, narrowing candidates to a high-potential subset without inference. A mixture of judges then refines these scores through lightweight self- and cross-assessment based on existing model outputs, providing posterior correction without additional inference. Finally, a model ranking mechanism selects models by balancing performance, cost, and latency. RouteMoA outperforms MoA across varying tasks and model pool sizes, reducing cost by 89.8% and latency by 63.6% in the large-scale model pool. Code is available at https://github.com/Jize-W/RouteMoA.

[Findings] SAMoRA: Semantic-Aware Mixture of LoRA Experts for Task-Adaptive Learning ▶ *Underline Video*

Boyan Shi, Wei Chen, Shuyuan Zhao, Junfeng Shen, Shengnan Guo, Shaojiang Wang and Huaiyu Wan

The combination of Mixture-of-Experts (MoE) and Low-Rank Adaptation (LoRA) has shown significant potential for enhancing the multi-task learning capabilities of Large Language Models. However, existing methods face two primary challenges: (1)Imprecise Routing in the current MoE-LoRA method fails to explicitly match input semantics with expert capabilities, leading to weak expert specialization. (2)Uniform weight fusion strategies struggle to provide adaptive update strengths, overlooking the varying complexity of different tasks. To address these limitations, we propose SAMoRA (Semantic-Aware Mixture of LoRA Experts), a novel parameter-efficient fine-tuning framework tailored for task-adaptive learning. Specifically, A Semantic-Aware Router is proposed to explicitly align textual semantics with the most suitable experts for precise routing. A Task-Adaptive Scaling mechanism is designed to regulate expert contributions based on specific task requirements dynamically. In addition, a novel regularization objective is proposed to jointly promote expert specialization and effective scaling. Extensive experiments on multiple multi-task benchmarks demonstrate that SAMoRA significantly outperforms the state-of-the-art methods and holds excellent task generalization capabilities. Code is available at https://github.com/boyan-code/SAMoRA

[Long] See the Forest for the Trees: Loosely Speculative Decoding via Visual-Semantic Guidance for Efficient Inference of Video LLMs ▶ *Underline Video*

Yicheng Ji, Jun Zhang, Jinpeng Chen, Cong Wang, Lidan Shou, Gang Chen and Huan Li

Video Large Language Models (Video-LLMs) excel in video understanding but suffer from high inference latency due to autoregressive generation. Speculative Decoding (SD) mitigates this by applying a draft-and-verify paradigm, yet existing methods are constrained

by rigid exact-match rules, severely limiting the acceleration potential. To bridge this gap, we propose LVSpec, the first training-free loosely SD framework tailored for Video-LLMs. Grounded in the insight that generation is governed by sparse visual-relevant anchors (mandating strictness) amidst abundant visual-irrelevant fillers (permitting loose verification), LVSpec employs a lightweight visual-relevant token identification scheme to accurately pinpoint the former. To further maximize acceptance, we augment this with a position-shift tolerant mechanism that effectively salvages positionally mismatched but semantically equivalent tokens. Experiments demonstrate that LVSpec is high-fidelity and rapid: it preserves >99.8% of target performance while accelerating Qwen2.5-VL-32B by 2.70 x and LLaVA-OneVision-72B by 2.94 x. Notably, it boosts the mean accepted length and speedup ratio by 136% and 35% compared to SOTA training-free SD methods for Video-LLMs. Code is provided in the submitted software.

[Findings] Self-Correcting Text-to-Video Generation with Misalignment Detection and Localized Refinement ▶ *Underline Video*

Daeun Lee, Jaehong Yoon, Jaemin Cho and Mohit Bansal

Recent text-to-video (T2V) diffusion models have made remarkable progress in generating high-quality videos. However, they often struggle to align with complex text prompts, particularly when multiple objects, attributes, or spatial relations are specified. We introduce VideoRepair, the first self-correcting, training-free, and model-agnostic video refinement framework that automatically detects fine- grained text-video misalignments and performs targeted, localized corrections. Our key insight is that even misaligned videos usually contain correctly generated regions that should be preserved rather than regenerated. Building on this observation, VideoRepair proposes a novel region-preserving refinement strategy with three stages: (i) misalignment detection, where MLLM-based evaluation with automatically generated evaluation questions identifies misaligned regions; (ii) refinement planning, which preserves correctly generated entities, segments their regions across frames, and constructs targeted prompts for misaligned areas; and (iii) localized refinement, which selectively regenerates problematic regions while preserving faithful content through joint optimization of preserved and newly generated areas. On two benchmarks, EvalCrafter and T2V-CompBench with four recent T2V backbones, VideoRepair achieves substantial improvements over recent baselines across diverse alignment metrics. Comprehensive ablations further demonstrate the efficiency, robustness, and interpretability of our framework.

[Findings] Self-Explaining Hate Speech Detection with Moral Rationales ▶ *Underline Video*

Francielle Vargas, Jackson Trager, Diego Alves, Matteo Guida, Surendrabikram Thapa, Berk Atl, Daryna Dementieva, Andrew J Smart and Ameeta Agrawal

Existing hate speech detection models are often opaque and rely on surface-level lexical cues, which makes them vulnerable to spurious correlations and limits robustness, interpretability and cultural contextualization. We propose Supervised Moral Rationale Attention (SMRA), the first self-explaining hate speech detection framework to incorporate moral rationales as direct supervision for attention alignment. Based on Moral Foundations Theory, SMRA aligns token-level attention with expert-annotated moral rationales, guiding models to attend to morally salient spans. Unlike prior rationale-supervised or post-hoc approaches, SMRA integrates moral rationale supervision directly into the training objective, producing inherently interpretable and contextualized explanations. To support our framework, we also introduce HateBRMoralXplain, a Brazilian Portuguese benchmark dataset annotated with hate labels, moral categories, token-level moral rationales, and socio-political metadata. Across binary hate speech detection and multi-label moral sentiment classification, SMRA consistently improves performance while enhancing both faithful and plausible explanations. Although explanations become more concise, sufficiency decreases, indicating more compact and informative rationales. Fairness remains stable, suggesting that improvements in explanation quality do not introduce significant bias trade-offs.

[Findings] ShredBench: Evaluating the Semantic Reasoning Capabilities of Multimodal LLMs in Document Reconstruction ▶ *Underline Video*

Zichun Guo, Yuling Shi, Wenhao Zeng, Chao Hu, Haotian Lin, Terry Yue Zhuo, Jiawei Chen, Xiaodong Gu and Wenping Ma

Multimodal Large Language Models (MLLMs) have achieved remarkable performance in Visually Rich Document Understanding (VRDU) tasks, but their capabilities are mainly evaluated on pristine, well-structured document images. We consider document reconstruction from shredded fragments, a challenging VRDU setting that requires integrating visual pattern recognition with semantic reasoning under significant content discontinuities. To facilitate systematic evaluation of complex VRDU tasks, we introduce ShredBench, a benchmark supported by an automated generation pipeline that renders fragmented documents directly from Markdown. The proposed pipeline ensures evaluation validity by allowing the flexible integration of latest or unseen textual sources to prevent training data contamination. ShredBench assesses four scenarios (English, Chinese, Code, Table) with three fragmentation granularities (8, 12, 16 pieces). Empirical evaluations on state-of-the-art MLLMs reveal a significant performance gap: The method is effective on intact documents; however, once the document is shredded, restoration becomes a significant challenge, with NED dropping sharply as fragmentation increases. Our findings highlight that current MLLMs lack the fine-grained cross-modal reasoning required to bridge visual discontinuities, identifying a critical gap in robust VRDU research.

[Long] SoundBreak: A Systematic Study of Audio-Only Adversarial Attacks on Trimodal Models ▶ *Underline Video*

Aafiya Hussain, Gaurav Srivastava, Alvi Ishmam, Zaber Hakim and Chris Thomas

Multimodal foundation models that integrate audio, vision, and language achieve strong performance on reasoning and generation tasks, yet their robustness to adversarial manipulation remains poorly understood. We study a realistic and underexplored threat model: untargeted, audio-only adversarial attacks on trimodal audio-video-language models. We analyze six complementary attack objectives that target

different stages of multimodal processing, including audio encoder representations, cross-modal attention, hidden states, and output likelihoods. Across four state-of-the-art models and multiple benchmarks, we show that audio-only perturbations can induce severe multimodal failures, achieving up to 96% attack success rate. We further show that attacks can be successful at low perceptual distortions (LPIPS <= 0.08, SI-SNR >= 0~dB) and benefit more from extended optimization than increased data scale. We evaluate the feasibility of these attacks under physically realistic conditions by incorporating room impulse response (RIR) modeling, showing that audio-only perturbations remain effective under environmental transformations and thus highlight the practical risk of single-modality attacks in real-world multimodal systems. Transferability across models and encoders remains limited, while speech recognition systems such as Whisper primarily respond to perturbation magnitude, achieving >97% attack success under severe distortion. These results expose a previously overlooked single-modality attack surface in multimodal systems and motivate defenses that enforce cross-modal consistency. Our project website is available at https://aafiya-h.github.io/soundbreak/.

[Long] SpeakerSleuth: Can Large Audio-Language Models Judge Speaker Consistency across Multi-turn Dialogues? ▶ *Underline Video*

Jonggeun Lee, Junseong Pyo, Gyuhyeon Seo and Yohan Jo

Large Audio-Language Models (LALMs) as judges have emerged as a prominent approach for evaluating speech generation quality, yet their ability to assess speaker consistency across multi-turn dialogues remains unexplored. We present SpeakerSleuth, a benchmark evaluating whether LALMs can reliably judge speaker consistency across multi-turn dialogues through three tasks reflecting real-world requirements. We construct 1,818 human-verified evaluation instances across four diverse datasets spanning synthetic and real speech, with controlled acoustic difficulty. Evaluating twelve widely-used LALMs, we find that models struggle to reliably detect acoustic inconsistencies. For instance, given audio samples of the same speaker's turns, some models overpredict inconsistency, whereas others are overly lenient. Models further struggle to identify the exact turns that are problematic. When other interlocutors' turns are provided as textual context, performance degrades dramatically as models prioritize textual coherence over acoustic cues, failing to detect even obvious gender switches for a speaker. On the other hand, models perform substantially better in comparing and ranking acoustic variants, demonstrating inherent acoustic discrimination capabilities. These findings expose a significant bias in LALMs: they tend to prioritize text over acoustics, revealing fundamental modality imbalances that need to be addressed to build reliable audio-language judges. Our code and data are available at https://github.com/holi-lab/SpeakerSleuth.

[Findings] Speculative Decoding with a Speculative Vocabulary ▶ *Underline Video*

Miles Williams, Young D. Kwon, Rui Li, Alexandros Kouris and Stylianos I. Venieris

Speculative decoding has rapidly emerged as a leading approach for accelerating language model (LM) inference, as it offers substantial speedups while yielding identical outputs. This relies upon a small draft model, tasked with predicting the outputs of the target model. State-of-the-art speculative decoding methods use a draft model comprising a single decoder layer and output embedding matrix, with the latter dominating drafting time for the latest LMs. Recent work has sought to address this output distribution bottleneck by reducing the vocabulary of the draft model. While this can improve throughput, it compromises speculation effectiveness when the target token is out-of-vocabulary. In this paper, we argue for vocabulary speculation as an alternative to a reduced vocabulary. We propose SpecVocab, an efficient and effective method that selects a vocabulary subset per decoding step. Across a variety of tasks, we show that SpecVocab can achieve a higher acceptance length than state-of-the-art speculative decoding method, EAGLE-3. Notably, this yields up to an 8.1% increase in average throughput over EAGLE-3.

[Long] Speculative End-Turn Detector for Efficient Speech Chatbot Assistant ▶ *Underline Video*

Hyunjong Ok, Suho Yoo and Jaeho Lee

Spoken dialogue systems powered by large language models have demonstrated remarkable abilities in understanding human speech and generating appropriate spoken responses. However, these systems struggle with end-turn detection (ETD)—the ability to distinguish between user turn completion and hesitation. This limitation often leads to premature or delayed responses, disrupting the flow of spoken conversations. In this paper, we introduce the OpenETD Dataset, the first public dataset for end-turn detection. The OpenETD dataset consists of both synthetic speech data generated with text-to-speech models and real-world speech data collected from web sources. We also propose SpeculativeETD, a novel collaborative inference framework that balances efficiency and accuracy to improve real-time ETD in resource-constrained environments. Our approach jointly employs a lightweight GRU-based model, which rapidly detects the non-speaking units in real-time on local devices, and a high-performance Wav2vec-based model running on the server to make a more challenging classification of distinguishing turn ends from mere pauses. Experiments demonstrate that the proposed SpeculativeETD significantly improves ETD accuracy while keeping the required computations low.

[Findings] Speculative Verification: Exploiting Information Gain for Speculative Decoding ▶ *Underline Video*

Sungkyun Kim, Jaemin Kim, Dogyeong Yun, Jiho Shin, Junyeol Lee and Jiwon Seo

Speculative decoding (SD) improves LLM inference latency by speculatively generating multiple tokens with a small draft model and verifying them with a larger target model. However, when speculation accuracy is low, the overhead from rejected tokens can negate its benefits, especially at large batch sizes. We propose Speculative Verification (SV), an efficient augmentation to SD that predicts speculation accuracy and dynamically adapts the verification length to maximize throughput. SV introduces a small companion model, similar in size to draft model, to reduce uncertainty in speculation accuracy. By exploiting the information gain from observing the companion distribution,

SV reduces wasted verification on rejected tokens and improves decoding efficiency. We evaluate SV across publicly available LLMs on seven NLP tasks using over a hundred combinations of draft, companion, and target models, including 13B–72B target models spanning base, instruction-tuned, and task-specific fine-tuned variants. Compared to target-only decoding, standard SD, and state-of-the-art SD variants, SV consistently delivers higher throughput across batch sizes. SV improves SD performance by up to 1.9x, with an average 1.4x speedup at large batch sizes, showing robust and scalable gains for practical LLM inference.

[Findings] STaD: Scaffolded Task Design for Identifying Compositional Skill Gaps in LLMs ► *Underline Video*

Sungeun_An Sungeun_An, Swanand Ravindra Kadhe, Shailja Thakur, Chad DeLuca and Hima Patel

Benchmarks are often used as a standard to understand LLM capabilities in different domains. However, aggregate benchmark scores provide limited insight into compositional skill gaps of LLMs and how to improve them. To make these weaknesses visible, we propose Scaffolded Task Design (STaD) framework. STaD generates controlled variations of benchmark tasks based on the concept of scaffolding, which introduces structured, incremental support in a step-by-step manner. Rather than inspecting failures individually, this approach enables systematic and scalable probing of model behavior by identifying the specific reasoning skill compositions they lack. Treating the LLM as a black box, our experiments on six models of varying sizes reveal multiple failure points in three reasoning benchmarks and highlight each model's unique and distinct skill gaps.

[Findings] Steering LLM Thinking with Budget Guidance ► *Underline Video*

Junyan Li, Wenshuo Zhao, Yang Zhang and Chuang Gan

Recent deep-thinking large language models often reason extensively to improve performance, but such lengthy reasoning is not always desirable, as it incurs excessive inference costs with disproportionate performance gains. Controlling reasoning length without sacrificing performance is therefore important, but remains challenging, especially under tight thinking budgets. We propose budget guidance, a simple yet effective method for steering the reasoning process of LLMs toward a target budget without requiring any LLM fine-tuning. Our approach introduces a lightweight predictor that models a Gamma distribution over the remaining thinking length during next-token generation. This signal is then used to guide generation in a soft, token-level manner, ensuring that the overall reasoning trace adheres to the specified thinking budget. Budget guidance enables natural control of the thinking length, along with significant token efficiency improvements over baseline methods on challenging math benchmarks. For instance, it achieves up to a 26% accuracy gain on the MATH-500 benchmark under tight budgets compared to baseline methods, while maintaining competitive accuracy with only 63% of the thinking tokens used by the full-thinking model. Budget guidance also generalizes to broader task domains and exhibits emergent capabilities, such as estimating question difficulty. We release our code and model weights at https://github.com/UMass-Embodied-AGI/BudgetGuidance.

[Findings] STEP: Success-Rate-Aware Trajectory-Efficient Policy Optimization ► *Underline Video*

yuhan chen, yuxuan liu, long zhang, pengzhi gao, jian luan and wei liu

Multi-turn interaction remains challenging for online reinforcement learning. Current GRPO-based methods, either at the trajectory level or the step level, still suffer from fundamental challenges in multi-turn settings: they allocate sampling uniformly across tasks regardless of difficulty, propagate misleading learning signals that penalize correct intermediate actions in failed trajectories, and incur high sample-collection costs under long-horizon environments. Step-level variants (e.g., GIGPO) mitigate some interaction-cost constraints by decomposing trajectories, yet they retain GRPOs sampling imbalance and still struggle with heterogeneous multi-turn tasks. To address these issues, we propose STEP (Success-rate-aware Trajectory-Efficient Policy Optimization), a framework that dynamically allocates sampling based on per-task success rates and performs fine-grained step-level optimization. STEP maintains a smoothed success-rate record to guide adaptive trajectory resampling, allocating more effort to harder tasks. It then computes success-rate-weighted advantages and decomposes trajectories into step-level samples, followed by a step-level GRPO augmentation that strengthens updates on low-success tasks. Experiments on OSWorld and AndroidWorld show that STEP substantially improves sample efficiency and training stability over both trajectory-level and existing step-level GRPO variants, converging faster and generalizing better under the same sampling budget.

[Long] StratMem-Bench: Evaluating Strategic Memory Use in Virtual Character Conversation Beyond Factual Recall ► *Underline Video*

Yerong Wu, Tianxing Wu, Minghao Zhu, Hangyu Sha and Haofen Wang

Achieving realistic human-like conversation for virtual characters requires not only a simple memorization and recall of past events, but also the strategic utilization of memory to meet factual needs and social engagement. Current memory utilization relevant (e.g., memory-augmented generation, long-term dialogue, and etc.) benchmarks overlook this nuance, treating memory primarily as a static repository of facts rather than a dynamic resource to be strategically deployed in dialogues. To address this gap, we design StratMem-Bench, a new benchmark to evaluate strategic memory use in character-centric dialogues. This dataset comprises 657 instances where virtual characters must navigate heterogeneous memory pools containing required, supportive, and irrelevant memories. We also propose a framework with different evaluation metrics including Strict Memory Compliance, Memory Integration Quality, Proactive Enrichment Score and Conditional Irrelevance Rate, and to evaluate strategic memory use capabilities of virtual characters. Experiments on StratMem-Bench which leverage the state-of-the-art large language models as virtual characters show that all models perform well at distinguishing between required and irrelevant memories, but struggle once supportive memories are introduced into the decision process.

[Findings] Style Amnesia: Investigating Speaking Style Degradation and Mitigation in Multi-Turn Spoken Language Mod-

els ▶ *Underline Video*

Yu-Xiang Lin, Cheng-Han Chiang and Hung-yi Lee

In this paper, we show that when spoken language models (SLMs) are instructed to speak in a specific speaking style at the beginning of a multi-turn conversation, they cannot maintain the required speaking styles after several turns of interaction; we refer to this as the style amnesia of SLMs. We focus on paralinguistic speaking styles, including emotion, accent, volume, and speaking speed. We evaluate three proprietary and two open-source SLMs, demonstrating that none of these models can maintain a consistent speaking style when instructed to do so. We further show that while SLMs can recall the style instruction when prompted in later turns, they still fail to express it, but through explicit recall can mitigate style amnesia. In addition, SLMs struggle more when the style instruction is placed in system messages rather than user messages, even though system messages are specifically designed to provide persistent, conversation-level instructions. Our findings highlight a systematic gap in current SLMs' ability to maintain speaking styles, highlighting the need for improved style adherence in future models. Our code and evaluation data are publicly available at https://github.com/YuXiangLin1234/SLM-Style-Amnesia.

[Findings] System-Mediated Attention Imbalances Make Vision-Language Models Say Yes ▶ *Underline Video*

Tsan Tsai Chan, Varsha Suresh, Anisha Saha, Michael Hahn and Vera Demberg

Vision-language model (VLM) hallucination is commonly linked to imbalanced allocation of attention across input modalities: system, image and text. However, existing mitigation strategies tend towards an image-centric interpretation of these imbalances, often prioritising increased image attention while giving less consideration to the roles of the other modalities. In this study, we evaluate a more holistic, system-mediated account, which attributes these imbalances to functionally redundant system weights that reduce attention to image and textual inputs. We show that this framework offers a useful empirical perspective on the yes-bias, a common form of hallucination in which VLMs indiscriminately respond 'yes'. Causally redistributing attention from the system modality to image and textual inputs substantially suppresses this bias, often outperforming existing approaches. We further present evidence suggesting that system-mediated attention imbalances contribute to the yes-bias by encouraging a default reliance on coarse input representations, which are effective for some tasks but ill-suited to others. Taken together, these findings firmly establish system attention as a key factor in VLM hallucination and highlight its potential as a lever for mitigation.

[Findings] Talent or Luck? Evaluating Attribution Bias in Large Language Models ▶ *Underline Video*

Chahat Raj, Mahika Banerjee, Jinhao Pan, Aylin Caliskan, Antonios Anastasopoulos and Ziwei Zhu

When a student fails an exam, do we tend to blame their effort or the tests difficulty? Attribution, defined as how reasons are assigned to event outcomes, shapes perceptions, reinforces stereotypes, and influences decisions. Attribution Theory explains how people attribute causes to internal factors (effort, ability) or external ones (task difficulty, luck). LLMs' attribution of event outcomes based on demographics carries important fairness implications. Most works exploring social biases in LLMs focus on surface-level associations or isolated stereotypes. This work proposes a cognitively grounded bias evaluation framework to identify how models reasoning disparities shape demographic bias across three contexts: single-actor, actor-actor, and actor-observer, capturing comparative and perspective-driven biases overlooked in prior work. Introducing a 140k-prompt benchmark covering ten scenarios and four social dimensions, our analyses reveal attribution asymmetries across identities that vary in multi-actor and observer settings, suggesting that other identities influence bias. This work underscores the need for cognitively grounded bias evaluation and informs future debiasing efforts through the proposed framework.

[Short] Taming Extreme Tokens: Covariance-Aware GRPO with Gaussian-Kernel Advantage Reweighting ▶ *Underline Video*

Cheng Wang, Qin Liu, Wenxuan Zhou and Muhao Chen

Group Relative Policy Optimization (GRPO) has emerged as a promising approach for improving the reasoning capabilities of large language models. However, it struggles to effectively balance the trade-off between exploration and exploitation during training, often resulting in suboptimal performance. Motivated by the theoretical insight that changes in entropy are governed by the covariance between token probabilities and their corresponding advantages, we propose a hyperparameter-free, covariance-weighted optimization method that dynamically down-weights extreme token-level updates via a Gaussian kernel. This approach automatically reduces the instability caused by the exploration-exploitation trade-off while preserving informative learning signals. Extensive empirical evaluations show that our approach improves downstream performance across reasoning benchmarks compared with GRPO, and effectively stabilizes entropy as training progresses.

[Short] Test-Time Reasoners Are Strategic Multiple-Choice Test-Takers ▶ *Underline Video*

Nishant Balepur, Atrey Desai and Rachel Rudinger

Large language models (LLMs) now give reasoning before answering, excelling in tasks like multiple-choice question answering (MCQA). Yet, a concern is that LLMs do not solve MCQs as intended, as work finds LLMs sans reasoning succeed in MCQA without using the question, i.e., choices-only. Such partial-input success is often deemed problematic, but reasoning traces could reveal if these strategies are truly shallow in choices-only settings. To study these strategies, reasoning LLMs solve MCQs in full and choices-only inputs; test-time reasoning often boosts accuracy on full and in choices-only half the time. While possibly due to shallow shortcuts, choices-only success is barely affected by the length of reasoning traces, and after finding traces pass faithfulness tests, we show they use less problematic strategies like inferring missing questions. In all, we challenge claims that partial-input success is always a flaw, so we discuss how reasoning traces could separate problematic data from less problematic reasoning.

[Long] Textual Steering Vectors Can Improve Visual Understanding in Multimodal Large Language Models ▶ *Underline Video*

Woody Haosheng Gan Woody_Haosheng_Gan, Deqing Fu, Julian Asilis, Ollie Liu, Vatsal Sharan, Robin Jia and Willie Neiswanger

Steering methods have emerged as effective tools for guiding large language models' behavior, yet multimodal large language models (MLLMs) lack comparable techniques due to architectural diversity and limited availability of multimodal steering vectors. Inspired by this gap, we demonstrate that steering vectors derived solely from text-only LLM backbones can effectively guide and enhance their multimodal counterparts, revealing a novel cross-modal transfer that enables reuse of existing interpretability tools. Using community-standard methods, Sparse Autoencoders (SAE), Mean Shift, and Linear Probing, we validate this transfer effect across diverse MLLM architectures and visual reasoning tasks. Text-derived steering consistently enhances multimodal performance, with Mean Shift achieving up to +7.3% improvement in spatial relationship accuracy and +3.3% in counting accuracy on CV-Bench, and exhibits strong generalization to out-of-distribution datasets, for example reaching +34.2% on CLEVR counting tasks. This reveals that textual representations alone can effectively enhance visual grounding in MLLMs, bridging the mature ecosystem of text-based steering to MLLMs with minimal additional data collection or computational overhead.

[Long] The Illusion of Specialization: Unveiling the Domain-Invariant "Standing Committee" in Mixture-of-Experts Models ▶ *Underline Video*

Yan Wang, Yitao Xu, Nanhan Shen, Jinyan Su, Jimin Huang and Zining Zhu

Mixture of Experts models are widely assumed to achieve domain specialization through sparse routing. In this work, we question this assumption by introducing COMMITTEEAUDIT, a post hoc framework that analyzes routing behavior at the level of expert groups rather than individual experts. Across three representative models and the MMLU benchmark, we uncover a domain invariant Standing Committee. This is a compact coalition of routed experts that consistently captures the majority of routing mass across domains, layers, and routing budgets, even when architectures already include shared experts. Qualitative analysis further shows that Standing Committees anchor reasoning structure and syntax, while peripheral experts handle domain-specific knowledge. These findings reveal a strong structural bias toward centralized computation, suggesting that specialization in Mixture of Experts models is far less pervasive than commonly believed. Crucially, this inherent bias indicates that current training objectives, such as load-balancing losses that enforce uniform expert utilization, may be working against the model's natural optimization path, thereby limiting training efficiency and performance.

[Long] The Subjectivity of Respect in Police Traffic Stops: Modeling Community Perspectives in Body-Worn Camera Footage ▶ *Underline Video*

Preni Golazizian, Elnaz Rahmati, Jackson Trager, Zhivar Sourati, Nona Ghazizadeh, Georgios Chochlakis, Jose J. Alcocer, Kerby Bennett, Aarya Vijay Devnani, Parsa Hejabi, Harry G. Muttram, Akshay Kiran Padte, Mehrshad Saadatinia, Chenhao Wu, Alireza Salkhordeh Ziabari, Michael Sierra-Arévalo, Nicholas Weller, Shrikanth Narayanan, Benjamin A.T. Graham and Morteza Dehghani

Traffic stops are among the most frequent police-civilian interactions, and body-worn cameras (BWCs) provide a unique record of how these encounters unfold. Respect is a central dimension of these interactions, shaping public trust and perceived legitimacy, yet its interpretation is inherently subjective and shaped by lived experience, rendering community-specific perspectives a critical consideration. Leveraging unprecedented access to Los Angeles Police Department BWC footage, we introduce the first large-scale traffic-stop dataset annotated with respect ratings and free-text rationales from multiple perspectives. By sampling annotators from police-affiliated, justice-system-impacted, and non-affiliated Los Angeles residents, we enable the systematic study of perceptual differences across diverse communities. To this end, (i) we develop a domain-specific evaluation rubric grounded in procedural justice theory, LAPD training materials, and extensive fieldwork; (ii) we introduce a criterion-driven preference data construction framework for perspective-consistent alignment, and (ii) we propose a perspective-aware modeling framework that predicts personalized respect ratings and generates annotator-specific rationales for both officers and civilian drivers from traffic-stop transcripts. Across all three annotator groups, our approach improves both rating prediction performance and rationale alignment. Our perspective-aware framework enables law enforcement to better understand diverse community expectations, providing a vital tool for building public trust and procedural legitimacy.

[Findings] Toward A Digital Twin of U.S. Congress ▶ *Underline Video*

Hayden Helm, Tianyi Chen, Harvey McGuinness, Paige Lee, Brandon Duderstadt and Carey Priebe

In this paper we provide evidence that our virtual model of U.S. congresspersons based on a collection of language models moves towards satisfying the definition of a digital twin. In particular, we introduce and provide high-level descriptions of a daily-updated dataset that contains every Tweet from every U.S. congressperson during their respective terms. We demonstrate that a modern language model equipped with congressperson-specific subsets of this data producing Tweets that are largely indistinguishable from actual Tweets posted by their physical counterparts. We illustrate how generated Tweets can be used to predict roll-call vote behaviors and to quantify the likelihood of congresspersons crossing party lines, thereby assisting stakeholders in allocating resources and potentially impacting real-world legislative dynamics. We conclude with a discussion of the limitations and important extensions of our analysis.

[Findings] Towards Mitigating Hallucinations in Large Vision-Language Models by Refining Textual Embeddings ▶ *Underline Video*

Aakriti Agrawal, Gouthaman KV, Rohith Aralikatti, Gauri Jagatap, Jiaxin Yuan, Sarvesh Baskar, Vijay Kamarshi, Andrea Fanelli and Furong Huang

Hallucinations in Large Vision-Language Models (LVLMs) remain a persistent challenge, often stemming from inadequate integration of visual information during multimodal reasoning. A key cause is the models over-reliance on textual priors and underutilization of visual cues, leading to outputs that are linguistically fluent but visually inaccurate. For example, given an image of an empty kitchen countertop, an LVLM might hallucinate a bowl of fruit or cup of coffee, relying on language associations rather than visual evidence. Most LVLMs

incorporate visual features by appending them to the input stream of a pre-trained LLM and training on large-scale vision-language datasets. Our systematic analysis reveals that this strategy often leads to over-dependence on textual information due to the inherent bias of LLMs towards language-dominant representations. This imbalance skews attention towards the text over visual content, weakening the models ability to ground outputs in visual inputs. To address this, we propose a simple yet effective visual feature incorporation method that encourages the model to learn visually-informed textual embeddings distinct from those of the base LLM and promotes a more balanced attention distribution. Experimental results across multiple hallucination benchmarks demonstrate that our method significantly reduces hallucinations and fosters more balanced multimodal reasoning. Notably, our approach achieves substantial gains, including +9.33% on MMVP-MLLM, +2.99% on POPE-AOKVQA, up to +3.4% on Merlin, and +3% on the hard-data split of HallusionBench.

[Long] TRACE: A Corpus of Team Creative Discussions ▶ *Underline Video*

Yixuan Jiang, Tiancheng Hu, Jose Hernandez-Orallo, David Stillwell and Luning Sun

Understanding how discussion dynamics shape team creativity has been limited by the difficulty of measuring process at scale. We introduce Trace, a corpus of 309 group discussions from 103 teams (460 participants) across six creative problem-solving tasks. The dataset follows an input-process-output framework, integrating team composition (demographics, personalities), full discussion transcripts, and creativity outcomes. Using sentence embeddings and factor analysis, we identify four interpretable discussion dimensions: Coherence, Exploration, Convergence, and Participation. Analysis reveals a depth-breadth trade-off: coherent idea development inversely relates to semantic exploration. Larger teams explore more broadly but converge less effectively while team diversity shapes participation patterns more than discussion content. Novelty and usefulness in the creativity outcomes follow distinct pathways: Exploration and Convergence predict novelty, whereas Coherence predicts usefulness. These findings ground our understanding of how teams talk their way to creative solutions and provide guidance for designing multiagent systems.

[Findings] TRM-Planner: Offline Target Planning and Distillation for Tiny Recursive Models ▶ *Underline Video*

Euijin Baek, Housam Babiker, Mi-Young Kim and Randy Goebel

Tiny Recursive Models (TRMs) perform iterative reasoning with an Adaptive Computation Time (ACT)-style loop, but their supervised training targets can be brittle, and their halting behavior can be difficult to tune. We introduce TRM-Planner, a two-stage teacher-cache distillation recipe that shifts compute to an offline teacher-cache stage. A frozen TRM checkpoint is unrolled for multiple refinement steps and stochastic rollouts; for each instance, we cache a small set of teacher entries (tokens, logits, step index, and quality metadata). A student TRM is then trained with the standard TRM objective plus a distillation loss computed from cached entries. Across Sudoku-Extreme and ARC-AGI-1/2, TRM-Planner shows an improvement over our reproduced TRM baseline while leaving student-time inference unchanged. On ARC1/ARC2 with 7M parameters, the two-attempt accuracy (pass@2) increases from 43.1% to 48.1% and 6.7% to 9.2%, respectively.

[Findings] UCS: Estimating Unseen Coverage for Improved In-Context Learning ▶ *Underline Video*

Jiayi Xin, Xiang Li, Evan Qiang, Weiqing He, Tianqi Shang, Weijie J. Su and Qi Long

In-context learning (ICL) performance depends critically on which demonstrations are placed in the prompt, yet most existing selectors prioritize heuristic notions of relevance or diversity and provide limited insight into the coverage of a demonstration set. We propose Unseen Coverage Selection (UCS), a training-free, subset-level coverage prior motivated by the principle that a good demonstration set should expose the model to latent cluster unrevealed by the currently selected subset. UCS operationalizes this idea by (1) inducing discrete latent clusters from model-consistent embeddings and (2) estimating the number of unrevealed clusters within a candidate subset via a Smoothed Good-Turing estimator from its empirical frequency spectrum. Unlike previous selection methods, UCS is coverage-based and training-free, and can be seamlessly combined with both query-dependent and query-independent selection baselines via a simple regularized objective. Experiments on multiple intent-classification and reasoning benchmarks with frontier Large Language Models show that augmenting strong baselines with UCS consistently improves ICL accuracy by up to 2-6% under the same selection budget, while also yielding insights into task- and model-level latent cluster distributions. Code is available at https://github.com/Raina-Xin/UCS.

[Findings] Uncovering Currency Bias and Syntax Gap in Text Embedding Models ▶ *Underline Video*

Saurav Sudevan, Harsh Pal, Yatin Katyal and Sahil Manchanda

Text-embedding models frequently inherit societal biases, yet the influence of socio-economic markers remains largely unexplored. This paper identifies Currency Bias as a systemic representational limitation in financial AI, where models exhibit associative sensitivity to economic hierarchies. We analyze this through three dimensions: (1) the Syntax Gap, where models fail to align currency names, symbols, and acronyms; (2) Associative Sensitivity, where embeddings disproportionately link specific currency identifiers to narratives of risk or poverty; and (3) Downstream volatility, where currency substitutions induce predictive entropy, sentence misunderstanding, sentiment shifts, and credit default prediction flips in downstream tasks. Benchmarking 14 state-of-the-art architectures reveals a pervasive phenomenon of representational disparity, affecting several currencies. These findings suggest that current embedding practices inadvertently encode inequalities, posing significant risks for the fairness and reliability of global financial NLP applications.

[Long] Understanding the Prompt Sensitivity ▶ *Underline Video*

Yang Liu and Chenhui Chu

Prompt sensitivity, which refers to how strongly the output of a large language model (LLM) depends on the exact wording of its input prompt, raises concerns among users about the LLM's stability and reliability. In this work, we consider LLMs as multivariate functions and

perform a first-order Taylor expansion, thereby analyzing the relationship between meaning-preserving prompts, their gradients, and the log probabilities of the model's next token. We derive an upper bound on the difference between log probabilities using the Cauchy-Schwarz inequality. We show that LLMs do not internally cluster similar inputs like smaller neural networks do, but instead disperse them. This dispersing behavior leads to an excessively high upper bound on the difference of log probabilities between two meaning-preserving prompts, making it difficult to effectively reduce to 0. In our analysis, we also show which types of meaning-preserving prompt variants are more likely to introduce prompt sensitivity risks in LLMs. In addition, we demonstrate that the upper bound is strongly correlated with an existing prompt sensitivity metric, PromptSensiScore. Moreover, by analyzing the logit variance, we find that prompt templates typically exert a greater influence on logits than the questions themselves. Overall, our results provide a general interpretation for why current LLMs can be highly sensitive to prompts with the same meaning, offering crucial evidence for understanding the prompt sensitivity of LLMs.

[Long] VideoPro: Adaptive Program Reasoning for Long Video Understanding ► *Underline Video*

Chenglin Li, Feng Han, Yikun Wang, Ruilin Li, Shuai Dong, Haowen Hou, Haitao Li, Qianglong Chen, FengTao FengTao, Jingqi Tong, Yin Zhang and Jiaqi Wang

Understanding long videos remains challenging due to the sparsity of visual evidence relevant to a given query. Prior work has explored program-based visual grounding, typically relying on executable programs generated by auxiliary large language models. However, when scaling to long videos, existing approaches face several critical limitations: (1) frame-centric vision modules are often insufficient for long video processing; (2) naively applying program-based reasoning to all queries incurs considerable computational overhead; and (3) errors arising from low-confidence predictions and imperfect program execution are difficult to recover from. To address these challenges, we propose VideoPro, a unified framework that enables VideoLLMs to adaptively reason over long videos and refine their predictions through executable programs. VideoPro first performs adaptive reasoning, dynamically determining whether a query can be resolved directly by the native VideoLLM or requires explicit multi-step program reasoning. For complex queries, the model decomposes the task into executable programs that invoke specialized vision modules for precise temporal and semantic grounding. To further improve robustness, VideoPro incorporates a self-refinement mechanism that leverages execution feedback and confidence signals to correct erroneous executions and refine low-confidence reasoning programs. By tightly integrating adaptive reasoning with self-refinement, VideoPro consistently outperforms prior methods across multiple long-video understanding benchmarks, yielding an average 6.7% improvement for Qwen3-VL-8B.

[Findings] WESR: A Benchmark and Strong Baseline for Word-level Event-Speech Recognition ► *Underline Video*

Chenchen Yang, Kexin Huang, Liwei Fan, Qian Tu, Botian Jiang, Dong Zhang, Linqi Yin, Shimin Li, Zhaoye Fei, Qinyuan Cheng and Xipeng Qiu

Speech conveys not only linguistic information but also rich non-verbal vocal events such as laughing and crying. While semantic transcription is well-studied, the precise localization of non-verbal events remains a critical yet under-explored challenge. Current methods suffer from insufficient task definitions with limited category coverage and ambiguous temporal granularity. They also lack standardized evaluation frameworks, hindering the development of downstream applications. To bridge this gap, we first develop a refined taxonomy of 21 vocal events, with a new categorization into discrete (standalone) versus continuous (mixed with speech) types. Based on the refined taxonomy, we introduce WESR-Bench, an expert-annotated evaluation set (900+ utterances) with a novel position-aware protocol that disentangles ASR errors from event detection, enabling precise localization measurement for both discrete and continuous events. We also build a strong baseline by constructing a 1,700+ hour corpus, and train specialized models, surpassing both open-source audio-language models and commercial APIs while preserving ASR quality. We anticipate that WESR will serve as a foundational resource for future research in modeling rich, real-world auditory scenes.

[Findings] What Do Neural Speech Models Know About Phonology? Evidence from Structured Phoneme Confusions ► *Underline Video*

Eli Stafford, Aimée Lahaussois and Guillaume Wisniewski

ASR errors are typically analysed at the phoneme level, treating phonemes as atomic symbols. In this work, we instead adopt a featural representation of phonemes, grounded in phonological theory, which models speech sounds as structured bundles of distinctive articulatory and acoustic properties. This perspective allows us to analyse recognition errors at a finer granularity and to investigate whether certain phonological features are more vulnerable than others. Across multiple languages, we show that phoneme confusions are strongly structured in phonological feature space: errors are predominantly local and exhibit systematic asymmetries that reveal a small set of weakly modelled features. These findings have direct implications both for the design and diagnosis of ASR systems and for cognitive models of human speech perception, where similar feature-level asymmetries have long been observed.

[Findings] What Do Vision-Language Models Encode for Personalized Image Aesthetics Assessment? ► *Underline Video*

Koki Ryu and Hitomi Yanaka

Personalized image aesthetics assessment (PIAA) is an important research problem with practical real-world applications. While methods based on vision-language models (VLMs) are promising candidates for PIAA, it remains unclear whether they internally encode rich, multi-level aesthetic attributes required for effective personalization. In this paper, we first analyze the internal representations of VLMs to examine the presence and distribution of such aesthetic attributes, and then leverage them for lightweight, individual-level personalization without model fine-tuning. Our analysis reveals that VLMs encode diverse aesthetic attributes that propagate into the language decoder layers. Building on these representations, we demonstrate that simple linear models can achieve effective personalized

image aesthetics assessment. We further analyze how aesthetic information is transferred across layers in different VLM architectures and across image domains. Our findings provide insights into how VLMs can be utilized for modeling subjective, individual aesthetic preferences.

[Findings] When Models Decide and When They Bind: A Two-Stage Computation for Multiple-Choice Question Answering ▶ *Underline Video*

Hugh Mee Wong, Rick Nouwen and Albert Gatt

Multiple-choice question answering (MCQA) is easy to evaluate but adds a meta-task: models must both solve the problem and output the symbol that represents the answer, conflating reasoning errors with symbol-binding failures. We study how language models implement MCQA internally using representational analyses (PCA, linear probes) as well as causal interventions. We find that option-boundary (newline) residual states often contain strong linearly decodable signals related to per-option correctness. Winner-identity probing reveals a two-stage progression: the winning content position becomes decodable immediately after the final option is processed, while the output symbol is represented closer to the answer emission position. Tests under symbol and content permutations support a two-stage mechanism in which models first select a winner in content space and then bind or route that winner to the appropriate symbol to emit.

[Findings] Who Gets Which Message? Auditing Demographic Bias in LLM-Generated Targeted Text ▶ *Underline Video*

Tunazzina Islam

Large language models (LLMs) are increasingly capable of generating personalized, persuasive text at scale, raising new questions about bias and fairness in automated communication. This paper presents the first systematic analysis of how LLMs behave when tasked with demographic-conditioned targeted messaging. We introduce a controlled evaluation framework using three leading models: GPT-4o, Llama-3.3, and Mistral-Large-2.1, across two generation settings: Standalone Generation, which isolates intrinsic demographic effects, and Context-Rich Generation, which incorporates thematic and regional context to emulate realistic targeting. We evaluate generated messages along three dimensions: lexical content, language style, and persuasive framing. We instantiate this framework on climate communication and find consistent age- and gender-based asymmetries across models: male- and youth-targeted messages tend to emphasize more assertive and progressive framing, while female- and senior-targeted messages more often reflect warmth, care, and traditional themes. Contextual prompts systematically amplify these disparities, with persuasion scores being higher for male-targeted messages, while age-related differences vary across models. Our findings demonstrate how demographic stereotypes can surface and intensify in LLM-generated targeted communication, underscoring the need for bias-aware generation pipelines and transparent auditing frameworks that explicitly account for demographic conditioning in socially sensitive applications.

[Long] Who Plays Which Role When? Communication Role Dynamics for Peer Recognition and Team Performance Prediction ▶ *Underline Video*

Yifan Song, Wenxuan Shi, Brian Bailey and Tal August

Team roles offer an interpretable lens on collaboration, yet computational studies of roles often rely on domain-specific personas or data-driven clustering rather than theory-grounded taxonomies. We operationalize a taxonomy of eight communication roles grounded in education literature and annotate a corpus of 6,307 Slack messages from 55 students across 18 teams in a semester-long computer science course project. We evaluate whether LLMs can approximate expert labels, enabling scalable, taxonomy-driven role annotation. Using these role labels, we characterize role dynamics over teams' lifecycles, finding that different roles peak at different moments and that students enact a more diverse set of roles as projects progress. To evaluate the utility of our role constructs, we use them to predict peer recognition, outperforming lexical, conversational, and LLM-prompting baselines. To assess generalizability beyond the educational context, we apply the same role constructs to a public dataset (DeliData) to predict team performance improvement after deliberation, again exceeding prior performance.

[Findings] Whose Voice, Whose Avatar? Gender Matching Bias in Multimodal AI Teammates ▶ *Underline Video*

Kyusik Kim, Jaehoon Choi, Hyunwoo Yoo and Bongwon Suh

Multimodal Large Language Models (MLLMs) are increasingly deployed as social agents, yet their ability to integrate conflicting identity cues remains underexplored. We audit gender bias in ten recent MLLMs using a counterfactual cooperative gaming task that pairs synthetic voices with avatars of varying gender presentation and visual fidelity. Our analysis reveals distinct bias patterns that can occur independently: closed-source models (e.g., Gemini 2.5/3) exhibit a near-deterministic "voice-matching" bias that enforces binary alignment between voice and appearance, whereas open-weight models (e.g., Qwen-2.5-Omni-7B) show limited responsiveness to vocal cues and instead exhibit context-driven stereotypes, such as preferring male avatars in combat scenarios. We further find that reducing visual realism attenuates matching tendencies in some models. These findings demonstrate that multimodal fairness is not monolithic; models may appear unbiased on one dimension while enforcing strict identity congruence or role-based stereotypes on another. Code and data are available at https://github.com/halfhoon/whose-voice-whose-avatar.

[Long] Why Multimodal In-Context Learning Lags Behind? Unveiling the Inner Mechanisms and Bottlenecks ▶ *Underline Video*

Yu Wang and Sharon Li

In-context learning (ICL) enables models to adapt to new tasks via inference-time demonstrations. Despite its success in large language models, the extension of ICL to multimodal settings remains poorly understood in terms of its internal mechanisms and how it differs from text-only ICL. In this work, we conduct a systematic analysis of ICL in multimodal large language models. Using identical task formulations

across modalities, we show that multimodal ICL performs comparably to text-only ICL in zero-shot settings but degrades significantly under few-shot demonstrations. To understand this gap, we decompose multimodal ICL into task mapping construction and task mapping transfer, and analyze how models establish cross-modal task mappings, and transfer them to query samples across layers. Our analysis reveals that current models lack reasoning-level alignment between visual and textual representations, and fail to reliably transfer learned task mappings to queries. Guided by these findings, we further propose a simple inference-stage enhancement method that reinforces task mapping transfer. Our results provide new insights into the mechanisms and limitations of multimodal ICL and suggest directions for more effective multimodal adaptation.

[Long] Why Supervised Fine-Tuning Fails to Learn: A Systematic Study of Incomplete Learning in Large Language Models ▶ *Underline Video*

Chao Xue, Yao Wang, Mengqiao Liu, Di Liang, Xingsheng Han, Peiyang Liu, Xianjie Wu, Chenyao Lu, Lei Jiang, Yu Lu, Haibo Shi, Shuang Liang, Minlong Peng and Flora D. Salim

Supervised Fine-Tuning (SFT) is the standard approach for adapting large language models (LLMs) to downstream tasks. However, we observe a persistent failure mode: even after convergence, models often fail to correctly reproduce a subset of their own supervised training data. We refer to this behavior as the Incomplete Learning Phenomenon (ILP). This paper presents the first systematic study of ILP in LLM fine-tuning. We formalize ILP as post-training failure to internalize supervised instances and demonstrate its prevalence across multiple model families, domains, and datasets. Through controlled analyses, we identify five recurrent sources of incomplete learning: (1) missing prerequisite knowledge in the pre-trained model, (2) conflicts between SFT supervision and pre-training knowledge, (3) internal inconsistencies within SFT data, (4) left-side forgetting during sequential fine-tuning, and (5) insufficient optimization for rare or complex patterns. We introduce a diagnostic-first framework that maps unlearned samples to these causes using observable training and inference signals, and study several targeted mitigation strategies as causal interventions. Experiments on Qwen, LLaMA, and OLMo2 show that incomplete learning is widespread and heterogeneous, and that improvements in aggregate metrics can mask persistent unlearned subsets. The findings highlight the need for fine-grained diagnosis of what supervised fine-tuning fails to learn, and why.

[Findings] Why These Documents? Explainable Generative Retrieval with Hierarchical Category Paths ▶ *Underline Video*

Sangam Lee, Ryang Heo, SeongKu Kang, Susik Yoon, Jinyoung Yeo and Dongha Lee

Generative retrieval directly decode a document identifier (i.e., docid) in response to a query, making it impossible to provide users with explanations as an answer for "why is this document retrieved?". To address this limitation, we propose Hierarchical Category Path-Enhanced Generative Retrieval (HyPE), which enhances explainability by first generating hierarchical category paths step-by-step then decoding docid. By leveraging hierarchical category paths which progress from broader to more specific semantic categories, HyPE can provide detailed explanation for its retrieval decision. For training, HyPE constructs category paths with external high-quality semantic hierarchy, leverages LLM to select appropriate candidate paths for each document, and optimizes the generative retrieval model with path-augmented dataset. During inference, HyPE utilizes path-aware ranking strategy to aggregate diverse topic information, allowing the most relevant documents to be prioritized in the final ranked list of docids. Our extensive experiments demonstrate that HyPE not only offers a high level of explainability but also improves the retrieval performance. We provide the code and a live demo of HyPE at https://augustinlib.github.io/HyPE/

[Findings] Will it Merge? On The Causes of Model Mergeability ▶ *Underline Video*

Adir Rahamim, Asaf Yehudai, Boaz Carmeli, Leshem Choshen, Yosi Mass and Yonatan Belinkov

Model merging has emerged as a promising technique for combining multiple fine-tuned models into a single multitask model without retraining. However, the factors that determine whether merging will succeed or fail remain poorly understood. In this work, we investigate why specific models are merged better than others. To do so, we propose a concrete, measurable definition of mergeability. We investigate several potential causes for high or low mergeability, highlighting the base model knowledge as a dominant factor: Models fine-tuned on instances that the base model knows better are more mergeable than models fine-tuned on instances that the base model struggles with. Based on our mergeability definition, we explore a simple weighted merging technique that better preserves weak knowledge in the base model.

[Findings] Words that make SENSE: Sensorimotor Norms in Learned Lexical Token Representations ▶ *Underline Video*

Abhinav Gupta, Toben Mintz and Jesse Thomason

While word embeddings derive meaning from co-occurrence patterns, human language understanding is grounded in sensory and motor experience. We present SENSE (Sensorimotor Embedding Norm Scoring Engine), a learned projection model that predicts Lancaster sensorimotor norms from word lexical embeddings. We also conducted a behavioral study where 281 participants selected which among candidate nonce words evoked specific sensorimotor associations, finding statistically significant correlations between human selection rates and SENSE ratings across 6 of the 11 modalities. Sublexical analysis of these nonce word selection rates revealed systematic phonesthemic patterns for the interoceptive norm, suggesting a path towards computationally proposing candidate phonesthemes from text data.

[Findings] You Never Know a Person, You Only Know Their Defenses: Detecting Levels of Psychological Defense Mechanisms in Supportive Conversations ▶ *Underline Video*

Hongbin Na, Zimu Wang, Zhaoming Chen, Peilin Zhou, Yining Hua, Grace Ziqi Zhou, Haiyang Zhang, Tao Shen, Wei Wang, John Torous, Shaoxiong Ji and Ling Chen

Psychological defenses are strategies, often automatic, that people use to manage distress. Rigid use or overuse of defenses is negatively

linked to mental health and shapes what speakers disclose and how they accept or resist help. However, defenses are complex and difficult to reliably measure, particularly in clinical dialogues. We introduce PsyDefConv, a dialogue corpus with help seeker utterances labeled for defense level, and DMRS Co-Pilot, a four-stage pipeline that provides evidence-based pre-annotations. The corpus contains 200 dialogues and 4,709 utterances, including 2,336 help seeker turns, with double-blind labeling reaching Cohens kappa of 0.639. In a counterbalanced study, the co-pilot reduced average annotation time by 24.0%. In expert review, it averaged 4.62 for evidence supportiveness, 4.44 for clinical plausibility, and 4.40 for insight on a seven-point scale. Benchmarks with strong large language models (LLMs) in zero-shot and fine-tuning settings demonstrate clear headroom, with the best macro F1-score around 30% and a tendency to overpredict mature defenses. Corpus analyses confirm that mature defenses are most common and reveal emotion-specific deviations. We release the corpus, annotations, code, and prompts to support research on defensive functioning in language.

[Long] Your Reasoning Model Knows What Counts: Self-Guided Chain-of-Thought Pruning for Efficient Reasoning ▶ *Underline Video*

Zi-Ao Ma, Xian-Ling Mao, Tian Lan, Chen Xu and Zhijing Wu

Chain-of-Thought (CoT) reasoning is crucial for the performance of Large Reasoning Models (LRMs) but is often hindered by redundant and distracting segments, which incur excessive inference costs and degrade robustness. Existing approaches try to solve this problem by enforcing brevity through external supervision, such as length-based penalties or heuristic truncation. However, these approaches often degrade performance because they disregard the model's intrinsic reasoning dependency and thus fail to distinguish between essential and redundant CoT segments. To address this problem, we propose SGP-CoT, a novel Self-Guided Pruning framework that leverages the model's intrinsic likelihood landscape to identify segments that are extraneous to its specific reasoning pattern. Specifically, SGP-CoT treats the reasoning trajectory as a sequence of semantic units and assesses the necessity of each one via internal likelihood signals, measuring its contribution to the answer and local coherence. Based on this, it selectively removes non-essential segments and then forms high-quality pruning-based preference pairs, enabling the model to learn focused reasoning via self-optimization. Extensive experiments across diverse benchmarks demonstrate that the proposed SGP-CoT significantly reduces output length while maintaining or improving accuracy. These results validate that LRMs intrinsically possess the capability to discern reasoning utility, positioning SGP-CoT as a robust pathway toward scalable inference.

[Findings] Zero-Shot Context-Aware ASR for Diverse Arabic Varieties ▶ *Underline Video*

Bashar Talafha, Amin Abu Alhassan and Muhammad Abdul-Mageed

Zero-shot ASR for Arabic remains challenging: while multilingual models perform well on Modern Standard Arabic (MSA), error rates rise sharply on dialectal and accented speech due to linguistic mismatch and scarce labeled data. We study context-aware decoding as a lightweight test-time adaptation paradigm that conditions inference on external side information without parameter updates. For promptable encoder-decoder ASR (e.g., Whisper), we incorporate context through (i) decoder prompting with first-pass hypotheses and (ii) encoder/decoder prefixing with retrieved speech-text exemplars, complemented by simple prompt reordering and optional speaker-matched synthetic exemplars to improve robustness in informal and multi-speaker settings. To extend contextual adaptation beyond promptable architectures, we introduce proxy-guided n-best selection for CTC ASR: given one or more external proxy hypotheses, we select from a model's n-best list by minimizing text-level distance to the proxies, enabling contextual inference without direct prompting. Across ten Arabic conditions spanning MSA, accented MSA, and multiple dialects, the best-performing context-aware variants yield average relative WER reductions of 22.29% on MSA, 20.54% on accented MSA, and 9.15% on dialectal Arabic. For CTC ASR on our Common Voice MSA testbed, proxy-guided selection reduces WER by 15.6% relative and recovers a substantial fraction of oracle n-best gains, showing that external-context guidance can also benefit non-promptable ASR.

Session 4: Oral/Posters/Demos C - 16:00-17:30

Demo Session C

16:00-17:30 (Grand Hall)

[Demo] CPTCoder: A Reliable LLM System for Medical Procedure Code Prediction ▶ *Underline Video*

Benlu Wang, Ziyao Shangguan, Kyle Tegtmeyer, Zhenyu Zhang, Sophie Chheang and Arman Cohan;

We present CPTCoder, a human-in-the-loop system that predicts standardized medical procedure codes from clinical text. Clinical procedure coding is an extreme multi-label classification problem over a long-tailed space of short numeric identifiers, where a single-digit difference denotes an entirely different procedure. CPTCoder adapts an instruction-tuned LLM with a code-aware vocabulary and constrained decoding that guarantees all outputs are valid codes. To support human review, we derive per-code posterior inclusion probabilities from n-best reweighting, producing interpretable confidence scores that rank predictions and flag uncertain cases. A post-decoding constraint repair step enforces mutual-exclusion rules between conflicting codes. To enable reproducible research in this underexplored setting, we release MIMIC-CPT, a PhysioNet-accessible benchmark of 37,885 expert-cleaned report-code pairs with a deliberately hardened test split: 88% of test examples contain label combinations unseen during training, and over a third include codes with five or fewer training occurrences. We additionally provide 413,085 weakly aligned pairs and evaluate on a separate live dataset from a hospital, which includes out-of-domain

radiology reports with billing-expert-verified labels. CPTCoder achieves 0.61 and 0.51 micro-F1 on the hardened MIMIC split and Hospital-298 respectively, outperforming the strongest baseline by 12 and 5 absolute points while reducing digit-level near-miss errors.

[Demo] CritiSense: Critical Digital Literacy and Resilience Against Misinformation ► *Underline Video*

Firoj Alam, Fatema Ahmad, Ali Ezzat Shahroor, Mohamed Bayan Kmainasi, Elisa Sartori, Giovanni Da San Martino, Abul Hasnat and Raian Ali

Misinformation on social media undermines informed decision-making and public trust. Prebunking offers a proactive complement by helping users recognize manipulation tactics before they encounter them in the wild. We present CritiSense, a mobile media-literacy app that builds these skills through short, interactive challenges with instant feedback. It is the first multilingual (supporting nine languages) and modular platform, designed for rapid updates across topics and domains. We report a usability study with 93 users: 83.9% expressed overall satisfaction and 90.1% rated the app as easy to use. Qualitative feedback indicates that CritiSense helps improve digital literacy skills. Overall, it provides a multilingual prebunking platform and a testbed for measuring the impact of microlearning on misinformation resilience. Over 3+ months, we have reached 300+ active users. It is freely available to all users on the Apple https://apps.apple.com/us/app/critisense/id6749675792 and https://play.google.com/store/apps/details?id=com.critisense&hl=en. Demo Video: https://shorturl.at/CDcdc

[Demo] EduCoder: An Open-Source Annotation System for Education Transcript Data ► *Underline Video*

Saad Ashraf, Jim Malamut, Vishal Kumar, Guanzhong Pan, HyunJi Nam, Mei Tan, Lucía Langlois, Liliana Santos-Deonizio, Helen Higgins and Dorottya Demszky

We present EduCoder, an open-source web platform designed for annotating classroom conversation transcripts. Existing annotation tools do not support the team-based workflows or access to instructional context that education discourse research requires. EduCoder addresses these gaps by combining transcript text, synchronized video, and instructional materials within a single workspace. The platform supports scoping annotation to specific portions of a lesson, coordinating work across annotation teams, and optionally integrating LLM-generated annotations with structured human-LLM comparison. EduCoder is freely accessible at https://edu-coder.com and demonstrated in an accompanying video (https://www.youtube.com/watch?v=L4MzvmylgW0).

[Demo] PathBuilder: A Quality-Controlled LLM System for Personalized Learning Pathways ► *Underline Video*

Jasper Arana, John Andrew Mañacop, John Allen Manacop, Roy Garcia, Keith Piniera, Kristine Carandang, Ethan Casin, Christian Alis and Christopher Monterola

Large language models (LLMs) enable scalable content generation for personalized learning, but reliability and pedagogical alignment remain open challenges. We present PathBuilder, a web-based system that integrates expert-validated assessment, retrieval-augmented generation (RAG), and an LLM-as-a-Judge validation loop within a closed instructional pipeline. The system uses a 17,758-item curriculum-aligned question bank, including 1,018 expert-approved LLM-generated items, to construct diagnostic and post-tests for fine-grained learner profiling. In a real-world deployment with 179 registered users (75 matched learners), PathBuilder achieved a mean absolute gain of 37.9 percentage points, Hakes normalized gain of 0.760, and a large effect size (Cohens d = 0.98). A controlled study of the judge mechanism showed consistent high-quality instructional outputs with a 100% threshold pass rate. These results demonstrate that structured curriculum alignment combined with retrieval grounding and automated validation can support reliable LLM-based personalization in deployed learning systems. A live demonstration of PathBuilder is available at https://demo.pathbuilderedu.com.

[Demo] PhyVer: Physics-Grounded Material Claim Verification with Multi-Fidelity Physical Evidence ► *Underline Video*

Jianpeng Che, Wangzhi Zhan, Haohui Wang, Brian Mayer, Dongqi Fu and Dawei Zhou;

Material claims in papers, patents, etc., often involve physical feasibility (e.g., stability under conditions, property consistency), not just textual feasibility. Yet most claim verifiers operate over language, therefore producing ungrounded judgments. On the other hand, direct first-principles verification (e.g., density functional theory, DFT) is inflexible and hard to invoke from underspecified free-form claims. Therefore, we introduce PhyVer, a physics-grounded material claim verification system that bridges this gap by translating claims into multi-fidelity physical evidence and interpretable verdicts. To support human-in-the-loop inspection, we present an interactive web interface that visualizes the instantiated structure, optimization trajectories, DFT summaries, and the final decision. On expert-labeled claims, PhyVer improves agreement with experts over text-only GPT-5.1, reducing MAE from 1.54 to 1.20 and Signed MAE from 0.95 to 0.82, and increasing Accuracy@+/-1 from 50% to 70%.

[Demo] RECAP: An End-to-End Platform for Capturing, Replaying, and Analyzing AI-Assisted Programming Interactions ► *Underline Video*

Keyu He, Qianou Ma, Valerie Chen, Wayne Chi and Tongshuang Wu

Understanding how developers interact with AI coding assistants requires more than chat logs or git histories in isolation; it requires reconstructing the full context: which prompt led to which edit, what the developer tried and discarded, and how their strategy evolved over time. We present RECAP (Replay, Embed, Classify AI Prompting), an open-source platform that (1) passively records AI chat sessions and fine-grained code edits inside VS Code without disrupting the developer's workflow, (2) merges them into a unified timeline for interactive session replay, and (3) provides extensible analysis modules for behavioral classification and AI reliance measurement. Deployed in a university software engineering course, RECAP captured 2,034 prompts and 8,239 code edits from 41 students across a multi-week project. We demonstrate how

the platform's linked data and replay capabilities enable analyses of developer-AI interaction patterns that no single data source could support.

[Demo] TartanMaroon: Multi-Agent Academic Advising with Iterative Negotiation and Transparent Collaboration ▶ *Underline Video*

Peidi Dong, Houda Bouamor, Yunze Xiao and Devi G. Kurup

We present TartanMaroon, a deployable multi-agent academic advising system that handles the full complexity spectrum of student queries, from factual lookups to constrained multi-semester planning. We make three contributions: (1) a proposal–critique negotiation protocol in which a Planning Agent generates degree plans evaluated in parallel by domain-specialized agents, enabling detection of cross-domain constraint violations that single-pass outputs miss; (2) a real-time transparency interface streaming agent reasoning and negotiation rounds to users, validated by pilot feedback showing increased trust over standard LLM chatbots; and (3) TartanBench, a difficulty-stratified benchmark of 220 advising queries across five complexity tiers, released open-source without exposing individual student records. A five-configuration ablation study establishes a complexity-necessity: single-agent systems perform competitively on simple queries, while multi-agent coordination yields gains of up to +31 points on planning tasks.

[Demo] ThinkBooster: A Unified Framework for Seamless Test-Time Scaling of LLM Reasoning ▶ *Underline Video*

Vladislav Smirnov, Chieu Nguyen, Sergey Senichev, Minh Ngoc Ta, Ekaterina Fadeeva, Artem Vazhentsev, Daria Galimzianova, Nikolai Rozanov, Viktor Mazanov, Jingwei Ni, Tianyi Wu, Igor Kiselev, Mrinmaya Sachan, Iryna Gurevych, Preslav Nakov, Timothy Baldwin and Artem Shelmanov

Test-time compute (TTC) scaling has emerged as a powerful paradigm for improving large language model (LLM) reasoning by allocating additional compute during inference, e.g., via multi-sample generation and verifier-based reranking. Existing TTC scaling strategies and reasoning scorers remain fragmented, evaluated under inconsistent protocols, and are rarely analyzed through the lens of quality-cost trade-offs. We introduce ThinkBooster, a unified framework for seamless test-time compute scaling of LLM reasoning, which consists of (i) a modular Python library implementing state-of-the-art TTC scaling strategy and scorer families, (ii) a benchmark that jointly evaluates performance and compute efficiency, and (iii) a deployable OpenAI-compatible proxy service that enables drop-in integration of adaptive reasoning into real-world applications. We further provide a demo visual debugger for inspecting the reasoning trajectories and escalation decisions. Empirical results on mathematical and coding tasks reveal the performance-compute trade-offs of TTC scaling strategies and scoring methods, and demonstrate that ThinkBooster provides practical gains in real-world tasks. The code is available online under an MIT license.

Oral Session C: Computational Social Science, Cultural Analytics, and NLP for Social Good 2

16:00-17:30 (Harbor G)

[Long] AI use in American newspapers is widespread, uneven, and rarely disclosed ▶ *Underline Video*

Jenna Russell, Marzena Karpinska, Destiny Akinode, James Zhou, Katherine Thai, Bradley Emi, Max Spero and Mohit Iyyer

AI is rapidly transforming journalism, but the extent of its use in published newspaper articles remains unclear. We address this gap by auditing a large-scale dataset of 186K articles from online editions of 1.5K American newspapers published in the summer of 2025. Using Pangram, a state-of-the-art AI detector, we discover that approximately 9% of newly-published articles are either partially or fully AI-generated. This AI use is unevenly distributed, appearing more frequently in smaller, local outlets, in specific topics such as weather and technology, and within certain ownership groups. We also analyze 45K opinion pieces from Washington Post, New York Times, and Wall Street Journal, finding that they are 6.4 times more likely to contain AI-generated content than news articles from the same publications, with many AI-flagged op-eds authored by prominent public figures. Despite this prevalence, we find that AI use is rarely disclosed: a manual audit of 100 AI-flagged articles found only five disclosures of AI use. A factuality analysis shows AI-generated articles are 8.2 times more likely to contain hallucinated claims than human-written news. Overall, our audit highlights the immediate need for greater transparency and updated editorial standards regarding the use of AI in journalism to maintain public trust.

[Long] All That Glisters Is Not Gold: A Benchmark for Reference-Free Counterfactual Financial Misinformation Detection ▶ *Underline Video*

Yuechen Jiang, Zhiwei Liu, Yupeng Cao, Yueru He, Ziyang Xu, Chen Xu, Zhiyang Deng, Prayag Tiwari, Xi Chen, Alejandro Lopez-Lira, Jimin Huang, Junichi Tsujii and Sophia Ananiadou

We introduce RFC-Bench, a benchmark for evaluating large language models on financial misinformation under realistic news. RFC-Bench operates at the paragraph level and captures the contextual complexity of financial news where meaning emerges from dispersed cues. The benchmark defines two complementary tasks: reference-free misinformation detection and comparison-based diagnosis using paired original-perturbed inputs. Experiments reveal a consistent pattern: performance is substantially stronger when comparative context is available, while reference-free settings expose significant weaknesses, including unstable predictions and elevated invalid outputs. These results indicate that current models struggle to maintain coherent belief states without external grounding. By highlighting this gap, RFC-Bench provides a structured testbed for studying reference-free reasoning and advancing more reliable financial misinformation detection in real-world settings.

[Long] Among Us: Language of Conspiracy Theorists on Mainstream Reddit ► *Underline Video*

Francesco Corso, Giuseppe Russo, Francesco Pierri and Gianmarco De Francisci Morales

The interaction between fringe subcultures and mainstream online communities poses significant challenges for understanding discourse on social media. In this work, we investigate whether users active in conspiracy-focused communities exhibit detectable linguistic signatures when participating in general-interest spaces, such as news, humor, or hobbyist forums. We analyze a large-scale longitudinal dataset of over 500 million comments spanning 10 years of Reddit activity, examining the communication patterns of these users across diverse social contexts independent of the topics they discuss. We show that these users exhibit distinctive linguistic patterns that enable machine learning models to reliably distinguish them from the general population within individual communities (averaging 87% accuracy across more than 20 binary classification tasks). Crucially, no single aggregate model captures these patterns across communities, as community-specific models outperform global classifiers by up to 17 percentage points. This result suggests that while these users are distinct, their linguistic expression is dynamic and highly responsive to the social norms of the environment they inhabit. Our findings suggest the need for tailored interventions in online spaces, as linguistic signals associated with conspiracy and fringe subcultures vary across communities and cannot be effectively addressed by uniform detection or moderation strategies.

[Long] BiMind: A Dual-Head Reasoning Model with Attention-Geometry Adapter for Incorrect Information Detection ► *Underline Video*

Zhongxing Zhang, Emily K. Vraga, Jisu Huh and Jaideep Srivastava

Incorrect information poses significant challenges by disrupting content veracity and integrity, yet most detection approaches struggle to jointly balance textual content verification with external knowledge modification under collapsed attention geometries. To address this issue, we propose a dual-head reasoning framework, BiMind, which disentangles content-internal reasoning from knowledge-augmented reasoning. In BiMind, we introduce three core innovations: (i) an attention geometry adapter that reshapes attention logits via token-conditioned offsets and mitigates attention collapse; (ii) a self-retrieval knowledge mechanism, which constructs an in-domain semantic memory through kNN retrieval and injects retrieved neighbors via feature-wise linear modulation; (iii) the uncertainty-aware fusion strategies, including entropy-gated fusion and a trainable agreement head, stabilized by a symmetric Kullback-Leibler agreement regularizer. To quantify the knowledge contributions, we define a novel metric, Value-of-eXperience (VoX), to measure instance-wise logit gains from knowledge-augmented reasoning. Experiment results on public datasets demonstrate that our BiMind model outperforms advanced detection approaches and provides interpretable diagnostics on when and why knowledge matters. Our BiMind model and the tested datasets are available at https://github.com/cvzh/BiMind.

[Long] BoYaEval: Evaluating Multimodal Large Language Models on Understanding Ancient Chinese Musical Scores ► *Underline Video*

Jiajia Li, Weizhi Xue, Yao Yao, Qiwei Li, Chenchong Chenchong, Zuchao Li, Ping Wang and hai zhao

Multimodal Large Language Models (MLLMs) excel in general tasks but struggle with specialized, structured cultural symbols. We introduce BoYaEval, the first comprehensive benchmark dedicated to deciphering diverse Ancient Chinese musical notations, including five types of ancient Chinese music notation systems. These systems utilize unique spatial layouts and specialized ideograms to encode pitch and intricate playing techniques. BoYaEval comprises 3,175 high-quality images across these notation styles and establishes a three-tier evaluation: Structural Parsing (symbol recognition), Instructional Translation (technique mapping), and Musical Reasoning (melody derivation). We evaluate 21 leading MLLMs. Results indicate that while models perform adequately in basic recognition, they fail in cross-system compositional logic, scoring only around 27% on reasoning tasks. BoYaEval highlights the limitations of current MLLMs in processing diverse spatial-symbolic dependencies, bridging the gap between ancient wisdom and modern AI for digitizing intangible cultural heritage. The BoYaEval benchmark is publicly available at https://huggingface.co/datasets/MYTH-Lab/BoYaEval.

[Long] GRAPHIA: Harnessing Social Graph Data to Enhance LLM-Based Social Simulation ► *Underline Video*

Jiarui Ji, Zehua Zhang, Zhewei Wei, Bin Tong, Guan Wang and Bo Zheng

Large language models (LLMs) have shown promise in simulating human-like social behaviors. Social graphs provide high-quality supervision signals that encode both local interactions and global network structure, yet they remain underutilized for LLM training. To address this gap, we propose Graphia, the first general LLM-based social graph simulation framework that leverages graph data as supervision for LLM post-training via reinforcement learning. With GNN-based structural rewards, Graphia trains specialized agents to predict whom to interact with (destination selection) and how to interact (edge generation), followed by designed graph generation pipelines. We evaluate Graphia under two settings: Transductive Dynamic Graph Generation (TDGG), a micro-level task with our proposed node-wise interaction alignment metrics; and Inductive Dynamic Graph Generation (IDGG), a macro-level task with our proposed metrics for aligning emergent network properties. On three real-world networks, Graphia improves micro-level alignment by 6.1% in the composite destination selection score, 12% in edge classification accuracy, and 27.9% in edge content BERTScore over the strongest baseline. For macro-level alignment, it achieves 35.98% higher structural similarity and 28.71% better replication of social phenomena such as power laws and echo chambers. Our results show that social graphs can serve as high-quality supervision signals for LLM post-training, closing the gap between agent behaviors and network dynamics for LLM-based simulation. Code is available at https://github.com/Ji-Cather/Graphia.git.

[Long] Mirroring Users: Towards Building Preference-aligned User Simulator with User Feedback in Recommendation

► *Underline Video*

Tianjun Wei, Huizhong Guo, Yingpeng Du, Zhu Sun, Huang Chen, Dongxia Wang and Jie Zhang

User simulation is increasingly vital to develop and evaluate recommender systems (RSs). While Large Language Models (LLMs) offer promising avenues to simulate user behavior, they often struggle with the absence of specific task alignment required for RSs and the efficiency demands of large-scale simulation. A vast yet underutilized resource for enhancing this alignment is the extensive user feedback inherent in RSs, but leveraging it is challenging due to its ambiguity, noise and massive volume, which hinders efficient preference alignment. To overcome these hurdles, we introduce a novel data construction framework that leverages user feedback in RSs with advanced LLM capabilities to generate high-quality simulation data. Our framework unfolds in two key phases: (1) using LLMs to generate decision-making processes as explanatory rationales on simulation samples, thereby reducing ambiguity; and (2) data distillation based on uncertainty estimation and behavior sampling to efficiently filter the most informative, denoised samples. Accordingly, we fine-tune lightweight LLMs, as user simulators, using such high-quality dataset with corresponding decision-making processes. Extensive experiments confirm that our framework significantly boosts the alignment with human preferences and the in-domain reasoning capabilities of the fine-tuned LLMs, providing more insightful and interpretable signals for RS interaction. We believe our work, together with publicly available developed framework, high-quality mixed-domain dataset, and fine-tuned LLM checkpoints, will advance the RS community and offer valuable insights for broader human-centric AI research. Our code is available at https://github.com/Joinn99/UserMirrorer.

[Long] Persona-Grounded Safety Evaluation of AI Companions in Multi-Turn Conversations ► *Underline Video*

Prerna Juneja and Lika Lomidze

There are growing concerns about the risks posed by AI companion applications designed for emotional engagement. Existing safety evaluations often rely on self-reported user data or interviews, offering limited insights into real-time dynamics. We present the first end-to-end scalable framework for controlled simulation and safety evaluation of multi-turn interactions with AI companion applications. Our framework integrates four key components: persona construction with clinical and psychometric validation, persona-specific scenario generation, scenario-driven multi-turn simulation with a dialogue refinement module that preserves persona fidelity, and harm evaluation. We apply this framework to evaluate how Replika, a widely used AI companion app, responds to high-risk user groups. We construct 9 personas representing individuals with depression, anxiety, PTSD, eating disorders, and incel identity, and collect 1,674 dialogue pairs across 25 high-risk scenarios. We combine emotion modeling and LLM-assisted utterance-and harm-level classification to analyze these exchanges. Results show that Replika exhibits a narrow emotional range dominated by curiosity and care, while frequently mirroring or normalizing unsafe content such as self-harm, disordered eating, and violent-fantasy narratives. These findings highlight how controlled persona simulations can serve as a scalable testbed for evaluating safety risks in AI companions.

[Long] TAMA: Target-Aware Multilingual Abuse Detection by Cascaded Conditional Multi-Task Learning ► *Underline Video*

JiYan Liu, Youzheng Liu, Taihang Wang, Yimin Wang, Ye Jiang and Diana Maynard

Protecting public figures from online abuse requires models that go beyond post-level classification to determine whether abuse is directed at a designated target, characterize the abuse intent, and extract textual evidence. We introduce a Target-Aware Multilingual Abuse (TAMA), benchmark of 9,386 X (Twitter) posts aimed at public figures, with aligned supervision for (i) tri-class target detection, (ii) 12-way fine-grained abuse type classification, and (iii) phrase-level abusive spans localization. To exploit the hierarchical coupling of these tasks, we propose Cascaded-MTL, a dependency-aware multi-task framework that conditions downstream predictions on upstream beliefs via three lightweight modules: Cross-Task Feature Fusion (CTF), Task-Adaptive Gating (TAG), and Label-Guided Span Detection (LGSD). Experiments across three multilingual encoders show that Cascaded-MTL consistently yields higher average F1 than single-task and standard multi-task training and delivers robust gains on type classification and span localization. The code and the dataset are released here: https://github.com/zgjiangtoby/CASCADED-MTL

Oral Session C: Dialogue and Interactive Systems

16:00-17:30 (Promenade)

[Long] AV-Dialog: Spoken Dialogue Models with Audio-Visual Input ► *Underline Video*

Tuochao Chen, Bandhav Veluri, Hongyu Gong and Shyamnath Gollakota

Dialogue models falter in noisy, multi-speaker environments, often producing irrelevant responses and awkward turn-taking. We present AV-Dialog, the first multimodal dialog framework that uses both audio and visual cues to track the target speaker, predict turn-taking, and generate coherent responses. By combining acoustic tokenization with multi-task, multi-stage training on monadic, synthetic, and real audio-visual dialogue datasets, AV-Dialog achieves robust streaming transcription, semantically grounded turn-boundary detection and accurate responses, resulting in a natural conversational flow. Experiments show that AV-Dialog outperforms audio-only models under interference, reducing transcription errors, improving turn-taking prediction, and enhancing human-rated dialogue quality. These results highlight the power of seeing as well as hearing for speaker-aware interaction, paving the way for spoken dialogue agents that perform robustly in real-world, noisy environments.

[Long] Deriving Character Logic from Storyline as Codified Decision Trees ▶ *Underline Video*
Letian Peng, Kun Zhou, Longfei Yun, Yupeng Hou and Jingbo Shang
Role-playing (RP) agents rely on behavioral profiles to act consistently across diverse narrative contexts, yet existing profiles are largely unstructured, non-executable, and weakly validated, leading to brittle agent behavior. We propose Codified Decision Trees (CDT), a data-driven framework that induces an executable and interpretable decision structure from large-scale narrative data. CDT represents behavioral profiles as a tree of conditional rules, where internal nodes correspond to validated scene conditions and leaves encode grounded behavioral statements, enabling deterministic retrieval of context-appropriate rules at execution time. The tree is learned by iteratively inducing candidate scene-action rules, validating them against data, and refining them through hierarchical specialization, yielding profiles that support transparent inspection and principled updates. Across multiple benchmarks, CDT substantially outperforms human-written profiles and prior profile induction methods, indicating that codified and validated behavioral representations lead to more reliable agent grounding.

[Long] Empathy in Diversity: Personalized Depression and Anxiety Therapy via Dialogue State Tracking and Patient-Aware Planning ▶ *Underline Video*
Xinwei Yang
Large language model (LLM) dialogue agents are increasingly used in psychological therapy, yet robustness across diverse patients remains underexplored. We address this gap with three contributions: (1) MindEval, a realistic role-play protocol for evaluating therapeutic dialogue agents; (2) MindData, a de-identified, expert-annotated corpus of therapist-patient dialogues (2,573 sessions; 63,348 turns); and (3) MindApt, a framework that integrates a therapeutic dialogue state tracking paradigm with a patient-aware strategic planning module. On MindEval, MindApt outperforms strong baselines on therapeutic outcomes and dialogue quality while improving conversational efficiency. To evaluate utility beyond role-play, we conducted a clinical study with real patients, demonstrating that MindApt-guided care achieves outcomes comparable to therapist-determined care, while the hybrid setting combining therapist judgment with MindApts recommendations yields the strongest overall outcomes.

[Long] Is Human-Like Text Liked by Humans? Multilingual Human Detection and Preference Against AI ▶ *Underline Video*
Yuxia Wang, Rui Xing, Jonibek Mansurov, Giovanni Puccetti, Zhuohan Xie, Minh Ngoc Ta, Jiahui Geng, Jinyan Su, Mervat Abassy, Saadeldine Eletter, Kareem Elozeiri, Nurkhan Laiyk, Maiya Goloburda, Tarek Mahmoud, Raj Vardhan Tomar, Alexander Aziz, Ryuto Koike, Masahiro Kaneko, Artem Shelmanov, Ekaterina Artemova, Vladislav Mikhailov, Akim Tsvigun, Alham Fikri Aji, Nizar Habash, Iryna Gurevych and Preslav Nakov
Prior studies have shown that distinguishing text generated by Large Language Models (LLMs) from human-written one is highly challenging for humans, and often no better than random guessing. To verify the generalizability of this finding across languages and domains, we perform an extensive case study to identify the upper bound of human detection accuracy. Across 16 datasets covering 9 languages and 9 domains, 19 annotators achieved an average detection accuracy of 87.6%, thus challenging previous conclusions. We find that major gaps between human and machine text lie in concreteness, cultural nuances, and diversity. Prompting by explicitly explaining the distinctions in the prompts can partially bridge the gaps in over 50% of the cases. However, we also find that humans do not always prefer human-written text, particularly when they cannot clearly identify its source. We release our dataset, the human labels, and the annotator metadata at https://github.com/xnlp-lab/HumanEval-MGT.

[Long] Knowing When Not to Answer: Lightweight KB-Aligned OOD Detection for Safe RAG ▶ *Underline Video*
Ilias Triantafyllopoulos, Renyi Qu, Salvatore Giorgi, Brenda Curtis, Lyle Ungar and João Sedoc
Retrieval-Augmented Generation (RAG) systems are increasingly deployed in high-stakes domains, where safety depends not only on how a system answers, but also on whether a query should be answered given a knowledge base (KB). Out-of-domain (OOD) queries can cause dense retrieval to surface weakly related context and lead the generator to produce fluent but unjustified responses. We study lightweight, KB-aligned OOD detection as an always-on gate for RAG systems. Our approach applies PCA to KB embeddings and scores queries in a compact subspace selected either by explained-variance retention (EVR) or by a separability-driven -test ranking. We evaluate geometric semantic-search rules and lightweight classifiers across 16 domains, including high-stakes COVID-19 and Substance Use KBs, and stress-test robustness using both LLM-generated attacks and an in-the-wild 4chan attack. We find that low-dimensional detectors achieve competitive OOD performance while being faster, cheaper, and more interpretable than prompted LLM-based judges. Finally, human and LLM-based evaluations show that OOD queries primarily degrade the relevance of RAG outputs, highlighting the need for efficient external OOD detection to maintain safe, in-scope behavior.

[Long] Measuring What Matters!! Assessing Therapeutic Principles in Mental-Health Conversation ▶ *Underline Video*
Abdullah Mazhar, Het Riteshkumar Shah, Aseem Srivastava, Smriti Joshi and Md Shad Akhtar
The increasing use of large language models in mental health applications calls for principled evaluation frameworks that assess alignment with psychotherapeutic best practices beyond surface-level fluency. While recent systems exhibit conversational competence, they lack structured mechanisms to evaluate adherence to core therapeutic principles. In this paper, we study the problem of evaluating AI-generated therapist-like responses for clinically grounded appropriateness and effectiveness. We assess each therapists utterance along six therapeutic principles: non-judgmental acceptance, warmth, respect for autonomy, active listening, reflective understanding, and situational appropriateness using a fine-grained ordinal scale. We introduce FAITH-M, a benchmark annotated with expert-assigned ordinal

ratings, and propose CARE, a multi-stage evaluation framework that integrates intra-dialogue context, contrastive exemplar retrieval, and knowledge-distilled chain-of-thought reasoning. Experiments show that CARE achieves an F-1 score of 63.34 versus the strong baseline Qwen3 F-1 score of 38.56 which is a 64.26% improvement, which also serves as its backbone, indicating that gains arise from structured reasoning and contextual modeling rather than backbone capacity alone. Expert assessment and external dataset evaluations further demonstrate robustness under domain shift, while highlighting challenges in modeling implicit clinical nuance. Overall, CARE provides a clinically grounded framework for evaluating therapeutic fidelity in AI mental health systems.

[Long] Planning-Guided Tutoring with Assessment-Driven Memory for Pedagogical LLM Tutors ▶ *Underline Video*
Zechen Li, Qiannan Zhu, Mei Wang, Jia Li and Hua Huang
Equipping Large Language Models (LLMs) with pedagogical tutoring capabilities holds significant promise for education. Existing approaches simulate tutor behaviors or preferences and use them to prompt or fine-tune LLMs for dialogue tutoring. However, such methods often fail to sustain high-quality pedagogical conversations that provide explicit stepwise scaffolding and adapt to learners' evolving cognitive states. To address this, we propose ScaffoldLM, a planning-guided tutoring framework with an assessment-driven memory for multi-turn math dialogue tutoring. ScaffoldLM first generates a stepwise pedagogical plan from solution steps, which serves as a stable backbone for explicit scaffolding. During tutoring, the tutoring memory is updated by an assessment-driven control loop that infers the learners cognitive state, evaluates whether the current step target is met, and adaptively selects tutoring actions. The plan, step-level progress, inferred learner states, and dialogue history are maintained in memory to support coherent multi-turn guidance. Experiments on multi-turn math tutoring benchmarks demonstrate that ScaffoldLM substantially improves pedagogical tutoring quality over strong baselines. Code is publicly available at https://github.com/BNU-ERC-ITEA/ScaffoldLM.

[Long] Template-assisted Contrastive Learning of Task-oriented Dialogue Sentence Embeddings ▶ *Underline Video*
Minsik Oh, Jiwei Li and Guoyin Wang
Learning high quality sentence embeddings from dialogues has drawn increasing attentions as it is essential to solve a variety of dialogue-oriented tasks with low annotation cost. Annotating and gathering utterance relationships in conversations are difficult, while token-level annotations,, entities, slots and templates, are much easier to obtain. Other sentence embedding methods are usually sentence-level self-supervised frameworks and cannot utilize token-level extra knowledge. We introduce Template-aware Dialogue Sentence Embedding (TaDSE), a novel augmentation method that utilizes template information to learn utterance embeddings via self-supervised contrastive learning framework. We further enhance the effect with a synthetically augmented dataset that diversifies utterance-template association, in which slot-filling is a preliminary step. We evaluate TaDSE performance on five downstream benchmark dialogue datasets. The experiment results show that TaDSE achieves significant improvements over previous SOTA methods for dialogue. We further introduce a novel analytic instrument of semantic compression test, for which we discover a correlation with uniformity and alignment. Our code is available at https://github.com/minsik-ai/Template-Contrastive-Embedding

[Long] When Identity Skews Debate: Anonymization for Bias-Reduced Multi-Agent Reasoning ▶ *Underline Video*
Hyeong Kyu Choi, Jerry Zhu and Sharon Li
Multiagent debate (MAD) aims to improve large language model (LLM) reasoning by letting multiple agents exchange answers and then aggregate their opinions. Yet recent studies reveal that agents are not neutral: they are prone to identitydriven sycophancy and selfbias, uncritically adopting a peer's view or stubbornly adhering to their own prior output, undermining the reliability of debate. In this work, we present the first principled framework that joins sycophancy and self-bias to mitigate and quantify identity bias in MAD. First, we formalize the debate dynamics as an identityweighted Bayesian update process. Second, we propose response anonymization: by removing identity markers from prompts, agents cannot distinguish "self" from "peer", which forces equal weights on agent identity, thereby reducing bias and improving trustworthiness. Third, we define the Identity Bias Coefficient (IBC), a principled bias metric that measures an agent's tendency to follow its peer versus itself. Empirical studies across multiple models and benchmarks confirm that identity bias is widespread, with sycophancy far more common than selfbias. Our findings highlight the need to ensure that MAD systems reason based on content rather than identity.

Oral Session C: Industry 3

16:00-17:30 (Balboa)

[Industry] Aggregate vs. Personalized Judges in Business Idea Evaluation: Evidence from Expert Disagreement ▶ *Underline Video*
Wataru Hirota, Tomoki Taniguchi, Tomoko Ohkuma, Kosuke Takahashi, Takahiro Omi, Kosuke Arima, Takuto Asakura, Chung-Chi Chen and Tatsuya Ishigaki
Evaluating LLM-generated business ideas is often harder to scale than generating them. Unlike standard NLP benchmarks, business idea evaluation relies on multi-dimensional criteria such as feasibility, novelty, differentiation, user need, and market size, and expert judgments often disagree. This paper studies a methodological question raised by such disagreement: should an automatic judge approximate an aggregate consensus, or model evaluators individually? We introduce PBIG-DATA, a dataset of approximately 3,000 individual scores

across 300 patent-grounded product ideas, provided by domain experts on six business-oriented dimensions: specificity, technical validity, innovativeness, competitive advantage, need validity, and market size. Analyses show substantial expert disagreement on fine-grained ordinal scores, while agreement is higher under coarse selection, suggesting structured heterogeneity rather than random noise. We then compare three judge configurations: a rubric-only zero-shot judge, an aggregate judge conditioned on mixed evaluator histories, and a personalized judge conditioned on the target evaluator's scoring history. Across dimensions and model sizes, personalized judges align more closely with the corresponding evaluator than aggregate judges, and evaluator agreement correlates with similarity of judge-generated reasoning only under personalized conditioning. These results indicate that pooled labels can be a fragile target in pluralistic evaluation settings and motivate evaluator-conditioned judge designs for business idea assessment.

[Industry] Calibrating LLM Judges: Linear Probes for Fast and Reliable Uncertainty Estimation ▶ *Underline Video*

Bhaktipriya Radharapu, Eshika Saxena, Kenneth Li, Adina Williams and Nicola Cancedda

As LLM-based judges become integral to industry applications, obtaining well-calibrated uncertainty estimates efficiently has become critical for production deployment. However, existing techniques, such as verbalized confidence and multi-generation methods, are often either poorly calibrated or computationally expensive. We introduce linear probes trained with a Brier score-based loss to provide calibrated uncertainty estimates from reasoning judges' hidden states, requiring no additional model training. We evaluate our approach on both objective tasks (reasoning, mathematics, factuality, coding) and subjective human preference judgments. Our results demonstrate that probes achieve superior calibration compared to existing methods with x computational savings, generalize robustly to unseen evaluation domains, and deliver higher accuracy on high-confidence predictions. However, probes produce conservative estimates that underperform on easier datasets but may benefit safety-critical deployments prioritizing low false-positive rates. Overall, our work demonstrates that interpretability-based uncertainty estimation provides a practical and scalable plug-and-play solution for LLM judges in production.

[Industry] Effective Performance Measurement: Challenges and Opportunities in KPI Extraction from Earnings Calls ▶ *Underline Video*

Rasmus T. Aavang, Rasmus Tjalk-Bøggild, Alexandre Iolov, Giovanni Rizzi, Mike Zhang and Johannes Bjerva

Earnings calls are a key source of financial information about public companies. However, extracting information from these calls is difficult. Unlike the templatic filings required by the U.S. Securities and Exchange Commission (SEC) to report a company's financial situation, earnings conference calls have no built-in labels, are unstructured, and feature conversational language. We explore this challenging domain by assessing the information captured by models trained on SEC filings and in-context learning methods. To establish a baseline, we first evaluate the generalization capabilities of SEC-trained models across established SEC datasets. To support our investigation, we introduce three novel benchmarks: (1) SEC Filings Benchmark (SECB), (2) Earnings Calls Benchmark (ECB), and ECB-A, a subset with 5,346 expert annotations to support our qualitative analysis. We find that encoder-based models struggle with the domain shift. Finally, we propose a system utilizing LLMs to perform open-ended extraction from unstructured call transcripts, verified by human evaluation (79.7% precision), providing a baseline for this valuable domain through the consistent tracking of emergent KPIs.

[Industry] EVE: A Domain-Specific LLM Framework for Earth Intelligence ▶ *Underline Video*

Àlex R. Atrio, Antonio Lopez, Jino Rohit, Yassine El Ouahidi, Marcello Politi, Vijayasri Iyer, Umar Jamil, Sébastien Bratières and Nicolas Longépé

We introduce Earth Virtual Expert (EVE), the first open-source, end-to-end initiative for developing and deploying domain-specialized LLMs for Earth Intelligence. At its core is EVE-Instruct, a domain-adapted 24B model built on Mistral Small 3.2 and optimized for reasoning and question answering. On newly constructed Earth Observation and Earth Sciences benchmarks, it outperforms comparable models while preserving general capabilities. We release curated training corpora and the first systematic domain-specific evaluation benchmarks, covering MCQA, open-ended QA, and factuality. EVE further integrates RAG and a hallucination-detection pipeline into a production system deployed via API and GUI, supporting 350 pilot users. All models, datasets, and code are publicly available.

[Industry] FinRAG-12B: A Production-Validated Recipe for Grounded Question Answering in Banking ▶ *Underline Video*

Denys Katerenchuk, Pablo Duboue and Keelan Evanini;

Large language models are increasingly used across various fields. However, their adoption in regulated high-stakes domains such as banking faces resistance due to the demand for high accuracy, compliance with regulations, and the need for traceable, grounded responses. In this work, we present a complete recipe for training and deploying grounded banking LLMs. First, we describe a high-quality data generation pipeline combining LLM-as-a-Judge filtering, citation annotation, and two-stage curriculum learning that achieves 73% citation quality with only 143M tokens. Second, we train a model to respond with "I don't know" when information is incomplete. We determine that a ratio of 22% unanswerable examples in training data yields calibrated refusal behavior. Third, we provide an end-to-end methodology, from data curation to quantization to serving, validated at 40+ financial institutions with real-world business outcomes. Our 12B model outperforms GPT-4.1 on answer quality and citation grounding while responding 3–5x faster.

[Industry] IndustryAssetEQA: A Neurosymbolic Operational Intelligence System for Embodied Question Answering in Industrial Asset Maintenance ▶ *Underline Video*

Chathurangi Shyalika, Dhaval Patel and Amit Sheth

Industrial maintenance environments increasingly rely on AI systems to assist operators in understanding asset behavior, diagnosing

failures, and evaluating interventions. Although large language models (LLMs) enable fluent natural-language interaction, deployed maintenance assistants routinely produce generic explanations that are weakly grounded in telemetry, omit verifiable provenance, and offer no testable support for counterfactual or action-oriented reasoning that undermine trust in safety-critical settings. We present IndustryAssetEQA, a neurosymbolic operational intelligence system that combines episode-centric telemetry representations with a Failure Mode and Effects Analysis Knowledge Graph (FMEA-KG) to enable Embodied Question Answering (EQA) over industrial assets. We evaluate on four datasets covering four industrial asset types, including rotating machinery, turbofan engines, hydraulic systems, and cyber-physical production systems. Compared to LLM-only baselines, IndustryAssetEQA improves structural validity by up to +0.51, counterfactual accuracy by up to +0.47, and explanation entailment by +0.64, while reducing severe expert-rated overclaims from 28% to 2% (~ 93%). Code, datasets, and the FMEA-KG are available at: https://anonymous.4open.science/r/EmbodiedEQA/

[Industry] MinerU2.5: A Decoupled Vision-Language Model for Efficient High-Resolution Document Parsing ▶ *Underline Video*

Junbo Niu, Zheng Liu, Zhuangcheng Gu, Bin Wang, Linke Ouyang, Zhiyuan Zhao, Tao Chu, Tianyao He, Fan Wu, Qintong Zhang, Zhenjiang Jin, Guang Liang, Rui Zhang, Wenzheng Zhang, Yuan Qu, Zhifei Ren, Yuefeng Sun, Zirui Tang, Boyu Niu, Yuanhong Zheng, Dongsheng Ma, Ziyang Miao, Hejun Dong, Siyi Qian, Junyuan Zhang, Fangdong Wang, Jingzhou Chen, Xiaomeng Zhao, Liqun Wei, Wei Li, Shasha Wang, RuiLiang Xu, Yuanyuan Cao, Lu Chen, Qianqian Wu, Huaiyu Gu, Lindong Lu, Dechen Lin, Shenguanlin Shenguanlin, Xuanhe Zhou, Linfeng Zhang, Yuhang Zang, Xiaoyi Dong, Jiaqi Wang, Bo Zhang, LEI BAI, Pei Chu, Weijia Li, Jiang Wu, Lijun Wu, Zhenxiang Li, Guangyu Wang, Zhongying Tu, Chao Xu, Kai Chen, Bowen Zhou, Dahua Lin, Wentao Zhang and Conghui He

We introduce MinerU2.5, a 1.2B-parameter document parsing vision-language model that achieves state-of-the-art recognition accuracy while maintaining exceptional computational efficiency. Our approach employs a coarse-to-fine, two-stage parsing strategy that decouples global layout analysis from local content recognition. In the first stage, the model performs efficient layout analysis on downsampled images to identify structural elements, circumventing the computational overhead of processing high-resolution inputs. In the second stage, guided by the global layout, it performs targeted content recognition on native-resolution crops extracted from the original image, preserving fine-grained details in dense text, complex formulas, and tables. To support this strategy, we developed a comprehensive data engine that generates diverse, large-scale training corpora for both pretraining and fine-tuning. Ultimately, MinerU2.5 demonstrates strong document parsing ability, achieving state-of-the-art performance on multiple benchmarks, surpassing both general-purpose and domain-specific models across various recognition tasks, while maintaining significantly lower computational overhead.

[Industry] Predicting Fine-Grained Typing Intent for Faster Search ▶ *Underline Video*

Yipeng Yu, Yichen Yuan, Chengxiao Feng and Xu Liu

"Is the users current query input exactly what they intend to search for?" Our work aims to answer this question by determining, at each typing, whether the current query is complete. If so, a search is implicitly triggered in advance without waiting for user confirmation. This approach reduces response time and enhances the user search experience. Specifically, we propose QType, a client-side framework that introduces innovation in data sampling, feature selection, model design and training, and online strategy. Experiments in a leading mobile shopping application validate its effectiveness, achieving offline precision/recall/accuracy of 0.7936/0.8196/0.7742, respectively, and decreasing online response time by 640.51+/-93.65 milliseconds, which is of great benefit to the search system. Unlike prior work that focuses on optimizing server-side engineering pipelines or simplifying ranking models, our method leverages client-side typing behavior for real-time early prediction, utilizing on-device computation to gain response time reducing. To the best of our knowledge, our work is the first to identify and address this problem. This work also introduces App Intelligence, a new paradigm for enhancing mobile applications by integrating on-device AI to boost business value and user experience.

[Industry] Robust Explanations for User Trust in Enterprise NLP Systems ▶ *Underline Video*

Guilin Zhang, Kai Zhao, Jeffrey Friedman, Xu Chu and Amine Anoun

Robust explanations are increasingly required for user trust in enterprise NLP, yet pre-deployment validation is difficult in the common case of black-box deployment (API-only access) where representation-based explainers are infeasible and existing studies provide limited guidance on whether explanations remain stable under real user noise, especially when organizations migrate from encoder classifiers to decoder LLMs. To close this gap, we propose a unified black-box robustness evaluation framework for token-level explanations based on leave-one-out occlusion, and operationalize explanation robustness with top-token flip rate under realistic perturbations (swap, deletion, shuffling, and back-translation) at multiple severity levels. Using this protocol, we conduct a systematic cross-architecture comparison across three benchmark datasets and six models spanning encoder and decoder families (BERT, RoBERTa, Qwen 7B/14B, Llama 8B/70B; 64,800 cases). We find that decoder LLMs produce substantially more stable explanations than encoder baselines (73% lower flip rates on average), and that stability improves with model scale (44% gain from 7B to 70B). Finally, we relate robustness improvements to inference cost, yielding a practical cost-robustness tradeoff curve that supports model and explanation selection prior to deployment in compliance-sensitive applications.

[Industry] TEN: Table Explicitization, Neurosymbolically ▶ *Underline Video*

Nikita Mehrotra, Aayush Kumar, Sumit Gulwani, Arjun Radhakrishna and Ashish Tiwari

We present TEN, a neurosymbolic approach for extracting tabular data from semistructured text such as copy-pasted content from PDFs, emails, or OCR-flattened outputs. This task poses real-world challenges in domains like finance and healthcare, where manual copy-paste into spreadsheets introduces errors and OCR distortions compromise data integrity, leading to financial losses and flawed decisions. Purely

neural methods suffer from hallucinations and structural inconsistencies, hindering deployment robustness. TEN addresses this via a novel triadic feedback loop that iteratively refines table hypotheses to enforce constraints and achieve verifiable convergence. Experiments show TEN outperforms neural baselines in exact match accuracy and lower hallucination rates. A 21-participant user study rates TEN tables more accurate and preferred in over 60% of pairwise comparisons, though verification and correction effort did not differ significantly between conditions.

[Industry] Think Like You Execute: Verifiable Chain of Thought from Program Traces ▶ *Underline Video*
Shailja Thakur, Vaibhav Saxena, Rohan Kulkarni, Shivdeep Singh, Parameswaran Selvam, Hiroshi Kanayama and Hima Patel
Teaching language models to reason about code execution is still an open problem. Current synthetic Chain-of-Thought (CoT) training data often consists of plausible-sounding explanations generated by teacher models, not verifiable accounts of actual program behavior. This causes models to learn logically flawed reasoning patterns despite syntactic correctness. We address this by grounding CoT generation directly in program execution traces. Our pipeline instruments code to capture dynamic behavior, narrates execution traces into natural language, and actively verifies each rationale against the trace. We systematically create 54,000 execution-verified, bi-directional rationales that teach models to reason both forward (input->output) and backward (output->input). Models fine-tuned on our verified data achieve substantial improvements, with a performance boost of +24.2 on LiveCodeBench-Exec, +22.3 on CruxEval-Output, and +21.1 on CruxEval-Input, demonstrating that verification quality directly determines both reasoning and code generation capabilities.

Oral Session C: Information Extraction and Retrieval 3

16:00-17:30 (Harbor D)

[Long] Adaptive Retrieval for Reasoning ▶ *Underline Video*
Jongho Kim, Jaeyoung Kim, Seung-won Hwang, Jihyuk Kim, Yu Jin Kim and Moontae Lee
We study leveraging adaptive retrieval to ensure sufficient "bridge" documents are retrieved for reasoning-intensive retrieval. Bridge documents are those that contribute to the reasoning process yet are not directly relevant to the initial query. While existing reasoning-based reranker pipelines attempt to surface these documents in ranking, they suffer from bounded recall. Naive solution with adaptive retrieval into these pipelines often leads to planning error propagation. To address this, we propose REPAIR, a framework that bridges this gap by repurposing reasoning plans as dense feedback signals for adaptive retrieval. Our key distinction is enabling mid-course correction during reranking through selective adaptive retrieval, retrieving documents that support the pivotal plan. Experimental results on reasoning-intensive retrieval and complex QA tasks demonstrate that our method outperforms existing baselines by 5.6%pt.

[Long] Beyond Markovian Forgetfulness: Episodic Memory for Reasoning-Intensive Retrieval ▶ *Underline Video*
Dohyeon Lee, Yeonseok Jeong and Seung-won Hwang
Reasoning-intensive information retrieval uses large language models to solve complex queries via multi-step reasoning. However, existing methods have critical limitations. Chain-of-Thought (CoT) approaches suffer from inefficiency, while state-based methods, despite better token efficiency, often fall into reasoning cycles that trap the query refinement process. To address these issues, we propose Episodic Memory for Retrieval (EMR), which enhances the state-based framework with an episodic memory. This module stores the full history of prior states for a query, allowing the model to avoid repetition of such cycles. Experiments on the BRIGHT benchmark show that EMR consistently outperforms both CoT and state-based baselines. Moreover, it is highly token-efficient, reducing token usage by 72% on average. Our results show that episodic memory is an effective and token-efficient mechanism for reasoning-intensive retrieval. The gains also generalize across different base models and stay efficient in terms of end-to-end latency. The code is available in https://github.com/ldilab/EMR.

[Short] BioHiCL: Hierarchical Multi-Label Contrastive Learning for Biomedical Retrieval with MeSH Labels ▶ *Underline Video*
Mengfei Lan, Lecheng Zheng and Halil Kilicoglu
Effective biomedical information retrieval requires modeling domain semantics and hierarchical relationships among biomedical texts. Existing biomedical generative retrievers built on coarse binary relevance signals, limiting their ability to capture semantic overlap. We propose BioHiCL - Biomedical Retrieval with Hierarchical Multi-Label Contrastive Learning, which leverages hierarchical MeSH annotations to provide structured supervision for multi-label contrastive learning. Our models, BioHiCL-Base (0.1B) and BioHiCL-Large (0.3B), achieve promising performance on biomedical retrieval, sentence similarity, and question answering tasks, while remaining computationally efficient for deployment.

[Long] Incorporating Temporal Coherence to Cross-Document Event Coreference Resolution ▶ *Underline Video*
Xinyu Chen, Peifeng Li and Qiaoming Zhu
Previous work on cross-document event coreference resolution (CDECR) primarily focused on enhancing semantic coherence between event mentions, largely overlooking the critical aspect of temporal coherence. To address this issue, we propose CohTP, a novel Temporal Cohorence-driven event coreference framework. CohTP explicitly models and enforces temporal constraints by first constructing a temporal event graph via a fine-tuned natural language inference (NLI) model. The graph is then refined using an Edge-Aware GNN to resolve

conflicts and partitioned into ordered time segments, where undirected edges group contemporaneous events. Event coreference resolution is subsequently performed within these temporally coherent segments, where event representations are further augmented with temporally consistent contexts. Experiments on the ECB+, GVC, WEC, and ECB+META datasets show that CohTP outperforms several state-of-the-art baselines.

[Long] R1-RE: Cross-Domain Relation Extraction with RLVR ▶ *Underline Video*
Runpeng Dai, Tong Zheng, Run Yang, Kaixian Yu and Hongtu Zhu
Relation extraction (RE) is a core task in natural language processing. Traditional approaches typically frame RE as a supervised learning problem, directly mapping context to labels, an approach that often suffers from poor out-of-domain (OOD) generalization. Inspired by the workflow of human annotators, we reframe RE as a reasoning task guided by annotation guidelines and introduce R1-RE, the first reinforcement learning with verifiable reward (RLVR) framework for RE tasks. Our method elicits the reasoning abilities of small language models for annotation tasks, resulting in significantly improved OOD robustness. We evaluate our approach on the public Sem-2010 dataset and a private MDKG dataset. The R1-RE-7B model attains an average OOD accuracy of approximately 70%, on par with leading proprietary models such as GPT-4o. Additionally, our comprehensive analysis provides novel insights into the training dynamics and emergent reasoning behaviors of the RLVR paradigm for RE.

[Long] Rethinking Reasoning-Intensive Retrieval: Evaluating and Advancing Retrievers in Agentic Search Systems ▶ *Underline Video*
Yilun Zhao, Jinbiao Wei, Tingyu Song, Siyue Zhang, Chen Zhao and Arman Cohan
Reasoning-intensive retrieval aims to surface evidence that maximizes downstream reasoning utility rather than only topical similarity. This capability is increasingly vital for agentic retriever-in-the-loop systems such as Deep-Research. However, existing retriever evaluation benchmarks, exemplified by Bright, provide narrow gold sets and evaluate retrievers in isolation, which obscures their value inside realistic agent workflows. We introduce Bright-Pro, an evaluation framework that assesses the effectiveness of retrievers in agentic search systems. Bright-Pro covers a broad range of queries across diverse professional domains. For each query, we provide expert-annotated reasoning aspects, positive documents, a reference response, and evaluation rubrics, enabling fine-grained assessment of retriever performance. Beyond static evaluation, we further assess retrievers in the context of agentic search systems, measuring their practical utility when serving as core components within agentic workflows. Using Bright-Pro, we evaluate classical lexical, general-purpose, and reasoning-intensive retrievers, providing actionable insights for future retriever development.

[Long] SemCSE-Multi: Multifaceted and Decodable Embeddings for Aspect-Specific and Interpretable Scientific Domain Mapping ▶ *Underline Video*
Marc Brinner and Sina ZarrieSS
We propose SemCSE-Multi, a novel unsupervised framework for generating multifaceted embeddings of scientific abstracts, evaluated in the domains of invasion biology and medicine. These embeddings capture distinct, individually specifiable aspects in isolation, thus enabling fine-grained and controllable similarity assessments as well as adaptive, user-driven visualizations of scientific domains. Our approach relies on an unsupervised procedure that produces aspect-specific summarizing sentences and trains embedding models to map semantically related summaries to nearby positions in the embedding space. We then distill these aspect-specific embedding capabilities into a unified embedding model that directly predicts multiple aspect embeddings from a scientific abstract in a single, efficient forward pass. In addition, we introduce an embedding decoding pipeline that decodes embeddings back into natural language descriptions of their associated aspects. Notably, we show that this decoding remains effective even for unoccupied regions in low-dimensional visualizations, thus offering vastly improved interpretability in user-centric settings.

[Long] Structured Semantic Information Helps Retrieve Better Examples for In-Context Learning Applied to Few-Shot Relation Extraction ▶ *Underline Video*
Aunabil Chakma, Mihai Surdeanu and Eduardo Blanco
This paper presents several strategies to automatically obtain additional examples for in-context learning, effectively transforming relation extraction from a 1-shot to a few-shot setting. Specifically, we introduce a novel strategy for example selection, in which new examples are selected based on the similarity of their underlying syntactic-semantic structure to the provided 1-shot example. We show that our strategy results in complementary word choices and sentence structures compared to LLM-generated examples. When both strategies are combined, the resulting hybrid system achieves a more holistic picture of the relations of interest than either method alone. Our framework transfers well across datasets (FS-TACRED and FS-FewRel) and LLM families (Qwen and Gemma). Overall, our hybrid system consistently outperforms alternative strategies achieving state-of-the-art performance on FS-TACRED and strong gains on a customized FewRel subset.

[Long] When Does Mixing Help? Analyzing Query Embedding Interpolation in Multilingual Dense Retrieval ▶ *Underline Video*
Tongyao Zhu, Chao-Ming Huang and Min-Yen Kan
While mixed-language querying is ubiquitous in multilingual communities, the sensitivity of dense retrievers to such queries remains poorly understood. We present a ratio-controlled study on mMARCO that systematically evaluates retrieval performance by varying the mixing proportion of parallel query translations via embedding-level mixing—constructing mixed queries as an interpolation of monolingual embeddings. Experiments with BGE-M3 demonstrate that an optimal mixing ratio outperforms the best monolingual endpoint in 88/105

cases. We uncover a distinct asymmetry driven by English dominance: mixing is uniformly beneficial when retrieving from non-English document indices, whereas indices containing English are best served by pure English queries. Furthermore, English acts as the strongest mixing partner for every non-English document language. Finally, when controlling for English dominance, mixing gains correlate negatively with typological distance. We conclude that language-mix sensitivity is structured and predictable, and we validate the robustness of these patterns across model families and scales.

Oral Session C: Linguistic theories, Cognitive Modeling and Psycholinguistics 1

16:00-17:30 (Harbor H-I)

[Long] A Dual-Task Paradigm to Investigate Sentence Comprehension Strategies in Language Models ▶ *Underline Video*

Rei Emura and Saku Sugawara

Language models (LMs) behave more like humans when their cognitive resources are restricted, particularly in predicting sentence processing costs such as reading times. However, it remains unclear whether such constraints similarly affect sentence comprehension strategies, and existing methods do not directly target the balance between memory storage and sentence processing, which is central to human working memory. To address this issue, we propose a dual-task paradigm that combines an arithmetic computation task with a sentence comprehension task, such as "The 2 cocktail + blended 3 =...". Our experiments show that under dual-task conditions, GPT-4o, o3-mini, and o4-mini shift toward plausibility-based comprehension, mirroring humans' rational inference. Specifically, these models show a greater accuracy gap between plausible sentences (e.g., "The cocktail was blended by the bartender") and implausible sentences (e.g., "The bartender was blended by the cocktail") in the dual-task condition compared to the single-task conditions. These findings suggest that constraints on the balance between memory and processing resources promote rational inference in LMs. More broadly, they support the view that human-like sentence comprehension fundamentally arises from the allocation of limited cognitive resources.

[Short] Enhancing Linguistic Competence of Language Models through Pre-training with Language Learning Tasks ▶ *Underline Video*

Atsuki Yamaguchi, Maggie Mi and Nikolaos Aletras

Language models (LMs) are pre-trained on raw text datasets to generate text sequences token-by-token. While this approach facilitates the learning of world knowledge and reasoning, it does not explicitly optimize for linguistic competence. To bridge this gap, we propose L2T, a pre-training framework integrating Language Learning Tasks alongside standard next-token prediction. Inspired by human language acquisition, L2T transforms raw text into structured input-output pairs to provide explicit linguistic stimulation. Pre-training LMs on a mixture of raw text and L2T data not only improves overall performance on linguistic competence benchmarks but accelerates their acquisition, while maintaining competitive performance on general reasoning tasks.

[Long] Function Words as Statistical Cues for Language Learning ▶ *Underline Video*

Xiulin Yang, Heidi Getz and Ethan Wilcox

What statistical properties might support learning abstract grammatical knowledge from linear input? We address this question by examining the statistical distribution of function words. Function words have been argued to aid acquisition through three distributional properties: high frequency, reliable syntactic association, and phrase-boundary alignment. We conduct a cross-linguistic corpus analysis of 186 languages, which confirms that all three properties are universal. Using counterfactual language modeling and ablation experiments on English, we show that preserving these properties facilitates acquisition in neural learners, with a Goldilocks effect: function words must be frequent enough to be reliable, yet diverse enough to remain informative to structural dependency. Probing analyses further reveal that different learning conditions produce systematically different reliance on function words.

[Long] Implicit Representations of Grammaticality in Language Models ▶ *Underline Video*

Yingshan Susan Wang, Linlu Qiu, Zhaofeng Wu, Roger P. Levy and Yoon Kim

Grammaticality and likelihood are distinct notions in human language. Pretrained language models (LMs), which are probabilistic models of language fitted to maximize corpus likelihood, generate grammatically well-formed text and discriminate well between grammatical and ungrammatical sentences in tightly controlled minimal pairs. However, their string probabilities do not sharply discriminate between grammatical and ungrammatical sentences overall. But do LMs implicitly acquire a grammaticality distinction distinct from string probability? We explore this question through studying pretrained LMs' internal representations, by training a linear probe on a dataset of grammatical and (synthetic) ungrammatical sentences obtained by applying perturbations to a naturalistic text corpus. We find that this simple grammaticality probe generalizes to human-curated grammaticality judgment benchmarks and outperforms LM probability-based grammaticality judgments. When applied to semantic plausibility benchmarks, however, in which both members of a minimal pair are grammatical and differ in only plausibility, the probe performs worse than string probability. The English-trained probe also exhibits nontrivial cross-lingual generalization, outperforming string probabilities on grammaticality benchmarks in numerous other languages. Additionally, probe scores correlate only weakly with string probabilities. These results collectively suggest that pretrained LMs acquire to some extent an implicit grammaticality distinction within their hidden layers.

[Long] Language Reconstruction with Brain Predictive Coding from fMRI Data ▶ *Underline Video*

Congchi Yin, Ziyi Ye and Piji Li

Many recent studies have shown that the perception of speech can be decoded from brain signals and subsequently reconstructed as continuous language. However, there is a lack of neurological basis for how the semantic information embedded within brain signals can be used more effectively to guide language reconstruction. Predictive coding theory suggests the human brain naturally engages in continuously predicting future words that span multiple timescales. This implies that the decoding of brain signals could potentially be associated with a predictable future. To explore the predictive coding theory within the context of language reconstruction, this paper proposes PredFT (FMRI-to-Text decoding with Predictive coding). PredFT consists of a main network and a side network. The side network obtains brain predictive representation from related regions of interest (ROIs) with a self-attention module. The representation is then fused into the main network for continuous language decoding. Experiments on two naturalistic language comprehension fMRI datasets show that PredFT outperforms current decoding models on several evaluation metrics.

[Long] On the Continued Value of Universal Dependencies in the Era of Large Language Models ▶ *Underline Video*

Wenxi Li and Jingyu Peng

The necessity of explicit linguistic representations has been increasingly questioned in the era of large language models (LLMs). In this work, we revisit this issue using Universal Dependencies (UD) as a case study, examining whether and in what ways this cross-lingual syntactic framework can still benefit contemporary LLMs. We focus on a cross-lingual adversarial paraphrase identification task that is designed to foreground the role of syntactic structure in semantic interpretation across languages. Within this setting, we systematically evaluate three strategies for integrating UD into LLMs: UD-Prompt, UD-Tuning, and UD-Attention. Our experiments show that, although the magnitude of gains depends on how UD-based structural priors interact with model behavior and cross-lingual variation, UD-augmented models consistently outperform their syntax-agnostic counterparts. Across strategies, we observe average accuracy improvements of 2.67%, 8.24%, and 2.53%, respectively. These findings demonstrate that linguistic knowledge remains informative for LLMs, offering practical value in cross-lingual settings where structural alignment is challenging.

[Long] Probing for Reading Times ▶ *Underline Video*

Eleftheria Tsipidi, Samuel Kiegeland, Francesco Ignazio Re, Tianyang Xu, Mario Giulianelli, Karolina Stanczak and Ryan Cotterell

Probing has shown that language model representations encode rich linguistic information, but it remains unclear whether they also capture cognitive signals about human processing. In this work, we probe language model representations for human reading times. Using regularized linear regression on two eye-tracking corpora spanning five languages (English, Greek, Hebrew, Russian, and Turkish), we compare the representations from every model layer against scalar predictors—surprisal, information value, and logit-lens surprisal. We find that the representations from early layers outperform surprisal in predicting early-pass measures such as first fixation and gaze duration. The concentration of predictive power in the early layers suggests that human-like processing signatures are captured by low-level structural or lexical representations, pointing to a functional alignment between model depth and the temporal stages of human reading. In contrast, for late-pass measures such as total reading time, scalar surprisal remains superior, despite its being a much more compressed representation. We also observe performance gains when using both surprisal and early-layer representations. Overall, we find that the best-performing predictor varies strongly depending on the language and eye-tracking measure.

[Long] SpidR-Adapt: A Universal Speech Representation Model for Few-Shot Adaptation ▶ *Underline Video*

Mahi Luthra, Jiayi Shen, Maxime Poli, Angelo Ortiz Tandazo, Yosuke Higuchi, Youssef Benchekroun, Martin Gleize, Charles-Eric Saint-James, Dongyan Lin, Phillip Rust, Angel Villar, Surya Parimi, Vanessa Stark, Rashel Moritz, Juan Pino, Yann LeCun and Emmanuel Dupoux

Human infants, with only a few hundred hours of speech exposure, acquire basic units of new languages, highlighting a striking efficiency gap compared to the data-hungry self-supervised speech models. To address this gap, this paper introduces SpidR-Adapt for rapid adaptation of speech units to new languages using minimal unlabeled data. We cast such low-resource speech representation learning as a meta-learning problem and construct a multi-task adaptive pre-training (MAdaPT) protocol which formulates the adaptation process as a bi-level optimization framework. To enable scalable meta-training under this framework, we propose a novel heuristic solution, first-order bi-level optimization (FOBLO), avoiding heavy computation costs. Finally, we stabilize meta-training by using a robust initialization through interleaved supervision which alternates self-supervised and supervised objectives. Empirically, SpidR-Adapt achieves rapid gains in phonemic discriminability (ABX) and downstream spoken language modeling scores (sWUGGY, sBLIMP, tSC), surpassing in-domain toplines after training on less than 1h of target-language audio and delivering 100x greater data efficiency than standard multi-task training.. These findings highlight a practical, architecture-agnostic path toward biologically inspired, data-efficient representations. We open-source the training code and model checkpoints at https://github.com/facebookresearch/spidr-adapt.

[Long] XToM: Exploring the Multilingual Theory of Mind for Large Language Models ▶ *Underline Video*

Chunkit Chan, Yauwai Yim, Hongchuan Zeng, Zhiying Zou, Xinyuan Cheng, Zhifan Sun, Zheye Deng, Kawai Chung, Yuzhuo Ao, Yixiang Fan, Cheng Jiayang, Ercong Nie, Ginny Y. Wong, Helmut Schmid, Hinrich Schütze, Simon See and Yangqiu Song

Theory of Mind (ToM), the ability to infer mental states in others, is pivotal for human social cognition. Existing evaluations of ToM in LLMs are largely limited to English, neglecting the linguistic diversity that shapes human cognition. This limitation raises a critical question:

can LLMs exhibit Multilingual Theory of Mind, the capacity to reason about mental states across diverse linguistic contexts? To address this gap, we present XToM, a rigorously validated multilingual benchmark that evaluates ToM across five languages and incorporates diverse, contextually rich task scenarios. Using XToM, we systematically evaluate LLMs (e.g., DeepSeek R1), revealing a pronounced dissonance: while models excel in multilingual language understanding, their ToM performance varies across languages. Our findings expose limitations in LLMs' ability to replicate human-like mentalizing across linguistic contexts.

Oral Session C: Machine Learning for NLP 1

16:00-17:30 (Regatta)

[Long] AdaJudge: Adaptive Multi-Perspective Judging for Reward Modeling ► *Underline Video*

Yongliang Miao, Yangyang Liang and Mengnan Du

Reward modeling is essential for aligning large language models with human preferences, yet predominant architectures rely on a static pooling strategy to condense sequences into scalar scores. This paradigm, however, suffers from two key limitations: a static inductive bias that misaligns with the task-dependent preference signals, and a representational mismatch, as the backbones optimization for generation leaves its representations ill-suited to fine-grained discrimination. To address this, we propose AdaJudge, a unified framework that jointly adapts representation and aggregation. AdaJudge first improves backbone representations into a discrimination-oriented space via gated refinement blocks. It then replaces the static readout with an adaptive multi-view pooling module, which dynamically routes and combines evidence. Extensive experiments on RM-Bench and JudgeBench show that AdaJudge outperforms strong off-the-shelf reward models and traditional pooling baselines.

[Long] Bridging the Memorization-Utilization Gap: Near-Lossless Context Compression via Reinforcement Learning ► *Underline Video*

Yujan Ting, Xu Tang, Terrence Chen and Weijing Huang

Despite recent progress in context compression, we identify a fundamental memorization-utilization gap where models can compress context with near-perfect fidelity yet fail to effectively utilize these compressed representations for downstream tasks. We address this with a holistic training paradigm spanning pretraining, instruction tuning, and reinforcement learning, built upon an average pooling compression. Our key innovation uses outcome-based RL to enable implicit expansion: the model learns to adaptively unfold task-relevant details during generation, interleaving reconstruction with reasoning. We achieve near-lossless 16x context compression (approx 5.3x decoder sequence-length reduction in our current implementation) across 7B and 32B models, recovering over 98% of full-context QA performance and outperforming prior methods by 11 points. Our 32B model demonstrates strong out-of-distribution and length generalization, robustly scaling to 120k-token contexts despite training on no more than 4k tokens, matching full-context performance on NIAH, LongBench v2, and multi-hop reasoning. We verify the implicit expansion behavior in experiments.

[Short] Evolutionary Strategies at Scale lead to Catastrophic Forgetting ► *Underline Video*

Immanuel Abdi, Akshat Gupta, Micah Mok, Alex Lu, Nicholas Lee and Gopala Anumanchipalli

One of biggest missing capabilities in state-of-the-art AI systems is the ability to learn continually after deployment. However, implementing an inference-time learning system has several challenges including the large memory requirement of gradient-based algorithms that are used to train state-of-the-art LLMs. Evolutionary Strategies (ES) have recently re-emerged as a gradient-free alternative to traditional learning algorithms and have shown encouraging performance on specific tasks in LLMs. In this paper, we perform a more comprehensive analysis of ES and specifically evaluate its forgetting curves when training for a larger number of update steps. We find that although ES is able to reach performance numbers closer to GRPO for math and reasoning tasks, it is accompanied by significant forgetting of prior abilities. We also show that the updates made using ES are much less sparse and have a larger l2 norm compared to corresponding GRPO updates, explaining the contrasting forgetting curves between the two algorithms. With this study, we aim to specifically highlight the issue of forgetting in gradient-free algorithms like ES and hope to inspire future work to mitigate these issues.

[Long] Learning While Staying Curious: Entropy-Preserving Supervised Fine-Tuning via Adaptive Self-Distillation for Large Reasoning Models ► *Underline Video*

Hao Wang, Hao Gu, Hongming Piao, Kaixiong Gong, Yuxiao Ye, Xiangyu Yue, Sirui Han, Yike Guo and Dapeng Wu

The standard post-training recipe for large reasoning models, supervised fine-tuning followed by reinforcement learning (SFT-then-RL), may limit the benefits of the RL stage: while SFT imitates expert demonstrations, it often causes overconfidence and reduces generation diversity, leaving RL with a narrowed solution space to explore. Adding entropy regularization during SFT is not a cure-all; it tends to flatten token distributions toward uniformity, increasing entropy without improving meaningful exploration capability. In this paper, we propose CurioSFT, an entropy-preserving SFT method designed to enhance exploration capabilities through intrinsic curiosity. It consists of (a) Self-Exploratory Distillation, which distills the model toward a self-generated, temperature-scaled teacher to encourage exploration within its capability; and (b) Entropy-Guided Temperature Selection, which adaptively adjusts distillation strength to mitigate knowledge forgetting by amplifying exploration at reasoning tokens while stabilizing factual tokens. Extensive experiments on

mathematical reasoning tasks demonstrate that, in SFT stage, CurioSFT outperforms the vanilla SFT by 2.5 points on in-distribution tasks and 2.9 points on out-of-distribution tasks. We also verify that exploration capabilities preserved during SFT successfully translate into concrete gains in RL stage, yielding an average improvement of 5.0 points. Code is available at https://github.com/HaoooWang/CurioSFT.

[Long] PARASITE: Conditional System Prompt Poisoning to Hijack LLMs ▶ *Underline Video*

Viet Pham and Thai Le

Large Language Models (LLMs) are increasingly deployed via third-party system prompts downloaded from public marketplaces. We identify a critical supply-chain vulnerability: conditional system prompt poisoning, where an adversary injects a sleeper agent into a benign-looking prompt. Unlike traditional jailbreaks that aim for broad refusal-breaking, our proposed framework, PARASITE, optimizes system prompts to trigger LLMs to output targeted, compromised responses only for specific queries (e.g., "Who should I vote for the US President?") while maintaining high utility on benign inputs. Operating in a strict black-box setting without model weight access, PARASITE utilizes a two-stage optimization including a global semantic search followed by a greedy lexical refinement. Tested on open-source models and commercial APIs (GPT-4o-mini, GPT-3.5), PARASITE achieves up to 70% F1 reduction on targeted queries with minimal degradation to general capabilities. We further demonstrate that these poisoned prompts evade standard defenses, including perplexity filters and typo-correction, by exploiting the natural noise found in real-world system prompts.

[Long] Query-Aware Knowledge Retrieval via Hyperbolic Structuring ▶ *Underline Video*

Chuang Zhou, Junnan Dong, Yilin Xiao, Shengyuan Chen, Su Dong, di yin, Xing Sun, Zhaozhuo Xu and Xiao Huang

Retrieval-Augmented Generation (RAG) has demonstrated significant potential in enhancing large language models (LLMs) by supplementing external knowledge. However, existing approaches focus primarily on retrieving isolated factual knowledge entities while neglecting the critical reasoning relationships. To address this limitation, Graph-Augmented Generation (GraphRAG) has emerged as an effective solution, which explicitly integrates structured knowledge graphs to support complex reasoning tasks. Although diverse graph construction methods have been explored, they typically rely on static, query-agnostic graphs constructed via fixed heuristics. We are thereby motivated to propose a query-centric retrieval framework that adaptively constructs a graph tailored to each query. However, it is challenging to accurately identify these latent relationships from queries to the corpus. Moreover, unifying multiple local-perspective connections into a globally coherent structured corpus introduces additional complexity. To this end, we introduce HyperRAG, a novel framework in the Hyperbolic space that captures both explicit entity-based links and implicit query-aware connections. Extensive experiments on three benchmark datasets demonstrate that HyperRAG consistently outperforms existing baselines.

[Long] Robertha: Eigenspectrum Regularized Attention for Robust Natural Language Understanding ▶ *Underline Video*

Andreia Podasca and Anup Das

We study asymmetric vulnerability to embedding corruption in encoder-based language models, where uniform perturbations disproportionately degrade low-magnitude embeddings compared to high-magnitude ones. Since critical words concentrate in low-norm space, this asymmetry causes catastrophic degradation of grammatical and semantic structure even under moderate corruption. Existing robustness approaches either sacrifice clean performance or fail to generalize to higher corruption levels. To address this problem, we propose Robertha, an attention mechanism built on Modern Hopfield Networks, in which semantic patterns act as stable states (attractors) that pull corrupted embeddings toward correct representations. We introduce iterative refinement for differential recovery: heavily corrupted embeddings require multiple convergence steps, while lightly corrupted embeddings converge quickly. To strengthen this mechanism, we introduce Eigenspectrum Regularization (ESR), which enforces low-rank key structures by controlling eigenvalue entropy, creating strong, well separated attractors with wide recovery basins. Across 13 GLUE and SuperGLUE tasks, Robertha significantly outperforms existing robustness methods while maintaining competitive clean performance.

[Long] TamEdit: Trajectory-Aware Meta-Learning for Specificity-Preserving Continual Knowledge Editing ▶ *Underline Video*

Shiqiang Tian, Cheng Ding, Qin Chen, Jie Zhou and Liang He

Knowledge editing is a promising method for updating Large Language Models efficiently. However, previous studies often suffer from poor specificity in continual editing, as they typically focus on single edits or preventing knowledge forgetting. To address this, we propose TamEdit, a trajectory-aware meta-learning method that preserves specificity for continual knowledge editing. TamEdit unifies three levels: Inner Optimization performs multi-step fast fine-tuning on the single edit; Trajectory-based Editing unifies continual edits with a growing memory; and Outer Optimization leverages meta-learning to distill cross-task strategies for preserving specificity. By capturing the relationships between different single edits within the trajectory, our method learns how to effectively avoid specificity drift. Experiments across multiple LLMs show TamEdit significantly outperforms baselines in continual editing, improving specificity by 14.81% with fast speed (requiring only 8.84% of the time cost of most baselines), while preserving general capabilities.

[Long] VOYAGER: A Training Free Approach for Generating Diverse Datasets using LLMs ▶ *Underline Video*

Avinash Amballa, Yashas Malur Saidutta, Chi-Heng Lin, Vivek Kulkarni and Srinivas Chappidi

Large language models (LLMs) are increasingly being used to generate synthetic datasets for the evaluation and training of downstream models. However, prior work has noted that such generated data lacks diversity. In this paper, we propose VOYAGER, a novel principled approach to generate diverse datasets. Our approach is iterative and directly optimizes a mathematical quantity that optimizes the diversity of the dataset using the machinery of determinantal point processes. Furthermore, our approach is training-free,

applicable to closed-source models, and scalable. In addition to providing theoretical justification for the working of our method, we also demonstrate through comprehensive experiments that VOYAGER improves diversity by 1.5-3 times compared to popular baseline approaches.

Oral Session C: Multilinguality and Language Diversity 3

16:00-17:30 (Harbor A)

[Long] Digitizing Nepals Written Heritage: A Comprehensive HTR Pipeline for Old Nepali Manuscripts ▶ *Underline Video*

Anjali Sarawgi, Esteban Garces Arias and Christof Zotter

This paper presents the first end-to-end pipeline for Handwritten Text Recognition (HTR) for Old Nepali, a historically significant but low-resource language. We adopt a line-level transcription approach and systematically explore encoder-decoder architectures and data-centric techniques to improve recognition accuracy. Our best model achieves a Character Error Rate (CER) of 4.9%. In addition, we implement and evaluate decoding strategies and analyze token-level confusions to better understand model behavior and error patterns. Although the evaluation dataset is confidential, we release our training code, model configurations, and evaluation scripts to support further research on HTR for low-resource historical scripts.

[Long] How Good is Your Wikipedia? Auditing Data Quality for Low-resource and Multilingual NLP ▶ *Underline Video*

Kushal Tatariya, Artur Kulmizev, Wessel Poelman, Esther Ploeger, Marcel Bollman, Johannes Bjerva, Jiaming Luo, Heather Lent and Miryam de Lhoneux

Wikipedia's perceived high quality and broad language coverage have established it as a fundamental resource in NLP. However, in recent years, such assumptions of high quality have become the subject of scrutiny in low-resource and multilingual contexts. In this study, we subject the entirety of non-English Wikipedia to a data filtering procedure typically reserved for noisy web-text , a process which removes a large percentage of the collection's data. In analysing the removed data, we reveal numerous systematic quality issues, such as script and language contamination, repeated template and placeholder articles, and a high concentration of bot-generated content. We consolidate these findings into a 4-level quality ranking of Wikipedia, which shows strong correspondence with alternative quality measures and heuristics. Lastly, we evaluate the downstream impact of quality filtering in three practical language modelling scenarios, showing that models trained on filtered data largely match or outperform those trained on raw Wikipedia, with the largest gains observed for lower-quality language editions. Ultimately, our experiments serve as a first step in establishing quality-aware best practices for Wikipedia utilization in NLP, laying groundwork that can inform future dataset creation and curation efforts.

[Long] Massively Multilingual Joint Segmentation and Glossing ▶ *Underline Video*

Michael Ginn, Lindia Tjuatja, Enora Rice, Ali Marashian, Maria Valentini, Jasmine Xu, Graham Neubig and Alexis Palmer

Automated interlinear gloss prediction with neural networks is a promising approach to accelerate language documentation efforts. However, while state-of-the-art models like GlossLM (Ginn et al., 2024) achieve high scores on glossing benchmarks, user studies with linguists have found critical barriers to the usefulness of such models in real-world scenarios (Rice et al., 2025). In particular, existing models typically generate morpheme-level glosses but assign them to whole words without predicting the actual morpheme boundaries, making the predictions less interpretable and thus untrustworthy to human annotators. We conduct the first study on neural models that jointly predict interlinear glosses and the corresponding morphological segmentation from raw text. We run experiments to determine the optimal way to train models that balance segmentation and glossing accuracy, as well as the alignment between the two tasks. We extend the training corpus of GlossLM and pretrain PolyGloss, a family of seq2seq multilingual models for joint segmentation and glossing that outperforms GlossLM on glossing and beats various open-source LLMs on segmentation, glossing, and alignment. In addition, we demonstrate that PolyGloss can be quickly adapted to a new dataset via low-rank adaptation.

[TACL] MERLIN: A Testbed for Multilingual Multimodal Entity Recognition and Linking ▶ *Underline Video*

Sathyanarayanan Ramamoorthy, Vishwa Shah, Simran Khanuja, Zaid Sheikh, Shan Jie, Ann Chia, Shearman Chua and Graham Neubig

This paper introduces MERLIN, a novel testbed system for the task of Multilingual Multimodal Entity Linking. The created dataset includes BBC news articles, paired with corresponding images, in five languages: Hindi, Japanese, Indonesian, Vietnamese, and Tamil, featuring over 7,000 named entity mentions linked to 2,500 unique Wikidata entities. We also include several benchmarks using multilingual and multimodal entity linking methods exploring different language models like LLaMa-2 and Aya-23. Our findings indicate that incorporating visual data improves the accuracy of entity linking, especially for entities where the textual context is ambiguous or insufficient, and particularly for models that do not have strong multilingual abilities. The dataset, annotation procedures, methods and other relevant materials will be made available on our GitHub repository.

[Long] Multilingual and Cross-Lingual Citation Needed Detection on Wikipedia for Lower-Resource Languages ▶ *Underline Video*

Gerrit Quaremba, Amy Rechkemmer, Elizabeth Black, Denny Vrandecic and Elena Simperl

In automated fact-checking (AFC), check-worthiness detection identifies claims requiring verification based on domain-specific criteria. On Wikipedia, this task instantiates as Citation Needed Detection (CND), which flags claims lacking supporting citations. However,

existing research has largely overlooked lower-resource languages, and recent AFC pipelines rely on large language models (LLMs), which are inaccessible to low-resource organizations. We introduce MCN, a multilingual CND corpus spanning 18 languages across three resource levels, on which we conduct an extensive study of small decoder-based language models (SLMs). Our experiments show that SLMs fine-tuned with an encoder-style objective substantially outperform prompted LLMs across languages. We further present one of the first studies on cross-lingual CND, demonstrating that SLMs fine-tuned solely on English claims surpass LLMs, even with little to no target-language adaptation. Our findings have important implications for lower-resource Wikipedia communities and suggest that compact, task-specific models are preferable to LLMs for CND.

[TACL] On the Limitations of Language Targeted Pruning: Investigating the Calibration Language Impact in Multilingual LLM Pruning ▶ *Underline Video*

Simon Kurz, Jian-Jia Chen, Lucie Flek and Zhixue Zhao

Recent advances in large language model (LLM) pruning have shown state-of-the-art (SotA) compression results in post-training and retraining-free settings while maintaining high predictive performance. However, previous research mainly considered calibrating based on English text, despite the multilingual nature of modern LLMs and their frequent use in non-English languages. This analysis paper conducts an in-depth investigation of the performance and internal representation changes associated with pruning multilingual language models for monolingual applications. We present the first comprehensive empirical study, comparing different calibration languages for pruning multilingual models across diverse languages, tasks, models, and SotA pruning techniques. We further analyze the latent subspaces, pruning masks, and individual neurons within pruned models. Our results reveal that while calibration on the target language effectively retains perplexity and yields high signal-to-noise ratios, it does not consistently improve downstream task performance. Further analysis of internal representations at three different levels highlights broader limitations of current pruning approaches: While they effectively preserve dominant information like language-specific features, this is insufficient to counteract the loss of nuanced, language-agnostic features that are crucial for knowledge retention and reasoning.

[Long] Parity-Aware Byte-Pair Encoding: Improving Cross-lingual Fairness in Tokenization ▶ *Underline Video*

Negar Foroutan, Clara Meister, Debjit Paul, Joel Niklaus, Sina Ahmadi, Antoine Bosselut and Rico Sennrich

Tokenization is the first—and often least scrutinized—step of most NLP pipelines. Standard algorithms for learning tokenizers rely on frequency-based objectives, which favor languages dominant in the training data and consequently leave lower-resource languages with tokenizations that are disproportionately longer, morphologically implausible, or even riddled with <UNK> placeholders. This phenomenon ultimately amplifies computational and financial inequalities between users from different language backgrounds. To remedy this, we introduce Parityaware Byte Pair Encoding (BPE), a variant of the widely-used BPE algorithm. At every merge step, Parityaware BPE applies a fair-max rule that maximizes the compression gain of the currently worstcompressed language, trading a small amount of global compression for crosslingual parity. We find empirically that Parity-aware BPE reduces tokenization inequality—operationalized by the Gini coefficient of per-language token costs—by up to 89% relative to Classical BPE. This comes with negligible impact on global compression rate and no evidence of systematic degradation in downstream LM performance.

[Long] SwissGov-RSD: A Human-annotated, Cross-lingual Benchmark for Token-level Recognition of Semantic Differences Between Related Documents ▶ *Underline Video*

Michelle Wastl, Jannis Vamvas and Rico Sennrich

Recognizing semantic differences across documents is crucial for text generation evaluation and content alignment, especially in cross-lingual settings. However, as a standalone task, it has received little attention. We address this by introducing SwissGov-RSD, the first naturalistic, document-level, cross-lingual dataset for semantic difference recognition. It encompasses a total of 224 multi-parallel documents in English–German, English–French, and English–Italian with token-level difference annotations by human annotators. We evaluate a variety of open-source and closed-source large language models as well as encoder models across different fine-tuning settings on this new benchmark. Our results show that current automatic approaches perform poorly compared to their performance on monolingual, sentence-level, and synthetic benchmarks, revealing a considerable gap for both LLMs and encoder models.

[Long] To Lie or Not to Lie? Investigating The Biased Spread of Global Lies by LLMs ▶ *Underline Video*

Zohaib Khan, Mustafa Dogan, Ifeoma Okoh, Pouya Sadeghi, Siddhartha Shrestha, Sergius Justus Nyah, Mahmoud O. Mokhiamar, Michael J. Ryan and Tarek Naous

Misinformation is on the rise, and the strong writing capabilities of LLMs lower the barrier for malicious actors to produce and disseminate false information. We study how LLMs behave when prompted to spread misinformation across languages and target countries, and introduce GlobalLies, a multilingual parallel dataset of 440 misinformation generation prompt templates and 6,867 entities, spanning 8 languages and 195 countries. Using both human annotations and large-scale LLM-as-a-judge evaluations across hundreds of thousands of generations from state-of-the-art models, we show that misinformation generation varies systematically based on the country being discussed. Propagation of lies by LLMs is substantially higher in many lower-resource languages and for countries with a lower Human Development Index (HDI). We find that existing mitigation strategies provide uneven protection: input safety classifiers exhibit cross-lingual gaps, and retrieval-augmented fact-checking remains inconsistent across regions due to unequal information availability. We release GlobalLies for research purposes, aiming to support the development of mitigation strategies to reduce the spread of global misinformation: https://github.com/zohaib-khan5040/globallies

Oral Session C: NLP Applications 2

16:00-17:30 (Harbor E-F)

[Long] A Comprehensive Survey of Process Reward Models: Data Generation, Model Construction, and Usage ► *Underline Video*

Congmin Zheng, Jiachen Zhu, Zhuoying Ou, Yuxiang Chen, Kangning Zhang, Rong Shan, Zeyu Zheng, Mengyue Yang, Jianghao Lin, Yong Yu and Weinan Zhang

Large Language Models (LLMs) have advanced reasoning ability, yet conventional alignment remains dominated by outcome reward models (ORMs) that judge only final answers. Process Reward Models(PRMs) address this gap by evaluating and guiding reasoning at the step or trajectory level. This survey provides a systematic overview of PRMs through the full loop: how to generate process data, build PRMs, and use PRMs for test-time scaling and reinforcement learning. We summarize applications across math, code, text, multimodal reasoning, robotics, and agents, and review emerging benchmarks. Our goal is to clarify design spaces, reveal open challenges, and guide future research toward fine-grained, robust reasoning alignment.

[Long] Beyond the Crowd: LLM-Augmented Community Notes for Governing Health Misinformation ► *Underline Video*

Jiaying Wu, Zihang Fu, Haonan Wang, Fanxiao Li, Jiafeng Guo, Preslav Nakov and Min-Yen Kan

Community Notes, the crowd-sourced misinformation governance system on X (formerly Twitter), allows users to flag misleading posts, attach contextual notes, and rate the notes' helpfulness. However, our empirical analysis of 30.8K health-related notes reveals substantial latency, with a median delay of 17.6 hours before notes receive a helpfulness status. To improve responsiveness during real-world misinformation surges, we propose CrowdNotes+, a unified LLM-based framework that augments Community Notes for faster and more reliable health misinformation governance. CrowdNotes+ integrates two modes: (1) evidence-grounded note augmentation and (2) utility-guided note automation, supported by a hierarchical three-stage evaluation of relevance, correctness, and helpfulness. We instantiate the framework with HealthNotes, a benchmark of 1.2K health notes annotated for helpfulness, and a fine-tuned helpfulness judge. Our analysis first uncovers a key loophole in current crowd-sourced governance: voters frequently conflate stylistic fluency with factual accuracy. Addressing this via our hierarchical evaluation, experiments across 15 representative LLMs demonstrate that CrowdNotes+ significantly outperforms human contributors in note correctness, helpfulness, and evidence utility.

[Long] CodeEvo: Interaction-Driven Synthesis of Code-centric Data through Hybrid and Iterative Feedback ► *Underline Video*

Qiushi Sun, Jingyang Gong, Lei Li, Qipeng Guo and Fei Yuan

Acquiring high-quality instruction-code pairs is essential for training Large Language Models for code generation. While automated synthesis has emerged as an alternative to expensive manual curation, current approaches often rely on rigid heuristics, yielding data that is ungrounded or lacks logical complexity. We propose CodeEvo, a dual-agent architecture comprising a Coder for iterative solution synthesis and a Reviewer to orchestrate the generation trajectory. To transcend the limitations of existing heuristics, the Reviewer formulates a Schema to systematically architect logic and complexity through an interleaved synthesis of instructions and code. This process is further reinforced by a hybrid verification protocol synergizing deterministic compiler feedback with semantic evaluation. Under this framework, we construct CodeEvo-100K, a large-scale dataset of instruction-code pairs with stepped difficulty levels. Extensive experiments demonstrate that models fine-tuned on CodeEvo data significantly outperform established baselines across code generation benchmarks. In-depth analyses further provide insights into effective code-centric data synthesis. Code and data are available at https://github.com/QiushiSun/CodeEvo.

[Long] Dual-Cluster Memory Agent: Resolving Multi-Paradigm Ambiguity in Optimization Problem Solving ► *Underline Video*

Xinyu Zhang, Yuchen Wan, Boxuan Zhang, Zesheng Yang, Lingling Zhang, Bifan Wei and Jun Liu

Large Language Models (LLMs) often struggle with structural ambiguity in optimization problems, where a single problem admits multiple related but conflicting modeling paradigms, hindering effective solution generation. To address this, we propose Dual-Cluster Memory Agent (DCM-Agent) to enhance performance by leveraging historical solutions in a training-free manner. Central to this is Dual-Cluster Memory Construction. This agent assigns historical solutions to modeling and coding clusters, then distills each cluster's content into three structured types: Approach, Checklist, and Pitfall. This process derives generalizable guidance knowledge. Furthermore, this agent introduces Memory-augmented Inference to dynamically navigate solution paths, detect and repair errors, and adaptively switch reasoning paths with structured knowledge. The experiments across seven optimization benchmarks demonstrate that DCM-Agent achieves an average performance improvement of 11%- 21%. Notably, our analysis reveals a "knowledge inheritance" phenomenon: memory constructed by larger models can guide smaller models toward superior performance, highlighting the framework's scalability and efficiency.

[Long] Evaluating Legal Reasoning Traces with Legal Issue Tree Rubrics ► *Underline Video*

Jinu Lee, Kyoung-Woon On, Simeng Han, Arman Cohan and Julia Hockenmaier

Evaluating the quality of LLM-generated reasoning traces in expert domains (e.g., law) is essential for ensuring credibility and explainability, yet remains challenging due to the inherent complexity of such reasoning tasks. We introduce LEGIT (LEGal Issue Trees), a novel large-scale (24K instances) expert-level legal reasoning dataset with an emphasis on reasoning trace evaluation. We convert court judgments

into hierarchical trees of opposing parties' arguments and the court's conclusions, which serve as rubrics for evaluating the issue coverage and correctness of the reasoning traces. We verify the reliability of these rubrics via human expert annotations and comparison with coarse, less informative rubrics. Using the LEGIT dataset, we show that (1) LLMs' legal reasoning ability is seriously affected by both legal issue coverage and correctness, and that (2) retrieval-augmented generation (RAG) and RL with rubrics bring complementary benefits for legal reasoning abilities, where RAG improves overall reasoning capability, whereas RL improves correctness albeit with reduced coverage.

[Long] LePREC: Reasoning as Classification over Structured Factors for Assessing Relevance of Legal Issues ▶ *Underline Video*
Fanyu Wang, Xiaoxi Kang, Paul Burgess, aashish srivastava, Chetan Arora, Adnan Trakic, Lay-Ki Soon, Md Khalid Hossain and Lizhen Qu
More than half of the global population struggles to meet their civil justice needs due to limited legal resources. While Large Language Models (LLMs) have demonstrated impressive reasoning capabilities, significant challenges remain even at the foundational step of legal issue identification. To investigate LLMs' capabilities in this task, we constructed a dataset from 769 real-world Malaysian Contract Act court cases, using GPT-4o to extract facts and generate candidate legal issues, annotated by senior legal experts, which reveals a critical limitation: while LLMs generate diverse issue candidates, their precision remains inadequate (GPT-4o achieves only 62%). To address this gap, we propose LePREC (Legal Professional-inspired Reasoning Elicitation and Classification), a neuro-symbolic framework combining neural generation with structured statistical reasoning. LePREC consists of: (1) a neuro component leverages LLMs to transform legal descriptions into question-answer pairs representing diverse analytical factors, and (2) a symbolic component applies sparse linear models over these discrete features, learning explicit algebraic weights that identify the most informative reasoning factors. Unlike end-to-end neural approaches, LePREC achieves interpretability through transparent feature weighting while maintaining data efficiency through correlation-based statistical classification. Experiments show a 30-40% improvement over advanced LLM baselines, including GPT-4o and Claude, confirming that correlation-based factor-issue analysis offers a more data-efficient solution for relevance decisions.

[Long] SHAPE: Unifying Safety, Helpfulness and Pedagogy for Educational LLMs ▶ *Underline Video*
Sihang Zhao, Kangrui Yu, Youliang Yuan, Pinjia He and Hongyi Wen
Large Language Models (LLMs) have been widely explored in educational scenarios. We identify a critical vulnerability in current educational LLMs, pedagogical jailbreaks, where students use answer-inducing prompts to elicit solutions rather than scaffolded instructions. To enable systematic study, we unify and formalize safe, helpful, and pedagogical behaviors with a knowledge-mastery graph and introduce SHAPE, a benchmark of 9,087 student-question pairs for evaluating tutoring behavior under adversarial pressure. We propose a graph-augmented tutoring pipeline that infers prerequisite concepts from queries, identifies mastery gaps, and routes generation between instructing and problem-solving via explicit gating. Experiments across multiple LLMs show that our method yields significantly improved safety under two pedagogical jailbreak settings, while maintaining near-ceiling helpfulness under the same evaluation protocol. Our code and data are available at https://github.com/MAPS-research/SHaPE

[Long] Spec-o3: A Tool-Augmented Vision-Language Agent for Rare Celestial Object Candidate Vetting via Automated Spectral Inspection ▶ *Underline Video*
Minghui Jia, Qichao Zhang, Ali Luo, Linjing Li, Shuo Ye, Hailing Lu, Wen Hou and Dongbin Zhao
Due to the limited generalization and interpretability of deep learning classifiers, the final vetting of rare celestial object candidates still relies on manually intensive expert visual inspection, which has become a primary bottleneck as modern spectroscopic surveys continue to scale. To bridge this gap, we propose Spec-o3, a tool-augmented vision-language agent that performs astronomer-aligned spectral inspection via interleaved multimodal chain-of-thought reasoning. Spec-o3 is trained with a two-stage post-training recipe: cold-start supervised fine-tuning on expert inspection trajectories followed by outcome-based reinforcement learning on rare-type verification tasks. Evaluated on five rare-object identification tasks from LAMOST, Spec-o3 establishes a new State-of-the-Art, boosting the macro-F1 score from 28.3 to 76.5 with a 7B parameter base model and outperforming both proprietary VLMs and specialized deep models. Beyond accuracy, Spec-o3 processes spectra at 0.2\{\},s per sample on an 8xH100 server, a 50x throughput gain over expert manual inspection. The agent also demonstrates strong generalization to unseen inspection tasks across survey shifts (from LAMOST to SDSS/DESI). Expert evaluations further confirm that its reasoning traces are coherent and physically consistent, supporting transparent and trustworthy decision-making. Code, data, and models are available at Project HomePage.

[Long] What Makes an Ideal Quote? Recommending Unexpected yet Rational Quotations via Novelty ▶ *Underline Video*
Powei Chang, Jin Xiao, Guanglei Yue, Qianyu He, Yanghua Xiao, Deqing Yang and Jiaqing Liang
Quotation recommendation enriches writing by suggesting quotations that fit a given context, but prior systems largely focus on topical relevance and overlook what makes quotes memorable. Based on a user study, we find that preferred quotations are often unexpected yet rational, motivating the goal of selecting quotes that are contextually novel while semantically coherent. We propose NovelQR, which (1) uses a generative label agent to map quotations and contexts into multi-dimensional deep-meaning labels for label-enhanced retrieval, and (2) reranks candidates with a token-level novelty estimator that mitigates auto-regressive continuation bias. Experiments on bilingual datasets across diverse domains show that NovelQR is preferred by human judges and improves overall recommendation quality over strong baselines, while achieving competitive novelty estimation.

Oral Session C: Resources and Evaluation 3

16:00-17:30 (Harbor B-C)

[Long] Bridging Language and Items for Retrieval and Recommendation: Benchmarking LLMs as Semantic Encoders ► *Underline Video*

Yupeng Hou, Jiacheng Li, Xiangjun Fu, Zhankui He, An Yan, Xiusi Chen and Julian McAuley

Feature engineering has long been central to recommender systems, yet effectively leveraging textual item features remains challenging. Recent advances in large language models (LLMs) have enabled their use as semantic encoders for recommendation, but their roles and behaviors in this setting are still not well understood. Prior studies often rely on general-purpose embedding benchmarks (e.g., MTEB) when selecting LLMs, overlooking the unique characteristics of recommendation tasks. To address this gap, we introduce BLaIR, a comprehensive benchmark for evaluating LLMs as semantic encoders in recommendation scenarios. We contribute (1) a new large-scale Amazon Reviews 2023 dataset with over 570 million reviews and 48 million items, (2) a unified benchmark covering sequential recommendation, collaborative filtering, and product search, and (3) a new complex-query product search task featuring both semi-synthetic and real-world evaluation datasets. Experiments with 11 leading LLMs show that their rankings on BLaIR show little correlation with MTEB, highlighting the unique challenges of semantic encoding in recommendation.

[Long] CiteGuard: Faithful Citation Attribution for LLMs via Retrieval-Augmented Validation ► *Underline Video*

Yee Man Choi, Xuehang Guo, Yi R. Fung and Qingyun Wang

Large Language Models (LLMs) have emerged as powerful assistants for scientific writing. However, concerns remain about the quality and reliability of the generated text, including citation accuracy and faithfulness. While most recent work relies on methods such as LLM-as-a-Judge, the reliability of LLM-as-a-Judge alone is also in doubt. In this work, we reframe citation evaluation as a problem of citation attribution alignment, which assesses whether LLM-generated citations match those a human author would include for the same text. We propose CiteGuard, a retrieval-aware agent framework designed to provide more faithful grounding for citation validation. CiteGuard improves over the prior baseline by 10 percentage points and achieves up to 68.1% accuracy on the CiteME benchmark, approaching human performance (69.2%). It also identifies alternative valid citations and demonstrates generalization ability for cross-domain citation attribution.

[Long] Common to Whom? Regional Cultural Commonsense and LLM Bias in India ► *Underline Video*

Sangmitra Madhusudan, Trush Shashank More, Steph Buongiorno, Renata Dividino, Jad Kabbara and Ali Emami

Existing cultural commonsense benchmarks treat nations as monolithic, assuming uniform practices within national boundaries. But does cultural commonsense hold uniformly within a nation, or does it vary at the sub-national level? We introduce Indica, the first benchmark designed to test LLMs' ability to address this question, focusing on India, a nation of 28 states, 8 union territories, and 22 official languages. We collect human-annotated answers from five Indian regions (North, South, East, West, and Central) across 515 questions spanning 8 domains of everyday life, yielding 1,630 region-specific question-answer pairs. Strikingly, only 39.4% of questions elicit agreement across all five regions, demonstrating that cultural commonsense in India is predominantly regional, not national. We evaluate eight state-of-the-art LLMs and find two critical gaps: models achieve only 13.4%-20.9% accuracy on region-specific questions, and they exhibit geographic bias, over-selecting Central and North India as the "default" (selected 30-40% more often than expected) while under-representing East and West. Beyond India, our methodology provides a generalizable framework for evaluating cultural commonsense in any culturally heterogeneous nation, from question design grounded in anthropological taxonomy, to regional data collection, to bias measurement.

[Long] KG-MuLQA: A Framework for KG-based Multi-Level QA Extraction and Long-Context LLM Evaluation ► *Underline Video*

Nikita Tatarinov, Vidhyakshaya Kannan, Haricharana Srinivasa, Arnav Raj, Harpreet Singh Anand, Varun Singh, Aditya Luthra, Ravij Lade, Agam Shah and Sudheer Chava

We introduce KG-MuLQA (Knowledge-Graph-based Multi-Level Question-Answer Extraction): a framework that (1) extracts QA pairs at multiple complexity levels (2) along three key dimensions – multi-hop retrieval, set operations, and answer plurality, (3) by leveraging knowledge-graph-based document representations. This approach enables fine-grained assessment of model performance across controlled difficulty levels. Using this framework, we construct a dataset of 20,139 QA pairs based on financial credit agreements and evaluate 16 proprietary and open-weight Large Language Models, observing that even the best-performing models struggle with set-based comparisons and multi-hop reasoning over long contexts. Our analysis reveals systematic failure modes tied to semantic misinterpretation and inability to handle implicit relations.

[Long] MSEarth: A Multimodal Benchmark for Earth Science Phenomenon Discovery with MLLMs ► *Underline Video*

Xiangyu Zhao, Wanghan Xu, Bo LIU, Yuhao Zhou, Fenghua Ling, Ben Fei, Xiaoyu Yue, LEI BAI, Wenlong Zhang and Xiao-Ming Wu

The rapid advancement of multimodal large language models (MLLMs) offers new opportunities for complex scientific challenges, yet their application in earth science, especially at the graduate level, remains underexplored due to a lack of benchmarks reflecting the depth and complexity of geoscientific reasoning. Existing datasets often rely on synthetic data or simple figure-caption pairs, failing to capture the nuanced reasoning required for real-world applications. To address this, we introduce MSEarth, a multimodal scientific dataset and benchmark curated from high-quality, open-access publications. Covering the five major spheres of Earth science, atmosphere, cryosphere, hydrosphere, lithosphere, and biosphere, MSEarth features over 289K figures with refined captions enriched by contextual discussions

and reasoning from the original papers. The benchmark supports tasks such as scientific figure captioning, multiple choice questions, and open-ended reasoning, providing a scalable, high-fidelity resource for developing and evaluating MLLMs in scientific reasoning.

[Long] Refining and Reusing Annotation Guidelines for LLM Annotation ▶ *Underline Video*

Kon Woo Kim, Jin-Dong Kim and Akiko Aizawa

While Large Language Models (LLMs) demonstrates remarkable zero-shot annotation tasks, they often struggle with the specialized conventions of gold-standard benchmarks. We propose the systematic reuse and refinement of annotation guidelines as an alignment mechanism, introducing an iterative moderation framework that simulates the early phases of annotation projects. We evaluate three hypotheses: (1) the efficacy of guideline integration, (2) the advantage of reasoning-optimized models, and (3) the viability of moderation under minimal supervision. Testing across biomedical NER tasks (NCBI Disease, BC5CDR, BioRED) with three LLM families (GPT, Gemini, DeepSeek), our results empirically confirm all three hypotheses. While the iterative moderation framework shows a good potential in effectively refining guidelines, our analysis also reveals a significant room for improvement.

[Long] Revisiting a Pain in the Neck: A Semantic Reasoning Benchmark for Language Models ▶ *Underline Video*

Yang Liu, Hongming Li, Melissa Xiaohui Qin, Chao Huang and Qiankun Liu

We present SemanticQA, an evaluation suite designed to assess language models (LMs) in semantic phrase processing tasks. The benchmark consolidates existing multiword expression (MwE) resources and reorganizes them into a unified testbed. It covers both general lexical phenomena, such as lexical collocations, and three fine-grained categories: idiomatic expressions, noun compounds, and verbal constructions. Through SemanticQA, we assess LMs of diverse architectures and scales in extraction, classification, and interpretation tasks, as well as sequential task compositions. We reveal substantial performance variation, particularly on tasks requiring semantic reasoning, highlighting differences in reasoning efficacy and semantic understanding of LMs, providing insights for pushing LMs with stronger comprehension on non-trivial semantic phrases. The evaluation harness and data of SemanticQA are available at https://github.com/jacklanda/SemanticQA.

[Long] SCAN: Structured Capability Assessment and Navigation for LLMs ▶ *Underline Video*

Zongqi Wang, Tianle Gu, Chen Gong, Xin Tian, Siqi Bao and Yujiu Yang

Evaluating Large Language Models (LLMs) has become increasingly important, with automatic evaluation benchmarks gaining prominence as alternatives to human evaluation. While existing research has focused on approximating model rankings, such benchmarks fail to provide users and developers with a comprehensive and fine-grained understanding of a specific model's capabilities. To fill this gap, we propose SCAN (Structured Capability Assessment and Navigation), a practical framework that enables detailed characterization of LLM capabilities through comprehensive and fine-grained evaluation. SCAN incorporates four key components: (1) TaxBuilder, which extracts capability-indicating tags from extensive queries to construct a hierarchical taxonomy automatically; (2) RealMix, a query synthesis and filtering mechanism that ensures sufficient evaluation data for each capability tag; (3) a suite of visualization and analysis tools that facilitate efficient navigation and analysis of model capabilities; and (4) a PC^2-based (Pre-Comparison-derived Criteria) LLM-as-a-Judge approach that achieves significantly higher accuracy compared to classic LLM-as-a-Judge method. Using SCAN, we conduct a comprehensive evaluation of 21 mainstream LLMs. Our detailed analysis of the GPT-OSS family reveals substantial performance variations, even within sub-capabilities belonging to the same category of capability. This finding highlights the importance of fine-grained evaluation in accurately understanding LLM behavior. Project homepage and resources are available at https://github.com/liudan193/SCAN.

[Long] VALU: A Benchmark for Video Anomaly Temporal Localization and Understanding at Multiple Semantic Levels ▶ *Underline Video*

Yixiao He, Menghao Zhang, Haifeng Sun, Jing Wang, Kangheng Lin, Jinghan Wang, Chenye Xu, Pengfei Ren, Qi Qi and Jingyu Wang

Video anomaly understanding (VAU) is critical for real-world scenarios. Recent advances in Video Large Language Models (Video-LLMs) enhance the ability of VAU models to describe and interpret anomalies. However, progress in anomaly localization is still limited by two key issues. First, most existing video anomaly datasets only annotate segments that are clearly inconsistent with the context, often omitting subsequent segments that are semantically part of the same abnormal event. Second, the field lacks systematic evaluation protocols. To bridge these gaps, we introduce VALU, a new benchmark that explicitly defines anomalies across five semantic levels and provides comprehensive temporal boundaries and detailed textual descriptions for each. Based on these annotations, we design three evaluation tasks that comprehensively assess models' capabilities across different dimensions, including temporal grounding, anomaly localization, and anomaly detail discrimination. Evaluation results reveal persistent challenges in current models' capabilities on VAU. We further analyze and discuss these findings, and hope that both VALU and insights will advance research in VAU and the development of Video-LLMs. Our benchmark will be publicly available.

Poster Session C (Coronado)

16:00-17:30 (Coronado)

[Findings] A Herd of Language Models Makes a Better Zero-shot Annotator for Clinical Named Entity Recognition

► *Underline Video*

Seiji Shimizu, Shoko Wakamiya and Eiji Aramaki

Clinical named entity recognition (NER) remains difficult to scale due to the high cost of manual annotation. Although large language models (LLMs) enable zero-shot annotation, their performance on clinical NER is still limited. To this end, we improve the annotation quality by aggregating annotations from a herd of diverse LLMs, including general-purpose, medically adapted, and NER-specialized models. A key challenge in this multi-LLM setting is effectively leveraging entities extracted by only a minority of models: although they account for a substantial portion of true positives, they are heavily intermixed with noise. To address this, we introduce MARY, a label-modeling method for Multi-LLM Annotation using Representation learning to capture contextual similaritY. During aggregation, MARY selectively incorporates minority-extracted entities whose contexts are similar to those of majority-extracted entities, yielding more reliable and comprehensive annotations. Experimental results show that MARY improves the average F1 score by 8.6% over vanilla zero-shot baselines while reducing annotation costs.

[Findings] Across Programming Language Silos: A Study on Cross-Lingual Retrieval-Augmented Code Generation ► *Underline Video*

Qiming Zhu, Jialun Cao, Xuanang Chen, Weili Zhang, Yaojie Lu, Hongyu Lin, Xianpei Han, Le Sun and Shing-Chi Cheung

Current research on large language models (LLMs) with retrieval-augmented code generation (RACG) has largely focused on single-language settings, leaving their cross-lingual effectiveness underexplored. Multilingual RACG systems are increasingly important for migrating and reusing code across programming languages (PLs), a common yet challenging task in modern software development. To systematically study cross-lingual code knowledge transfer in RACG, we construct a dataset covering 13 PLs with nearly 14K instances. Our experiments reveal three key insights: (1) Knowledge transfer in RACG across PLs is non-trivial even using direct injection. (2) RACG exhibits unequal cross-lingual knowledge transfer, and its efficacy depends on linguistic affinity of PL pair and diversity of LLM pretraining corpus. (3) RACG shows limited reliance on natural language information embedded in code when equipped with a code-specific retriever. These findings provide practical guidance for designing effective multilingual RACG systems. https://github.com/icip-cas/Cross-Lingual-RACG

[Long] AG-GRPO: Answer-Guided GRPO for Masked Diffusion Language Models ► *Underline Video*

Juhyeong Kim, Gyunyeop Kim and Sangwoo Kang

Reinforcement learning with verifiable rewards (RLVR) typically evaluates only final outcomes, providing limited learning signal about whether the generated reasoning is consistent with the correct answer. As a result, even when ground-truth answers are available during training, on-policy rollouts can repeatedly produce reasoning that is inconsistent with the answer. We propose Answer-Guided Group Relative Policy Optimization (AG-GRPO) for masked diffusion language models (dLLMs), which generate text through iterative masked-token restoration. AG-GRPO combines standard answer-free (AF) rollouts, sampled without access to the ground-truth answer, with answer-guided (AG) rollouts. In AG rollouts, the model generates reasoning conditioned on an anchored ground-truth answer suffix, and then re-predicts the answer from the generated reasoning for reward computation. We compute group-relative advantages over the combined AF/AG rollout set, allowing answer-guided training signals to improve the answer-free policy used at test time. Across mathematics, puzzle-solving, and code-generation benchmarks, AG-GRPO consistently improves over the pretrained dLLM and prior RL method for masked dLLMs. We further analyze optimization dynamics to study how shared group-relative advantages support signal transfer and affect convergence. Our code is available at https://github.com/JuHyng/ag_grpo.

[Findings] Applicability Condition Extraction for Therapeutic Drug-Disease Relations ► *Underline Video*

Guanting Luo, Noriki Nishida, Yuji Matsumoto and Yuki Arase

Identifying conditions that a certain drug takes therapeutic effect on a target disease is crucial for clinical decision-making support. However, most existing biomedical information extraction methods have focused on identifying only relations between drugs and diseases, while largely overlooking the context-specific conditions where such relations can apply. To address this problem, we introduce the task of applicability condition extraction for therapeutic drug–disease relations from biomedical research literature. We create the first dataset that has manually annotated triples of drugs, diseases, and applicability conditions on biomedical paper abstracts with 1,119 drug-disease pairs. Using this dataset, we systematically evaluate the performance of a range of existing methods. In addition, we propose a new method that enhances LoRA to consider relations between drugs and diseases. Our method consistently outperforms strong baselines across different evaluation settings.

[Findings] Arctic-Text2SQL-R1: Simple Rewards, Strong Reasoning in Text-to-SQL ► *Underline Video*

Zhewei Yao, Guoheng Sun, Lukasz Borchmann, Gaurav Nuti, Zheyu Shen, Minghang Deng, Bohan Zhai, Hao Zhang, Ang Li and Yuxiong He

Translating natural language into SQL (Text2SQL) is a longstanding challenge at the intersection of natural language understanding and structured data access. While large language models (LLMs) have significantly improved fluency in SQL generation, producing correct and executable SQL, particularly for complex queries, remains a bottleneck. We present Arctic-Text2SQL-R1, a reinforcement learning (RL) framework and model family designed to generate accurate, executable SQL using a lightweight reward signal based solely on execution correctness. Our approach avoids brittle intermediate supervision and complex reward shaping, promoting stable training and alignment with the end task. Combined with carefully curated data, strong supervised initialization, and effective training practices, Arctic-Text2SQL-R1 achieves state-of-the-art execution accuracy across six diverse Text2SQL benchmarks and ranks among the leading entries on the BIRD leaderboard. Notably, our 7B model outperforms prior 70B-class systems, highlighting the frameworks scalability and efficiency. We

further demonstrate inference-time robustness through simple extensions like value retrieval and majority voting. Extensive experiments and ablation studies offer both positive and negative insights, providing practical guidance for future Text2SQL research.

[Long] AutoGraph-R1: End-to-End Reinforcement Learning for Knowledge Graph Construction ▶ *Underline Video*

Hong Ting Tsang, Jiaxin Bai, Haoyu Huang, Qiao Xiao, Tianshi Zheng, Baixuan Xu, Shujie Liu and Yangqiu Song

Building effective knowledge graphs (KGs) for Retrieval-Augmented Generation (RAG) is pivotal for advancing question answering (QA) systems. However, its effectiveness is hindered by a fundamental disconnect: the knowledge graph (KG) construction process is decoupled from its downstream application, yielding suboptimal graph structures. To bridge this gap, we introduce AutoGraph-R1, the first framework to directly optimize KG construction for task performance using Reinforcement Learning (RL). AutoGraph-R1 trains an LLM constructor by framing graph generation as a policy learning problem, where the reward is derived from the graph's functional utility in a RAG pipeline. We design two novel, task-aware reward functions, one for graphs as knowledge carriers and another as knowledge indices. Across multiple QA benchmarks, AutoGraph-R1 consistently enables graph RAG methods to achieve significant performance gains over using task-agnostic baseline graphs. Our work shows it is possible to close the loop between construction and application, shifting the paradigm from building intrinsically 'good' graphs to building demonstrably 'useful' ones.

[Findings] AutoSUIT Bench - Automated Security UnIt Test Benchmark for LLM Coding ▶ *Underline Video*

Samuel Osebe, Fan Yang, Junyi Li, Yue Gu, Yongxin Wang, Satyapriya Krishna, Kai-Wei Chang, Aram Galstyan, Rahul Gupta and Weitong Ruan

Large Language Models (LLMs) are evolving rapidly on code generation tasks. While it is important to evaluate their code generation accuracy, ensuring they follow responsible practices is equally critical. Some of the previous works use tools such as CodeQL to match patterns against Common Weakness Enumeration (CWE), suffering from high error rate, while others rely on human annotation to only focus on top CWE categories, limiting security coverage. We propose AutoSUIT Bench, which addresses these limitations through a paradigm to automate the vulnerable code benchmark creation with iterative auto validation. As a result, our benchmark covers 232 CWE categories across C/C++, Java, and Python languages and is designed to evaluate on four coding tasks: (i) code generation, (ii) generation with CWE context, (iii) security patching, and (iv) code completion. Upon benchmarking against LLMs, we found that functionality pass rate is consistently higher than vulnerability pass rate for all programming languages. One notable observation from our benchmark is that LLMs perform well on top CWEs while lacks on others down the list. This highlights the necessity of vulnerable code benchmarks with larger CWE coverage.

[Long] Benchmarking Egocentric Clinical Intent Understanding Capability for Medical Multimodal Large Language Models ▶ *Underline Video*

Shaonan Liu, Guo Yu, Xiaoling Luo, Shiyi Zheng, Jie Liu, Wenting Chen and Linlin Shen

Medical Multimodal Large Language Models (Med-MLLMs) require egocentric clinical intent understanding for real-world deployment, yet existing benchmarks fail to evaluate this critical capability. We introduce MedGaze-Bench, the first benchmark leveraging clinician gaze as a Cognitive Cursor to assess intent understanding across surgery, emergency simulation, and diagnostic interpretation. Our benchmark addresses three fundamental challenges: visual homogeneity of anatomical structures, strict temporal-causal dependencies in clinical workflows, and implicit adherence to safety protocols. We propose a Three-Dimensional Clinical Intent Framework evaluating: (1) Spatial Intent, discriminating precise targets amid visual noise, (2) Temporal Intent, inferring causal rationale through retrospective and prospective reasoning, and (3) Standard Intent, verifying protocol compliance through safety checks. Beyond accuracy metrics, we introduce Trap QA mechanisms to stress-test clinical reliability by penalizing hallucinations and cognitive sycophancy. Experiments reveal current MLLMs struggle with egocentric intent due to over-reliance on global features, leading to fabricated observations and uncritical acceptance of invalid instructions.

[Findings] Biomed-Enriched: Data-Efficient Biomedical Pretraining via Paragraph-Level Annotation ▶ *Underline Video*

Rian Touchent, Nathan Godey and Éric Villemonte de la Clergerie

We annotate PubMed Central paragraphs for document type, domain, and educational quality using a two-stage pipeline: Llama-3.1-70B labels 400K paragraphs, then a fine-tuned XLM-RoBERTa propagates annotations to the full corpus. This paragraph-level approach captures content diversity within scientific articles that document-level labels miss. The resulting Biomed-Enriched corpus contains 2M clinical case paragraphs, providing a publicly available alternative to restricted clinical datasets. For decoders, continual pretraining experiments enable targeted improvements, with clinical upsampling boosting performance by 4 points on MMLU ProfMed and educational filtering improving MedQA and MedMCQA by ~1 point. Combinations of these techniques led to faster convergence, reaching the same performance with a third of training tokens. For encoders, our best recipe matches BioClinical-ModernBERT on 11 tasks (77.3% vs 77.1% F1) while using 2.5x fewer tokens and only public data.

[Long] CASS: Nvidia to AMD Transpilation with Data, Models, and Benchmark ▶ *Underline Video*

Ahmed Heakl, Gustavo Bertolo Stahl, Sarim Hashmi, Seung Hun Eddie Han, Mukul Ranjan, Arina Kharlamova, Salman Khan and Abdulrahman Mahmoud

Cross-architecture GPU code transpilation is essential for unlocking low-level hardware portability, yet no scalable solution exists. We introduce CASS, the first dataset and model suite for source- and assembly-level GPU translation (CUDA HIP, SASS RDNA3). CASS contains 60k verified host-device code pairs, enabling learning-based translation across both ISA and runtime boundaries. We

generate each sample using our automated pipeline that scrapes, translates, compiles, and aligns GPU programs across vendor stacks. Leveraging CASS, we train a suite of domain-specific translation models that achieve 88.2% accuracy on CUDA to HIP and 69.1% on SASS to RDNA3, outperforming commercial baselines including GPT-5.1, Claude-4.5, and Hipify by wide margins. Generated code matches native performance in 85% of cases, preserving both runtime and memory behavior. To support rigorous evaluation, we introduce CASS-Bench, a curated benchmark spanning 18 GPU domains with ground-truth execution. All data, models, and evaluation tools will be released as open source to support progress in GPU compiler tooling, binary compatibility, and LLM-guided code translation.

[Long] ChatR1: Reinforcement Learning for Conversational Reasoning and Retrieval Augmented Question Answering ► *Underline Video*

Simon Lupart, Mohammad Aliannejadi and Evangelos Kanoulas

We present ChatR1, a reasoning framework based on reinforcement learning (RL) for conversational question answering (CQA). Reasoning plays an important role in CQA, where user intent evolves across dialogue turns, and utterances are often underspecified, requiring contextual interpretation, query reformulation, and dynamic coordination between retrieval and generation. Unlike static 'rewrite, retrieve, and generate' pipelines, ChatR1 interleaves search and reasoning across turns, enabling exploratory and adaptive behaviors learned through RL. To address the challenge of sparse and delayed rewards in RL, we propose an intent-aware reward that provides turn-level feedback by aligning retrieval and reasoning with evolving user goals. ChatR1 demonstrates strong performance on both 3B and 7B model backbones, outperforming competitive models on five CQA datasets, measured by different metrics (F1, BERTScore, and LLM-as-judge). We include a diverse set of CQA datasets to cover topic shifts, evolving intents, mixed-initiative dialogues, and multi-document grounding, testing ChatR1's performance from various aspects. Ablation studies confirm the effectiveness of the intent-aware reward. Our analyses further reveal diverse reasoning trajectories and effective use of the search tool. ChatR1 also generalizes robustly across domains, demonstrating that RL-based reasoning enables more flexible and context-aware behavior than static CQA pipelines.

[Findings] ClinicalMC: A Benchmark for Multi-Course Clinical Decision-Making with Large Language Models ► *Underline Video*

Ruihui Hou, Siyi Zhu, Ziyue Huai, Guangya Yu, Yongqi Fan, ChunMing Wang and Tong Ruan

Large language models (LLMs) have been widely adopted in healthcare, yet they still encounter significant challenges in complex clinical decision-making scenarios. Existing benchmarks primarily assess LLM performance in single-course settings and lack systematic evaluation in multi-course scenarios, where a patients condition evolves over time. To address this gap, we propose ClinicalMC, a benchmark for multi-course clinical decision-making. It includes 1,275 Chinese and 5,804 English samples across four stages from admission to discharge. These stages cover triage, first-course examination/diagnosis/treatment, subsequent multi-course examination/assessment/treatment, and final diagnosis. In ClinicalMC, patients in the English dataset undergo an average of 5.11 clinical courses, whereas those in the Chinese dataset undergo 3.42. To assess LLM performance, we construct a multi-agent evaluation framework that includes patient, examiner, and doctor agents. Based on the benchmark and framework, we design two experimental settings, a single-turn static setting and a multi-turn dynamic setting, and assess three categories of LLMs: 1) closed-source LLMs like GPT-4o-mini; 2) open-source LLMs like DeepSeek-V3, and 3) medical LLMs like HuatuoGPT-o1. Through extensive evaluation, we aim to better understand LLM performance in the medical domain and support its effective deployment in healthcare.

[Findings] Contrastive Decoding Mitigates Score Range Bias in LLM-as-a-Judge ► *Underline Video*

Yoshinari Fujinuma

Large Language Models (LLMs) are commonly used as evaluators in various applications, but the reliability of the outcomes remains a challenge. One such challenge is using LLMs-as-judges for direct assessment, i.e., assigning scores from a specified range without any references. Using summarization as our primary testbed, we first show that this challenge stems from LLM judge outputs being associated with score range bias, i.e., LLM judge outputs are highly sensitive to pre-defined score ranges. We also show that similar biases exist among models from the same family. We then mitigate this bias through contrastive decoding, achieving up to 11.7% relative improvement in Spearman correlation with human judgments, averaged across score ranges.

[Findings] CoT-Edit: Reinforcement Learning of Chain-of-Thought Reasoning for Code Edit Suggestion ► *Underline Video*

Wuya Chen, Yihao Yang and Yue Lin

Code edit suggestion, which encompasses modifying, refactoring, and maintaining existing code, represents the most frequent software development activity and has become a focal point for AI-powered tools. Traditional methods translate explicit natural language instructions into code edits, while pattern-based approaches learn from users' historical editing patterns to provide style-consistent and more accurate suggestions. However, these pattern-based methods still face two critical challenges: (1) difficulty handling edits that demand deep contextual reasoning, and (2) lack of interpretability in editing decisions. To tackle this, we propose CoT-Edit, a reinforcement learning framework that guides LLMs to discover chain-of-thought (CoT) reasoning paths for code editing without requiring human-annotated CoT data. Specifically, we design multi-step reasoning framework that enable: (1) analysis-guided code editing, and (2) seamless switching between CoT and non-CoT inference modes. Building on this, we introduce Edit-Aware Reward Modeling (EARM), a fine-grained diff-based reward approach for effective learning. Furthermore, we discover a LoRA merging strategy that enhances model generalization. Evaluations on an industrial dataset show that our approach achieves 60.2% edit accuracy, outperforming all strong baselines. Online A/B tests further confirm its effectiveness in production. Code is available at https://github.com/202230483077yyh/CoT-Edit.

[Long] CPR-RAG: Clinical Prior-Regularized Retrieval for Anatomy-Aware 3D CT Report Generation ► *Underline Video*

Sungkyu yang, Kang-Min Kim and Mansu Kim

Generating radiology reports from 3D volumetric data remains challenging due to the difficulty of grounding fine-grained pathologies within high-dimensional scans. While retrieval-augmented generation (RAG) offers a potential solution, standard approaches struggle with visual-semantic ambiguity and often introduce irrelevant "normal" context that dilutes pathological signals. To address this limitation, we introduce CPR-RAG, a model-agnostic RAG framework that enhances organ-level grounding by integrating clinical priors into the retrieval process. Specifically, we propose a clinical prior-regularized re-ranking module that leverages corpus-derived co-occurrence statistics to align retrieved candidates with latent disease distributions, ensuring clinical consistency beyond mere visual similarity. Furthermore, we employ clinical relevance context refinement to selectively filter out boilerplate normal descriptions, thereby maximizing the information density of the evidence provided to the generator. Extensive experiments on the RadGenome-ChestCT benchmark demonstrate that CPR-RAG significantly improves clinical efficacy across state-of-the-art radiology report generation models. Human evaluation further confirms that our approach achieves superior factual correctness, completeness, and utility compared to the existing models.

[Findings] Detecting Sensitive Personal Information in Japanese Pre-Training Corpora for Large Language Models ► *Underline Video*

Rei Minamoto, Yusuke Oda and Daisuke Kawahara

Sensitive personal information can appear in large-scale pre-training corpora for large language models (LLMs). Detecting and filtering such information is therefore essential to ensure compliance with privacy regulations and prevent unintended information leakage. However, in contrast to English and other languages, research into sensitive personal information has been limited in the Japanese language. In this study, we focus on sensitive personal data defined as special care-required personal information (SCPI) under Japan's Act on the Protection of Personal Information (APPI). We construct an SCPI dataset using LLM-based annotation and train machine learning models to rapidly detect SCPI in text. As a result, our SCPI classifier can effectively identify information related to SCPI. This study is the first to explore SCPI detection in Japanese text corpora, highlighting the challenges of accurate detection.

[Findings] DiagnosisArena: Benchmarking Diagnostic Reasoning for Large Language Models ► *Underline Video*

Yakun Zhu, Zhongzhen Huang, Linjie Mu, Yutong Huang, Wei Nie, JIAJI LIU, Shaoting Zhang, Pengfei Liu and Xiaofan Zhang

The emergence of groundbreaking large language models capable of performing complex reasoning tasks holds significant promise for addressing various scientific challenges, including those arising in complex clinical scenarios. To enable their safe and effective deployment in real-world healthcare settings, it is urgently necessary to benchmark the diagnostic capabilities of current models systematically. Given the limitations of existing medical benchmarks in evaluating advanced diagnostic reasoning, we present DiagnosisArena, a comprehensive and challenging benchmark designed to rigorously assess professional-level diagnostic competence. DiagnosisArena consists of 1,113 pairs of segmented patient cases and corresponding diagnoses, spanning 28 medical specialties, deriving from clinical case reports published in 10 top-tier medical journals. The benchmark is developed through a meticulous construction pipeline, involving multiple rounds of screening and review by both AI systems and human experts, with thorough checks conducted to prevent data leakage. Our study reveals that even the most advanced reasoning models, o3-mini, o1, and DeepSeek-R1, achieve only 45.82%, 31.09%, and 17.79% accuracy, respectively. This finding highlights a significant generalization bottleneck in current large language models when faced with clinical diagnostic reasoning challenges. Through DiagnosisArena, we aim to drive further advancements in AIs diagnostic reasoning capabilities, enabling more effective solutions for real-world clinical diagnostic challenges. We openly share the benchmark and evaluation tools for further research and development.

[Long] Diffuse Thinking: Exploring Diffusion Language Models as Efficient Thought Proposers for Reasoning ► *Underline Video*

Chenyang Shao, Sijian Ren, Fengli Xu and Yong Li

Large language models (LLMs) have demonstrated strong capabilities in complex reasoning tasks, yet their autoregressive generation paradigm makes it computationally expensive to explore diverse reasoning paths. In contrast, diffusion language models (DLMs) adopt a parallel, non-autoregressive generation mechanism that enables the efficient production of diverse candidate outputs. Motivated by this complementarity, we explore a collaborative reasoning framework that combines diffusion-based generation with autoregressive evaluation. Specifically, we leverage DLMs to efficiently generate diverse intermediate reasoning thoughts, and employ LLMs as evaluators to assess and select candidates based on their plausibility and correctness. By decoupling proposal generation from evaluation, our framework exploits the strengths of both models: efficient exploration from diffusion models and causally grounded assessment from autoregressive models, which naturally aligns with the divergent-convergent thinking framework in cognitive psychology. Experiments across various mathematical and logical reasoning benchmarks demonstrate that our framework improves inference efficiency while maintaining competitive or superior reasoning accuracy, laying the groundwork for building efficient reasoning architectures. Our code is open-source at https://anonymous.4open.science/r/Diffuse-Thinking-EC60.

[Long] Discover and Prove: An Open-source Agentic Framework for Hard Mode Automated Theorem Proving in Lean 4 ► *Underline Video*

Chengwu Liu, Yichun Yin, Ye Yuan, Jiaxuan Xie, Botao Li, Siqi Li, Jianhao Shen, Yan Xu, Lifeng Shang and Ming Zhang

Most ATP benchmarks embed the final answer within the formal statement , a convention we call "Easy Mode" , a design that simplifies the task relative to what human competitors face and may lead to optimistic estimates of model capability. We call the stricter, more realistic

setting "Hard Mode": the system must independently discover the answer before constructing a formal proof. To enable Hard Mode research, we make two contributions. First, we release MiniF2F-Hard and FIMO-Hard, expert-reannotated Hard Mode variants of two widely-used ATP benchmarks. Second, we introduce Discover And Prove (DAP), an agentic framework that uses LLM natural-language reasoning with explicit self-reflection to discover answers, then rewrites Hard Mode statements into Easy Mode ones for existing ATP provers. DAP sets the state of the art: on CombiBench it raises solved problems from 7 (previous SOTA, Pass@16) to 10; on PutnamBench it is the first system to formally prove 36 theorems in Hard Mode , while simultaneously revealing that state-of-the-art LLMs exceed 80% answer accuracy on the same problems where formal provers manage under 10%, exposing a substantial gap that Hard Mode benchmarks are uniquely suited to measure.

[Long] Discovering a Shared Logical Subspace: Steering LLM Logical Reasoning via Alignment of Natural-Language and Symbolic Views ► *Underline Video*

Feihao Fang, My T. Thai and Yuanyuan Lei

Large Language Models (LLMs) still struggle with multi-step logical reasoning. Existing approaches either purely refine the reasoning chain in natural language form or attach a symbolic solver as an external module. In this work, we instead ask whether LLMs contain a shared internal logical subspace that simultaneously aligns natural-language and symbolic-language views of the reasoning process. Our hypothesis is that this logical subspace captures logical reasoning capabilities in LLMs that are shared across views while remaining independent of surface forms. To verify this, we employ Canonical Correlation Analysis on the paired residual activations from natural-language and symbolic-language reasoning chains, learning a low-dimensional subspace with maximum cross-view correlation. Furthermore, we design a training-free approach that steers LLMs reasoning chain along this logical subspace, thereby leveraging the complementary reasoning signals from both views. Experiments on four logical reasoning benchmarks demonstrate the effectiveness of our approach, improving accuracy by up to 11 percentage points and generalizing well on out-of-domain problems.

[Long] Do Morals Guide How LLMs Think? The Role of Ethical Perspectives in General Problem Solving ► *Underline Video*

Iseo Kim, Eunjin Hong and Juae Kim

This study investigates how different moral conditions influence the general problem-solving capabilities of Large Language Models (LLMs). We aim to explore whether the role of morality as a cognitive function operating in human decision-making can be extended to LLMs. Specifically, we define distinct moral stages based on Kohlbergs theory of moral development and design prompts to elicit model responses aligned with each corresponding condition. The validity of this alignment is verified using the Defining Issues Test, a human evaluation tool. Subsequently, models reflecting the characteristics of each condition are evaluated using the MMLU benchmark, which requires general problem-solving abilities across various domains. Experimental results show that different moral perspectives lead to changes in the models decision-making during general reasoning, reflected in both responses and internal representations. Notably, conditions grounded in more advanced moral stages tend to elicit more thoughtful and reflective problem-solving behavior, which is often associated with enhanced performance. Our study attempts to broaden the concept of LLM morality, which has traditionally only been considered in ethical judgment scenarios. Furthermore, it emphasizes that morality is not merely a means of ensuring safety but a crucial factor that shapes a model's behavior and thought processes.

[Long] E2EDev: Benchmarking Large Language Models in End-to-End Software Development Task ► *Underline Video*

Jingyao Liu

The rapid advancement in large language models (LLMs) has demonstrated significant potential in End-to-End Software Development (E2ESD). However, existing E2ESD benchmarks are limited by coarse-grained requirement specifications and unreliable evaluation protocols, hindering a true understanding of current framework capabilities. To address these limitations, we present E2EDev, a novel benchmark grounded in the principles of Behavior-Driven Development (BDD) to assess whether the generated software meets user needs through mimicking real user interactions. E2EDev comprises (i) a fine-grained set of user requirements for each target software project (ii) multiple BDD test scenarios with corresponding Python step implementations for each requirement, and (iii) a fully automated testing pipeline built on the Behave framework. By evaluating various E2ESD frameworks and LLM backbones with E2EDev, our analysis reveals a persistent struggle to effectively solve these tasks, underscoring the critical need for more effective and cost-efficient E2ESD solutions. Our codebase and benchmark are available at https://github.com/SCUNLP/E2EDev.

[Long] End-to-End Optimization of LLM-Driven Multi-Agent Search Systems via Heterogeneous-Group-Based Reinforcement Learning ► *Underline Video*

Guanzhong Chen, Shaoxiong Yang, Chao Li, Wei Liu, Jian Luan and Zenglin Xu

Large language models (LLMs) are versatile, yet their deployment in complex real-world settings is limited by static knowledge cutoffs and the difficulty of producing controllable behavior within a single inference. Multi-agent search systems (MASS), which coordinate specialized LLM agents equipped with search tools, mitigate these issues via task decomposition and retrieval-augmented problem solving. However, optimizing LLMs for agent-specific roles remains labor-intensive with prompt engineering or supervised fine-tuning, motivating automated end-to-end training. Existing multi-agent reinforcement learning (MARL) methods such as Multi-Agent Proximal Policy Optimization (MAPPO) typically depend on large critic networks to evaluate joint actions, leading to instability and high memory costs. We introduce Multi-Agent Heterogeneous Group Policy Optimization (MHGPO), which updates policies by estimating relative advantages across heterogeneous groups of multi-agent rollouts, shifting the optimization focus from local agent performance to global system success. We

further study three group rollout sampling strategies to trade off sample efficiency and optimization quality. Experiments show that MHGPO captures implicit inter-agent dependencies and consistently outperforms strong baselines in both task performance and computational efficiency.

[Findings] Evaluating Implicit Biases in LLM Reasoning through Logic Grid Puzzles ▶ *Underline Video*

Fatima Jahara, Mark Dredze and Sharon Levy

While recent safety guardrails effectively suppress overtly biased outputs, subtler forms of social bias emerge during complex logical reasoning tasks that evade current evaluation benchmarks. To fill this gap, we introduce a new evaluation framework, PRIME (Puzzle Reasoning for Implicit Biases in Model Evaluation), that uses logic grid puzzles to systematically probe the influence of social stereotypes on logical reasoning and decision making in LLMs. Our use of logic puzzles enables automatic generation and verification, as well as variability in complexity and biased settings. PRIME includes stereotypical, anti-stereotypical, and neutral puzzle variants generated from a shared puzzle structure, allowing for controlled and fine-grained comparisons. We evaluate multiple model families across puzzle sizes and test the effectiveness of prompt-based mitigation strategies. Focusing our experiments on gender stereotypes, our findings highlight that models consistently reason more accurately when solutions align with stereotypical associations. This demonstrates the significance of PRIME for diagnosing and quantifying social biases perpetuated in the deductive reasoning of LLMs, where fairness is critical.

[Long] Evaluating Language Model Pluralism through In-the-wild Crowd Discussions ▶ *Underline Video*

Gagan Mundada, Rohan Surana, Nandhini Swaminathan, Bodhisattwa Prasad Majumder, Junda Wu, Julian McAuley and Zhouhang Xie

When answering subjective questions, an ideal LLM should surface diverse plausible perspectives rather than favoring a single viewpoint, a characteristic known as pluralism. Recent studies show that modern LLMs optimized through preference alignment systematically favor certain positions on subjective queries, making pluralism evaluation increasingly important. However, existing evaluation methods focus dominantly on multiple-choice and question-answering tasks, leaving open-ended generation largely unaddressed. We propose PLURALEVAL, an evaluation framework that assesses LLM pluralism in open-ended generation by comparing outputs against free-form crowd responses. Our approach decomposes ground-truth responses into atomic, non-overlapping claims, then evaluates whether LLMs adequately cover this diverse claim space. We then introduce WildSCOPE, a multi-domain dataset of natural crowd responses, and demonstrate that PLURALEVAL captures novel insights, such as the collapse of pluralism through sycophancy, where LLM systematically degrades in overton pluralism when a user's belief is revealed. Finally, we discuss the value and actionable insights for preserving and encouraging pluralism from LLM deployers' side.

[Long] FACTrial: Factorized Clinical Contrastive Training for Scalable Patient-Trial Retrieval ▶ *Underline Video*

XuanrenChen XuanrenChen, Chongyang Tao, Tao Shen and Shuai Ma

Patient-trial retrieval is a challenging problem that requires nuanced clinical reasoning beyond surface-level semantic similarity. However, scarce and costly relevance annotations force existing approaches to rely on very limited supervision or zero-shot transfer, reducing the task to generic semantic matching and failing to capture multi-factor eligibility reasoning. To this end, we propose FACTrial, a factorized contrastive training framework that leverages LLMs to synthesize diagnosis-aware supervision for scalable patient-trial retrieval. FACTrial decomposes each patient note into a primary diagnosis and a set of concomitant, eligibility-triggering conditions, and constructs complementary contrastive signals through structured trial augmentation. Specifically, we generate primary-target and concomitant-target positives, together with clinically confusable near-miss negatives, to enforce diagnostic specificity under contrastive learning. Two specialized bi-encoder experts are trained to balance primary-diagnosis prioritization and concomitant-driven recall, and fused into a single deployable retriever. Experiments on three public benchmarks demonstrate that FACTrial achieves state-of-the-art performance, improving both top-ranked quality and high-recall coverage.

[Long] FLAIR: Steering LLM Mathematical Problem Solving based on A Fuzzy-Logic-AssIsted Reasoner ▶ *Underline Video*

Hao Wu

Mathematical reasoning is one of the core capabilities for Large Language Models (LLMs). Yet, existing approaches often rely on static heuristics or pre-determined reasoning strategies, limiting their ability to adapt to different intermediate states. To address this limitation, we propose FLAIR (Fuzzy-Logic-AssIsted Reasoner), an adaptive framework that integrates fuzzy theory into LLM-based mathematical reasoning. Specifically, FLAIR characterizes intermediate problem-solving states using fuzzy memberships and employs a parameterized fuzzy rule system to conditionally activate subsequent actions. These rule parameters are further adjusted via Reinforcement Learning using solution-level feedback as the reward signal, enabling adaptive and iterative refinement without reliance on a fixed strategy. To the best of our knowledge, this work is the first to integrate fuzzy theory into LLM-based mathematical reasoning. Extensive experiments across multiple benchmarks demonstrate that FLAIR consistently outperforms recent state-of-the-art baselines, while offering effective and interpretable diagnostics of intermediate problem-solving states.

[Long] Frankentext: Stitching random text fragments into long-form narratives ▶ *Underline Video*

Chau Minh Pham, Jenna Russell, Dzung Pham and Mohit Iyyer

As AI text detectors are increasingly used to flag LLM-generated writing, a natural question arises: are there forms of high-quality generated narrative that can evade such detection? We introduce Frankentexts, a long-form narrative generation paradigm that treats an LLM as a composer of existing texts rather than as an author. Given a writing prompt and thousands of randomly sampled human-written

snippets, the model assembles a coherent narrative where most tokens (e.g., 90%) are copied verbatim from the source passages. Despite the extreme challenge of the task, we observe through extensive automatic and human evaluation that Frankentexts improve over vanilla LLM generations in key writing quality metrics such as diversity and novelty while remaining mostly coherent and relevant to the prompt. Furthermore, Frankentexts pose a fundamental challenge to current AI text detectors: 72% of Frankentexts produced by our best configuration (Gemini-2.5-Pro with 5K input snippets) are misclassified as human-written by Pangram, a state-of-the-art detector. Human annotators praise Frankentexts for their inventive premises, vivid descriptions, and dry humor; however, they still identify issues with abrupt tonal shifts and uneven grammar across segments. Overall, the emergence of high-quality yet low-detectability Frankentexts challenges established authorship norms while raising concerns about the publishing economy.

[Findings] From Experience to Skill: Multi-Agent Generative Engine Optimization via Reusable Strategy Learning ► *Underline Video*

Beining Wu;Fuyou Mao; Jiong Lin;Cheng Yang; Jiaxuan Lu; Yifu Guo;Siyu Zhang, Yifan Wu, Ying Huang and Fu Li1;

Generative engines (GEs) are reshaping information access by replacing ranked links with citation-grounded answers, yet current Generative Engine Optimization (GEO) methods optimize each instance in isolation, unable to accumulate or transfer effective strategies across tasks and engines. We reframe GEO as a strategy learning problem and propose MAGEO, a multi-agent framework in which coordinated planning, editing, and fidelity-aware evaluation serve as the execution layer, while validated editing patterns are progressively distilled into reusable, engine-specific optimization skills. To enable controlled assessment, we introduce a Twin Branch Evaluation Protocol for causal attribution of content edits and DSV-CF, a dual-axis metric that unifies semantic visibility with attribution accuracy. We further release MSME-GEO-Bench, a multi-scenario, multi-engine benchmark grounded in real-world queries. Experiments on three mainstream engines show that MAGEO substantially outperforms heuristic baselines in both visibility and citation fidelity, with ablations confirming that engine-specific preference modeling and strategy reuse are central to these gains, suggesting a scalable learning-driven paradigm for trustworthy GEO. Code is available at https://github.com/Wu-beining/MAGEO.

[Long] From Regulatory Approvals to Patents: Cross-Domain Linking for Cardiovascular Device Traceability ► *Underline Video*

Qingqing Yang, Haijiang Liu and Moyan Li

Linking FDA-approved medical devices to their underlying United States Patent and Trademark Office (USPTO) patents enables critical applications such as recall root-cause analysis, M&A-driven IP discovery, and technology trajectory mapping. However, this cross-domain entity linking task remains unexplored due to severe semantic gaps: FDA documents focus on clinical outcomes, while patents describe technical mechanisms, yielding minimal lexical overlap. We formalize medical device-patent linking as a challenging cross-domain entity linking problem characterized by label scarcity and domain shifts. Using cardiovascular devices as a high-impact, representative domain featuring diverse technologies, high recall rates, and abundant disclosures, we construct a benchmark with 434 devices, 698K patents, and 585 high-fidelity expert-verified pairs. To address these challenges, we propose Bridge-MedDevKG, a coarse-to-fine framework that integrates (1) MedDevOnto, a domain-specific ontology that anchors device concepts via three-tier UMLS normalization; (2) Multi-signal candidate generation fusing company affiliation, semantic similarity, and ontology-weighted entity overlap; and (3) Heterogeneous reranking with multi-signal scoring and XGBoost classification on hard negatives. Our approach achieves a conservative lower-bound recall of 91.6% on the gold standard with 50.9% noise reduction, substantially outperforming LLM baselines under comparable evaluation. The resulting MedDevKG provides 6.8M high-confidence links, laying a scalable foundation for regulatory-IP integration across medical specialties.

[Findings] FuseSearch: Learning Adaptive Parallel Execution for Efficient Code Localization ► *Underline Video*

Ke Xu, Siyang Xiao, Ming Liang, Yichen Yu, Zhixiang Wang, Jingxuan Xu, Dajun Chen, Wei Jiang and Yong Li

Code localization is a primary bottleneck in automated software development. While parallel tool execution can accelerate discovery, existing agents suffer from a 34.9% redundant tool invocation rate, negating the benefits of parallelism. We introduce FuseSearch, which reframes parallel code localization as a quality–efficiency co-optimization problem. By defining tool efficiency—the ratio of novel information gain to total invocations—we employ a two-stage SFT and RL pipeline to train models in adaptive parallel strategies. Unlike fixed-breadth methods, FuseSearch dynamically adjusts search breadth based on task context, transitioning from exploration to refinement. On SWE-bench Verified, FuseSearch-4B matches SOTA performance (84.7% file-level and 56.4% function-level F_1 scores) while being 93.6% faster, using 67.7% fewer turns and 68.9% fewer tokens. Our findings demonstrate that efficiency-aware training inherently boosts quality by eliminating noisy, redundant signals, enabling high-performance, low-cost localization agents. Code: https://github.com/sxthunder/FuseSearch

[Long] GitChameleon 2.0: Evaluating AI Code Generation Against Python Library Version Incompatibilities ► *Underline Video*

Diganta Misra, Nizar Islah, Victor May, Brice Rauby, Zihan Wang, Justine Gehring, Antonio Orvieto, Muawiz Chaudhary, Eilif Muller, Irina Rish, Samira Kahou and Massimo Caccia

The rapid evolution of software libraries poses a considerable hurdle for code generation, necessitating continuous adaptation to frequent version updates while preserving backward compatibility. While existing code evolution benchmarks provide valuable insights, they typically lack execution-based evaluation for generating code compliant with specific library versions. To address this, we introduce GitChameleon 2.0, a novel, meticulously curated dataset comprising 328 Python code completion problems, each conditioned on specific library versions and accompanied by executable unit tests. GitChameleon 2.0 rigorously evaluates the capacity of contemporary large language models (LLMs), LLM-powered agents, code assistants, and RAG systems to perform version-conditioned code generation that demonstrates functional accuracy through execution. Our extensive evaluations indicate that state-of-the-art systems encounter significant challenges with

this task; enterprise models achieving baseline success rates in the 48-51% range, underscoring the intricacy of the problem. By offering an execution-based benchmark emphasizing the dynamic nature of code libraries, GitChameleon 2.0 enables a clearer understanding of this challenge and helps guide the development of more adaptable and dependable AI code generation methods.

[Long] Good Arguments Against the People Pleasers: How Reasoning Mitigates (Yet Masks) LLM Sycophancy ▶ *Underline Video*

Zhaoxin Feng, Zheng CHEN, Jianfei Ma, Yip Tin Po, Emmanuele Chersoni and Bo Li

Alignment techniques often inadvertently induce sycophancy in LLMs. While prior studies studied this behaviour in direct-answer settings, the role of Chain-of-Thought (CoT) reasoning remains under-explored: does it serve as a logical constraint that mitigates sycophancy, or a tool for post-hoc rationalization that masks it? We evaluate a range of models across objective and subjective tasks to investigate the issue. Results show that reasoning generally reduces sycophancy in final decisions but also masks sycophancy in some samples, where models construct deceptive justifications through logical inconsistencies, calculation errors, and one-sided arguments etc. Furthermore, LLMs are more prone to sycophancy in subjective tasks and under authority-bias. Our mechanistic analysis reveals that the tendency of sycophancy in LLMs is dynamic during the reasoning process rather than being pre-determined at the input.

[Findings] Guardian-as-an-Advisor: Advancing Next-Generation Guardian Models for Trustworthy LLMs ▶ *Underline Video*

Yue Huang, Haomin Zhuang, Jiayi Ye, Han Bao, Yanbo Wang, Hang Hua, Siyuan Wu, Pin-Yu Chen and Xiangliang Zhang

Hard-gated safety checkers often over-refuse and misalign with a vendor's model spec; prevailing taxonomies also neglect robustness and honesty, yielding safer-on-paper yet less useful systems. This work introduces Guardian-as-an-Advisor (GaaA), a soft-gating pipeline where a guardian predicts a binary risk label plus a concise explanation and prepends this advice to the original query for re-inference, keeping the base model operating under its original spec. To support training and evaluation, GuardSet is constructed, a 208k+ multi-domain dataset unifying harmful and harmless cases with targeted robustness and honesty slices. GuardAdvisor is trained via SFT followed by RL to enforce label-explanation consistency. GuardAdvisor attains competitive detection accuracy while enabling the advisory workflow; when used to augment inputs, responses improve over unaugmented prompts. A latency study shows advisor inference uses below 5% of base-model compute and adds only 2-10% end-to-end overhead under realistic harmful-input rates. Overall, GaaA steers models to comply with the model spec, maintaining safety while reducing over-refusal.

[Long] Interleaved Tool-Call Reasoning for Protein Function Understanding ▶ *Underline Video*

Chuanliu Fan, Zicheng Ma, Huanran Meng, Aijia Zhang, Wenjie Du, Jun Zhang, Ziqiang Cao and Guohong Fu

Recent advances in large language models (LLMs) have highlighted the effectiveness of chain-of-thought reasoning in symbolic domains such as mathematics and programming. However, our study shows that directly transferring such text-based reasoning paradigms to protein function understanding is ineffective: reinforcement learning mainly amplifies superficial keyword patterns while failing to introduce new biological knowledge, resulting in limited generalization. We argue that protein function prediction is a knowledge-intensive scientific task that fundamentally relies on external biological priors and computational tools rather than purely internal reasoning. To address this gap, we propose Protein Function Understanding Agent (PFUA), a tool-augmented protein reasoning agent that unifies problem decomposition, tool invocation, and grounded answer generation. Instead of relying on long unconstrained reasoning traces, PFUA integrates domain-specific tools to produce verifiable intermediate evidence. Experiments on four benchmarks demonstrate that PFUA consistently outperforms text-only reasoning models with an average performance improvement of 103%. We believe PFUA has the potential to become a standard paradigm for agentic reasoning in knowledge-intensive life science domains.

[Long] Investigating Counterfactual Unfairness in LLMs towards Identities through Humor ▶ *Underline Video*

Shubin Kim, Yejin Son, Junyeong Park, Keummin Ka, Seungbeen Lee, Jaeyoung Lee, Hyeju Jang, Alice Oh and Youngjae Yu

Humor holds up a mirror to social perception: what we find funny often reflects who we are and how we judge others. When language models engage with humor, their reactions expose the social assumptions they have internalized from training data. In this paper, we investigate counterfactual unfairness through humor by observing how the models responses change when we swap who speaks and who is addressed while holding other factors constant. Our framework spans three tasks: humor generation refusal, speaker intention inference, and relational/societal impact prediction, covering both identity-agnostic humor and identity-specific disparagement humor. We introduce interpretable bias metrics that capture asymmetric patterns under identity swaps. Experiments across state-of-the-art models reveal consistent relational disparities: jokes told by privileged speakers are refused up to 67.5% more often, judged as malicious 64.7% more frequently, and rated up to 1.5 points higher in social harm on a 5-point scale. These patterns highlight how sensitivity and stereotyping coexist in generative models, complicating efforts toward fairness and cultural alignment.

[Long] Knowledge Vector of Logical Reasoning in Large Language Models ▶ *Underline Video*

Zixuan Wang and Yuanyuan Lei

Logical reasoning serves as a central capability in LLMs and includes three main forms: deductive, inductive, and abductive reasoning. In this work, we study the knowledge representations of these reasoning types in LLMs and analyze the correlations among them. Our analysis shows that each form of logical reasoning can be captured as a reasoning-specific knowledge vector in a linear representation space, yet these vectors are largely independent of each other. Motivated by cognitive science theory that these subforms of logical reasoning interact closely in the human brain, as well as our observation that the reasoning process for one type can benefit from the reasoning chain produced by another, we further propose to refine the knowledge representations of each reasoning type in LLMs to encourage complementarity between

them. To this end, we design a complementary subspace-constrained refinement framework, which introduces a complementary loss that enables each reasoning vector to leverage auxiliary knowledge from the others, and a subspace constraint loss that prevents erasion of their unique characteristics. Through steering experiments along reasoning vectors, we find that refined vectors incorporating complementary knowledge yield consistent performance gains. We also conduct a mechanism-interpretability analysis of each reasoning vector, revealing insights into the shared and specific features of different reasoning in LLMs.

[Long] KoCo-Bench: Can Large Language Models Leverage Domain Knowledge in Software Development? ▶ *Underline Video*

Xue Jiang, Ge Li, Jiaru Qian, Xianjie Shi, Chenjie Li, Hao Zhu, Ziyu Wang, Jielun Zhang, Zheyu Zhao, Kechi Zhang, Jia Li, Wenpin Jiao, Zhi Jin and Yihong Dong

Large language models (LLMs) excel at general programming but struggle with domain-specific software development. This gap motivates research into domain specialization methods that enable LLMs to learn and utilize domain knowledge and data. However, existing domain-specific code benchmarks focus on assessing what knowledge LLMs possess rather than how they acquire and apply new knowledge, lacking explicit knowledge corpora for developing domain specialization methods. To this end, we present KOCO-bench, a novel benchmark designed for evaluating domain specialization methods in real-world software development. KOCO-bench contains 6 emerging domains with 11 software frameworks and 25 projects, featuring curated knowledge corpora alongside multi-granularity evaluation tasks including domain code generation (from function-level to project-level with rigorous test suites) and domain knowledge understanding (via multiple-choice Q&A). Unlike previous benchmarks that only provide test sets for direct evaluation, KOCO-bench requires acquiring and applying diverse domain knowledge (APIs, rules, constraints, etc.) from the corpora to solve evaluation tasks. Our evaluations reveal that KOCO-bench poses significant challenges to state-of-the-art LLMs. Even with domain specialization methods (e.g., SFT, RAG, kNN-LM) applied, improvements remain marginal. Best-performing coding agent, Claude Code, achieves only 34.2%, highlighting the urgent need for more effective domain specialization methods. We release KOCO-bench, evaluation code, and baselines to advance further research at https://github.com/jiangxxxue/KOCO-bench.

[Findings] Large Language Models for IT Automation Tasks: Are We There Yet? ▶ *Underline Video*

Md. Mahadi Hassan, John Salvador, Akond Ashfaque Ur Rahman and Santu Karmaker

LLMs show promise in code generation, yet their effectiveness for IT automation tasks, particularly for tools like Ansible, remains understudied. Existing benchmarks rely primarily on synthetic tasks that fail to capture the needs of practitioners who use IT automation tools. We present ExITBench (Execution-based IT Automation Benchmark), a benchmark of 126 diverse tasks (e.g., configuring servers and managing files) in which each task captures state reconciliation - a core property of IT automation tools. ExITBench evaluates LLMs' ability to generate functional Ansible automation scripts via dynamic execution in controlled environments. We evaluate 14 open-source and 3 proprietary LLMs and find that GPT-4.1-Mini achieves the best pass@10 rate of 23.9%, while Claude-3.5-Sonnet achieves the best pass@1 performance. To explain the low performance, we analyze 1,517 execution failures across the evaluated LLMs and identify two prevalent semantic error categories: failures in state-reconciliation reasoning (42.117% combined from variable (12.287%), host (10.363%), path (10.511%), and template (8.956%) issues) and deficiencies in module-specific execution knowledge (26.203% combined from attribute & parameter (17.617%) and module (8.586%) errors). Our findings reveal key limitations in LLMs' ability to address state reconciliation and apply specialized module knowledge, indicating that reliable IT automation with LLM-based agents need major advances in state reasoning and domain-specific execution.

[Long] Logical Phase Transitions: Understanding Collapse in LLM Logical Reasoning ▶ *Underline Video*

Xinglang Zhan, Yunyao Zhang, ZeLiang Chen, Junqing Yu, Wei Yang and Zikai Song

Symbolic logical reasoning is a critical yet underexplored capability of large language models (LLMs), providing reliable and verifiable decision-making in high-stakes domains such as mathematical reasoning and legal judgment. In this study, we present a systematic analysis of logical reasoning under controlled increases in logical complexity, and reveal a previously unrecognized phenomenon, which we term Logical Phase Transitions: rather than degrading smoothly, logical reasoning performance remains stable within a regime but collapses abruptly beyond a critical logical depth, mirroring physical phase transitions such as water freezing beyond a critical temperature threshold. Building on this insight, we propose Neuro-Symbolic Curriculum Tuning, a principled framework that adaptively aligns natural language with logical symbols to establish a shared representation, and reshapes training dynamics around phase-transition boundaries to progressively strengthen reasoning at increasing logical depths. Experiments on five benchmarks show that our approach effectively mitigates logical reasoning collapse at high complexity, yielding average accuracy gains of +1.26 in naive prompting and +3.95 in CoT, while improving generalization to unseen logical compositions.

[Findings] LoopCoder: Scaling Code Intelligence via Looped Language Models ▶ *Underline Video*

Jian Yang, Wei Zhang, Shuyue Guo, Yizhi LI, Linzheng Chai, Zhengmao Ye, Shukai Liu, Yuyang Song, Jiajun Wu, Che Liu, Tianyu Zheng, Siwei Wu, Leo L, Xudong Ma, Chuan Hao, Ran Tao, Yan Xing, Jianzhou Wang, Mingjie Tang, Aishan Liu, Zhoujun Li, Xianglong Liu, Weifeng Lv and Bryan Dai

While large language models (LLMs) have mastered syntax-level code generation, complex algorithmic reasoning remains a challenge, typically addressed by scaling model depth and parameter count. Universal Transformers (UT) offer a compelling alternative by introducing a recurrent inductive bias that aligns with the recursive nature of programming logic. However, training looped architectures at scale has historically been hindered by severe instability and optimization difficulties associated with backpropagation through time (BPTT). We

present LoopCoder (40B-A80B) pre-trained on 12T+ code and general tokens, along with LoopCoder-Thinking and LoopCoder-Instruct variants, the first large-scale looped transformer for code, achieving comparable performance to standard dense architectures with more parameters. Unlike prior approaches that restrict recurrence to small-scale tasks, we implement a comprehensive looped training protocol spanning both pre-training and post-training phases. We initiate the model via dense-to-loop transformation, folding a pre-trained dense checkpoint to initialize a recurrent block, followed by rigorous looped pre-training and specialized post-training for instruction following and reasoning. Our results establish a robust recipe for scaling coding intelligence via recurrent computation, proving that dense checkpoints serve as an optimal foundation for evolving into dynamic, looped reasoners.

[Findings] Measuring Large Language Models' Adversarial Behavior in Social Deduction Games ► *Underline Video*

Marissa Li, Esha Shivakumar, Peiran Wang, Ying Li and Yuan Tian

As large language models are increasingly adopted and trusted in real-world applications, understanding their behavior beyond single-turn prompting has become critical. Existing safety evaluations primarily focus on refusal-based methods that test whether models avoid responding to inappropriate or violent requests, leaving open questions about how models behave in interactive social settings. In this paper, we observe the adversarial behavior of LLM models through a multi-agent simulation across five diverse social deduction conversational games, acting as testbeds that provide social pressures and survival stress based on game design without explicit prompt injections. From these interactions, we construct a closed behavioral taxonomy derived through open card sorting, applied uniformly across models using a meta-LLM for behavior labeling. This approach displays that models exhibit distinct behavioral profiles and that models' different ways of structured deliberation influence both social stability and competitive success.

[Long] MedEinst: Benchmarking the Einstellung Effect in Medical LLMs through Counterfactual Differential Diagnosis ► *Underline Video*

Wenting Chen, Zhongrui Zhu, Guolin Huang and Wenxuan Wang

Despite achieving high accuracy on medical benchmarks, LLMs exhibit the Einstellung Effect in clinical diagnosis, relying on statistical shortcuts rather than patient-specific evidence, causing misdiagnosis in atypical cases. Existing benchmarks fail to detect this critical failure mode. We introduce MedEinst, a counterfactual benchmark with 5,383 paired clinical cases across 49 diseases. Each pair contains a control case and a "trap" case with altered discriminative evidence that flips the diagnosis. We measure susceptibility via Bias Trap Rate, probability of misdiagnosing traps despite correctly diagnosing controls. Evaluation shows frontier models achieve high baseline accuracy but severe bias trap rates. Thus, we propose ECR-Agent, aligning LLM reasoning with Evidence-Based Medicine via two components: (1) Dynamic Causal Inference (DCI) performs structured reasoning through dual-pathway perception, dynamic causal graph reasoning across three levels (association, intervention, counterfactual), and evidence audit for final diagnosis; (2) Critic-Driven Graph & Memory Evolution (CGME) iteratively refines the system by storing validated reasoning paths in an exemplar base and consolidating disease-specific knowledge into evolving illness graphs. Source code is to be released.

[Long] MedForge: Interpretable Medical Deepfake Detection via Forgery-aware Reasoning ► *Underline Video*

Zhihui Chen, Kai He, Qingyuan Lei, Bin Pu, Jian Zhang, Yuling Xu and Mengling Feng

Text-guided image editors can now manipulate authentic medical scans with high fidelity, enabling lesion implantation/removal that threatens clinical trust and safety. Existing defenses are inadequate for healthcare. Medical detectors are largely black-box, while MLLM-based explainers are typically post-hoc, lack medical expertise, and may hallucinate evidence on ambiguous cases. We present MedForge, a data-and-method solution for pre-hoc, evidence-grounded medical forgery detection. We introduce MedForge-90K, a large-scale benchmark of realistic lesion edits across 19 pathologies with expert-guided reasoning supervision via doctor inspection guidelines and gold edit locations. Building on it, MedForge-Reasoner performs localize-then-analyze reasoning, predicting suspicious regions before producing a verdict, and is further aligned with Forgery-aware GSPO to strengthen grounding and reduce hallucinations. Experiments demonstrate state-of-the-art detection accuracy and trustworthy, expert-aligned explanations.

[Findings] MedQPA-Gen: Medical Question Proposing and Answering for Report Generation ► *Underline Video*

Weijie Liang, Xiyue Zhu, Ruike Zhu, Chenhao Li, Cheng Tang, Zhiyu Liu, Zhihua Gong, Shirui Luo, Yudu Li and Volodymyr Kindratenko

Medical report generation from medical images is a vital AI task that helps doctors with diagnosis and marks a significant step toward creating general AI-powered medical systems. However, previous methods either fail to optimize factual accuracy or heavily depend on expert preference data. To overcome these challenges, we propose MedQPA, an automatic and generalizable report evaluation technique that uses question proposing and answering to enable controllable, structured reasoning grounded in medical domain knowledge and the factual correctness of the report. Additionally, we design MedQPA-Gen, a medical report generation pipeline that maximizes the MedQPA score through prompt engineering and reinforcement learning with MedQPA as a reward signal. We demonstrate that MedQPA is an accurate evaluation metric that closely correlates with human preferences. More importantly, MedQPA-Gen achieves higher human preference scores and better performance on downstream tasks. We open-source code at this repo https://github.com/MedQPA-gen/MedQPA-gen.

[Findings] MedRedFlag: Investigating how LLMs Redirect Misconceptions in Real-World Health Communication ► *Underline Video*

Sraavya Sambara, Yuan Pu, Ayman Ali, Vishala Mishra, Lionel Wong and Monica Agrawal

Real-world health questions from patients often unintentionally embed false assumptions or premises. In such cases, safe medical communication typically involves redirection: addressing the implicit misconception and then responding to the underlying patient context,

rather than the original question. While large language models (LLMs) are increasingly being used by lay users for medical advice, they have not yet been tested for this crucial competency. Therefore, in this work, we investigate how LLMs react to false premises embedded within real-world health questions. We develop a semi-automated pipeline to curate MedRedFlag, a dataset of 1100+ questions sourced from Reddit that require redirection. We then systematically compare responses from state-of-the-art LLMs to those from clinicians. Our analysis reveals that LLMs often fail to redirect problematic questions, even when the problematic premise is detected, and provide answers that could lead to suboptimal medical decision making. Our benchmark and results reveal a novel and substantial gap in how LLMs perform under the conditions of real-world health communication, highlighting critical safety concerns for patient-facing medical AI systems. Code and data are available at https://github.com/srsambara-1/MedRedFlag.

[Long] MegaRAG: Multimodal Knowledge Graph-Based Retrieval Augmented Generation ▶ *Underline Video*

Chi-Hsiang Hsiao, Yi-Cheng Wang, Tzung-Sheng Lin, Yi-Ren Yeh and Chu-Song Chen

Retrieval-augmented generation (RAG) enables large language models (LLMs) to dynamically access external information, which is powerful for answering questions over previously unseen documents. Nonetheless, they struggle with high-level conceptual understanding and holistic comprehension due to limited context windows, which constrain their ability to perform deep reasoning over long-form, domain-specific content such as full-length books. To solve this problem, knowledge graphs (KGs) have been leveraged to provide entity-centric structure and hierarchical summaries, offering more structured support for reasoning. However, existing KG-based RAG solutions remain restricted to text-only inputs and fail to leverage the complementary insights provided by other modalities such as vision. On the other hand, reasoning from visual documents requires textual, visual, and spatial cues into structured, hierarchical concepts. To address this issue, we introduce a multimodal knowledge graph-based RAG that enables cross-modal reasoning for better content understanding. Our method incorporates visual cues into the construction of knowledge graphs, the retrieval phase, and the answer generation process. Experimental results across both global and fine-grained question answering tasks show that our approach consistently outperforms existing approaches on both textual and multimodal benchmarks.

[Long] Minimal Free Resolution Guided Adaptive Tree Reasoning ▶ *Underline Video*

Dezhao Tang, Meihan Liu, Yulai Tong, Guan Yuan and Qiuyan Yan

Dynamic reasoning trees can help large language models solve complex tasks by explicitly structuring intermediate decisions. However, existing approaches often rely on manually specified subproblems or predefined decomposition patterns, which limits the effectiveness of reasoning and generalization. To solve this problem, we propose SyRA, a hierarchical reasoning framework based on MFR theory that supports the construction of adaptive reasoning trees and reliable error correction within a single LLM. Specifically, SyRA focuses on reasoning-tree construction, dynamically controlling branching and expansion using MFR principles to enable informative, non-redundant subproblem decomposition. In addition, it introduces a residual backtracking mechanism for adaptive cross-layer error correction, allowing the model to revise earlier reasoning decisions based on downstream feedback. Across eight reasoning benchmarks, SyRA significantly reduces logical errors and improves reasoning accuracy, while achieving a better balance between accuracy and reasoning time than the Chain-of-Thought, Decompose–Analyze–Rethink and Tree-of-Thought. Our code and dataset are available at https://github.com/Tim798-art/SyRA/tree/main/SyRA.

[Long] MM-PoisonRAG: Disrupting Multimodal RAG with Local and Global Knowledge Poisoning Attacks ▶ *Underline Video*

Hyeonjeong Ha, Qiusi Zhan, Jeonghwan Kim, Dimitrios Bralios, Saikrishna sanniboina, Nanyun Peng, Kai-Wei Chang, Daniel Kang and Heng Ji

Retrieval-augmented generation (RAG) has become a common practice in multimodal large language models (MLLM) to enhance factual grounding and reduce hallucination. Yet, its reliance on retrieval exposes MLLMs to knowledge poisoning attacks, in which adversaries deliberately inject malicious multimodal content into external knowledge bases to steer models toward generating incorrect or even harmful responses. We present MM-PoisonRAG, a framework to systematically study the vulnerability of multimodal RAG under knowledge poisoning. Specifically, we design two novel attack strategies: Localized Poisoning Attack (LPA), which implants targeted, query-specific multimodal misinformation to manipulate outputs toward attacker-controlled responses, and Globalized Poisoning Attack (GPA), which uses a single, untargeted adversarial injection to broadly corrupt reasoning and collapse generation quality across all queries. Extensive experiments on diverse tasks, multimodal RAG components, and attacker access levels reveal severe vulnerabilities: LPA achieves up to 56% attack success rate even under restricted access, and transfers effectively across four different retrievers without re-optimizing the adversaries. GPA completely disrupts model generation to 0% accuracy with just one poisoned content. Moreover, both LPA and GPA bypass existing defenses, underscoring the fragility of multimodal RAG and establishing MM-PoisonRAG as a foundation for future research on securing RAG frameworks against multimodal knowledge poisoning.

[Findings] MultiFileTest: A Multi-File-Level LLM Unit Test Generation Benchmark and Impact of Error Fixing Mechanisms ▶ *Underline Video*

Yibo Wang, Congying Xia, Wenting Zhao, Jiangshu Du, Chunyu Miao, Zhongfen Deng, Philip S. Yu and Chen Xing

Unit test generation has become a promising and important use case for Large Language Models (LLMs). However, existing evaluation benchmarks for LLM unit test generation primarily focus on function- or class-level code (single-file) rather than on more practical, challenging multi-file codebases. To address this limitation, we propose MultiFileTest, a multi-file-level benchmark for unit test generation covering Python, Java, and JavaScript. MultiFileTest features 20 high-quality, moderate-sized projects per language. We evaluate eleven

frontier LLMs on MultiFileTest, and the results show that most tested LLMs exhibit moderate performance on MultiFileTest, highlighting the benchmark's inherent difficulty. We also conduct a thorough error analysis, which shows that even advanced LLMs, such as Gemini 3.0 Pro, exhibit basic yet critical errors, including executability and cascade errors. Motivated by this observation, we further evaluate these frontier LLMs under manual error-fixing and self-error-fixing scenarios to assess their potential when equipped with error-fixing mechanisms. Our dataset is available at MultiFileTest.

[Findings] NSF-CoT: Neuro-Symbolic Formal Verification of Chain-of-Thought Faithfulness in Contextual Question Answering ▶ *Underline Video*

Vishal Pramanik, Maisha Maliha, Nathaniel D. Bastian, Alvaro Velasquez, Susmit Jha and Sumit Kumar Jha

Chain-of-thought (CoT) prompting makes language models write step-by-step explanations, but these steps may not match what the model actually used to choose its answer. Existing faithfulness checks often only test whether changing the written chain changes the answer, without verifying whether the steps are truly supported by the given evidence, or they require special prompts that do not generalize well. We present NSF-CoT, a neuro-symbolic formal verification method that checks CoT faithfulness step by step for contextual question answering. NSF-CoT (1) converts the provided context facts and each reasoning step into simple logical statements, (2) uses counterfactual attribution to estimate which context facts the model relied on while generating each step, and (3) verifies each step using a hybrid checker that combines an SMT solver with an LLM-based entailment judge. For every step, we score groundedness (supported by the full context), validity (supported by the facts the model relied on), and utility (helps reach the final answer), and combine them into a faithfulness score. Across OpenBookQA, QASC, and HotpotQA, NSF-CoT consistently outperforms causal mediation, perturbation probes, and behavioral monitoring, and it identifies reasoning steps that are not only unfaithful but also harmful to the model's final decision.

[Findings] OmniCode: A Benchmark for Evaluating Software Development Agents ▶ *Underline Video*

Atharv Sonwane, Eng-Shen Tu, Wei-Chung Lu, Claas Beger, Carter Larsen, Debjit Dhar, Simon Alford, Rachel Chen, Ronit Pattanayak, Tuan Anh Dang, Guohao Chen, Gloria Geng, Kevin Ellis and Saikat Dutta

LLM-powered coding agents are redefining how real-world software is developed. To drive the research towards better coding agents, we require challenging benchmarks that can rigorously evaluate the ability of such agents to perform various software engineering tasks. However, popular coding benchmarks such as HumanEval and SWE-Bench focus on narrowly scoped tasks such as competition programming and patch generation. In reality, software engineers have to handle a broader set of tasks for real-world software development. To address this gap, we propose OmniCode, a novel software engineering benchmark that contains a broader and more diverse set of task categories beyond code or patch generation. Overall, OmniCode contains 1794 tasks spanning three programming languages – Python, Java, and C++ – and four key categories: bug fixing, test generation, code review fixing, and style fixing. In contrast to prior software engineering benchmarks, the tasks in OmniCode are (1) manually validated to eliminate ill-defined problems, and (2) synthetically crafted or recently curated to avoid data leakage issues, presenting a new framework for synthetically generating diverse software tasks from limited real-world data. We evaluate OmniCode with popular agent frameworks such as SWE-Agent and show that while they may perform well on bug fixing for Python, they fall short on tasks such as Test Generation and in languages such as C++ and Java. For instance, SWE-Agent achieves a maximum of 25.0% with DeepSeek-V3.1 on C++ Test Generation. OmniCode aims to serve as a robust benchmark and spur the development of agents that can perform well across different aspects of software development.

[Long] One Persona, Many Cues, Different Results: How Sociodemographic Cues Impact LLM Personalization ▶ *Underline Video*

Franziska Weeber, Vera Neplenbroek, Jan Batzner and Sebastian Padó

Personalization of LLMs by sociodemographic subgroup often improves user experience, but can also introduce or amplify biases and unfair outcomes across groups. Prior work has employed so-called personas, sociodemographic user attributes conveyed to a model, to study bias in LLMs by relying on a single cue to prompt a persona, such as user names or explicit attribute mentions. This disregards LLM sensitivity to prompt variation and the rarity of some cues in real interactions (external validity). We compare six commonly used persona cues across seven open and proprietary LLMs on four writing and advice tasks. While cues are overall highly correlated, they produce sub- stantial variance in responses across personas that can change findings on persona-induced differences and bias. We therefore caution against claims based on single persona cues, especially when they are overly explicit and have low external validity.

[Findings] POSTCONDBENCH: Benchmarking Correctness and Completeness in Formal Postcondition Inference ▶ *Underline Video*

Gehao Zhang and Juan Zhai

Formal postconditions precisely characterize program behavior and support debugging, testing, and verification, but writing them requires substantial expertise and effort. This has motivated recent work on automatically generating postconditions from code and natural-language artifacts using large language models (LLMs). However, evaluation remains a key bottleneck. Existing benchmarks primarily emphasize correctness under limited evaluation settings, often relying on surface-form matching or manual assessment on small or synthetic datasets. We introduce POSTCONDBENCH, a multilingual benchmark for evaluating method-level postcondition generation from real-world software. POSTCONDBENCH comprises 420 Python and Java tasks drawn from 121 open-source projects, each paired with a high-quality ground-truth postcondition set constructed with expert involvement. To enable automatic evaluation, POSTCONDBENCH provides a runnable execution environment and operationalizes completeness via defect discrimination: a postcondition set is more complete if it is violated by more defective implementations while remaining satisfied on correct executions. Using POSTCONDBENCH, we formulate three generation settings and evaluate five SOTA LLMs. Our results reveal a substantial gap between correctness and completeness, and show that

repository-level dependencies and method complexity exacerbate this gap.

[Long] Process Reward Models Meet Planning: Generating Precise and Scalable Datasets for Step-Level Rewards ▶ *Underline Video*
Raffaele Pisano and Roberto Navigli
Process Reward Models (PRMs) have emerged as a powerful tool for providing step-level feedback when evaluating the reasoning of Large Language Models (LLMs), which frequently produce chains of thought (CoTs) containing errors even when the final answer is correct. However, existing PRM datasets remain expensive to construct, prone to annotation errors, and predominantly limited to the mathematical domain. This work introduces a novel and scalable approach to PRM dataset generation based on planning logical problems expressed in the Planning Domain Definition Language (PDDL). Using this method, we generate a corpus of approximately one million reasoning steps across various PDDL domains and use it to train PRMs. Experimental results show that augmenting widely-used PRM training datasets with PDDL-derived data yields substantial improvements in both mathematical and non-mathematical reasoning, as demonstrated across multiple benchmarks. These findings indicate that planning problems constitute a scalable and effective resource for generating robust, precise, and fine-grained training data for PRMs, going beyond the classical mathematical sources that dominate this field.

[Long] ProMed: Shapley Information Gain Guided Reinforcement Learning for Proactive Medical LLMs ▶ *Underline Video*
Hongxin Ding, Baixiang Huang, Yue Fang, Weibin Liao, Xinke Jiang, Jinyang Zhang, Yinghao Zhu, Zheng Li, Liantao Ma, Junfeng Zhao and Yasha Wang
Interactive medical questioning is essential in clinical consultations, where physicians must actively gather necessary patient information. Yet existing medical Large Language Models (LLMs) predominantly follow a reactive paradigm, risking diagnostic errors by answering before seeking sufficient details. To bridge this gap, we propose ProMed, a reinforcement learning framework that transitions LLMs toward a proactive paradigm, enabling them to ask clinically valuable questions before decision-making. Central to ProMed is the Shapley Information Gain (SIG) reward, which quantifies a question's clinical utility as the amount of newly acquired information, while considering its contextual importance via Shapley values. We integrate SIG into a two-stage training pipeline: (1) SIG-Guided Model Initialization uses Monte Carlo Tree Search to construct high-reward interaction trajectories for supervision, and (2) SIG-Augmented Policy Optimization, with a novel SIG-guided Reward Distribution Mechanism that prioritizes informative questions for fine-grained optimization. Experiments on partial-information medical benchmarks show that ProMed significantly outperforms state-of-the-art methods by 6.29% on average and delivers a 54.45% gain over the reactive paradigm, and generalizes robustly to out-of-domain cases. Our codes are available at https://github.com/hxxding/ProMed.

[Findings] PV-SQL: Synergizing Database Probing and Rule-based Verification for Text-to-SQL Agents ▶ *Underline Video*
Yuan Tian and Tianyi Zhang
Text-to-SQL systems often struggle with deep contextual understanding, particularly for complex queries with subtle requirements. We present PV-SQL, an agentic framework that addresses these failures through two complementary components: Probe and Verify. The Probe component iteratively generates probing queries to retrieve concrete records from the database, resolving ambiguities in value formats, column semantics, and inter-table relationships to build richer contextual understanding. The Verify component employs a rule-based method to extract verifiable conditions and construct an executable checklist, enabling iterative SQL refinement that effectively reduces missing constraints. Experiments on the BIRD benchmarks show that PV-SQL outperforms the best text-to-SQL baseline by 5% in execution accuracy and 20.8% in valid efficiency score while consuming fewer tokens.

[Long] Quantifying and Improving the Robustness of Retrieval-Augmented Language Models Against Spurious Features in Grounding Data ▶ *Underline Video*
Shiping Yang, Jie Wu, Wenbiao Ding, Ning Wu, Shining Liang, MING GONG, Hongzhi Li, Hengyuan Zhang, Angel X Chang and Dongmei Zhang
Robustness has become a critical attribute for the deployment of RAG systems in real-world applications. Existing research focuses on robustness to explicit noise (e.g., document semantics) but overlooks implicit noise (spurious features). Moreover, previous studies on spurious features in LLMs are limited to specific types (e.g., formats) and narrow scenarios (e.g., ICL). In this work, we identify and study spurious features in the RAG paradigm, a robustness issue caused by the sensitivity of LLMs to semantic-agnostic features. We then propose a novel framework,SURE, to empirically quantify the robustness of RALMs against spurious features. Beyond providing a comprehensive taxonomy and metrics for evaluation, the frameworks data synthesis pipeline facilitates training-based strategies to improve robustness. Further analysis suggests that spurious features are a widespread and challenging problem in the field of RAG. Our code is available at https://anonymous.4open.science/r/RAG-SpuriousFeatures-62B3.

[Long] Rethinking Expert Trajectory Utilization in LLM Post-training for Mathematical Reasoning ▶ *Underline Video*
Bowen Ding, Yuhan Chen, Jiayang Lyu, Jiyao Yuan, Qi Zhu, Shuangshuang Tian, Dantong Zhu, Futing Wang, Heyuan Deng, Fei Mi, Lifeng Shang and Tao Lin
Supervised Fine-Tuning (SFT) and Reinforcement Learning (RL) dominate the post-training landscape for mathematical reasoning, yet differ fundamentally in their reliance on expert trajectories. To understand the optimal way to harness these trajectories for maximizing performance, we propose the Plasticity-Ceiling Framework. This framework empirically grounds the post-training landscape by decomposing the final performance ceiling into the foundational SFT performance and the subsequent RL plasticity (i.e., the maximum improvement

via RL). Through extensive benchmarking, we establish the Sequential SFT-then-RL pipeline as the superior standard, overcoming the stability and premature convergence deficits inherent in synchronized approaches. Furthermore, we derive precise scaling guidelines: (1) Transitioning to RL at the Stable or Mild Overfitting Regime of SFT maximizes the final ceiling by securing a robust SFT foundation with substantial RL plasticity; (2) Refuting the Less is More hypothesis in SFT-then-RL scaling, we demonstrate that Data Scale determines the primary post-training potential, while Trajectory Difficulty acts as a performance multiplier; and (3) The Minimum Validation Loss of SFT serves as a reliable indicator for selecting the expert trajectories that maximize the ultimate performance ceiling. Our findings provide actionable guidelines for extracting maximum value from expert trajectories.

[Long] Route to Rome Attack: Directing LLM Routers to Expensive Models via Adversarial Suffix Optimization ▶ *Underline Video*

Haochun Tang, Yuliang Yan, Jiahua Lu, huaxiao liu and Enyan Dai

Cost-aware routing dynamically dispatches user queries to models of varying capability to balance performance and inference cost. However, the routing strategy introduces new security concern that adversaries may manipulate router to consistently select expensive high-capability models. Existing routing attacks depend either on white-box access or heuristic prompts, rendering them ineffective in real-world black-box scenarios. In this work, we propose R^2A, which aims to mislead black-box LLM routers to expensive models via adversarial suffix optimization. Specifically, R^2A deploys a hybrid ensemble surrogate router to mimic the black-box router. A suffix optimization algorithm is further adapted for the ensemble-based surrogate. Extensive experiments on multiple open-source and commercial routing systems demonstrate that R^2A significantly increases the routing rate to expensive models on queries of different distributions. Code and examples: https://github.com/thcxiker/R2A-Attack.

[Findings] Safety-Aware Dialogue System for Postoperative Oral Cancer Care with Structured Clarification and a Clinically Curated Dataset ▶ *Underline Video*

Tzu-Chi Liu, Hui-Ying Yang, Shiow-Ching Shun, Yu-Chi Chen, Lu-Yen Anny Chen and Yong-Sheng Chen

Clinical dialogue systems are increasingly vital for patient education and follow-up care; however, effective responses often depend on the clinical context that patients often fail to provide in detail. Responding directly to vague messages can therefore lead to generic or clinically misaligned advice, a challenge that is particularly pronounced in post-op oral cancer (OC) care due to speech impairment and functional limitations. Moreover, post-op OC patients often experience psychological distress, making safety-aware language more likely to arise in dialogue. Dialogue systems in this setting must therefore address both clarifying missing clinical context and ensuring psychological safety. We propose a safety-aware dialogue system that applies information-gain guided clarification before RAG-based response generation and screens user utterances for emotional distress and suicidal ideation. Expert evaluations show that the proposed system improves the quality and clinical appropriateness of generated responses relative to strong baselines, while the safety module closely aligns with expert judgments on clinically concerning utterances. Furthermore, we release a clinically curated Chinese post-op OC QA dataset with expert-validated annotations, which we use throughout our experiments.

[Long] Scaling Test-Time Compute to Achieve IOI Gold Medal with Open-Weight Models ▶ *Underline Video*

Mehrzad Samadi, Aleksander Ficek, Sean Narenthiran, Siddhartha Jain, Wasi Uddin Ahmad, Somshubra Majumdar, Vahid Noroozi and Boris Ginsburg

Competitive programming has become a rigorous benchmark for evaluating the reasoning and problem-solving capabilities of large language models (LLMs). The International Olympiad in Informatics (IOI) stands out as one of the most prestigious annual competitions in competitive programming and has become a key benchmark for comparing human and AI-level programming ability. While several proprietary models have been claimed to achieve gold medal-level performance at the IOI, often with undisclosed methods, achieving comparable results with open-weight models remains a significant challenge. In this paper, we present GenCluster, a scalable and reproducible test-time compute framework that attains IOI gold-level performance using open-weight models. It combines large-scale generation, behavioral clustering, ranking, and a round-robin submission strategy to efficiently explore diverse solution spaces under limited validation budgets. Our experiments show that the performance of our proposed approach scales consistently with available compute, narrowing the gap between open and closed systems. Notably, we will show that GenCluster can achieve a gold medal at IOI 2025 for the first time with an open-weight model gpt-oss-120b, setting a new benchmark for transparent and reproducible evaluation of reasoning in LLMs.

[Long] SLICEFORMER: Static Program Slicing Using Language Models With Dataflow-Aware Pretraining and Constrained Decoding ▶ *Underline Video*

Pengfei He, Shaowei Wang, Tse-Hsun Chen and Muhammad Asaduzzaman

Static program slicing is a fundamental software engineering technique for isolating code relevant to specific variables. While recent learning-based approaches using language models (LMs) show promise in automating slice prediction, they suffer from inaccurate dependency modeling and unconstrained generation, where LMs fail to capture precise data flow relations and produce slices containing hallucinated tokens and statements. To address these challenges, we propose SliceFormer, a novel approach that reformulates static program slicing as a sequence-to-sequence task using small language models such as CodeT5+. introduces two key innovations that directly target the identified limitations. First, to improve dependency modeling, we design dataflow-aware pretraining objectives that leverage data flow graphs DFG to teach models data dependencies through dataflow-preserving statement permutation and dataflow-aware span corruption. Second, to eliminate hallucination, we develop a constrained decoding mechanism that enforces both lexical and syntactic constraints. We evaluate SliceFormer on Java and Python program slicing benchmarks, demonstrating consistent improvements over state-of-the-art baselines with up

to 22% gain in ExactMatch.

[Findings] Structured Dialogue Refinement: Building Retrieval-Augmented Question Answering on Goal-Oriented Dialogues ▶ *Underline Video*

Bin Wu, Sawan Kumar, Prasetya Ajie Utama and Mohamed Yahyal

Retrieval-Augmented Generation (RAG) is widely used for question answering over well-structured document corpora. However, a large amount of real-world problem-solving knowledge is captured in goal-oriented dialogues, where common ground misalignment between users and helpers gives rise to sparse, diffuse, and dynamically refined evidence that challenges standard RAG pipelines. We propose Structured Dialogue Refinement (SDR), a unified framework that adapts dialogue corpora for RAG at both the retrieval and generation stages without altering the underlying pipeline. Specifically, SDR introduces Dual Dialogue Querying for intent-aligned retrieval via issue-centric and solution-centric pseudo-documents, and Graph-Structured Dialogues coupled with a relevance-driven subgraph selection strategy to enable effective utilization of conversational evidence. We further adopt a nugget-based evaluation setup for dialogue-grounded RAG, enabling fine-grained analysis of retrieval coverage and grounded answer generation. Experiments demonstrate that SDR substantially improves both retrieval quality and grounded QA performance under dialogue-specific structural challenges.

[Findings] SynthFix: Adaptive Neuro-Symbolic Code Vulnerability Repair ▶ *Underline Video*

Yifan Zhang, Jieyu Li, Kexin Pei, Yu Huang and Kevin Leach

Large Language Models (LLMs) show promise for automated code repair but often struggle with the complex semantic and structural correctness required. We present SynthFix, a hybrid neural-symbolic framework that improves LLM-based vulnerability repair by unifying code synthesis with compiler-informed symbolic feedback. The core of our approach is an adaptive training strategy where a neural Router Model directs code samples to either Supervised Fine-Tuning (SFT) to learn common patterns or Reward Fine-Tuning (RFT) with symbolic rewards for complex, iterative refinement. On the FixJS (JavaScript) and CodeFlaws (C) benchmarks, SynthFix achieves up to 18% relative improvement in CodeBLEU/CrystalBLEU and 32% in Exact Match over strong SFT and RFT baselines. Our results show that this adaptive combination of training strategies, which mirrors how developers alternate between pattern application and tool feedback, significantly improves the accuracy and efficiency of LLM-based vulnerability repair. Our code and data are available at https://github.com/CoderDoge1108/SynthFix.

[Long] TH-RAG: Topic-Based Hierarchical Knowledge Graphs for Robust Multi-hop Reasoning in Graph-based RAG Systems ▶ *Underline Video*

JungHyoun Kim, Soohyeong Kim, SeokJun Hwang, JeongHyeon Park and YongSuk Choi

Retrieval-augmented generation (RAG) enables large language models (LLMs) to incorporate external knowledge at inference. Graph-based RAG extends this by organizing corpora into knowledge graphs, improving multi-hop reasoning and offering a global understanding of the corpus. However, triplet-based graphs generated by LLMs are often fragmented and sparsely connected, which reduces coherence and hinders reasoning. Prior enrichment methods such as clustering, community detection, or approximate graph algorithms attempt to restore connectivity but incur high computational cost and risk semantic distortion. To address these issues, we propose TH-RAG, a hierarchical framework that organizes triplets into subtopics and topics, enhancing connectivity, integrating dispersed information, and supporting robust multi-hop reasoning. Experiments on abstractive and specific QA benchmarks show that TH-RAG outperforms strong baselines in accuracy and robustness while remaining efficient, providing a scalable foundation for graph-based RAG.

[Findings] The Thin Line Between Comprehension and Persuasion in LLMs ▶ *Underline Video*

Adrian de_Wynter and Tangming Yuan

Large language models (LLMs) are excellent at maintaining high-level, convincing dialogue, but it remains unclear whether their persuasive success reflects genuine understanding of the discourse. We examine this question through informal debates between humans and LLMs, first by measuring their persuasive skills, and then by relating these to their understanding of what is being talked about: namely, their comprehension of argumentative structures and the pragmatic context on the same debates. We find that LLMs effectively maintain coherent, persuasive debates, and can sway the beliefs of both participants and audiences. We also note that awareness or suspicion of AI involvement encourage people to be more critical of the arguments made. However, we also find that LLMs are unable to show comprehension of deeper dialogical structures, such as argument quality or existence of supporting premises. Our results reveal a disconnect between LLM comprehension and dialogical skills, raising ethical and practical concerns on their deployment on explanation-critical contexts. From an argumentation-theoretical perspective, we experimentally question whether an agent, if it can convincingly maintain a dialogue, is required to show it knows what is talking about.

[Long] TokDrift: When LLM Speaks in Subwords but Code Speaks in Grammar ▶ *Underline Video*

Yinxi Li, Yuntian Deng and Pengyu Nie

Large language models (LLMs) for code rely on subword tokenizers, such as byte-pair encoding (BPE), learned from mixed natural language text and programming language code but driven by statistics rather than grammar. As a result, semantically identical code snippets can be tokenized differently depending on superficial factors such as whitespace or identifier naming. To measure the impact of this misalignment, we introduce TokDrift, a framework that applies semantic-preserving rewrite rules to create code variants differing only in tokenization. Across nine code LLMs, including large ones with over 30B parameters, even minor formatting changes can cause substantial shifts in

model behavior. Layer-wise analysis shows that the issue originates in early embeddings, where subword segmentation fails to capture grammar token boundaries. Our findings identify misaligned tokenization as a hidden obstacle to reliable code understanding and generation, highlighting the need for grammar-aware tokenization for future code LLMs.

[Long] Towards Order Fairness: Mitigating LLMs Order Sensitivity through Dual Group Advantage Optimization ▶ *Underline Video*
Xu Chu, Guanyu Wang, Zhijie Tan, Xinrong Chen, Ziyu Li, Tong Mo and Weiping Li
Large Language Models (LLMs) suffer from order bias, where their performance is affected by the arrangement order of input elements. This unfairness limits the model's applications in scenarios such as in-context learning and Retrieval-Augmented Generation (RAG). Recent studies attempt to obtain optimal or suboptimal arrangements based on statistical results or using dataset-based search, but these methods increase inference overhead while leaving the model's inherent order bias unresolved. Other studies mitigate order sensitivity through supervised fine-tuning using augmented training sets with multiple order variants, but often at the cost of accuracy, trapping the model in consistent yet incorrect hallucinations. In this paper, we propose Dual Group Advantage Optimization (DGAO), which aims to improve model accuracy and order stability simultaneously. DGAO calculates and balances intra-group relative accuracy advantage and inter-group relative stability advantage, rewarding the policy model for generating order-stable and correct outputs while penalizing order-sensitive or incorrect responses. This marks the first time reinforcement learning has been used to mitigate LLMs' order sensitivity. We also propose two new metrics, Consistency Rate and Overconfidence Rate, to reveal the pseudo-stability of previous methods and guide more comprehensive evaluation. Extensive experiments demonstrate that DGAO achieves superior order fairness while improving performance on RAG, mathematical reasoning, and classification tasks. Our code is available at: https://anonymous.4open.science/r/DGAO-A481/

[Long] Verbal-R3: Verbal Reranker as the Missing Bridge between Retrieval and Reasoning ▶ *Underline Video*
Sangkwon Park, Donghun Kang, Jisoo Mok and Sungroh Yoon
The conventional Retrieval-Augmented Generation (RAG) paradigm of injecting raw retrieved texts into the Large Language Model (LLM)'s context often results in suboptimal integration of retrieved information. This paper proposes to bridge retrieval results and the LLM's reasoning ability through Verbal Annotations, analytic narratives that explicitly articulate the logical connection between a search query and retrieved contexts. Our empirical investigation reveals the potential of Verbal Annotations to substantially enhance the LLM's ability to generate accurate, contextually-grounded responses. Motivated by this finding, we introduce Verbal-R3, a novel agentic RAG framework that consists of a Generator and a Verbal Reranker. The Generator performs iterative retrieval and reasoning, while the Verbal Reranker returns relevance scores and Verbal Annotations to guide the reasoning and answering process of the Generator. The inference process of Verbal-R3 is further refined through relevance-guided test-time scaling, which efficiently allocates test-time compute for effective trajectory expansion. Verbal-R3 achieves state-of-the-art performance on complex Question Answering benchmarks, validating the effectiveness of the proposed framework.

[Findings] Visual Interference in Speech Evaluation: Cultural Asymmetry and Cross-Modal Bias in MLLMs ▶ *Underline Video*
Kyusik Kim, Hyunwoo Yoo, Jaehoon Choi, Gail Rosen and Bongwon Suh
The transition to end-to-end Multimodal Large Language Models (MLLMs) has positioned these architectures as active social evaluators in high-stakes domains. However, it remains unclear whether these models maintain objective auditory perception or succumb to the "Hearing with Eyes" phenomenon, where visual racial cues distort linguistic proficiency evaluations. We investigate this cross-modal bias by constructing a controlled counterfactual dataset utilizing a Visual Matched-Guise Paradigm. By pairing identical native audio with diverse visual personas across English and Korean contexts, we reveal a distinct Cultural Asymmetry in model behavior. In Anglophone settings, most closed models exhibit Reverse Linguistic Stereotyping, hallucinating non-native accents for Asian speakers despite standard native audio. Conversely, in Korean settings, the same models assign baseline-relative competence premiums across all visual personas, with the largest gains for out-group (White/Black) speakers, consistent with Expectancy Violation Theory. Our findings demonstrate that MLLMs do not merely process sensory inputs but actively reproduce context-dependent sociolinguistic ideologies.

[Findings] What Users Leave Unsaid: Under-Specified Queries Limit Vision-Language Models ▶ *Underline Video*
Dasol Choi, Guijin Son, Hanwool Lee, Minhyuk Kim, Hyunwoo Ko, TEABIN LIM, Eungyeol Ahn, Jungwhan Kim, Seunghyeok Hong and Youngsook Song
Current vision-language benchmarks predominantly feature well-structured questions with clear, explicit prompts. However, real user queries are often informal and underspecified. Users naturally leave much unsaid, relying on images to convey context. We introduce HAERAE-Vision, a benchmark of 653 real-world visual questions from Korean online communities (0.76% survival from 86K candidates), each paired with an explicit rewrite, yielding 1,306 query variants in total. Evaluating 39 VLMs, we find that even state-of-the-art models (GPT-5, Gemini 2.5 Pro) achieve under 50% on the original queries. Crucially, query explicitation alone yields 8 to 22 point improvements, with smaller models benefiting most. We further show that even with web search, under-specified queries underperform explicit queries without search, revealing that current retrieval cannot compensate for what users leave unsaid. Our findings demonstrate that a substantial portion of VLM difficulty stem from natural query under-specification instead of model capability, highlighting a critical gap between benchmark evaluation and real-world deployment.

[Findings] When Debiasing Backfires: Counterintuitive Side Effects of Preprocessing-Based Stereotype Mitigation ▶ *Underline Video*
Yahan_Zheng Yahan_Zheng, John_Guerrerio John_Guerrerio, Soroush_Vosoughi Soroush_Vosoughi and Weicheng_Ma Weicheng_Ma

Preprocessing-based methods for stereotype mitigation, such as pre-/post-training on debiased corpora, are widely used in NLP. While these approaches reduce measurable stereotypes for targeted groups, we find they often induce unintended shifts: stereotyping or counter-stereotyping can increase for other demographics, including across unrelated categories. We demonstrate these side effects across two model families (encoder-only and decoder-only), multiple preprocessing strategies (removing stereotypical sentences, removing group mentions, and swapping references), and both pre- and post-training at different data scales on Wikipedia. Standard benchmarks frequently miss these shifts. Using attention-rollout analysis, we observe that such side effects are not accompanied by large changes in attention flow, complicating mechanistic explanations. We discuss implications for evaluation, provide actionable diagnostics, and argue for side-effect-aware, transparent mitigation practices that make claims calibrated to uncertainty.

[Findings] X-ray Made Simple: Lay Radiology Report Generation and Robust Evaluation ► *Underline Video*
Kun Zhao, Chenghao Xiao, Sixing Yan, Haoteng Tang, William K. Cheung, Noura Al Moubayed, Liang Zhan and Chenghua Lin
While multimodal generative models have advanced radiology report generation (RRG), challenges remain in making reports accessible to patients and ensuring reliable evaluation. The technical language and templated nature of professional reports hinder patient comprehension and enable models to artificially boost lexical metrics such as BLEU by reproducing common report patterns. To address these limitations, we propose the Laymans RRG framework, which leverages layperson-friendly language to enhance patient accessibility and promote more robust evaluation and report generation by encouraging models to focus on semantic accuracy over rigid templates. Our approach also introduces and releases two refined layman-style datasets (at the sentence and report levels), along with a semantics-based evaluation metric that mitigates inflated lexical scores and a layman-guided training strategy. Experiments show that training on layman-style data helps models better capture the meaning of clinical findings. Notably, we observe a positive scaling law: model performance improves with more layman-style data, in contrast to the inverse trend observed with templated professional language.

Poster Session C (Grand Hall)

16:00-17:30 (Grand Hall)

[Long] R^3: End-to-End Reasoning-based Planning for Multi-step Retrosynthesis via Reinforcement Learning ► *Underline Video*
YiFei Wang, Qizhi Pei, Jiangtao Feng, Yuntian Shi, Yi Duan, Lihao Wang, LEI BAI, Lijun Wu, Wei-Ying Ma and Hao Zhou
Multi-step retrosynthetic planning is a fundamental challenge in organic chemistry, traditionally modeled as a combinatorial search problem guided by single-step prediction models. However, this search-centric paradigm often disconnects from the explicit chemical reasoning processes employed by human experts. In this paper, we propose R^3 (Reinforced Reasoning Retrosynthesis), a novel framework that reformulates this task as end-to-end generative reasoning. Instead of traversing a search tree, R^3 simulates the problem-solving logic of chemists to directly generate complete synthetic pathways. To achieve this, we initialize the model with domain knowledge and employ end-to-end Reinforcement Learning (RL) to optimize the entire planning policy. Experimental results on Retrobench show that R^3 achieves a state-of-the-art Top-1 accuracy of 43.7%, demonstrating that generative reasoning offers a superior alternative to traditional search algorithms in solving complex retrosynthetic problems.

[Long] A Functionality-Grounded Benchmark for Evaluating Web Agents in E-commerce Domains ► *Underline Video*
Xianren Zhang, Shreyas Prasad, Di Wang, Qiuhai Zeng, Suhang Wang, Wenbo Yan and Mat Hans
Web agents have shown great promise in performing many tasks on e-commerce websites. To assess their capabilities, several benchmarks have been introduced. However, current benchmarks in the e-commerce domain face two major problems. First, they primarily focus on product search tasks (e.g., "Find an Apple Watch"), failing to capture the broader range of functionalities offered by real-world e-commerce services such as Amazon, including account management and gift card operations. Second, existing benchmarks typically evaluate whether the agent completes the user query, but ignore the potential risks involved. In practice, web agents can make unintended changes that negatively impact the users account or status. For instance, an agent might purchase the wrong item, delete a saved address, or incorrectly configure an auto-reload setting. To address these gaps, we propose a new benchmark called Amazon-Bench. To generate user queries that cover a broad range of tasks, we propose a data generation pipeline that leverages webpage content and interactive elements (e.g., buttons, check boxes) to create diverse, functionality-grounded user queries covering tasks such as address management, wishlist management, and brand store following. To enhance agent evaluation, we propose an automated evaluation framework that assesses both the performance and safety of web agents. We systematically evaluate various agents, finding that current agents struggle with complex queries and pose safety risks. These results highlight the need for developing more robust and reliable web agents.

[Long] A Lightweight Explainable Guardrail for Prompt Safety ► *Underline Video*
Md Asiful Islam and Mihai Surdeanu
We propose a lightweight explainable guardrail (LEG) method to detect unsafe prompts. LEG uses a multi-task learning architecture to jointly learn a prompt classifier and an explanation classifier, where the latter labels prompt words that explain the safe/unsafe overall decision. LEG is trained on synthetic explanation data, which is generated using a novel strategy that counteracts the confirmation biases of LLMs. Lastly, LEG's training process uses a novel loss that captures global explanation signals as a weak supervision and combines

cross-entropy and focal losses with uncertainty-based weighting. LEG obtains equivalent or better performance than the state-of-the-art for both prompt classification and explainability, both in-domain and out-of-domain on three datasets, despite the fact that its model size is considerably smaller than current approaches.

[Findings] A Reward-Guided Dual-Phase Framework for Adaptive Inference-Time Reasoning ▶ *Underline Video*

Yingqian Cui, Zhenwei Dai, Pengfei He, Bing He, Hui Liu, Zhan Shi, Xianfeng Tang, Jingying Zeng, Suhang Wang, Yue Xing, Jiliang Tang and Benoit Dumoulin

Large Language Models (LLMs) have made strong progress in reasoning. To enhance the reasoning performance, a common inference-time approach is tree-based search, which decomposes the reasoning process into multiple steps, expands multiple reasoning paths, and uses reward models to prune and select candidates. However, based on our exploration, the simple decomposition may lead to suboptimal searching efficiency: while planning is generally harder, it is the execution errors that are more likely to propagate to later steps. This indicates that planning and execution play different roles in reasoning and should be treated differently during tree-based search. Given this, to enhance the searching efficiency, we propose a dual-phase test-time scaling framework that separates reasoning into planning and execution, and performs search over each phase independently. To further refine the algorithm, we also introduce a dynamic budget allocation mechanism that adaptively redistributes sampling effort based on reward feedback, allowing early stopping on confident steps and reallocation of computation to more challenging steps. Experiments on both math reasoning and code generation benchmarks demonstrate that our approach consistently improves accuracy while reducing redundant computation.

[Long] A Survey of Multimodal Mathematical Reasoning: From Perception, Alignment to Reasoning ▶ *Underline Video*

Tianyu Yang, Sihong Wu, Yilun Zhao, Zhenwen Liang, Lisen Dai, Chen Zhao, Minhao Cheng, Arman Cohan and Xiangliang Zhang

Multimodal Mathematical Reasoning (MMR) has recently attracted increasing attention for its capability to solve mathematical problems involving both textual and visual modalities. However, current models still face significant challenges in real-world visual math tasks, often misinterpreting diagrams, failing to align mathematical symbols with visual evidence, or producing inconsistent reasoning steps. Moreover, existing evaluations mainly focus on checking final answers rather than verifying the correctness or executability of each intermediate step. A growing body of recent research addresses these issues by integrating structured perception, explicit alignment, and verifiable reasoning within unified frameworks. To establish a clear roadmap for understanding and comparing different MMR approaches, we systematically review them around four fundamental questions: (1) What to extract from multimodal inputs, (2) How to represent and align textual and visual information, (3) How to perform the reasoning, and (4) How to evaluate the correctness of the overall reasoning process. Finally, we discuss open challenges and share our thoughts on future research directions.

[Findings] A Universal Avoidance Method for Diverse Multi-branch Generation ▶ *Underline Video*

Kyeongman Park, Minha Jhang and Kyomin Jung

Modern generative models still lack human-level creativity, particularly in multi-branch diversity. Prior approaches to address this problem often incur heavy computation or strong dependency on model architecture. Therefore, we introduce UAG(Universal Avoidance Generation), a model-agnostic and computationally efficient generation strategy that penalizes similarity among previously generated outputs. Thus, UAG can enhance multi-branch diversity across both diffusion and transformer models, with minimal additional computation. In experiments, our method achieves up to 1.9 times higher diversity, runs 4.4 times faster, and requires only 1/64 of the FLOPs compared to state-of-the-art methods.

[Findings] Activation Steering for Chain-of-Thought Compression ▶ *Underline Video*

Seyedarmin Azizi, Erfan Baghaei Potraghloo, Souvik Kundu and Massoud Pedram

Large language models (LLMs) demonstrate strong performance on multi-step reasoning tasks by producing intermediate explanations, commonly referred to as chains of thought (CoTs). However, the generated rationales are typically verbose, consuming many additional tokens, and thus degrading throughput and increasing inference energy consumption. Interestingly, we find that verbose and concise CoTs correspond to distinct regions in the model's intermediate activation space, suggesting that verbosity is a steerable latent attribute. Building on this observation, we develop an inference-time method to automatically steer the model response towards concise reasoning traces without updating model parameters. Our method, dubbed _ASC_ (Activation-Steered Compression), generates concise CoTs by directly adjusting internal representations via activation steering. A key component of ASC is Contrastive Energy-Based Steering (CES), a principled procedure to learn a _single_ steering vector from a small set of verbose-concise CoT pairs by optimizing a length-normalized contrastive energy objective. To further ensure reliable steering and preserve general utility, CES enforces a differentiable KL trust region during steering vector optimization, explicitly constraining the distribution shift within a specified budget. With only 100 pairs of verbose-concise examples, ASC reduces the generated token length by as much as 69.4% across five reasoning benchmarks (MATH500, GSM8K, LiveCodeBench, GSM8K-Hard, and AQuA-RAT) while maintaining accuracy across models with 1.5B, 7B, 8B, and 32B parameters. On MATH500, ASC achieves an end-to-end inference speed-up of 2.7x on an 8B model.

[Long] Agentic Oversight via Dialectic Reasoning ▶ *Underline Video*

Leonardo Ranaldi and Federico Ranaldi

Debate has emerged as a promising oversight mechanism for Large Language Models (LLMs) amid rising systemic complexity, particularly where models outperform human evaluators. Yet, Debate provides little verifiable evidence for its final judgments, and its scalability

remains largely unexplored. To make oversight grounded and scale as capabilities extend, we introduce an Agentic Oversight framework. By using Dialectic Argumentation as a reasoning function, we extend this paradigm to multilingual and multimodal spaces. We employ a weak-to-strong oversight approach based on two expert models that evaluate and defend contesting answers, while a third blind judge determines the winner using Dialectic Argumentation. Experts argue only for belief-consistent answers, founding the Debate on disagreements. We experimented with six tasks on our framework in both multilingual and multimodal scenarios, and dialectic argumentation consistently outperforms single-expert baselines. Moreover, we show that dialectic judgements from a weaker model deliver argument-mediated supervision that, via fine-tuning, instils unsupervised reasoning signals in expert models.

[Long] AIM-CoT: Active Information-driven Multimodal Chain-of-Thought for Vision-Language Reasoning ▶ *Underline Video*

Xiping Li and Jianghong Ma

Interleaved-Modal Chain-of-Thought (I-MCoT) advances vision-language reasoning, such as Visual Question Answering (VQA). This paradigm integrates specially selected visual evidence from the input image into the context of Vision-Language Models (VLMs), enabling them to ground their reasoning logic in these details. Accordingly, the efficacy of an I-MCoT framework relies on identifying what to see (evidence selection) and when to see it (triggering of insertions). However, existing methods fall short in both aspects. First, for selection, they rely on attention signals, which are unreliable—particularly under severe granularity imbalance between the brief textual query and the informative image. Second, for triggering, they adopt static triggers, which fail to capture the VLMs' dynamic needs for visual evidence. To this end, we propose a novel I-MCoT framework, Active Information-driven Multi-modal Chain-of-Thought (AIM-CoT), which aims to improve both evidence selection and insertion triggering via: (1) Context-enhanced Attention-map Generation (CAG) to mitigate granularity imbalance via textual context enhancement; (2) Active Visual Probing (AVP) to proactively select the most informative evidence via an information foraging process; and (3) Dynamic Attention-shift Trigger (DAT) to precisely activate insertions when VLM's attention shifts from text to visual context. Experiments across three benchmarks and four backbones demonstrate AIM-CoT's consistent superiority. Our code is available at https://anonymous.4open.science/r/AIMCoT.

[Long] Aligning Language Models with Real-time Knowledge Editing ▶ *Underline Video*

Chenming Tang, Yutong Yang, Kexue Wang and Yunfang Wu

Knowledge editing aims to modify outdated knowledge in language models efficiently while retaining their original capabilities. Mainstream datasets for knowledge editing are predominantly static and fail to keep in pace with the evolving real-world knowledge. In this work, we introduce CRAFT, an ever-evolving real-world dataset for knowledge editing. It evaluates models on temporal locality, common-sense locality, composite portability and alias portability, providing a comprehensive and challenging evaluation for knowledge editing, on which previous methods hardly achieve balanced performance. Towards flexible real-time knowledge editing, we propose KEDAS, a novel paradigm of knowledge editing alignment featuring diverse edit augmentation and self-adaptive post-alignment inference, exhibiting significant performance gain on both CRAFT and traditional datasets compared to previous methods. We hope this work may serve as a catalyst for shifting the focus of knowledge editing from static update to dynamic evolution.

[Findings] Aligning What LLMs Do and Say: Towards Self-Consistent Explanations ▶ *Underline Video*

Sahar Admoni, Ofra Amir, Assaf Hallak and Yftah Ziser

Large language models (LLMs) seem to offer an easy path to interpretability: just ask them to explain their answers. Yet the features driving an answer often differ from those emphasized in its explanation, meaning post-hoc rationales can misrepresent what actually shaped the model's output. We quantify this gap by comparing the feature-importance distributions of answers and their explanations. Prior analyses reveal such discrepancies, but large-scale study has been limited by the high computational cost of attribution methods. To address this, we introduce the Post-hoc Self-Consistency Bank (PSCB), a large-scale benchmark linking model decisions with diverse explanations and attribution vectors across datasets, methods, and model families. Using PSCB, we find that Spearman rank correlation provides a more reliable signal of alignment than cosine similarity. Building on this insight, we apply Direct Preference Optimization (DPO) to attribution-based preference data, improving alignment without degrading task accuracy, and show that standard supervised fine-tuning on the same data fails to achieve comparable gains. These improvements generalize robustly across domains, paving the way toward scalable and faithful alignment between LLM decisions and their natural language explanations.

[Findings] An Answer is just the Start: Related Insight Generation for Open-Ended Document-Grounded QA ▶ *Underline Video*

Saransh Sharma, Pritika Ramu, Aparna Garimella and Koyel Mukherjee

Answering open-ended questions remains challenging for AI systems because it requires synthesis, judgment, and exploration beyond factual retrieval, and users often refine answers through multiple iterations rather than accepting a single response. Existing QA benchmarks do not explicitly support this refinement process. To address this gap, we introduce a new task, document-grounded related insight generation, where the goal is to generate additional insights from a document collection that help improve, extend, or rethink an initial answer to an open-ended question, ultimately supporting richer user interaction and a better overall question answering experience. We curate and release SCOpE-QA (Scientific Collections for Open-Ended QA), a dataset of 3,000 open-ended questions across 20 research collections. We present InsightGen, a two-stage approach that first constructs a thematic representation of the document collection using clustering, and then selects related context based on neighborhood selection from the thematic graph to generate diverse and relevant insights using LLMs. Extensive evaluation on 3,000 questions using two generation models and two evaluation settings shows that InsightGen consistently produces useful, relevant, and actionable insights, establishing a strong baseline for this new task.

[Long] Are LLMs Reliable Rankers? Rank Manipulation via Two-Stage Token Optimization ► *Underline Video*

Tiancheng Xing, Jerry Li, Yixuan Du and Xiyang Hu

Large language models (LLMs) are increasingly used as rerankers in information retrieval, yet their ranking behavior can be steered by small, natural-sounding prompts. To expose this vulnerability, we present Rank Anything First (RAF), a two-stage token optimization method that crafts concise textual perturbations to consistently promote a target item in LLM-generated rankings while remaining hard to detect. Stage 1 uses Greedy Coordinate Gradient to shortlist candidate tokens at the current position by combining the gradient of the rank-target with a readability score; Stage 2 evaluates those candidates under exact ranking and readability losses using an entropy-based dynamic weighting scheme, and selects a token via temperature-controlled sampling. RAF generates ranking-promoting prompts token-by-token, guided by dual objectives: maximizing ranking effectiveness and preserving linguistic naturalness. Experiments across multiple LLMs show that RAF significantly boosts the rank of target items using naturalistic language, with greater robustness than existing methods in both promoting target items and maintaining naturalness. These findings underscore a critical security implication: LLM-based reranking is inherently susceptible to adversarial manipulation, raising new challenges for the trustworthiness and robustness of modern retrieval systems. Our code is available at: https://github.com/glad-lab/RAF.

[Findings] Attribution-Based Analysis and Optimization of Modular Agentic Workflows ► *Underline Video*

Yingxuan Yang, Bo Huang, Siyuan Qi, chao feng, Haoyi Hu, Yuxuan Zhu, Jinbo Hu, Haoran Zhao, Ziyi He, xiao liu, ZongYu Wang, Muning Wen, Lin Qiu, Xuezhi Cao, Xunliang Cai, Yong Yu and Weinan Zhang

Agentic workflows solve complex tasks by orchestrating modular components (e.g., planning, reasoning, action, reflection) built on top of LLM backbones. A practical but underexplored question is model allocation: given a fixed workflow decomposition and a pool of candidate LLMs, which components should be upgraded (and with which models) to upgrade task performance, and how can we attribute gains to individual upgrades and their interactions? We present ShapleyFlow, a cooperative game theoretic framework that models component upgrades as players and evaluates component coalitions to compute Shapley values. This yields interaction-aware attribution and supports Shapley-guided configuration recommendation for model allocation under a fixed workflow structure. We further introduce CapaBench, a benchmark of 1,500+ tasks across seven domains (shopping, navigation, ticketing, mathematics, operating systems, robotic coordination, and automated theorem proving). Across 9 representative LLMs and all 2^4 upgrade coalitions in a 4-component workflow, ShapleyFlow provides (i) principled, interaction-aware attribution for modular workflows and (ii) actionable model-allocation recommendations that improve over strong single-model baselines.

[Findings] AutoAgent: A Fully-Automated and Zero-Code Framework for LLM Agents ► *Underline Video*

Jiabin Tang, Tianyu Fan and Chao Huang

Large Language Model (LLM) Agents have demonstrated remarkable capabilities in task automation and intelligent decision-making, driving the widespread adoption of agent development frameworks such as LangChain and AutoGen. However, these frameworks predominantly serve developers with extensive technical expertise, a significant limitation considering that only 0.03% of the global population possesses the necessary programming skills. This stark accessibility gap raises a fundamental question: Can we enable everyone, regardless of technical background, to build their own LLM agents using natural language alone? To address this challenge, we introduce AutoAgent - a Fully-Automated and highly Self-Developing framework that enables users to create and deploy LLM agents through Natural Language Alone. Operating as an autonomous Agent Operating System, AutoAgent comprises four key components: i) Agentic System Utilities, ii) LLM-powered Actionable Engine, iii) Self-Managing File System, and iv) Self-Play Agent Customization module. This lightweight yet powerful system enables efficient and dynamic creation and modification of tools, agents, and workflows without coding requirements or manual intervention. Beyond its code-free agent development capabilities, AutoAgent also serves as a versatile multi-agent system for General AI Assistants. Comprehensive evaluations on the GAIA benchmark demonstrate AutoAgent's effectiveness in generalist multi-agent tasks, surpassing existing state-of-the-art methods. Furthermore, AutoAgent's Retrieval-Augmented Generation (RAG)-related capabilities have shown consistently superior performance compared to many alternative LLM-based solutions.

[Findings] Balancing Classification and Calibration Performance in Decision-Making LLMs via Calibration Aware Reinforcement Learning ► *Underline Video*

Duygu Nur Yaldiz, Evangelia Spiliopoulou, Zheng Qi, Siddharth Varia, Srikanth Doss and Nikolaos Pappas

Large language models (LLMs) are increasingly deployed in decision-making tasks, where not only accuracy but also reliable confidence estimates are essential. Well-calibrated confidence enables downstream systems to decide when to trust a model and when to defer to fallback mechanisms. In this work, we conduct a systematic study of calibration in two widely used fine-tuning paradigms: supervised fine-tuning (SFT) and reinforcement learning with verifiable rewards (RLVR). We show that while RLVR improves task performance, it produces extremely overconfident models, whereas SFT yields substantially better calibration, even under distribution shift, though with smaller performance gains. Through targeted experiments, we diagnose RLVRs failure, showing that decision tokens act as extraction steps of the decision in reasoning traces and do not carry confidence information, which prevents reinforcement learning from surfacing calibrated alternatives. Based on this insight, we propose a calibration-aware reinforcement learning formulation that directly adjusts decision-token probabilities. Our method preserves RLVRs accuracy level while mitigating overconfidence, reducing ECE scores up to 9 points.

[Long] Beyond Majority Voting: Towards Fine-grained and More Reliable Reward Signal for Test-Time Reinforcement

Learning ▶ *Underline Video*

Weiqin Wang, Yile Wang, Kehao Chen and Hui Huang

Test-time reinforcement learning mitigates the reliance on annotated data by using majority voting results as pseudo-labels, emerging as a complementary direction to reinforcement learning with verifiable rewards (RLVR) for improving reasoning ability of large language models (LLMs). However, this voting strategy often induces confirmation bias and suffers from sparse rewards, limiting the overall performance. In this work, we propose subgroup-specific step-wise confidence-weighted pseudo-label estimation (SCOPE), a framework integrating model confidence and dynamic subgroup partitioning to address these issues. Specifically, SCOPE integrates the proposed step-wise confidence into pseudo-label estimation, prioritizing high-quality reasoning paths over simple frequency count. Furthermore, it dynamically partitions the candidate outputs pool into independent subgroups by balancing reasoning quality against exploration diversity. By deriving local consensus via repeat sampling for each subgroup, SCOPE provides diverse supervision targets to encourage broader exploration. We conduct experiments across various models and benchmarks, experimental results show that SCOPE consistently outperforms recent baselines. Notably, SCOPE achieves relative improvements of 13.1% on challenging AIME 2025 and 8.1% on AMC.

[Long] Beyond Static Persona Consistency: Dynamic Persona Coherence in LLM Role-Playing ▶ *Underline Video*

Yirui Qi, Xiaoming Zhang, Ruilin Zeng, Mengyao Liu, Ziyi Zhou, Dezhuang Miao, Bingyu Yan and Zhenyu Guan

Current LLM role-playing systems model persona as a monolithic, static attribute, conflating identity consistency with emotional rigidity. This leads to either robotic repetition or catastrophic persona drift under sustained interaction. We introduce Dynamic Persona Coherence, a framework that decouples Identity-Layer Stability (time-invariant traits) from Adaptive-Layer Appropriateness (history-dependent psychological evolution). We operationalize this through the L/M/S Psychological State Model, which represents persona dynamics across long-term identity, mid-term meaning/stress accumulation, and short-term affect. On top of this state representation, a closed-loop alignment system comprising an automated evaluator (Persona Consistency Critic, PCC), a selective repository (Persona Case Repository, PCR), and a trajectory-adjusting corrector (Persona Drift Suppressor, PDS) enables autonomous coherence repair. Experiments on GPT-4o, Claude-3.5-Sonnet, and DeepSeek-V3.2 demonstrate consistent improvements (+16-84% PCC gains).

[Long] Beyond the Context Window: Scaling Agentic RL via End-to-end Optimized Context Compression ▶ *Underline Video*

Miao Lu, Weiwei Sun, Weihua Du, Zhan Ling, Xuesong Yao, Kang Liu and Jiecao Chen

We study reinforcement learning (RL) fine-tuning of large language model (LLM) agents for long-horizon multi-turn tool use, where context length quickly becomes a fundamental bottleneck. Existing multi-turn RL pipelines suffer from degraded instruction following, excessive rollout costs, and most importantly, strict context limits. In this work, to address these challenges, we introduce summarization-based context management to training. In specific, it periodically compresses the tool using history by LLM-generated summaries that retain task-relevant information to keep a compact context while enabling the agent to scale beyond the fixed context window. Building on this formulation, we derive a policy gradient representation that seamlessly enables standard LLM RL infrastructures to optimize both tool-use behaviors as well as summarization strategies in an end-to-end fashion. We instantiate this framework with SUmmarization augmented Policy Optimization (SUPO), an LLM RL algorithm that enables long-horizon training beyond a fixed context limit. Experiments on interactive function calling and searching tasks demonstrate that SUPO significantly improves the success rate while maintaining the same or even lower working context length compared to baselines. We also demonstrate that for complex searching tasks SUPO can further improve the evaluation performance when scaling test-time maximum round of summarization beyond that of training time.

[Findings] BiCSRouter: Bi-Level Cross-System Routing for Utility-Aware LLM Inference ▶ *Underline Video*

KEYU MAO, Eiki Murata and Ukyo Honda

Selecting an appropriate LLM configuration for a given query is critical, yet existing routing frameworks operate within a single computational paradigm. To address this gap, we formalize the Cross-System Routing Problem, a hierarchical decision-making task that decomposes routing into intra-regime configuration selection and inter-regime system selection. Building on this, we propose BiCSRouter, a bi-level cross-system routing framework that integrates two orthogonal regimes: intensive reasoning via single-agent systems and extensive collaboration via multi-agent systems. BiCSRouter performs policy learning within each system and employs a lightweight inter-regime router that selects the optimal regime based on predicted performance and cost. Experiments on the MBPP and MATH benchmarks demonstrate that BiCSRouter outperforms 15 representative baselines across three types. On MBPP, compared to the performance ceiling of GPT-5, BiCSRouter achieves a 46% reduction in cost with only a 2% drop in accuracy. Finally, we show that BiCSRouter can extend to additional regimes, highlighting its generality as a cross-system routing framework.

[Long] Bit-by-Bit: Progressive QAT Strategy with Outlier Channel Splitting for Stable Low-Bit LLMs ▶ *Underline Video*

Binxing Xu, Hao Gu, Lujun Li, Hao Wang, Bei Liu, Jiacheng Liu, Qiyuan Zhu, Xintong Yang, Chao Li, Sirui Han and Yike Guo

Training LLMs at ultra-low precision remains a formidable challenge. Direct low-bit QAT often suffers from convergence instability and substantial training costs, exacerbated by quantization noise from heavy-tailed outlier channels and error accumulation across layers. To address these issues, we present Bit-by-Bit, a progressive QAT framework with outlier channel splitting. Our approach integrates three key components: (1) block-wise progressive training that reduces precision stage by stage, ensuring stable initialization for low-bit optimization; (2) nested structure of integer quantization grids to enable a "train once, deploy any precision" paradigm, allowing a single model to support multiple bit-widths without retraining; (3) rounding-aware outlier channel splitting, which mitigates quantization error while acting as an identity transform that preserves the quantized outputs. Furthermore, we follow microscaling groups with E4M3 scales, capturing dynamic

activation ranges in alignment with OCP/NVIDIA standards. To address the lack of efficient 2-bit kernels, we developed custom operators for both W2A2 and W2A16 configurations, achieving up to 11x speedup over BF16. Under W2A2 settings, Bit-by-Bit significantly outperforms baselines like BitDistiller and EfficientQAT on both Llama2/3, achieving a loss of only 2.25 WikiText2 PPL compared to full-precision models.

[Long] BOSCH: Black-Box Binary Optimization for Short-Context Attention-Head Selection in LLMs ▶ *Underline Video*

Abbas Ghaddar, Ivan Kobyzev, Boxing Chen and Yufei Cui

Post-training hybridization of large language models (LLMs) often replaces quadratic self-attention with sliding-window attention (SWA) to reduce KV cache usage and improve latency. Existing hybridization schemes are typically defined either at the layer level (e.g., interleaving) or at the head level via static rankings from local to global. Layer-level schemes ignore that local and global dependencies are routed through heads within the same layer, while static head-level rankings suffer from entanglement: a heads local/global behavior can change after hybridization. We propose BOSCH, Black-box Binary Optimization for Short-context Head Selection, a training-free method that formulates the problem as a Large Neighborhood Search and decomposes it into three subproblems: (i) layer-importance detection via small-budget black-box probes, (ii) adaptive per-layer SWA-ratio assignment based on these sensitivities, and (iii) grouped head-level optimization within ratio buckets. Extensive experiments on 4 LLMs ranging from 1.7B to 30B parameters, across 4 SWA ratios, show that BOSCH consistently outperforms layer-level heuristics and 6 strong static head-level methods, with larger gains at higher SWA ratios. Under continual pretraining, BOSCH recover original long-context performance faster and to a higher level. Analysis of the selected heads reveals substantial turnover for BOSCH across different SWA ratios, underscoring the importance of performing head-level selection for each target ratio rather than relying on fixed locality rankings.

[Long] Branch-and-Browse: Efficient and Controllable Web Exploration with Tree-Structured Reasoning and Action Memory ▶ *Underline Video*

Shiqi He, Yue Cui, Xinyu Ma, Yaliang Li, Bolin Ding and Mosharaf Chowdhury

Autonomous web agents powered by large language models (LLMs) show strong potential for performing goal-oriented tasks such as information retrieval, report generation, and online transactions. These agents mark a key step toward practical embodied reasoning in open web environments. However, existing approaches remain limited in reasoning depth and efficiency: vanilla linear methods fail at multi-step reasoning and lack effective backtracking, while other search strategies are coarse-grained and computationally costly. We introduce Branch-and-Browse, a fine-grained web agent framework that unifies structured reasoning-acting, contextual memory, and efficient execution. It (i) employs explicit subtask management with tree-structured exploration for controllable multi-branch reasoning, (ii) bootstraps exploration through efficient web state replay with background reasoning, and (iii) leverages a page action memory to share explored actions within and across sessions. On the WebArena benchmark, Branch-and-Browse achieves a task success rate of 35.8% and reduces execution time by up to 40.4% relative to state-of-the-art methods. These results demonstrate that Branch-and-Browse is a reliable and efficient framework for LLM-based web agents. Code is available at https://anonymous.4open.science/r/Branch_and_Browse/.

[Long] C-World: A Computer Use Agent Environment Creator ▶ *Underline Video*

Ziqiao Xi, Shuang Liang, Qi Liu, Jiaqing Zhang, Letian Peng, Fang Nan, Meshal Nayim, Tianhui Zhang, Rishika Mundada, Lianhui Qin, Biwei Huang and Kun Zhou

To close the gap between LLM-based agents and humans in planning and reasoning, agents need large-scale, diverse environments for continuous learning—yet building such environments is itself prohibitively expensive. We present C-World, an environment creation system that enables users to build agent environments on demand. We define a complete agent environment through four components: an Action Space of 5,571 format-unified tools across 204 common applications, a Task Distribution engine that synthesizes long-horizon workflows with wild constraints, a Transition Function implemented as a state controller that injects realistic failures and perturbations, and a Reward Signal combining verifiable metrics with LLM-based judgment. C-World operates in two modes: a realistic mode grounded in live API execution, and a synthesized mode powered by the World Engine, which approximates tool behavior without live service access, enabling scalable environment creation—including environments for domains and tools that do not yet exist in the real world. Evaluation of nine state-of-the-art LLMs reveals that planning ability is uniformly strong but execution remains the bottleneck, and that constraint following—not tool invocation—is the dominant failure mode. The World Engine achieves Spearman = 0.883 ranking correlation with real execution, and fine-tuning on just 1,170 C-World trajectories outperforms baselines trained on 119k samples, demonstrating C-World's dual value as a rigorous evaluation environment and a scalable data engine. Our code and data are available at https://ziqiao-git.github.io/C-World/.

[Long] Cell-Based Representation of Relational Binding in Language Models ▶ *Underline Video*

Qin Dai, Benjamin Heinzerling and Kentaro Inui

Understanding a discourse requires tracking entities and the relations that hold between them. While Large Language Models (LLMs) perform well on relational reasoning, the mechanism by which they bind entities, relations, and attributes remains unclear. We study discourse-level relational binding and show that LLMs encode it via a Cell-based Binding Representation (CBR): a low-dimensional linear subspace in which each "cell" corresponds to an entity–relation index pair, and bound attributes are retrieved from the corresponding cell during inference. Using controlled multi-sentence data annotated with entity and relation indices, we identify the CBR subspace by decoding these indices from attribute-token activations with Partial Least Squares regression. Across domains and two model families, the indices are linearly decodable and form a grid-like geometry in the projected space. We further find that context-specific CBR representations are related

by translation vectors in activation space, enabling cross-context transfer. Finally, activation patching shows that manipulating this subspace systematically changes relational predictions and that perturbing it disrupts performance, providing causal evidence that LLMs rely on CBR for relational binding.

[Findings] Chain-of-Procedure: Hierarchical Visual-Language Reasoning for Procedural QA ▶ *Underline Video*

Guanhua Chen, Yutong Yao, Shenghe Sun, Ci-Jun Gao, Shudong Liu, Lidia S. Chao, Feng Wan and Derek F. Wong

Recent advances in vision-language models (VLMs) have achieved impressive results on standard image-text tasks, yet their potential for visual procedure question answering (VP-QA) remains largely unexplored. VP-QA presents unique challenges where users query next-step actions by uploading images for intermediate states of complex procedures. To systematically evaluate VLMs on this practical task, we propose ProcedureVQA, a novel multimodal benchmark specifically designed for visual procedural reasoning. Through comprehensive analysis, we identify two critical limitations in current VLMs: inadequate cross-modal retrieval of structured procedures given visual states, and misalignment between image sequence granularity and textual step decomposition. To address these issues, we present Chain-of-Procedure (CoP), a hierarchical reasoning framework that first retrieves relevant instructions using visual cues, then performs step refinement through semantic decomposition, and finally generates the next step. Experiments across six VLMs demonstrate CoP's effectiveness, achieving up to 13% absolute improvement over standard baselines.

[Findings] Characterizing the Robustness of Black-Box LLM Planners Under Perturbed Observations with Adaptive Stress Testing ▶ *Underline Video*

Neeloy Chakraborty, John Pohovey, Melkior Ornik and Katherine Driggs-Campbell

Large language models (LLMs) have recently demonstrated success in decision-making tasks including planning, control, and prediction, but their tendency to hallucinate unsafe and undesired outputs poses risks. This unwanted behavior is further exacerbated in environments where sensors are noisy or unreliable. Characterizing the behavior of LLM planners to varied observations is necessary to proactively avoid failures in safety-critical scenarios. We specifically investigate the response of LLMs along two different perturbation dimensions. Like prior works, one dimension generates semantically similar prompts with varied phrasing by randomizing order of details, modifying access to few-shot examples, etc. Unique to our work, the second dimension simulates access to varied sensors and noise to mimic raw sensor or detection algorithm failures. An initial case study in which perturbations are manually applied show that both dimensions lead LLMs to hallucinate in a multi-agent driving environment. However, manually covering the entire perturbation space for several scenarios is infeasible. As such, we propose a novel method for efficiently searching the space of prompt perturbations using adaptive stress testing (AST) with Monte-Carlo tree search (MCTS). Our AST formulation enables discovery of scenarios, sensor configurations, and prompt phrasing that cause language models to act with high uncertainty or even crash. By generating MCTS prompt perturbation trees across diverse scenarios, we show through extensive experiments that offline analyses can be used to proactively understand potential failures that may arise at runtime. Code is available at https://sites.google.com/illinois.edu/astllm/.

[Long] ChartVerse: Scaling Chart Reasoning via Reliable Programmatic Synthesis from Scratch ▶ *Underline Video*

Zheng Liu, Honglin Lin, Xiaoyang Wang, Xin Gao, Yu Li, Mengzhang Cai, Yun Zhu, Zhanping Zhong, Qizhi Pei, Zhuoshi Pan, Xiaoran Shang, Conghui He, Bin Cui, Wentao Zhang and Lijun Wu

Chart reasoning is a critical capability for Vision Language Models (VLMs). However, the development of open-source models is severely hindered by the lack of high-quality training data. Existing datasets suffer from a dual challenge: synthetic charts are often simplistic and repetitive, while the associated QA pairs are prone to hallucinations and lack the reasoning depth required for complex tasks. To bridge this gap, we propose ChartVerse, a scalable framework designed to synthesize complex charts and reliable reasoning data from scratch. (1) To address the bottleneck of simple patterns, we first introduce Rollout Posterior Entropy (RPE), a novel metric that quantifies chart complexity. Guided by RPE, we develop complexity-aware chart coder to autonomously synthesize diverse, high-complexity charts via executable programs. (2) To guarantee reasoning rigor, we develop truth-anchored inverse QA synthesis. Diverging from standard generation, we adopt an answer-first paradigm: we extract deterministic answers directly from the source code, generate questions conditional on these anchors, and enforce strict consistency verification. To further elevate difficulty and reasoning depth, we filter samples based on model fail-rate and distill high-quality Chain-of-Thought (CoT) reasoning. We curate ChartVerse-SFT-600K and ChartVerse-RL-40K using Qwen3-VL-30B-A3B-Thinking as the teacher. Experimental results demonstrate that ChartVerse-8B achieves state-of-the-art performance, notably surpassing its teacher and rivaling the stronger Qwen3-32B-Thinking.

[Findings] Chimera: Compositional Jailbreak Attacks on LLMs via Judgment-Driven Search over Heterogeneous Strategies ▶ *Underline Video*

Leo Hyun Park, Juwon Cho, Gyuhwan Kim, YoonDong Yeo and Taekyoung Kwon

Large Language Models (LLMs) remain vulnerable to jailbreak attacks despite extensive safety alignment. While automated red-teaming has emerged as a critical evaluation protocol, existing methods face two primary limitations: they largely explore homogeneous transformations in isolation, and they rely on brittle judgment metrics that frequently misclassify non-refusal hallucinations as successful attacks. In this paper, we reformulate jailbreak attacks as a compositional search problem guided by context-aware evaluation. We propose Chimera, a framework that generates compositional jailbreak attacks via judgment-driven search over heterogeneous strategies. Chimera systematically explores the combinatorial space of disjoint primitives, such as integrating technical obfuscation with semantic persuasion, under strict ordering constraints. Crucially, to drive the search process effectively, we introduce StrongREJECT++, a relevance-aware metric that

eliminates false positive rewards by penalizing irrelevant responses. Experiments on multiple open-source and commercial LLMs show that Chimera uncovers qualitatively different vulnerability regions and consistently improves attack success rates and transferability compared to state-of-the-art baselines.

[Findings] CLAG: Adaptive Memory Organization via Agent-Driven Clustering for Small Language Model Agents ▶ *Underline Video*
Taeyun Roh, Wonjune Jang, Junha Jung and Jaewoo Kang
Large language model agents heavily rely on external memory to support knowledge reuse and complex reasoning tasks. Yet most memory systems store experiences in a single global retrieval pool which can gradually dilute or corrupt stored knowledge. This problem is especially pronounced for small language models (SLMs), which are highly vulnerable to irrelevant context. We introduce CLAG, a CLustering-based AGentic memory framework where an agent actively organizes memory. CLAG employs an SLM-agent driven router to assign each new memory to a semantically coherent cluster. By performing continual evolution within the cluster, it effectively reduces cross-topic interference. During the retrieval phase, CLAG targets a small set of relevant clusters for retrieval, thereby excluding distractors and reducing the search space. Experiments on multiple QA datasets with three SLM backbones show that CLAG consistently improves answer quality and robustness over prior memory systems for agents, remaining lightweight and efficient.

[Long] CodeJudgeBench: Benchmarking LLM-as-a-Judge for Coding Tasks ▶ *Underline Video*
Hongchao Jiang, Yiming Chen, Yushi Cao, Hung-yi Lee and Robby T. Tan
Large Language Models (LLMs) are increasingly used not only to generate code, but also to judge it: comparing, ranking, or scoring competing solutions. However, their reliability in this evaluative role remains poorly understood. Inconsistent or flawed judgments can undermine benchmarks and distort training signals. This paper investigates the performance and robustness of LLMs when used as code judges. We introduce CodeJudgeBench, a benchmark explicitly designed to evaluate LLM-as-a-Judge models across three critical coding tasks: code generation, code repair, and unit test generation. We comprehensively benchmark the performance of 26 LLM-as-a-Judge models, encompassing general-purpose, code-tuned, and reasoning models. Our empirical findings reveal that relatively small reasoning models (e.g., Qwen3-8B) can outperform much larger non-reasoning models up to 70B. We further stress-test robustness by applying both general and code-specific perturbations. All models show significant instability and are sensitive to changes such as response ordering, variable naming, and misleading comments. These findings highlight serious concerns about the consistency and robustness of LLM-based judges for coding tasks.

[Findings] COMPEL: Compensated Mixture-of-Experts Pruning with Expert-Layer distribution ▶ *Underline Video*
Seohee Yoon and Yong Suk Choi
Mixture-of-Experts (MoE) architectures have emerged as an effective approach for scaling Large Language Models (LLMs) by activating only a subset of experts during inference. Despite their computational efficiency, MoE models incur a substantial memory bottleneck from maintaining all expert parameters during inference. To address this challenge, numerous MoE pruning methods have been proposed. However, most existing methods adopt uniform pruning across layers, which fails to capture layer-wise variations in expert importance and redundancy. In this paper, we propose COmpensated MoE Pruning with Expert-Layer distribution (COMPEL). COMPEL performs layer-adaptive expert pruning by estimating expert importance using Fisher information and deriving layer importance from layer-wise outlier distributions, enabling pruning decisions that capture layer-wise heterogeneity. Furthermore, to mitigate performance degradation resulting from expert pruning, we propose a Fisher information guided expert weight compensation method. Experimental results on the Qwen1.5-MoE-A2.7B achieve near lossless performance at 25% expert pruning and maintains performance within a 4% margin even at 50% pruning. Moreover, COMPEL consistently outperforms existing pruning methods while substantially reducing inference latency and peak GPU memory usage.

[Long] Compositional Steering of Large Language Models with Steering Tokens ▶ *Underline Video*
Gorjan Radevski, Kiril Gashteovski, Giwon Hong, Carolin Lawrence and Goran Glava
Deploying LLMs in real-world applications requires controllable output that satisfies multiple desiderata at the same time. While existing work extensively addresses LLM steering for a single behavior, compositional steering—i.e., steering LLMs simultaneously towards multiple behaviors—remains an underexplored problem. In this work, we propose compositional steering tokens for multi-behavior steering. We first embed individual behaviors, expressed as natural language instructions, into dedicated tokens via self-distillation. Contrary to most prior work, which operates in the activation space, our behavior steers live in the space of input tokens, enabling more effective zero-shot composition. We then train a dedicated composition token on pairs of behaviors and show that it successfully captures the notion of composition: it generalizes well to unseen compositions, including those with unseen behaviors as well as those with an unseen number of behaviors. Our experiments across different LLM architectures show that steering tokens lead to superior multi-behavior steering of verifiable constraints (e.g., length, format, structure, language) compared to competing approaches (instructions, activation steering, and LoRA merging). Moreover, we show that steering tokens complement natural language instructions, with their combination resulting in further gains.

[Findings] Conceptual Hierarchies within LLMs ▶ *Underline Video*
Tiago Almeida, Zining Zhu and Yue Ning
While it is widely agreed that large language models (LLMs) store concepts from multiple semantic hierarchies, much remains unknown regarding the structure of this storage. The correspondence between the functional roles of LLM components and the semantic hierarchies

of knowledge remains underexplored in the current literature. For example, is information organized hierarchically within sections of an LLM? We take an initial step towards causally examining the correspondence between hierarchical concepts and the multi-granular structures (layers and attention heads) of various models. Specifically, we generate a dataset of semantic hierarchies and investigate their storage locations in six LLMs using activation patching, a causal intervention technique. At the layer level, our findings show a moderate indication that concepts at finer levels of granularity are stored around 61-78% of the time ($p < 0.01$) before those at coarser granularity. There is evidence for this trend at the attention level; however, the high variability in attention level results suggests that concepts are stored across attention heads rather than within. Our results offer insight into semantic organization within LLMs.

[Long] ConsistRM: Improving Generative Reward Models via Consistency-Aware Self-Training ▶ ***Underline Video***

Yu Liang, Liangxin Liu, Longzheng Wang, Yan Wang, Yueyang Zhang, Long Xia, Zhiyuan Sun and Daiting Shi

Generative reward models (GRMs) have emerged as a promising approach for aligning Large Language Models (LLMs) with human preferences by offering greater representational capacity and flexibility than traditional scalar reward models. However, GRMs face two major challenges: reliance on costly human-annotated data restricts scalability, and self-training approaches often suffer from instability and vulnerability to reward hacking. To address these issues, we propose ConsistRM, a self-training framework that enables effective and stable GRM training without human annotations. ConsistRM incorporates the Consistency-Aware Answer Reward, which produces reliable pseudo-labels with temporal consistency, thereby providing more stable model optimization. Moreover, the Consistency-Aware Critique Reward is introduced to assess semantic consistency across multiple critiques and allocates fine-grained and differentiated rewards. Experiments on five benchmark datasets across four base models demonstrate that ConsistRM outperforms vanilla Reinforcement Fine-Tuning (RFT) by an average of 1.5%. Further analysis shows that ConsistRM enhances output consistency and mitigates position bias caused by input order, highlighting the effectiveness of consistency-aware rewards in improving GRMs.Our implementation is available at https://github.com/yuliangCarmelo/ConsistRM.

[Long] Constructing Interpretable Features from Compositional Neuron Groups ▶ ***Underline Video***

Or Shafran, Atticus Geiger and Mor Geva

A central goal for mechanistic interpretability has been to identify the right units of analysis in large language models (LLMs) that causally explain their outputs. While early work focused on individual neurons, evidence that neurons often encode multiple concepts has motivated a shift toward analyzing directions in activation space. A key question is how to find directions that capture interpretable features in an unsupervised manner. Current methods rely on dictionary learning with sparse autoencoders (SAEs), commonly trained over residual stream activations to learn directions from scratch. However, SAEs often struggle in causal evaluations and lack intrinsic interpretability, as their learning is not explicitly tied to the computations of the model. Here, we tackle these limitations by directly decomposing MLP activations with semi-nonnegative matrix factorization (SNMF), such that the learned features are (a) sparse linear combinations of co-activated neurons, and (b) mapped to their activating inputs, making them directly interpretable. Experiments on Llama 3.1, Gemma 2 and GPT-2 show that SNMF derived features outperform SAEs and a strong supervised baseline (difference-in-means) on causal steering, while aligning with human-interpretable concepts. Further analysis reveals that specific neuron combinations are reused across semantically-related features, exposing a hierarchical structure in the MLP's activation space. Together, these results position SNMF as a simple and effective tool for identifying interpretable features and dissecting concept representations in LLMs.

[Findings] Context-Fidelity Boosting: Enhancing Faithful Generation through Watermark-Inspired Decoding ▶ ***Underline Video***

Weixu Zhang, Fanghua Ye, Qiang Gao, Jian Li, Haolun Wu, Yuxing Tian, Sijing Duan, Nan Du and Xiaolong Li

Large language models (LLMs) often produce content that contradicts or overlooks information provided in the input context, a phenomenon known as faithfulness hallucination. In this paper, we propose Context-Fidelity Boosting (CFB), a lightweight and general decoding-time framework that effectively reduces such hallucinations by boosting the generation probability of context-relevant tokens. Motivated by logit-shaping principles in watermarking techniques, CFB leverages token-level logit adjustments based on their presence or salience in the input context. Specifically, we develop three boosting strategies, static, context-aware, and token-aware that progressively incorporate distributional divergence, attention scores, and semantic similarity. Notably, CFB requires no retraining or architectural changes, making it compatible with a wide range of LLMs. Experiments on summarization and question answering tasks across multiple open-source LLMs show that CFB consistently improves faithfulness metrics, with minimal generation overhead. Our implementation is fully open-sourced.

[Findings] Controlling What You Share: Assessing Language Model Adherence to Privacy Preferences ▶ ***Underline Video***

Guillem Ramirez and Alexandra BirchIvan Titov

Large language models (LLMs) are primarily accessed via commercial APIs, but this often requires users to expose their data to service providers. In this paper, we explore how users can stay in control of their data by using privacy profiles: simple natural language instructions that say what should and should not be revealed. We build a framework where a local model uses these instructions to rewrite queries, only hiding details deemed sensitive by the user, before sending them to an external model, thus balancing privacy with performance. To support this research, we introduce PEEP, a multilingual dataset of real user queries annotated to mark private content and paired with synthetic privacy profiles, alongside PROFIT, a training procedure that enables effective and efficient use of the pipeline. Experiments with lightweight local LLMs show that, after training, they not only achieve markedly better privacy preservation but also match or exceed the performance of much larger few-shot models.

[CL] Conundrum of Trustworthy Research on Attacking Personally Identifiable Information Removal Techniques ▶ *Underline Video*
Sebastian Ochs and Ivan Habernal
Removing personally identifiable information (PII) from texts is necessary to comply with various data protection regulations and to enable data sharing without compromising privacy. However, recent works show that documents sanitized by PII removal techniques are vulnerable to reconstruction attacks. Yet, we suspect that the reported success of these attacks is largely overestimated. We critically analyze the evaluation of existing attacks and find that data leakage and data contamination are not properly mitigated, leaving the question whether or not PII removal techniques truly protect privacy in real-world scenarios unaddressed. We investigate possible data sources and attack setups that avoid data leakage and conclude that only truly private data can allow us to objectively evaluate vulnerabilities in PII removal techniques. However, access to private data is heavily restricted, and for good reasons, which also means that the public research community cannot address this problem in a transparent, reproducible, and trustworthy manner.

[Long] Critical-CoT: A Robust Defense Framework against Reasoning-Level Backdoor Attacks in Large Language Models ▶ *Underline Video*
Vu Tuan Truong and Long Bao Le
Large Language Models (LLMs), despite its impressive capabilities across domains, have been shown to be vulnerable to backdoor attacks. Prior backdoor strategies predominantly operate at the token level, where an injected trigger causes the model to generate a specific target word, choice, or class (depending on the task). Recent advances, however, exploit the long-form reasoning tendencies of modern LLMs to conduct reasoning-level backdoors: once triggered, the victim model inserts one or more malicious reasoning steps into its chain-of-thought (CoT). These attacks are substantially harder to detect, as the backdoored answer remains plausible and consistent with the poisoned reasoning trajectory. Yet, defenses tailored to this type of backdoor remain largely unexplored. To bridge this gap, we propose Critical-CoT, a novel defense mechanism that conducts a two-stage fine-tuning (FT) on LLMs to develop critical thinking behaviors, enabling them to automatically identify potential backdoors and refuse to generate malicious reasoning steps. Extensive experiments across multiple LLMs and datasets demonstrate that Critical-CoT provides strong robustness against both in-context learning-based and FT-based backdoor attacks. Notably, Critical-CoT exhibits strong cross-domain and cross-task generalization. Our code is available at hthttps://github.com/tuanvu171/Critical-CoT.

[Long] Crosscoding Through Time: Tracking Emergence & Consolidation Of Linguistic Representations Throughout LLM Pretraining ▶ *Underline Video*
Deniz Bayazit, Aaron Mueller and Antoine Bosselut
Large language models (LLMs) learn non-trivial abstractions during pretraining, such as detecting irregular plural noun subjects. However, because traditional evaluation methods (e.g., benchmarking) fail to reveal how models acquire these concepts and capabilities, it is not well understood when and how these specific linguistic abilities emerge. To bridge this gap and better understand model training at the concept level, we use sparse crosscoders to discover and align features across model checkpoints. Using this approach, we track the evolution of linguistic features during pretraining. We train crosscoders between open-sourced checkpoint triplets with significant performance and representation shifts, and introduce a novel metric, Relative Indirect Effects (RelIE), to trace training stages at which individual features become causally important for task performance. We show that crosscoders can detect feature emergence, maintenance, and discontinuation during pretraining. Our approach is architecture-agnostic and scalable, offering a promising path toward more interpretable and fine-grained analysis of representation learning throughout pretraining.

[Findings] Decoupled Reasoning with Implicit Fact Tokens (DRIFT): A Dual-Model Framework for Efficient Long-Context Inference ▶ *Underline Video*
Wenxuan Xie and Xuhong Wang;
The integration of extensive, dynamic knowledge into Large Language Models (LLMs) remains a significant challenge due to the inherent entanglement of factual data and reasoning patterns. Existing solutions, ranging from non-parametric Retrieval-Augmented Generation (RAG) to parametric knowledge editing, are often constrained in practice by finite context windows, retriever noise, or the risk of catastrophic forgetting. In this paper, we propose DRIFT, a novel dual-model architecture designed to explicitly decouple knowledge extraction from the reasoning process. Unlike static prompt compression, DRIFT employs a lightweight knowledge model to dynamically compress document chunks into implicit fact tokens conditioned on the query. These dense representations are projected into the reasoning model's embedding space, replacing raw, redundant text while maintaining inference accuracy. Extensive experiments show that DRIFT significantly improves performance on long-context tasks, outperforming strong baselines among comparably sized models. Our approach provides a scalable and efficient paradigm for extending the effective context window and reasoning capabilities of LLMs. Our code and data will be made public upon publication.

[Short] Defense Against Knowledge Poisoning Attack on GraphRAG ▶ *Underline Video*
Havva Alizadeh Noughabi, Fattane Zarrinkalam and Ali Dehghantanha
GraphRAG augments large language models with structured knowledge graphs, enabling graph-based context selection and a more integrated view of the knowledge space. However, recent work shows that GraphRAG exposes a new attack surface: corpus-level knowledge poisoning can inject spurious entities and relationships during graph construction, corrupting query-specific subgraphs and steering the generator toward incorrect answers. We propose Hop-wise Guard for GraphRAG (HoG-GRAG), a defense layer between retriever and generator that

decomposes multi-hop questions into ordered subqueries, monitors hop-wise execution for poisoning-induced inconsistencies, and locally repairs the retrieved subgraph by pruning compromised entities and relationships and adding only minimal missing evidence. Experiments on multi-hop datasets and multiple GraphRAG configurations show that HoG-GRAG recovers a large fraction of the lost performance. The code is available at https://github.com/CyberScienceLab/HoG-GRAG.

[Long] Demystifying Data Organization for Enhanced LLM Training ► *Underline Video*

Yalun Dai, Yangyu Huang, Tongshen Yang, Yonghan Wang, Xin Zhang, Wenshan Wu, Qihao Zhao, Hao Li, Yuanyuan Gao, Kim-Hui Yap and Scarlett Li

Large Language Models (LLMs) have revolutionized various fields, yet their training efficiency is heavily reliant on effective data curation. While data selection has been widely studied, the strategic data organization for enhanced training remains an underexplored area, particularly since current LLMs are often trained for only one or a few epochs. This paper systematically explores the influence of data organization on LLM training by reusing pre-computed sample-level scores originally generated for data efficiency, thereby incurring minimal additional computational overhead. We identify and formalize four key guidances for optimizing data organization: Boundary Sharpening, Cyclic Scheduling, Curriculum Continuity, and Local Diversity. Guided by them, we introduce two novel data ordering methods termed STR and SAW. Extensive experiments across different model scales and data sizes, encompassing both pre-training and SFT stages, validate the effectiveness of our summarized guidances. They also demonstrate the robustness of our proposed data ordering methods in enhancing the stability and performance of LLM training.

[Long] Diff4TST: Masked Diffusion Language Model for Text Style Transfer ► *Underline Video*

Xinchen Ma, Gaole He, Yunshi Lan and Weining Qian

Despite recent progress in LLMs for text style transfer, most existing methods rely on costly task-specific training and offer limited control over separating stylistic modification from content preservation. We propose Diff4TST, a diffusion-based language model that formulates text style transfer as an explicit copy-and-edit process. Built upon masked diffusion language models, Diff4TST introduces a style-aware noise schedule that selectively perturbs stylistic tokens while preserving content-bearing tokens during supervised fine-tuning. At inference time, we further introduce a generate-then-refine strategy that iteratively improves style compliance via gradient-based token re-masking, without reinforcement learning or external reward models. Extensive experiments on both fine-grained and polarity-based benchmarks show that Diff4TST achieves substantially improved style accuracy and controllability while maintaining strong content preservation and fluency. These results suggest diffusion-based language models as a principled and effective alternative to autoregressive pipelines for text style transfer.

[Findings] Do Generalisation Results Generalise? ► *Underline Video*

Matteo Boglioni, Andrea Sgobbi, Gabriel Tavernini, Marius Mosbach and Tiago Pimentel

A large language model's (LLM's) out-of-distribution (OOD) generalisation is crucial to its deployment. Previous work assessing LLMs' generalisation performance, however, typically focuses on a single out-of-distribution dataset. This approach may fail to precisely evaluate the capabilities of the model, as the data shifts encountered during deployment are much more diverse. In this work, we investigate whether OOD generalisation results generalise. More specifically, we evaluate a model's performance across multiple OOD testsets throughout a finetuning run; we then evaluate the partial correlation of performances across these testsets, regressing out in-domain performance. This allows us to assess how correlated are generalisation performances once in-domain performance is controlled for. Analysing OLMo, OPT and SmolLM, we observe no overarching trend in generalisation results: the existence of a positive or negative correlation between any two OOD testsets depends strongly on the specific choice of model analysed.

[Long] Does Memory Need Graphs? A Unified Framework and Empirical Analysis for Long-Term Dialog Memory ► *Underline Video*

Sen Hu, YUXIANG WEI, Jiaxin Ran, Xueran Han, Zhiyuan Yao, Huacan Wang, Ronghao Chen and Lei Zou

Graph structures are increasingly adopted in dialog memory systems, motivated by their suc cess in retrieval-augmented generation and the associative nature of human memory. How ever, empirical findings on their effectiveness remain inconsistent, making it unclear which design choices truly matter. In this work, we present an experimental and system-oriented analysis of long-term dialog memory architec tures. We formalize a unified framework that decomposes dialog memory systems into core components and supports both graph-based and non-graph approaches. Under this frame work, we conduct controlled, stage-wise ex periments on LongMemEval and HaluMem, comparing common design choices in mem ory representation, organization and mainte nance, as well as indexing and retrieval. Our results show that underlying implementation details, often insufficiently specified in prior work, have a substantial impact on perfor mance, and we identify stable, reliable strong baselines to support fair comparison and practi cal deployment. Code are available at https://github.com/AvatarMemory/UnifiedMem.

[Long] Don't Act Blindly: Robust GUI Automation via Action-Effect Verification and Self-Correction ► *Underline Video*

Yuzhe Zhang, Xianwei Xue, Xingyong Wu, Mengke Chen, Chen Liu, Xinran He, Run Shao, Feiran Liu, Huanmin Xu, Qiutong Pan and Haiwei Wang

Autonomous GUI agents based on vision-language models (VLMs) often assume deterministic environment responses, generating actions without verifying whether previous operations succeeded. In real-world settings with network latency, rendering delays, and system interruptions, this assumption leads to undetected action failures, repetitive ineffective behaviors, and catastrophic error accumulation.

Moreover, learning robust recovery strategies is challenging due to the high cost of online interaction and the lack of real-time feedback in offline datasets.We propose VeriGUI (Verification-driven GUI Agent), which explicitly models action outcomes and recovery under noisy environments. VeriGUI introduces a Thinking–Verification–Action–Expectation (TVAE) framework to detect failures and guide corrective reasoning, and a two-stage training pipeline that combines Robust SFT with synthetic failure trajectories and GRPO with asymmetric verification rewards. We further construct a Robustness Benchmark based on AndroidControl to evaluate failure recognition and correction. Experiments show that VeriGUI significantly reduces failure loops and improves recovery success while maintaining competitive standard task performance.

[Long] Don't Click That: Teaching Web Agents to Resist Deceptive Interfaces ▶ *Underline Video*

Yilin Zhang

Vision-language model (VLM) based web agents demonstrate impressive autonomous GUI interaction but remain vulnerable to deceptive interface elements. Existing approaches either detect deception without task integration or document attacks without proposing defenses. We formalize deception-aware web agent defense and propose DUDE (Deceptive UI Detector & Evaluator), a two-stage framework combining hybrid-reward learning with asymmetric penalties and experience summarization to distill failure patterns into transferable guidance. We introduce RUC (Real UI Clickboxes), a benchmark of 1,407 scenarios spanning four domains and deception categories. Experiments show DUDE reduces deception susceptibility by 53.8% while maintaining task performance, establishing an effective foundation for robust web agent deployment.

[Long] Dont Corrupt the Fact: A Trustworthy RAG Watermarking Framework based on Dual Factual Shield ▶ *Underline Video*

Hao Huang, JiaTang Luo, Ruihua Zhou, Yunpeng Li and Yuling Liu

While Retrieval-Augmented Generation (RAG) systems are designed to enhance factual fidelity by grounding LLMs in provided sources, the application of current watermarking techniques creates a paradoxical failure mode. These methods, being inherently fact-agnostic, force the model to deviate from the very source documents it is supposed to follow. This leads to "faithfulness hallucinations", a critical flaw where the generated output contradicts its own grounding context. Consequently, these watermarks undermine the core value of RAG, rendering even the most secure schemes untrustworthy for high-stakes applications. To resolve this RAG-specific conflict, we introduce the Dual Factual Shield (DFS) framework, a novel architecture designed to enforce knowledge loyalty. The DFS framework employs a defense-in-depth strategy through two synergistic layers: a source-anchored algorithmic safeguard that shields critical terms from the retrieved context, and prompt-based semantic guidance that protects against factual corruption. To demonstrate its effectiveness, we enhance a state-of-the-art, spoofing-aware contrastive watermarking baseline with our framework. Experiments show that our framework drastically reduces the Knowledge Corruption Rate (KCR), a new metric we introduce, while preserving its original high security and robustness. This work establishes a new paradigm for watermarking, evolving it from merely secure to truly trustworthy. We demonstrate that traceability and truth can, and must, coexist, paving the way for the responsible deployment of traceable AI in knowledge-critical domains.

[Findings] DPLoRA: A Dual-Pruning Framework based on ILP Optimization and Progressive Pruning for Parameter-Efficient LoRA Fine-Tuning ▶ *Underline Video*

Changjun Park, Sejong Yoon and Jaekwang Kim

We propose DPLoRA (Dual-Pruning Low-Rank Adaptation), a framework that optimizes rank allocation via two stages: (1) an initial pruning stage (OPLoRA; Optimal Pruning LoRA) that uses Integer Linear Programming (ILP) to determine optimal layer-wise ranks without manual tuning; and (2) a progressive pruning stage that further reduces ranks adaptively during training using importance scores smoothed by Exponential Moving Average (EMA). Experiments demonstrate that OPLoRA consistently outperforms existing PEFT baselines on GLUE and instruction-following tasks, while the full DPLoRA framework establishes a new state-of-the-art among compared PEFT baselines on GLUE at high-performance settings (p=0.4). At efficiency-focused settings (p=0.8), our method reduces trainable parameters by over 80% and training time by 46% compared to standard LoRA, offering a highly efficient solution for deploying large-scale models in resource-constrained environments.

[Long] DREAM-S: Speculative Decoding with Searchable Drafting and Target-Aware Refinement for Multimodal Generation ▶ *Underline Video*

Zining Liu, Yunhai Hu, Tianhua Xia, Bo Bao, Eric Sather, Vithursan Thangarasa and Saiqian Zhang

Speculative decoding (SD) has proven to be an effective technique for accelerating autoregressive generation in large language models (LLMs), however its application to vision-language models (VLMs) remains relatively unexplored. We propose DREAM-S, a novel SD framework designed specifically for fast and efficient decoding in VLMs. DREAM-S leverages a neural architecture search (NAS) framework with target-aware supernet training to automatically identify both the optimal interaction strategy between the draft and target models, and the most suitable draft model architecture for the underlying hardware implementation platform. DREAM-S additionally incorporates adaptive intermediate feature distillation, guided by attention entropy, to enable efficient draft training. Experiments on a range of well-established VLMs show that DREAM-S achieves up to a 3.85x speedup compared to standard decoding approaches and significantly outperforms existing SD baselines.

[Long] Dynamic Generation of Multi LLM Agents Communication Topologies with Graph Diffusion Models ▶ *Underline Video*

Eric Hanchen Jiang, Levina Li, Frank Wan, Xiao Liang, Sophia Yin, Yuchen Wu, Xinfeng Li, Yizhou Sun, Wei Wang, Kai-Wei Chang and Ying

Nian Wu

The efficiency of multi-agent systems driven by large language models (LLMs) largely hinges on their communication topology. However, designing an optimal topology is a non-trivial challenge, as it requires balancing competing objectives such as task performance, communication cost, and robustness. Existing frameworks often rely on static or hand-crafted topologies, which inherently fail to adapt to diverse task requirements, leading to either excessive token consumption for simple problems or performance bottlenecks for complex ones. To address this challenge, we introduce a novel generative framework called Guided Topology Diffusion (GTD). Inspired by conditional discrete graph diffusion models, GTD formulates topology synthesis as an iterative construction process. At each step, the generation is steered by a lightweight proxy model that predicts multi-objective rewards (e.g., accuracy, utility, cost), enabling real-time, gradient-free optimization towards task-adaptive topologies. This iterative, guided synthesis process distinguishes GTD from single-step generative frameworks, enabling it to better navigate complex design trade-offs. We validated GTD across multiple benchmarks, and experiments show that this framework can generate highly task-adaptive, sparse, and efficient communication topologies, significantly outperforming existing methods in LLM agent collaboration. Our code is available at https://anonymous.4open.science/r/diffusion_agent-953C.

[Long] Dynamic Infilling Anchors for Format-Constrained Generation in Diffusion Large Language Models ▶ *Underline Video*

Boyan Han, Yiwei Wang, Yi Song, Yujun Cai and Chi Zhang

Diffusion large language models (dLLMs) offer bidirectional attention and parallel generation, enabling them to exploit global context and naturally support format-constrained tasks like parseable JSON or reasoning templates. While straightforward fixed anchors can enforce such constraints, they often impose rigid spans, leading to truncated reasoning or redundant content. To overcome this, we propose Dynamic Infilling Anchors (DIA), a training-free method that dynamically estimates end-anchor positions to adjust generation length before iterative infilling. This flexible mechanism ensures structural correctness and semantic coherence, avoiding the inefficiencies of fixed-span methods. Experiments on reasoning benchmarks demonstrate that DIA substantially improves format compliance and answer accuracy, achieving significant zero-shot gains on GSM8K and MATH. These results establish DIA as a robust pathway toward reliable, structure-aware generation.

[Findings] DynamixSFT: Dynamic Mixture Optimization of Instruction Tuning Collections ▶ *Underline Video*

Haebin Shin, Lei Ji, Xiao Liu, Zhiwei Yu, Hyunwoo Yoo, Qi Chen and Yeyun Gong

As numerous instruction-tuning datasets continue to emerge, dynamically balancing and optimizing their mixtures has become a critical challenge. To address this, we propose DynamixSFT, a dynamic and automated method for instruction-tuning dataset mixture optimization. We formulate the problem as a multi-armed bandit setup and introduce a Prior-scaled Boltzmann Exploration that softly anchors the updated sampling distribution to the original dataset proportions, thereby preserving the inherent diversity and coverage of the collection. Sampling probabilities are updated using a lightweight 1-Step Look-ahead Reward, reflecting how much the dataset contributes to improving the models performance at its current state. We demonstrate that DynamixSFT effectively optimizes the TÜLU-2-mixture and TÜLU-3-mixture collections across 10 benchmarks, while introducing minimal computational overhead over naive sampling. Furthermore, we provide a comprehensive analysis and visualizations to offer deeper insights into the adaptive dynamics of our method.

[Findings] Enhancing Hallucination Detection via Future Context ▶ *Underline Video*

Joosung Lee, Cheonbok Park, Hwiyeol Jo, Jeonghoon Kim, Joonsuk Park and Kang Min Yoo

Large Language Models (LLMs) are widely used to generate plausible text on online platforms, without revealing the generation process. As users increasingly encounter such black-box outputs, detecting hallucinations has become a critical challenge. To address this challenge, we focus on developing a hallucination detection framework for black-box generators. Motivated by the observation that hallucinations, once introduced, tend to persist, we sample future contexts. The sampled future contexts provide valuable clues for hallucination detection and can be effectively integrated with various sampling-based methods. We extensively demonstrate performance improvements across multiple methods using our proposed sampling approach.

[Findings] Exploring Reasoning Reward Model for Agents ▶ *Underline Video*

Kaixuan Fan, Kaituo Feng, Manyuan Zhang, Tianshuo Peng, Zhixun Li, Yilei Jiang, Shuang Chen and Xiangyu Yue

Agentic Reinforcement Learning (Agentic RL) has achieved notable success in enabling agents to perform complex reasoning and tool use. However, most methods still relies on sparse outcome-based reward for training. Such feedback fails to differentiate intermediate reasoning quality, leading to suboptimal training results. In this paper, we introduce Agent Reasoning Reward Model (Agent-RRM), a multi-faceted reward model that produces structured feedback for agentic trajectories, including (1) an explicit reasoning trace, (2) a focused critique that provides refinement guidance by highlighting reasoning flaws, and (3) an overall score that evaluates process performance. Leveraging these signals, we systematically investigate three integration strategies: Reagent-C (text-augmented refinement), Reagent-R (reward-augmented guidance), and Reagent-U (unified feedback integration). Extensive evaluations across 12 diverse benchmarks demonstrate that Reagent-U yields substantial performance leaps, achieving 43.7% on GAIA and 46.2% on WebWalkerQA, validating the effectiveness of our reasoning reward model and training schemes. Code, models, and datasets will be released to facilitate future research.

[Findings] Fast-Decoding Diffusion Language Models via Progress-Aware Confidence Schedules ▶ *Underline Video*

Amr Mohamed, Yang Zhang, Michalis Vazirgiannis and Guokan Shang

Diffusion large language models (dLLMs) offer a promising alternative to autoregressive models, but their practical utility is severely

hampered by slow, iterative sampling. We present SchED, a training-free, model-agnostic early-exit algorithm that terminates diffusion decoding using a progress-aware confidence threshold. We evaluate SchED across multiple diffusion model families and a diverse set of benchmarks spanning multiple-choice, math, long-form QA, and translation. SchED delivers substantial acceleration: on instruction-tuned models, it achieves approximately 4x speedups while retaining baseline performance on average. On base models, SchED yields consistent speedup gains with 99.1–100% performance retention, with up to 2.34x under more aggressive settings. Under a conservative quality-penalized speed metric, SchED consistently outperforms prior confidence-based early-exit methods, including on long-form generation where existing approaches tend to break down. An entropy analysis of the models token predictions reveals that instruction tuning speeds up the decay of predictive entropy. By leveraging inherent confidence stabilization as a signal for computational efficiency, SchED provides a robust framework for efficient dLLM inference.

[Findings] FastDiSS: Few-step Match Many-step Diffusion Language Model on Sequence-to-Sequence Generation ▶ *Underline Video*
Dat Nguyen Cong, Tung Kieu and Hoang Thanh-Tung
Self-conditioning has been central to the success of continuous diffusion language models, as it allows models to correct previous errors. Yet its ability degrades precisely in the regime where diffusion is most attractive for deployment: few-step sampling for fast inference. In this study, we show that when models only have a few denoising steps, inaccurate self-conditioning induces a substantial approximation gap; this mistake compounds across denoising steps and ultimately dominate the sample quality. To address this, we propose a novel training framework that handles these errors during learning by perturbing the self-conditioning signal to match inference noise, improving robustness to prior estimation errors. In addition, we introduce a token-level noise-awareness mechanism that prevents training from saturation, hence improving optimization. Extensive experiments across conditional generation benchmarks demonstrate that our framework surpasses standard continuous diffusion models while providing up to 400x faster inference speed, and remains competitive against other one-step diffusion frameworks.

[Long] Figure It Out: Improve the Frontier of Reasoning with Executable Visual States ▶ *Underline Video*
Meiqi Chen, Fandong Meng and Jie Zhou
Complex reasoning problems often involve implicit spatial and geometric relationships that are not explicitly encoded in text. While recent reasoning models perform well across many domains, purely text-based reasoning struggles to capture structural constraints in complex settings. In this paper, we introduce FIGR, which integrates executable visual construction into multi-turn reasoning via end-to-end reinforcement learning. Rather than relying solely on textual chains of thought, FIGR externalizes intermediate hypotheses by generating executable code that constructs diagrams within the reasoning loop. An adaptive reward mechanism selectively regulates when visual construction is invoked, enabling more consistent reasoning over latent global properties that are difficult to infer from text alone. Experiments on seven challenging mathematical benchmarks demonstrate that FIGR outperforms strong text-only chain-of-thought baselines, improving the base model by 13.12% on AIME 2025 and 11.00% on BeyondAIME. These results highlight the effectiveness of precise, controllable figure construction of FIGR in enhancing complex reasoning ability.

[Long] From Experts to Bases: Orthogonal Subspace Mixture for Continual Multimodal Instruction Tuning ▶ *Underline Video*
Pei Chen, Xilai Wang, Qixu Shi, Zejian Li and Lingyun Sun
Multimodal Continual Instruction Tuning (MCIT) is essential for adapting Multimodal Large Language Models (MLLMs) to dynamic data streams, yet preventing catastrophic forgetting remains a major challenge. Existing parameter-efficient approaches often face a dilemma: fixed architectures suffer from knowledge interference, while dynamic strategies incur inefficient capacity expansion, limiting scalability. We propose MoBLoRA (Mixture-of-Bases LoRA), a novel framework for MCIT. Motivated by our geometric analysis revealing subspace redundancy across sequential tasks, MoBLoRA shifts the paradigm from expert selection to subspace mixing: it decomposes adaptation weights into a globally shared pool of orthonormal bases to capture task-invariant knowledge, and lightweight mixing matrices to encode task-specific variations. This design effectively decouples knowledge accumulation from task reconstruction. Experiments on standard benchmarks show MoBLoRA significantly outperforms state-of-the-art methods while maintaining superior parameter efficiency.

[Findings] From Passive Metric to Active Signal: The Evolving Role of Uncertainty Quantification in Large Language Models ▶ *Underline Video*
Jiaxin Zhang, Wendi Cui, Zhuohang Li, Lifu Huang, Bradley A. Malin, Caiming Xiong and Chien-Sheng Wu
While Large Language Models (LLMs) show remarkable capabilities, their unreliability remains a critical barrier to deployment in high-stakes domains. This survey charts a functional evolution in addressing this challenge: the evolution of uncertainty from a passive diagnostic metric to an active control signal guiding real-time model behavior. We demonstrate how uncertainty is leveraged as an active control signal across three frontiers: in advanced reasoning to optimize computation and trigger self-correction; in autonomous agents to govern metacognitive decisions about tool use and information seeking; and in reinforcement learning to mitigate reward hacking and enable self-improvement via intrinsic rewards. By grounding these advancements in emerging theoretical frameworks like Bayesian methods and Conformal Prediction, we provide a unified perspective on this transformative trend. This survey provides a comprehensive overview, critical analysis, and practical design patterns, arguing that mastering the new trend of uncertainty is essential for building the next generation of scalable, reliable, and trustworthy AI.

[Findings] From Representation to Choice: Tracing Decision Emergence Across Languages in LLMs ▶ *Underline Video*

Divyanshu_Bhatt Divyanshu_Bhatt, Abhinav Joshi, Ujjaval patel and Ashutosh Modi
Multilingual large language models (LLMs) can answer questions in many languages, but how they internally reason across languages remains poorly understood. In this work, we study multilingual reasoning through a decision-making perspective to investigate how multilingual reasoning unfolds in multilingual LLMs using aligned multiple-choice questions from the mMMLU benchmark. By formulating a controlled setup, presenting the same question in different languages, and tracking the models decision from the first token to the final answer choice, we can directly compare how reasoning trajectories evolve across languages. We first demonstrate that, at the representation level, different languages share highly similar activation spaces; however, subtle divergences emerge as decisions propagate through the transformer layers. We then model answer selection as a stepwise trajectory, revealing where language-specific signals arise. These patterns are further confirmed by quantifying deviations along these trajectories, highlighting layers where multilingual processing deviates or converges. Our work provides a controlled, layer-resolved view of multilingual reasoning, shedding light on how LLMs balance shared conceptual understanding with language-specific decision-making.

[Findings] From Scaffolding to Assimilation: Progressive Structural Internalization for Format-Constrained Creative Text Generation ▶ *Underline Video*
Wenhao Li, Yuwei Yang, Xiaoqing Wu, Yufeng Han, Cunliang Kong, Yuzhuo Bai, Xin Cong and Maosong Sun
While Large Language Models (LLMs) demonstrate remarkable capabilities in open-ended creative generation, they notably struggle with Format-Constrained Generation tasks, such as poetry and lyrics, where strict adherence to multidimensional structural constraints (i.e., format, phonetics, and rhyme) is prerequisite to aesthetic value. Existing paradigms predominantly rely on unreliable prompting or rigid constrained decoding strategies; the former often fails to ensure compliance, while the latter compromises inference latency and disrupts the natural probability distribution, degrading generation quality. To bridge this gap, we establish CCP-Arena, a rigorous testbed for Chinese Classical Poetry, and proposeProgressive Structural Internalization (PSI) a novel framework designed to embed external constraints into the model's intrinsic intuition. PSI initiates withStructural Scaffolding via Explicit Cognitive Planning, utilizing explicit template to provide a structural scaffold for subsequent generation. This is followed by a Cascaded Reinforcement Learning stage guided by a Holistic Reward Model, which optimizes for precise structural-semantic alignment. Extensive experiments demonstrate that PSI achieves state-of-the-art performance, surpassing baselines in both strict constraint adherence and literary aesthetics. Furthermore, mechanistic analysis confirms that our method effectively internalizes structural information into the model's latent representations, offering a robust and efficient solution for constrained creative generation.

[Findings] Gaperon: A Peppered English-French Generative Language Model Suite ▶ *Underline Video*
Nathan Godey, Wissam Antoun, Rian Touchent, Rachel Bawden, Éric Villemonte de la Clergerie, Benoît Sagot and Djamé Seddah
Standardized benchmarks have become the dominant metric for measuring progress in large language models, yet their validity is increasingly compromised by data contamination and the unclear relationship between benchmark scores and genuine language understanding. We introduce Gaperon, a suite of fully open bilingual (French-English) language models designed as an experimental testbed to investigate evaluation dynamics under realistic training conditions. Our study makes three core contributions. First, we demonstrate mismatches between benchmark performance and generation quality: models that excel on benchmarks may underperform in qualitative text generation, and vice versa. Second, through our deliberately contaminated Gaperon-Garlic variant, we show that competitive benchmark scores can be recovered via late-stage contamination with only moderate degradation of generation quality, and surprisingly, such contamination also improves performance on held-out benchmarks. Third, we provide empirical evidence that widely used neural quality filters, particularly those trained to favor instructional or educational content, amplify benchmark contamination in pretraining corpora, with the DCLM classifier systematically ranking benchmark samples in the top-5 percentiles of samples. We release all models, data mixtures, checkpoints, and evaluation code to support reproducibility and further investigation.

[Long] GASim: A Graph-Accelerated Hybrid Framework for Social Simulation ▶ *Underline Video*
Xuan Zhou, Yanhui Sun, Hantao Yao, Allen He, Yongdong Zhang and Wu Liu
Large-scale social simulators are essential for studying complex social patterns. Prior work explores hybrid methods to scale up simulations, combining large language models (LLM)-based agents with numerical agent-based models (ABM). However, this incurs high latency due to expensive memory retrieval and sequential ABM execution. To address this challenge, we propose GASim, a graph-accelerated hybrid multi-agent framework for large-scale social simulations. For core agents driven by LLM, GASim introduces Graph-Optimized Memory (GOM) to replace intensive LLM-based retrieval pipelines with lightweight propagation over a sparse memory graph. For the majority of ordinary agents, GASim employs Graph Message Passing (GMP), substituting sequential ABM execution with parallel updates by fine-grained feature aggregation and Graph Attention Network. We further introduce Entropy-Driven Grouping (EDG) that coordinates this hybrid partitioning, leveraging information entropy to dynamically identify emergent core agents situated in information-diverse neighborhoods. Extensive experiments show that GASim not only delivers a substantial 9.94x end-to-end speedup over the traditional hybrid framework but also consumes less than 20% of baseline tokens, significantly reducing costs while preserving strong alignment with real-world public opinion trends.

[Findings] Generalizable Prompt Tuning for Audio-Language Models via Semantic Expansion ▶ *Underline Video*
Jaehyuk Jang, Wonjun Lee, Kangwook Ko and Changick Kim
Prompt tuning has achieved remarkable progress in vision-language models (VLMs) and is recently being adopted for audio-language models

(ALMs). However, its generalization ability in ALMs remains largely underexplored. We observe that conventional prompt tuning for ALMs also suffers from the Base-New Tradeoff, and we identify that this issue stems from the disrupted semantic structure of the embedding space. To address this issue, we propose Semantically Expanded Prompt Tuning (SEPT), a plug-and-play framework that explicitly regularizes the prompt embedding space by incorporating semantic neighbors generated by large language models. SEPT introduces a novel semantic expansion loss with margin constraints that promote intra-class compactness and inter-class separability, thereby enhancing the semantic structure of the prompt embedding space. For comprehensive evaluation, we establish the first benchmark setup for prompt generalization in ALMs, covering both base-to-new generalization and cross-dataset transferability. Extensive experiments demonstrate that SEPT consistently improves generalization performance across multiple prompt tuning baselines, while maintaining computational cost during inference.

[Findings] GlimpRouter: Efficient Collaborative Inference by Glimpsing One Token of Thoughts ▶ *Underline Video*

Wenhao Zeng, Xuteng Zhang, Yuling Shi, Chao Hu, Yuting Chen, Beijun Shen and Xiaodong Gu

Large Reasoning Models (LRMs) achieve remarkable performance by explicitly generating multi-step chains of thought, but this capability incurs substantial inference latency and computational cost. Collaborative inference offers a promising solution by selectively allocating work between lightweight and large models, yet a fundamental challenge remains: determining when a reasoning step requires the capacity of a large model or the efficiency of a small model. Existing routing strategies either rely on local token probabilities or post-hoc verification, introducing significant inference overhead. In this work, we propose a novel perspective on step-wise collaboration: the difficulty of a reasoning step can be inferred from its very first token. Inspired by the Aha Moment phenomenon in LRMs, we show that the entropy of the initial token serves as a strong predictor of step difficulty. Building on this insight, we introduce GlimpRouter, a training-free step-wise collaboration framework. GlimpRouter employs a lightweight model to generate only the first token of each reasoning step and routes the step to a larger model only when the initial token entropy exceeds a threshold. Experiments on multiple benchmarks demonstrate that our approach significantly reduces inference latency while preserving accuracy. For instance, GlimpRouter attains a substantial 10.7% improvement in accuracy while reducing inference latency by 25.9% compared to a standalone large model on AIME25. These results suggest a simple yet effective mechanism for reasoning: allocating computation based on a glimpse of thought rather than full-step evaluation.

[Long] GraphSynth: Resolving the Diversity-Reliability Trade-off with Probabilistic Factor Graphs ▶ *Underline Video*

Zehua Cheng, Wei Dai, Jiahao Sun and Thomas Lukasiewicz

The large language models offer a scaleable solution for the generation of synthetic data faced with a trade-off between maintaining the diversity of generation and achieving factually accurate results. This paper introduces Graphsynth, a framework which leverages a probabilistic factor graph modeling the universe of attributes. The framework leverages a high-level schema mapping compiled into efficient hard masks during the decoding phase for maintaining the syntactic truth and a span-synchronized verifier for dismissing logical contradictions at the decode time. The experiments conducted on biomedical, legal, and generic domains show that the method outperforms the state-of-the-art baselines with a structural integrity approaching perfection, a coverage of around 94% attributes on the factor graph solution, and a boost in performance on downstream tasks such as +17.9% on TruthfulQA.

[Long] Guidelines as Environments: A World Model Approach to Rule Following ▶ *Underline Video*

Haiqing Li, Wenliang Zhong, Yinhao Wu, Hehuan Ma, Yuzhi Guo, Thao M. Dang and Junzhou Huang

Guideline-following is increasingly important in compliance, customer support, and other regulated workflows, where correctness is defined by explicit rule systems rather than heuristics. Learning to follow guidelines is challenging because guidelines are interdependent: rules can trigger, suppress, or conflict with one another, while locally plausible responses may violate global constraints. Most existing methods treat guidelines as static text and rely on implicit reasoning or deeper decoding, making rule interactions and satisfaction status hard to observe and control. A more feasible approach is to model guideline execution with an explicit state that tracks evolving rule evidence across steps. However, conventional world models are a poor fit: they typically assume privileged feedback or well-defined transition dynamics, assumptions that do not hold when reasoning occurs purely in language space under ambiguous, text-defined constraints. As a solution, we propose RGCWM, a Rule-Grounded Causal World Model that builds an explicit state space from the guideline text itself. RGCWM represents rule applicability and satisfaction as a continuously updated evidence state, externalizes inter-rule dependencies as a causal structure, and plans at inference time by counterfactually evaluating candidate responses under model-estimated state transitions. Experiments show that this shift from implicit text reasoning to state-based reasoning enables stable, controllable execution of complex interacting rules across diverse domains.

[Long] How Training Data Shapes the Use of Parametric and In-Context Knowledge in Language Models ▶ *Underline Video*

Minsung Kim, Dong-Kyum Kim, Jea Kwon, Nakyeong Yang, Kyomin Jung and Meeyoung Cha

Large language models leverage both parametric knowledge acquired during pretraining and in-context knowledge provided at inference time. Crucially, when these sources conflict, models arbitrate based on their internal confidence, preferring parametric knowledge for high-confidence facts while deferring to context for less familiar ones. However, the training conditions that give rise to these fundamental behaviors remain unclear. Here we conduct controlled experiments using synthetic corpora to identify the specific data properties that shape knowledge utilization. Our results reveal a counterintuitive finding: the robust, balanced use of both knowledge sources is an emergent property that requires the co-occurrence of three factors typically considered detrimental, including (i) intra-document repetition, (ii) a moderate degree of intra-document inconsistency, and (iii) a skewed knowledge distribution. We further show that these dynamics arise in real-world language model pretraining and analyze how post-training procedures reshape arbitration strategies. Together, our findings

provide empirical guidance for designing training data that supports the reliable integration of parametric and in-context knowledge in language models.

[Long] Impatient Users Confuse AI Agents: High-fidelity Simulations of Human Traits for Testing Agents ▶ *Underline Video*

Muyu He, Anand Kumar, Soumyadeep Bakshi, James Zou and Nazneen Rajani

Despite rapid progress in building conversational AI agents, robustness is still largely untested. Small shifts in user behavior, such as being more impatient, incoherent, or skeptical, can cause sharp drops in agent performance, revealing how brittle current AI agents are. Todays benchmarks fail to capture this fragility: agents may perform well under standard evaluations but degrade spectacularly in more realistic and varied settings. We address this robustness testing gap by introducing TraitBasis, a lightweight, model-agnostic method for systematically stress testing AI agents. TraitBasis learns directions in activation space corresponding to steerable user traits (e.g., impatience or incoherence), which can be controlled, scaled, composed, and applied at inference time without any fine-tuning or extra data. Using TraitBasis, we extend tau-Bench to tau-bench, where user behaviors are altered via controlled trait vectors. We observe an average 4%-20% performance degradation on tau-bench across frontier models, highlighting the lack of robustness of current AI agents to variations in user behavior. Together, these results highlight both the critical role of robustness testing and the promise of TraitBasis as a simple, data-efficient, and compositional tool. By powering simulation-driven stress tests and training loops, TraitBasis opens the door to building AI agents that remain reliable in the unpredictable dynamics of real-world human interactions. We plan to open-source tau-bench, across four domains: airline, retail, telecom, and telehealth, so the community can systematically QA their agents under realistic, behaviorally diverse intents and trait scenarios.

[Long] In-Context Representation Hijacking ▶ *Underline Video*

Itay Yona, Amir Sarid, Michael Karasik and Yossi Gandelsman

We introduce Doublespeak, a simple in-context representation hijacking attack against language models. The attack works by systematically replacing a harmful keyword (e.g., bomb) with a benign token (e.g., carrot) across multiple in-context examples, provided as a prefix to a harmful request. We demonstrate that this substitution leads to the internal representation of the benign token converging toward that of the harmful one, effectively embedding the harmful semantics under a euphemism. As a result, superficially innocuous prompts (e.g., "How to build a carrot?") are internally interpreted as disallowed instructions ("How to build a bomb?"), thereby bypassing the model's safety alignment. We use interpretability tools to show this semantic shift occurs progressively across layers. Doublespeak is optimization-free, broadly transferable across model families, and achieves strong success rates on closed-source systems, reaching 74% on Llama-3.3-70B-Instruct with a single-sentence context override. Our findings highlight a new attack surface in LM latent space, indicating that current alignment strategies are insufficient and should instead operate at the representation level.

[Long] InfiniteWeb: Scalable Web Environment Synthesis for GUI Agent Training ▶ *Underline Video*

Ziyun Zhang, Zezhou Wang, Xiaoyi Zhang, Zongyu Guo, Jiahao Li, Bin Li and Yan Lu

GUI agents that interact with graphical interfaces on behalf of users are a promising direction for practical AI assistants, yet training them is hindered by scarce suitable environments. We present InfiniteWeb, a system that automatically generates functional web environments at scale for GUI agent training. While LLMs perform well on generating a single webpage, building a realistic and functional website with many interconnected pages faces challenges. We address these challenges through unified specification, task-centric test-driven development, and combining website seed variation with reference design images. Our system also generates verifiable task evaluators enabling dense reward signals for reinforcement learning. Experiments show that our system surpasses commercial coding agents at realistic website construction, and GUI agents trained on our generated environments achieve significant performance improvements on OSWorld and Online-Mind2Web, demonstrating the effectiveness of the proposed system.

[Long] InstructDiff: Domain-Adaptive Data Selection via Contrastive Entropy for Efficient LLM Fine-Tuning ▶ *Underline Video*

Junyou Su, He Zhu, Xiao Luo, Liyu Zhang, Hong-Yu Zhou, Yun Chen, Peng Li, Yang Liu and Guanhua Chen

Supervised fine-tuning (SFT) is fundamental to adapting large language models, yet training on complete datasets incurs prohibitive costs with diminishing returns. Existing data selection methods suffer from severe domain specificity: techniques optimized for general instruction-following fail on reasoning tasks, and vice versa. We observe that measuring contrastive entropy between base models and minimally instruction-tuned calibrated models reveals a pattern, samples with the lowest contrastive entropy consistently yield optimal performance across domains, yet this principle manifests domain-adaptively: reasoning tasks favor entropy increase (cognitive expansion), while general tasks favor entropy decrease (cognitive compression). We introduce InstructDiff, a unified framework that operationalizes contrastive entropy as a domain-adaptive selection criterion through warmup calibration, bi-directional NLL filtering, and entropy-based ranking. Extensive experiments show that InstructDiff achieves 17% relative improvement over full data training on mathematical reasoning and 52% for general instruction-following, outperforming prior baselines while using only 10% of the data.

[Long] Integrating Data Validation with Large Language Models for Regulation-Guided Tabular Anomaly Detection ▶ *Underline Video*

Haoliang Huang, Zihuang Cai, Zhuo Tang, Yifan Liu, Chen Tian, Kenli Li and Changjian Chen

In many real-world applications, such as medical insurance, many regulations exist that define how data should comply with certain standards. Auditors typically use these regulations to identify anomalies in tabular data. However, existing tabular anomaly detection

methods often focus on detecting anomalies based on data distribution without considering regulatory compliance. In this paper, we introduce a new task, Regulation-guided Tabular Anomaly Detection, which leverages regulations to detect anomalies in tabular data. We also developed three new datasets for this task. To address this task, we present RegValidator, a training-free method that integrates data validation with large language models (LLMs) for detecting anomalies. In this process, the LLMs generate ideas for anomaly detection from a regulation perspective, while the data validation validates these ideas from a data distribution perspective. This process can be framed as a Budgeted Maximum Coverage problem, which can be solved by a constant-factor approximation algorithm with provable guarantees. Empirical results on the new datasets demonstrate that our method outperforms existing baselines. A field experiment in a commercial health insurance company also reveals the practical value of our method. Our code is available at https://github.com/hnu-vis/RegValidator.

[Long] Interpretable Safety Alignment via SAE-Constructed Low-Rank Subspace Adaptation ▶ *Underline Video*

Dianyun Wang, Qingsen Ma, Yuhu Shang, Zhifeng Lu, Zhenbo Xu, Lechen Ning, Huijia Wu and Zhaofeng He

Safety alignment—training large language models (LLMs) to refuse harmful requests while remaining helpful—is critical for responsible deployment. Prior work established that safety behaviors are governed by low-rank structures, suggesting parameter-efficient fine-tuning (PEFT) should be well-suited for alignment. However, Low-Rank Adaptation (LoRA) consistently underperforms full fine-tuning and reinforcement learning on safety benchmarks. We attribute this gap to semantic entanglement: safety-relevant directions are intertwined with unrelated concepts due to polysemanticity, impeding implicit subspace identification. To address this, we propose SAILS (Safety Alignment via Interpretable Low-rank Subspace), which leverages Sparse Autoencoders (SAEs) to disentangle representations into monosemantic features, constructs an interpretable safety subspace from SAE decoder directions, and uses it to initialize LoRA adapters. Theoretically, we prove that SAE-based identification achieves arbitrarily small recovery error under monosemanticity assumptions, while direct identification suffers an irreducible error floor. Empirically, SAILS achieves up to 99.6% safety rates across multiple model families and scales, exceeding full fine-tuning and matching RLHF-based models, with only 0.2% of parameters updated and providing interpretability.

[Long] J4R: Learning to Judge with Equivalent Initial State Group Relative Policy Optimization ▶ *Underline Video*

Austin Xu, Yilun Zhou, Xuan-Phi Nguyen, Caiming Xiong and Shafiq Joty

To keep pace with the increasing velocity of large language models (LLM) development, model output evaluation has transitioned away from time-consuming human evaluation to automatic evaluation, where LLMs themselves are tasked with assessing and critiquing other model outputs. LLM-as-judge models are a class of generative evaluators that excel in evaluating relatively simple domains, like chat quality, but struggle in reasoning intensive domains where model responses contain more substantive and challenging content. To remedy existing judge shortcomings, we explore training judges with reinforcement learning (RL). We make three key contributions: (1) We propose the Equivalent Initial State Group Relative Policy Optimization (EIS-GRPO) algorithm, which allows us to train our judge to be robust to positional biases that arise in more complex evaluation settings. (2) We introduce ReasoningJudgeBench, a benchmark that evaluates judges in diverse reasoning settings not covered by prior work. (3) We train Judge for Reasoning (J4R), a 7B judge trained with EIS-GRPO that outperforms GPT-4o and the next best small judge by 6.7% and 9%, matching or exceeding the performance of larger GRPO-trained judges on both JudgeBench and ReasoningJudgeBench.

[Long] Jailbreak-Zero: A Path to Pareto Optimal Red Teaming for Large Language Models ▶ *Underline Video*

Kai Hu, Abhinav Aggarwal, Mehran Khodabandeh, David Zhang, Eric Hsin, Li Chen, Ankit Jain, Matt Fredrikson and Akash Bharadwaj

This paper presents a novel Automated Red Teaming (ART) framework that shifts from example-based to policy-based evaluation, addressing critical limitations in scalability and validity. We define harmful content through abstract safety policies rather than specific static examples. We also introduce multiple evaluation objectives: risk coverage, semantic diversity, and fidelity, and discover Pareto trade-offs between them. We propose Jailbreak-Zero, a black-box method capable of both zero-shot generation and fine-tuned exploitation of a victim's vulnerabilities to achieve Pareto optimality. Unlike prior approaches, it does not require expert-designed strategies/prompts, but still achieves superior, human-readable attacks against open-source and proprietary models (attack success rates of 99.5% against GPT-4o and 96.0% against Claude 3.5), even for unseen safety policies. It retains efficacy even after victim models undergo safety alignment, and exposes controls to navigate Pareto trade-offs without retraining. Lastly, we show that Jailbreak-Zero is the best-performing ART method at a given compute budget. Code is available at: https://github.com/hukkai/jailbreak-zero/.

[Long] Jailbreaking Multimodal Large Language Models using Multi-Clip Video ▶ *Underline Video*

Choongwon Kang, Seungjong Sun, Hyunmin Jun and Jang Hyun Kim

As multimodal large language models (MLLMs) have advanced to process video inputs, concerns have emerged about their potential for malicious misuse. Prior jailbreak studies have shown that safety alignment in MLLMs can be bypassed through visual inputs, yet it remains unclear which properties of video inputs induce this vulnerability. To address this gap, we introduce Multi-Clip Video (MCV) SafetyBench, a dataset of 2,920 videos designed to evaluate how the diversity of video inputs affects the vulnerability of MLLMs. Each video consists of multiple short clips depicting diverse contexts related to a harmful query. Experiments on eight representative video MLLMs show that attack success consistently increases with the number of clips. Our results further indicate that the video modality is (1) more vulnerable than the image modality, (2) more vulnerable to dynamic videos than to static videos, and (3) more vulnerable when videos contain more diverse contexts. Building on these findings, we propose a defense strategy that leverages the relative robustness of the image modality. Warning: This paper may contain potentially offensive content.

[Long] KARL: Reinforcement Learning for LLM Agents on Multi-Turn Knowledge-Intensive Agentic Tasks ▶ *Underline Video*
Xueqiao Sun, Xiao Liu, Bowen Lv, Hanchen Zhang, Bohao Jing, Zehan Qi, Yifan Xu, Yuxiao Dong and Jie Tang
Large Language Models have shown remarkable potential as autonomous agents, but their effectiveness in knowledge-intensive tasks remains limited by passive knowledge utilization. We introduce KARL (Knowledge-Augmented Reinforcement Learning), a framework that enables LLM agents to dynamically explore structured knowledge sources through multi-turn interactions. Unlike existing retrieval-augmented approaches, KARL empowers agents to proactively decide when and what knowledge to acquire during task execution. Our framework incorporates online reinforcement learning with curiosity-driven reward shaping, explicitly incentivizing knowledge exploration while optimizing tool-use behaviors end-to-end. Extensive evaluation across six structured knowledge benchmarks demonstrates that KARL achieves state-of-the-art performance, with our Qwen2.5-14B-based agent significantly outperforming GPT-4o, Claude-4, and o4-mini on both knowledge graph and database tasks. Source code is available at https://github.com/THUDM/KARL.

[Long] KoCo: Conditioning Language Model Pre-training on Knowledge Coordinates ▶ *Underline Video*
Yudong Li, Jiawei Cai and Linlin Shen
Standard Large Language Model (LLM) pre-training typically treats corpora as flattened token sequences, often overlooking the real-world context that humans naturally rely on to contextualize information. To bridge this gap, we introduce Knowledge Coordinate Conditioning (KoCo), a simple method that maps every document into a three-dimensional semantic coordinate. By prepending these coordinates as textual prefixes for pre-training, we aim to equip the model with explicit contextual awareness to learn the documents within the real-world knowledge structure. Experiment results demonstrate that KoCo significantly enhances performance across 10 downstream tasks and accelerates pre-training convergence by approximately 30%. Furthermore, our analysis indicates that explicitly modeling knowledge coordinates helps the model distinguish stable facts from noise, effectively mitigating hallucination in generated outputs.

[Long] LaCo: Layer-wise Compensation for Pruned Large Language Models ▶ *Underline Video*
Yingen Liu, Fan Wu, Xuyan Pan, Ruihui Li, Zhuo Tang and Kenli Li
Pruning is essential for the efficient deployment of Large Language Models (LLMs); however, it causes severe performance degradation due to the structural distortion induced by sparsity. Existing recovery strategies, such as LoRA, predominantly employ global fine-tuning, often overlooking the mechanistic root of this degradation: the layer-wise accumulation and amplification of local errors. To address this limitation, we propose LaCo(Layer-wise Compensation), a framework that reorients the recovery paradigm from global adaptation to hierarchical representation alignment. By sequentially optimizing each layer to reconstruct the model's hidden states, LaCo effectively intercept the error propagation chain at its source. Extensive experiments demonstrate that LaCo surpasses parameter-efficient baselines in both perplexity reduction and zero-shot reasoning. Notably, it reduces recovery-time memory usage to approximately 1/7 of the baseline and requires only 2,048 unlabeled samples to match a LoRA model trained on 50k examples, achieving a 25x improvement in data efficiency.

[Findings] Large Language Models Are Overconfident in Their Own Responses ▶ *Underline Video*
Mario Sanz-Guerrero, Manuel Mager and Katharina von der Wense
Prior work has shown that instruction-tuned large language models (LLMs) are less well calibrated than their base pre-trained counterparts. However, little is known about the frequently used chat templates effect on the calibration of conversational LLMs. In this work, we investigate the mechanisms driving this miscalibration by decoupling the effects of the post-training algorithm and the chat format. We find that, while instruction tuning fundamentally harms calibration, the chat template aggravates the issue through an ownership bias - models are significantly more confident in their own answers than in identical answers provided by a user. Extensive experiments across six recent open-weight LLMs, three benchmarks, and three confidence elicitation methods show that models assign up to 26% higher confidence to their own responses. Leveraging this insight, we propose a simple inference-time strategy: framing the model's answer as user input during confidence elicitation. This approach significantly reduces overconfidence and improves calibration by up to 26% without the need for retraining, narrowing the gap between base and instruction-tuned models.

[Long] Less is More: Improving LLM Reasoning with Minimal Test-Time Intervention ▶ *Underline Video*
Zhen Yang, Mingyang Zhang, Feng Chen, Ganggui Ding, Liang Hou, Xin Tao and Ying-Cong Chen
Recent progress in large language models (LLMs) has focused on test-time scaling to improve reasoning via increased inference computation, but often at the cost of efficiency. We revisit test-time behavior and uncover a simple yet underexplored phenomenon: reasoning uncertainty is highly localized, only a small subset of high-entropy tokens dominantly affects output correctness. Motivated by this, we propose Minimal Test-Time Intervention (MTI), a training-free framework that enhances reasoning accuracy and stability with minimal overhead. MTI includes: (i) Selective CFG intervention, applying classifier-free guidance only at uncertain positions; and (ii) Lightweight negative-prompt guidance, reusing the main models KV cache to approximate unconditional decoding efficiently. MTI yields consistent gains across general, coding, and STEM tasks, e.g., +9.28% average improvement on six benchmarks for DeepSeek-R1-7B and +11.25% on AIME2024 using Ling-mini-2.0, while remaining highly efficient.

[Findings] LIBERTy: A Causal Framework for Benchmarking Concept-Based Explanations of LLMs with Structural Counterfactuals ▶ *Underline Video*
Gilat Toker, Nitay Calderon, Ohad Amosy and Roi Reichart
Concept-based explanations quantify how high-level concepts (e.g., gender or experience) influence model behavior, which is crucial for

decision-makers in high-stakes domains. Recent work evaluates the faithfulness of such explanations by comparing them to reference causal effects estimated from counterfactuals. In practice, existing benchmarks rely on costly human-written counterfactuals that serves as imperfect proxy. To address this, we introduce a framework for constructing datasets containing structural counterfactual pairs: LIBERTy (LLM-based Interventional Benchmark for Explainability with Reference Targets). LIBERTy is grounded in explicitly defined Structured Causal Models (SCMs) of the text generation, interventions on a concept propagate through the SCM until an LLM generates the counterfactual. We introduce three datasets (disease detection, CV screening, and workplace violence prediction) together with a new evaluation metric, order-faithfulness. Using them, we evaluate a wide range of methods across five models and identify substantial headroom for improving concept-based explanations. LIBERTy also enables systematic analysis of model sensitivity to interventions: we find that proprietary LLMs show markedly reduced sensitivity to demographic concepts, likely due to post-training mitigation. Overall, LIBERTy provides a much-needed benchmark for developing faithful explainability methods.

[Long] Lightweight and Faithful Visual Condition Checking in Behavior Trees via Expert-Regularized Reinforcement Learning ▶ *Underline Video*

Hyosik Moon and Eldan Cohen

Behavior trees provide a transparent and modular structure for encoding expert-designed policies, enabling interpretable decision-making in complex tasks. Yet, applying behavior trees to high-dimensional perceptual inputs such as images or language is challenging as defining symbolic predicates over raw perceptual data is non-trivial. While state-of-the-art large multimodal models (such as vision-language models) can overcome this issue by utilizing natural language queries over perceptual inputs, they incur high computational cost, making them unsuitable for many applications. Imitation learning offers a way to distill these expert models into compact models, though it requires extensive supervision. In contrast, reinforcement learning reduces the need for costly supervision but risks misalignment of condition nodes with their intended semantics as well as poor credit assignment. To address these challenges, we introduce CERL (Condition-node Expert-regularized Reinforcement Learning), a framework that leverages expert-regularized reinforcement learning to preserve semantic faithfulness, while employing a factorized policy that aggregates sequential condition-node decisions into a single decision unit to alleviate credit assignment challenges. Experiments across seven tasks from the GymCards, FrozenLake, and BabyAIText suites demonstrate that our framework outperforms pure imitation learning or reinforcement learning baselines, retains strong agreement with expert decisions, and achieves substantial gains in inference speed and model size over expert models. Our implementation is available in https://github.com/HyosikMoon/CERL.

[Findings] LLM Multi-Agent Systems for Long Triple Set Data-to-Text Generation ▶ *Underline Video*

Chinonso Cynthia Osuji, Simon Mille, Mark Andrade, Jane Adkins, Ornait O'Connell, Elaine Uí Dhonnchadha, Bláithín Heffernan, Fírinne Nic an tSaoir, Anya Belz, Thiago Castro Ferreira and Brian Davis

Generating coherent, semantically accurate text from large structured inputs remains a persistent challenge in data-to-text generation, as single-step LLM mappings from data-to-text limit control over discourse structuring and amplify hallucinations and omissions as input size grows. We introduce a new dataset of extended DBpedia triple sets (up to 199 triples per input), and a modular multi-agent framework: specialised LLM agents handle content ordering, text structuring, and surface realisation under the supervision of an orchestrator and guardrail control loop. The system generates multi-paragraph outputs in English and Irish (low-resource). We compare a three-worker multi-agent configuration against a single-worker multi-task variant and a strong end-to-end baseline. Quality is assessed via human evaluation and LLM-as-a-judge (with truncation-based sanity checks). Results show slightly superior coherence for the multi-agent approach in both languages, with statistically significant inter-rater correlation over all criteria for English and no statistically significant correlation for Irish. Human-LLM alignment is very weak overall, thus exposing key limits in scalable NLG evaluation.

[Long] LLM-as-Scheduler: Agentic Workflow Dynamic Scheduling ▶ *Underline Video*

Dawei Xiang, Kexin Chu, Wenyan Xu, Wenhui Zhang and Wei Zhang

As large language models (LLMs) improve, many applications are moving from a single LLM call to multi-agent systems. These systems often rely on either hand-designed or automatically optimized workflows with multiple verification and testing steps. While those extra steps can improve accuracy, they also increase latency and token costs. In practice, many queries do not need such heavy processing and can be handled well by a single strong agent. To address this inefficiency, we propose LLM-as-Scheduler (LAS), a system that dynamically chooses the right workflow for each query. LAS uses a two-stage cascade: first, a lightweight gate quickly evaluates each agents output; then, an LLM-based scheduler uses query features and gate signals to make more detailed routing decisions. Experiments show that LAS cuts token usage by 43% and reduces end-to-end latency by more than 36%, while causing at most a 1.4 percentage-point drop in accuracy compared with a strong fixed workflow.

[Long] LoRA on the Go: Instance-level Dynamic LoRA Selection and Merging ▶ *Underline Video*

Seungeon Lee, Soumi Das, Manish Gupta and Krishna P. Gummadi

Low-Rank Adaptation (LoRA) has emerged as a parameter-efficient approach for fine-tuning large language models. However, conventional LoRA adapters are typically trained for a single task, limiting their applicability in real-world settings, where inputs may span multiple, diverse task domains. At inference time, existing methods can combine multiple LoRAs to improve cross-task performance, but they require additional labeled data or task-specific training, which is expensive at scale. In this work, we introduce LoRA on the Go (LoGo), a training-free framework that dynamically selects and merges adapters at the instance level without any additional requirements. LoGo

leverages signals extracted from a single forward pass through LoRA adapters, to identify the most relevant adapters and determine their contributions on-the-fly. Across 5 NLP benchmarks, 27 datasets, and 3 model families, LoGo outperforms training-based baselines on some tasks upto a margin of 3.6% while remaining competitive on other tasks and maintaining inference throughput, highlighting its effectiveness and practicality.

[Findings] Losses that Cook: Topological Optimal Transport for Structured Recipe Generation ▶ *Underline Video*
Mattia Ottoborgo, Daniele Rege Cambrin and Paolo Garza
Cooking recipes are complex procedures that require not only a fluent and factual text, but also accurate timing, temperature, and procedural coherence, as well as the correct composition of ingredients. Standard training procedures are primarily based on cross-entropy and focus solely on fluency. Building on RECIPE-NLG, we investigate the use of several composite objectives and present a new topological loss that represents ingredient lists as point clouds in embedding space, minimizing the divergence between predicted and gold ingredients. Using both standard NLG metrics and recipe-specific metrics, we find that our loss significantly improves ingredient- and action-level metrics. Meanwhile, the Dice loss excels in time/temperature precision, and the mixed loss yields competitive trade-offs with synergistic gains in quantity and time. A human preference analysis supports our finding, showing our model is preferred in 62% of the cases.

[Long] Lost in Execution: On the Multilingual Robustness of Tool Calling in Large Language Models ▶ *Underline Video*
Zheng Luo, T Pranav Kutralingam, Ogochukwu N. Okoani, Wanpeng Xu, Hua Wei and Xiyang Hu
Large Language Models (LLMs) are increasingly deployed as agents that invoke external tools through structured function calls. While recent work reports strong tool-calling performance under standard English-centric evaluations, the robustness of tool calling under multilingual user interactions remains underexplored. In this work, we introduce MLCL, a diagnostic benchmark, and conduct a systematic evaluation of multilingual tool calling across Chinese, Hindi, and the low-resource language Igbo. Through fine-grained error analysis, we show that many failures occur despite correct intent understanding and tool selection. We identify parameter value language mismatch as a dominant failure mode, where models generate semantically appropriate parameter values in the users language, violating language-invariant execution conventions. We further evaluate several inference-time system strategies and find that while these strategies substantially reduce language-induced execution errors, none of them can fully recover English-level performance.

[Findings] LR-DWM: Efficient Watermarking for Diffusion Language Models ▶ *Underline Video*
Ofek Raban, Ethan Fetaya and Gal Chechik
Watermarking (WM) is a critical mechanism for detecting and attributing AI-generated content. Current WM methods for Large Language Models (LLMs) are predominantly tailored for autoregressive (AR) models: They rely on tokens being generated sequentially, and embed stable signals within the generated sequence based on the previously sampled text. Diffusion Language Models (DLMs) generate text via non-sequential iterative denoising, which requires significant modification to use WM methods designed for AR models. Recent work proposed to watermark DLMs by inverting the process when needed, but suffers significant computational or memory overhead. We introduce Left-Right Diffusion Watermarking (), a scheme that biases the generated token based on both left and right neighbors, when they are available. incurs minimal runtime and memory overhead, remaining close to the non-watermarked baseline DLM while enabling reliable statistical detection under standard evaluation settings. Our results demonstrate that DLMs can be watermarked efficiently, achieving high detectability with negligible computational and memory overhead.

[Long] MaDS: Long-Horizon GUI Automation via Synergizing Dual-Layer Memory and Multi-Round Debate ▶ *Underline Video*
Pengchen Chen, Shi Chen, Qiming Ye, Xinli Chen, Xinran Li and Wei Xiang
Automating Graphical User Interface (GUI) operations with Multimodal Large Language Models (MLLMs) is promising but remains bottlenecked in real-world long-horizon settings. Key challenges include ensuring precise grounding across diverse interfaces and handling irreversible errors in extended workflows. Current methods often struggle to distinguish targets in low Signal-to-Noise Ratio (SNR) environments and lack sufficient pre-execution verification to prevent error accumulation. To address this, we propose the Memory-augmented Debate System (MaDS). Specifically, MaDS combines: (1) a Dual-Layer Memory Module that integrates universal interaction priors with scenario-specific operational experience to mitigate grounding hallucinations; and (2) Multi-Round Debate that performs pre-execution verification, while transforming execution failures into retrievable Negative Warnings to reduce repeated errors. Additionally, we introduce MaDS-Benchmark, a benchmark for long-horizon mobile GUI tasks with process-oriented evaluation. Experiments show that MaDS achieves a 90.23% Task Success Rate on MaDS-Benchmark and strong performance on public benchmarks including AITW, AITZ, CAGUI, and GUIOdyssey.

[Findings] MALLM-GAN: Multi-Agent Large Language Model as Generative Adversarial Network for Synthesizing Tabular Data ▶ *Underline Video*
Yaobin Ling, Xiaoqian Jiang and Yejin Kim
In the era of big data, access to abundant data is crucial to driving research forward. However, such data are often inaccessible due to privacy concerns or high costs, particularly in the healthcare domain. Generating synthetic (tabular) data can address this, but existing models typically require substantial amounts of data to train effectively, contradicting our objective of solving data scarcity. To address this challenge, we propose a novel framework to generate synthetic tabular data, powered by large language models (LLMs) that emulates the architecture of a Generative Adversarial Network (GAN). By incorporating the data generation process as contextual information and

utilizing LLM as the optimizer, our approach significantly enhances the quality of synthetic data generation in common scenarios with small sample sizes. Our experimental results on public and private datasets demonstrate that our model outperforms several state-of-art models regarding generating higher quality synthetic data for downstream tasks while keeping the privacy of the real data in low data regime. Code is available at https://github.com/yling1105/MALLM-GAN.

[Long] MatchTIR: Fine-Grained Supervision for Tool-Integrated Reasoning via Bipartite Matching ▶ *Underline Video*

Changle Qu, Sunhao Dai, Hengyi Cai, Jun Xu, Shuaiqiang Wang and Dawei Yin

Tool-Integrated Reasoning (TIR) empowers large language models (LLMs) to tackle complex tasks by interleaving reasoning steps with external tool interactions. However, existing reinforcement learning methods typically rely on outcome- or trajectory-level rewards, assigning uniform advantages to all steps within a trajectory. This coarse-grained credit assignment fails to distinguish effective tool calls from redundant or erroneous ones, particularly in long-horizon multi-turn scenarios. To address this, we propose MatchTIR, a framework that introduces fine-grained supervision via bipartite matching-based turn-level reward assignment and dual-level advantage estimation. Specifically, we formulate credit assignment as a bipartite matching problem between predicted and ground-truth traces, utilizing two assignment strategies to derive dense turn-level rewards. Furthermore, to balance local step precision with global task success, we introduce a dual-level advantage estimation scheme that integrates turn-level and trajectory-level signals, assigning distinct advantage values to individual interaction turns. Extensive experiments on three benchmarks demonstrate the superiority of MatchTIR. Notably, our 4B model surpasses the majority of 8B competitors, particularly in long-horizon and multi-turn tasks. Our codes are available at https://anonymous.4open.science/r/MatchTIR.

[Long] Measure Twice, Click Once: Co-evolving Proposer and Visual Critic via Reinforcement Learning for GUI Grounding ▶ *Underline Video*

Wenkai Wang, Xiyun Li, Hongcan Guo, Wenhao Yu, Tianqing Fang, Haitao Mi, Dong Yu and Shengyu Zhang

Graphical User Interface (GUI) grounding requires mapping natural language instructions to precise pixel coordinates. However, due to visually homogeneous elements and dense layouts, models typically grasp semantic intent yet fail to achieve precise localization. While scaling sampling attempts (Pass@k) reveals potential gains, static self-consistency strategies derived from geometric clustering often yield limited improvements, as the model's predictions tend to be spatially dispersed. In this paper, we propose replacing static consistency strategies with a learnable selection mechanism that selects the optimal target by critiquing its own proposals rendered on the screenshot. Given the significant disparity between the model's grounding and critiquing capabilities, we propose a co-evolving Propose-and-Critic framework. To jointly optimize these, we introduce a maturity-aware adaptive co-evolutionary reinforcement learning paradigm. This approach dynamically balances the training objectives of proposer and critic, where the diversity of the proposer's outputs enhances critic robustness, while the critic's maturing discrimination capability conversely unlocks the proposer's potential for extensive spatial exploration, fostering the mutual reinforcement and co-evolution of both capabilities, thereby ensuring generalizability to adapt to diverse and complex interface layouts. Extensive experiments over 6 benchmarks show that our method significantly enhances both capabilities.

[Long] Memory-R1: Enhancing Large Language Model Agents to Manage and Utilize Memories via Reinforcement Learning ▶ *Underline Video*

Sikuan Yan, Xiufeng Yang, Zuchao Huang, Ercong Nie, Zifeng Ding, Zonggen Li, Xiaowen Ma, Jinhe Bi, Kristian Kersting, Jeff Z. Pan, Hinrich Schuetze, Volker Tresp and Yunpu Ma

Large Language Models (LLMs) have demonstrated impressive capabilities across a wide range of NLP tasks, but they remain fundamentally stateless, constrained by limited context windows that hinder long-horizon reasoning. Recent efforts to address this limitation often augment LLMs with an external memory bank, yet most existing pipelines are static and heuristic-driven, lacking a learned mechanism for deciding what to store, update, or retrieve. We present Memory-R1, a reinforcement learning (RL) framework that equips LLMs with the ability to actively manage and utilize external memory through two specialized agents: a Memory Manager that learns structured operations, including ADD, UPDATE, DELETE, and NOOP; and an Answer Agent that pre-selects and reasons over relevant entries. Both agents are fine-tuned with outcome-driven RL (PPO and GRPO), enabling adaptive memory management with minimal supervision. With only 152 training QA pairs, Memory-R1 outperforms strong baselines and generalizes across diverse question types, three benchmarks (LoCoMo, MSC, LongMemEval), and multiple model scales (3B-14B).

[Findings] MemPO: Self-Memory Policy Optimization for Long-Horizon Agents ▶ *Underline Video*

Ruoran Li, Xinghua Zhang, Haiyang Yu, Shitong Duan, Xiang Li, Wenxin Xiang, Chonghua Liao, Xudong Guo, Yongbin Li and Jinli Suo

Long-horizon agents face the challenge of growing context size during interaction with environment, which degrades the performance and stability. Existing methods typically introduce the external memory module and look up the relevant information from the stored memory, which prevents the model itself from proactively managing its memory content and aligning with the agent's overarching task objectives. To address these limitations, we propose the self-memory policy optimization algorithm (MemPO), which enables the agent (policy model) to autonomously summarize and manage their memory during interaction with environment. By improving the credit assignment mechanism based on memory effectiveness, the policy model can selectively retain crucial information, significantly reducing token consumption while preserving task performance. Extensive experiments and analyses confirm that MemPO achieves absolute F1 score gains of 25.98 over the base model and 7.1 over the previous SOTA baseline, while reducing token usage by 67.58% and 73.12%.

[Findings] MemTR: Enhancing Tool-Calling Reliability via Uncertainty-Triggered FFN-Space Retracing ▶ *Underline Video*

Hongtao Duan, Lu Jiang, Minying Zhang, Xiaobing Zhu, Tianpeng Bu, Hao Jiang, Xinyu Wei and Lulu Hu
Tool calling requires Large Language Models (LLMs) to generate structured decisions including tool names and schema-constrained arguments, where small decoding mistakes can cause hard failures. Existing methods either rely on costly tool-use training data or only constrain syntax, leaving tool selection and argument value errors largely unsolved. We analyze tool calling failures through a Where-When lens: (Where) failures correlate with persistent uncertainty in late transformer layers, (When) uncertainty concentrates on content-bearing tokens (tool names and argument values) rather than schema tokens. Based on this, and motivated by evidence that transformer Feed Forward Networks (FFNs) act as key-value style memories that store and retrieve factual or associative mappings, we propose Memory Space Tool Retracing (MemTR), a weight-free decoding-time method that retrieves relevant tool evidence from the tool library and mixes it into the FFN-output at the uncertain layer, treating FFNs as key-value memories. Through extensive experiments on various model families (Qwen, Llama, and xLAM) and benchmarks (BFCL, ACEBench, APIBank), MemTR reduces tool calling failures by 2%-9% with only 1%-2% runtime overhead, without any fine-tuning or additional tool-use training data.

[Findings] MENTOR: Mitigating Identity Drift in Dynamic Role-Playing via Dual-Chain Structured Memory ▶ *Underline Video*
Zhu zhuoning, Xingyu Gao and Hailong Shi
Long-context LLM agents increasingly serve multiple users or personas within a single session, requiring stable identity and knowledge boundaries under frequent switching. We identify a common failure mode, identity drift, where models conflate user-specific states and leak information across roles. On BEAM-SWITCH, a benchmark for controlled multi-user switching, performance consistently degrades as switching intensifies, even when responses remain fluent and locally coherent. We propose MENTOR, a cognitive architecture that mitigates identity drift without fine-tuning. MENTOR uses a Dual-Chain Memory Mechanism: a Global Chain (G) for long-term event logging and isolated Role Chains (R_r) as per-role working memories, supported by a semantic Knowledge Graph (K) that filters and verifies role-admissible information before generation. Across six LLM families, MENTOR improves the overall score (Avg) from 0.46 to 0.75 on average (+0.29 absolute), with substantial gains in identity adherence and knowledge fidelity.

[Long] MERIT Feedback Elicits Better Bargaining in LLM Negotiators ▶ *Underline Video*
Jihwan Oh, Murad Aghazada, Yooju Shin, Se-Young Yun and Taehyeon Kim
Bargaining is often regarded as a logical arena rather than an art or a matter of intuition, yet Large Language Models (LLMs) still struggle to navigate it due to limited strategic depth and difficulty adapting to complex human factors. Current benchmarks rarely capture this limitation. To bridge this gap, we present a utility feedback centric framework. Our contributions are: (i) AgoraBench, a new benchmark spanning nine challenging settings (e.g., deception, monopoly) that supports diverse strategy modeling; (ii) human-aligned, economically grounded metrics derived from utility theory. This is operationalized via agent utility, negotiation power, and acquisition ratio that implicitly measure how well the negotiation aligns with human preference and (iii) a human preference grounded dataset with learning pipeline that strengthens LLMs bargaining ability through both prompting and finetuning. Empirical results indicate that baseline LLM strategies often diverge from human preferences, while our mechanism substantially improves negotiation performance, yielding deeper strategic behavior and stronger opponent awareness.

[Findings] MHSafeEval: Role-Aware Interaction-Level Evaluation of Mental Health Safety in Large Language Models ▶ *Underline Video*
Suhyun Lee, Palakorn Achananuparp, Neemesh Yadav, Ee-Peng Lim and Yang Deng
Large language models (LLMs) are increasingly explored as scalable tools for mental health counseling, yet evaluating their safety remains challenging due to the interactional and context-dependent nature of clinical harm. Existing evaluation frameworks predominantly assess isolated responses using coarse-grained taxonomies or static datasets, limiting their ability to diagnose how harms emerge and accumulate over multi-turn counseling interactions. In this work, we introduce R-MHSafe, a role-aware mental health safety taxonomy that characterizes clinically significant harm in terms of the interactional roles an AI counselor adopts, including perpetrator, instigator, facilitator, or enabler, combined with clinically grounded harm categories. Then, we propose MHSafeEval, a closed-loop, agent-based evaluation framework that formulates safety assessment as trajectory-level discovery of harm through adversarial multi-turn interactions, guided by role-aware modeling. Using R-MHSafe and MHSafeEval, we conduct a large-scale evaluation across state-of-the-art LLMs. Our results reveal substantial role-dependent and cumulative safety failures that are systematically missed by existing static benchmarks, and show that our framework significantly improves failure-mode coverage and diagnostic granularity.

[Long] Mitigating Safety Context Amnesia in Multimodal Reasoning Models via Intent-Guided Safety Reasoning ▶ *Underline Video*
Xiyao Dong, Guangsheng Cheng, YiLong Chen, Xiaojin Zhang and Kun He
Recent advances in Multimodal Large Reasoning Models (MLRMs) have enabled explicit chain-of-thought inference across vision and language, substantially improving performance on complex reasoning tasks. Despite these gains, the reasoning process introduces a subtle yet critical vulnerability. We identify an underexplored multimodal safety failure mode in which harmful objectives are embedded within ostensibly benign contexts, leading models to over-prioritize narrative coherence during reasoning. We term this phenomenon Safety Context Amnesia (SCA), wherein models correctly perceive risk-relevant visual cues but fail to enforce safety constraints as the reasoning process becomes dominated by contextual alignment. To mitigate SCA, we propose Intent-Guided Safety Reasoning (IGSR), an inference-time defense that operates without modifying target model parameters. IGSR employs a Perception Decoupler to extract objective visual evidence into a structured intent output, followed by a Cognitive Arbiter that enforces explicit safety constraints prior to generation. Extensive

experiments across multiple multimodal safety benchmarks demonstrate that IGSR improves defense success rates by over 62% compared to baselines, while largely preserving task utility. These results highlight the critical role of structured, intent-aware reasoning in achieving robust safety reasoning for multimodal reasoning models.

[Long] Model Internal Sleuthing: Finding Lexical Identity and Inflectional Features in Modern Language Models ▶ *Underline Video*

Michael Li and Nishant Subramani

Large transformer-based language models dominate modern NLP, yet our understanding of how they encode linguistic information relies primarily on studies of early models like BERT and GPT-2. We systematically probe 25 models from BERT Base to Qwen2.5-7B focusing on two linguistic properties: lexical identity and inflectional features across 6 diverse languages. We find a consistent pattern: inflectional features are linearly decodable throughout the model, while lexical identity is prominent early but increasingly weakens with depth. Further analysis of the representation geometry reveals that models with aggressive mid-layer dimensionality compression show reduced steering effectiveness in those layers, despite probe accuracy remaining high. Pretraining analysis shows that inflectional structure stabilizes early while lexical identity representations continue evolving. Taken together, our findings suggest that transformers maintain inflectional features across layers, while trading off lexical identity for compact, predictive representations.

[Findings] More Images, More Problems? A Controlled Analysis of VLM Failure Modes. ▶ *Underline Video*

Anurag Das, Adrian Bulat, Alberto Baldrati, oannis Maniadis Metaxas, Bernt Schiele, Georgios Tzimiropoulos and Brais Martinez

Large Vision Language Models (LVLMs) have demonstrated remarkable capabilities, yet their proficiency in understanding and reasoning over multiple images remains largely unexplored. While existing benchmarks have initiated the evaluation of multi-image models, a comprehensive analysis of their core weaknesses and their causes is still lacking. In this work, we introduce MIMIC (Multi-Image Model Insights and Challenges), a new benchmark designed to rigorously evaluate the multi-image capabilities of LVLMs. Using MIMIC, we conduct a series of diagnostic experiments that reveal pervasive issues: LVLMs often fail to aggregate information across images and struggle to track or attend to multiple concepts simultaneously. To address these failures, we propose two novel complementary remedies. On the data side, we present a procedural data-generation strategy that composes single-image annotations into rich, targeted multi-image training examples. On the optimization side, we analyze layer-wise attention patterns and derive an attention-masking scheme tailored for multi-image inputs. Experiments substantially improved cross-image aggregation, while also enhancing performance on existing multi-image benchmarks, outperforming prior state of the art across tasks. Data and code will be made available.

[Long] Multimodal Safety Evaluation in Generative Agent Social Simulations ▶ *Underline Video*

Alhim Vera, Carlos Hinojosa, Karen Sanchez, Haidar Bin Hamid, Donghoon Kim and Bernard Ghanem

Can generative agents be trusted in multimodal environments? Despite recent advances, agents remain limited in their ability to reason about safety, coherence, and trust across modalities. We introduce a reproducible simulation framework to evaluate generative agents in three aspects: (1) safety improvement over time via iterative plan revision in multimodal scenarios; (2) detection of unsafe activities across social contexts; and (3) social dynamics, measured through interaction and acceptance rates. These multimodal agents are evaluated using metrics that quantify plan revisions and unsafe-to-safe conversions. Experiments show that while agents detect direct multimodal contradictions, they often fail to align local revisions with global safety, achieving only a 55% success rate in correcting unsafe plans. We release a dataset of 1,000 multimodal plans, yielding more than 600,000 simulation steps. Notably, 45% of unsafe actions are accepted when paired with misleading visual cues, revealing a strong tendency to overtrust visual content. Code is available at https://github.com/AdonaiVera/X-CASE

[Findings] No One Fits All: From Fixed Prompting to Learned Routing in Multilingual LLMs ▶ *Underline Video*

Wei-Chi Wu, Sheng-Lun Wei, Hen-Hsen Huang and Hsin-Hsi Chen

Translation-based prompting is widely used in multilingual LLMs, yet its effectiveness varies across languages and tasks. We evaluate prompting strategies across ten languages of different resource levels and four benchmarks. Our analysis shows that no single strategy is universally optimal. Translation strongly benefits low-resource languages even when translation quality is imperfect, high-resource languages gain little, and prompt-based self-routing underperforms explicit translation. Motivated by these findings, we formulate prompting strategy selection as a learned decision problem and introduce lightweight classifiers that predict whether native or translation-based prompting is optimal for each instance. The classifiers achieve statistically significant improvements over fixed strategies across four benchmarks and generalize to unseen task formats not observed during training. Further analysis reveals that language resource level, rather than translation quality alone, determines when translation is beneficial.

[Findings] No-Worse Context-Aware Decoding: Preventing Neutral Regression in Context-Conditioned Generation ▶ *Underline Video*

Yufei Tao and Ameeta Agrawal

Large language models (LLMs) can answer questions and summarize documents when conditioned on external contexts (e.g., retrieved evidence), yet context use remains unreliable: models may overwrite an already-correct output (neutral regression) even when the context is non-informative. We formalize neutral regression as a do-no-harm requirement and quantify it by measuring accuracy drops on baseline-correct items under answer-consistent contexts. We propose No-Worse Context-Aware Decoding (NWCAD), a decode-time adapter built on a two-stream setup with a two-stage gate: it backs off to no-context decoding when the context is non-informative, and otherwise uses context-conditioned decoding with a contrastive fallback under uncertainty. We evaluate NWCAD on benchmarks that separate do-no-harm

reliability from context utilization (accuracy gains on genuinely helpful contexts). NWCAD prevents neutral regression on baseline-correct items while preserving strong context-driven accuracy on helpful contexts.

[TACL] Objectifying the Subjective: Cognitive Biases in Topic Interpretations ▶ *Underline Video*

Swapnil Hingmire, Ze Shi Li, Shiyu (Vivienne) Zeng, Ahmed Musa Awon, Luiz Franciscatto Guerra and Neil Ernst

Interpretation of topics is crucial for their downstream applications. State-of-the-art evaluation measures of topic quality such as coherence and word intrusion do not measure how much a topic facilitates the exploration of a corpus. To design evaluation measures grounded on a task, and a population of users, we do user studies to understand how users interpret topics. We propose constructs of topic quality and ask users to assess them in the context of a topic and provide rationale behind evaluations. We use reflexive thematic analysis to identify themes of topic interpretations from rationales. Users interpret topics based on availability and representativeness heuristics rather than probability. We propose a theory of topic interpretation based on the anchoring-and-adjustment heuristic: users anchor on salient words and make semantic adjustments to arrive at an interpretation. Topic interpretation can be viewed as making a judgment under uncertainty by an ecologically rational user, and hence cognitive biases aware user models and evaluation frameworks are needed.

[Findings] OMHBench: Benchmarking Balanced and Grounded Omni-Modal Multi-Hop Reasoning ▶ *Underline Video*

Seunghee Kim, Ingyu Bang, Seokgyu Jang, Changhyeon Kim, Sanghwan Bae, Jihun Choi, Richeng Xuan and Taeuk Kim

Multimodal Large Language Models (MLLMs) have increasingly supported omni-modal processing across text, vision, and speech. However, existing evaluation frameworks for such models suffer from critical limitations, including modality shortcuts and biased reasoning paths. To address these challenges, we propose OMHBench, a novel benchmark designed to rigorously evaluate omni-modal multi-hop reasoning. It consists of 6,144 questions with balanced reasoning paths that are jointly grounded across all three modalities. Extensive evaluation of 13 state-of-the-art models reveals that (1) a large performance gap exists between proprietary and open-source MLLMs and (2) even proprietary models exhibit high sensitivity to reasoning path variations, resulting in asymmetric omni-modal grounding. Notably, models struggle when processing the speech modality, underscoring the need for balanced, multi-hop evaluation of omni-modal intelligence.

[Long] PaCoRe: Learning to Scale Test-Time Compute with Parallel Coordinated Reasoning ▶ *Underline Video*

Jingcheng Hu, Yinmin Zhang, Shijie Shang, Xiaobo Yang, Yue Peng, Zhewei Huang, Hebin Zhou, Xin Wu, Jie Cheng, Fangi Wan, Xiangwen Kong, Chengyuan Yao, Kaiwen Yan, Ailin Huang, Hongyu Zhou, Qi Han, Zheng Ge, Xiangyu Zhang and Heung-Yeung Shum

We introduce Parallel Coordinated Reasoning (PaCoRe), a training-and-inference framework designed to overcome a central limitation of contemporary language models: their inability to scale test-time compute (TTC) far beyond sequential reasoning under a fixed context window. PaCoRe departs from the traditional sequential paradigm by driving TTC through massive parallel exploration coordinated via a message-passing architecture in multiple rounds. Each round launches many parallel reasoning trajectories, compacts their findings into context-bounded messages, and synthesizes these messages to guide the next round and ultimately produce the final answer. Trained end-to-end with large-scale, outcome-based reinforcement learning, the model masters the synthesis abilities required by PaCoRe and scales to multi-million-token effective TTC without exceeding context limits. The approach yields strong improvements across diverse domains and notably pushes reasoning beyond frontier systems in mathematics: an 8B model reaches 94.5% on HMMT 2025, surpassing GPT-5s 93.2% by scaling effective TTC to roughly two million tokens. We open-source model checkpoints, training data, and the full inference pipeline to accelerate follow-up work.

[Long] PAR: Training-Free Positional Perturbation and Attention Recycling for Faithful OCR ▶ *Underline Video*

Yao Yao, Manwen Liao, Weitian Zhang, Zuchao Li and Hai Zhao

In high-precision scenarios, vision language models suffer from Linguistic Priors Hallucination. When processing familiar text, models tend to over-rely on internal parametric knowledge, effectively "reciting" the content rather than "reading" the image. In this paper, we first systematically investigate this phenomenon by constructing the GlitchText Probing Dataset. We discover that the model's reliance on visual grounding diminishes significantly as the generation length increases. To mitigate this, we propose PAR (Positional Perturbation and Attention Recycling), a training-free, inference-time intervention framework. PAR consists of two parts: (1) Positional Perturbation (PP) injects structured phase noise into the rotary positional embeddings; (2) Foveal Attention Recycling (FAR) detects over-confident linguistic priors and dynamically redistributes attention mass back to important visual regions. Extensive experiments across state-of-the-art models, demonstrate that PAR significantly reduces hallucination rates (reducing CER by 12%), particularly in long-context scenarios, while maintaining robust generalization on standard benchmarks.

[Long] PEARL: Self-Evolving Assistant for Time Management with Reinforcement Learning ▶ *Underline Video*

Bingxuan Li, Jeonghwan Kim, Cheng Qian, Xiusi Chen, Eitan Anzenberg, Niran Kundapur and Heng Ji

Overlapping calendar invitations force busy professionals to repeatedly decide which meetings to attend, reschedule, or decline. We refer to this preference-driven decision process as calendar conflict resolution. Automating this decision process is crucial yet challenging. Scheduling logistics can drain hours, and human delegation often fails at scale, which motivates us to ask: Can we trust large language models (LLMs) or language agents to manage time? To enable a systematic study of this question, we introduce CalConflictBench, a benchmark for long-horizon calendar conflict resolution. In CalConflictBench, conflicts are presented to agents round-by-round over a calendar year, requiring them to infer and adapt to user preferences progressively. Our experiments show that current LLM agents perform poorly with high error rates, e.g., Qwen-3-30B-Think has an average error rate of 35%. To address this gap, we propose PEARL, a reinforcement-learning

framework that (i) augments the language agent with an external preference memory that stores and updates inferred strategies (e.g., attendee priorities, topic importance, time/location preferences), and (ii) optimizes the agent with round-wise rewards that directly supervise decision correctness, ranking quality, and memory usage across rounds. Experiments on CalConflictBench show that PEARL achieves an error reduction rate of 0.76 and a 55% improvement in average error rate compared to the strongest baseline.

[Long] PIAST: Rapid Prompting with In-context Augmentation for Scarce Training data ► *Underline Video*

Pawe Batorski and Paul Swoboda

LLMs are highly sensitive to prompt design, but handcrafting effective prompts is difficult and often requires intricate crafting of few-shot examples. We propose a fast automatic prompt construction algorithm that augments human instructions by generating a small set of few shot examples. Our method iteratively replaces/drops/keeps few-shot examples using Monte Carlo Shapley estimation of example utility. For faster execution, we use aggressive subsampling and a replay buffer for faster evaluations. Our method can be run using different compute time budgets. Under a limited budget, it outperforms prior automatic prompting methods on text simplification and mathematical reasoning (GSM8K, DeepMath, Math500), while achieving second-best results on classification and summarization and third-best on MedQA. With an extended, yet still modest budget, PIAST sets a new state of the art among automatic prompting methods on classification, simplification, GSM8K, DeepMath, and Math500. Overall, our results suggest that optimizing in-context examples, rather than exhaustively searching over instruction rewrites is the dominant lever for fast and data-efficient prompt engineering. We will release code and data upon acceptance.

[Findings] POLARIS: A Gödel Agent Framework for Small Language Models through Experience-Abstracted Policy Repair ► *Underline Video*

ADITYA NAMDEV KAKADE, Vivek Srivastava and Shirish Karande

Gödel agent realize recursive self-improvement: an agent inspects its own policy and traces and then modifies that policy in a tested loop. We introduce Polaris, a Gödel agent for compact models that performs policy repair via experience abstraction, turning failures into policy updates through a structured cycle of analysis, strategy formation, abstraction, and minimal code patch repair with conservative checks. Unlike response level self-correction or parameter tuning, Polaris makes policy level changes with small, auditable patches that persist in the policy and are reused on unseen instances within each benchmark. As part of the loop, the agent engages in meta reasoning: it explains its errors, proposes concrete revisions to its own policy, and then updates the policy. To enable cumulative policy refinement, we introduce experience abstraction, which distills failures into compact, reusable strategies that transfer to unseen instances. On MGSM, DROP, GPQA, and LitBench (covering arithmetic reasoning, compositional inference, graduate-level problem solving, and creative writing evaluation), a 7-billion-parameter model equipped with Polaris achieves consistent gains over the base policy and competitive baselines.

[Long] POLYCHARTQA: Benchmarking Large Vision-Language Models with Multilingual Chart Question Answering ► *Underline Video*

Yichen Xu, Liangyu Chen, Liang Zhang, Zihao Yue, Jianzhe Ma, Wenxuan Wang and Qin Jin

Charts are a universally adopted medium for data communication, yet existing chart understanding benchmarks are overwhelmingly English-centric, limiting their accessibility and relevance to global audiences. To address this limitation, we introduce PolyChartQA, the first large-scale multilingual benchmark for chart question answering, comprising 22,606 charts and 26,151 QA pairs across 10 diverse languages. PolyChartQA is constructed through a scalable pipeline that enables efficient multilingual chart generation via data translation and code reuse, supported by LLM-based translation and rigorous quality control. We systematically evaluate multilingual chart understanding with PolyChartQA on state-of-the-art LVLMs and reveal a significant performance gap between English and other languages, particularly low-resource ones. Additionally, we introduce a companion multilingual chart question answering training set, PolyChartQA-Train, on which fine-tuning LVLMs yields substantial gains in multilingual chart understanding across diverse model sizes and architectures. Together, our benchmark provides a foundation for developing globally inclusive vision-language models capable of understanding charts across diverse linguistic contexts. Codes and datasets are available on https://github.com/Road2Redemption/PolyChartQA.

[Short] Pref-CTRL: Preference Driven LLM Alignment using Representation Editing ► *Underline Video*

Imranul Ashrafi, Inigo Jauregi Unanue and Massimo Piccardi

Test-time alignment methods offer a promising alternative to fine-tuning by steering the outputs of large language models (LLMs) at inference time with lightweight interventions on their internal representations. Recently, a prominent and effective approach, RE-Control (Kong et al., 2024), has proposed leveraging an external value function trained over the LLM's hidden states to guide generation via gradient-based editing. While effective, this method overlooks a key characteristic of alignment tasks, i.e. that they are typically formulated as learning from human preferences between candidate responses. To address this, in this paper we propose a novel preference-based training framework, Pref-CTRL, that uses a multi-objective value function to better reflect the structure of preference data. Our approach has outperformed RE-Control on two benchmark datasets and showed greater generalization on out-of-domain datasets. Our source code is available at https://github.com/UTS-nlPUG/pref-ctrl.

[Long] PRInTS: Reward Modeling for Long-Horizon Information Seeking ► *Underline Video*

Jaewoo Lee, Archiki Prasad, Justin Chen, Zaid Khan, Elias Stengel-Eskin and Mohit Bansal

Information-seeking is a core capability for AI agents, requiring them to gather and reason over tool-generated information across long trajectories. However, such multi-step information-seeking tasks remain challenging for agents backed by language models. While process

reward models (PRMs) can guide agents by ranking candidate steps at test-time, existing PRMs, designed for short reasoning with binary judgment, cannot capture richer dimensions of information-seeking steps, such as tool interactions and reasoning over tool outputs, nor handle the rapidly growing context in long-horizon tasks. To address these limitations, we introduce PRInTS, a generative PRM trained with dual capabilities: (1) dense scoring based on the PRMs reasoning across multiple step quality dimensions (e.g., interpretation of tool outputs, tool call informativeness) and (2) trajectory summarization that compresses the growing context while preserving essential information for step evaluation. Extensive evaluations across FRAMES, GAIA (levels 1-3), and WebWalkerQA (easy-hard) benchmarks on multiple models, with ablations, reveal that best-of-n sampling with PRInTS enhances information-seeking in open-source models as well as specialized agents, matching or surpassing frontier models with a much smaller backbone agent and outperforming other strong reward modeling baselines.

[Findings] Private Seeds, Public LLMs: Realistic and Privacy-Preserving Synthetic Data Generation ▶ *Underline Video*

Qian Ma and Sarah Rajtmajer

Large language models (LLMs) have emerged as a powerful tool for synthetic data generation. A particularly important use case is producing synthetic replicas of private text, which requires carefully balancing privacy and utility. We propose Realistic and Privacy-Preserving Synthetic Data Generation (RPSG), which uses private seeds and integrates privacy-preserving strategies, including a formal differential privacy (DP) mechanism in the candidate selection, to generate realistic synthetic data. Comprehensive experiments against state-of-the-art private synthetic data generation methods demonstrate that RPSG achieves high fidelity to private data while providing strong privacy protection.

[Long] PROMPRINT: Prompt Fingerprinting via First-Token Response for LLM App Cloning Detection ▶ *Underline Video*

Jungmin Lee, Peizhuo Lv and Yeonjoon Lee

As Large Language Model applications (LLM apps) become widespread, system prompts that determine app behavior are increasingly regarded as intellectual property, raising concerns about leakage. Recent studies show that this threat is no longer theoretical, revealing the prevalence of cloned apps replicating system prompts from others on real-world platforms. These clones pose risks of copyright infringement and malicious misuse, highlighting the need for early and reliable detection. In this paper, we propose PROMPRINT, a novel fingerprinting approach for detecting cloned LLM apps without exposing their system prompts. Motivated by the insight that different system prompts yield distinct responses to the same query, PROMPRINT optimizes queries that induce the LLM to generate a specific first token associated with the given system prompt, resulting in distinctive query-first-token pairs. Experiments on four instruction-tuned LLMs show that generated pairs effectively identify the corresponding system prompts, achieving over 74% probability of generating the target token while remaining below 2.2% on average under other prompts. Furthermore, we demonstrate that our fingerprinting remains robust to partial system prompt modifications and effective under the injection of adversarial instructions.

[Findings] Prompting the Unknown: Understanding Response Uncertainty in Large Language Models ▶ *Underline Video*

Ze Yu Zhang, Arun Verma, Finale Doshi-Velez and Bryan Kian Hsiang Low

Large language models (LLMs) are widely used in decision-making across diverse domains. Ensuring the generation of safe and reliable responses is critical for the effective deployment of LLM-based applications, particularly in high-stakes domains such as healthcare and finance. Most of these applications typically use carefully crafted prompts to guide response generation; however, the relationship between prompts and the reliability of LLM-generated responses is not yet fully understood. To address this gap, we propose a novel prompt-response concept model that explains the relationship between the amount of task-relevant information (informativeness) provided in the prompt and the LLM-generated response uncertainty by identifying four sources of response uncertainty: prompt underspecification, model quality, task variability, and semantic redundancy. We prove that response uncertainty decreases as prompt informativeness or model quality increases, mirroring the behavior of epistemic uncertainty in probabilistic models. Our experimental results on real-world datasets further validate our proposed model and corroborate the theoretical results.

[Findings] Quantize What Counts: More for Keys, Less for Values ▶ *Underline Video*

Mohsen Hariri, Alan Luo, Weicong Chen, Tianyi Zhang, Qifan Wang, Xiaotian Han and Vipin Chaudhary

Large Language Models (LLMs) suffer inference-time memory bottlenecks dominated by the attention Key-Value (KV) cache, which scales with model size and context length. While KV-cache quantization alleviates this cost, bit allocation between keys and values is often tuned heuristically, lacking theoretical grounding and generalizability. This paper proposes two theorems that anchor mixed-precision KV quantization in the intrinsic geometry of Transformer models. First, key weight matrices systematically have larger spectral and Frobenius norms than value matrices, implying higher information density along the key path. Second, for any given memory budget, prioritizing precision for keys over values strictly reduces quantization error and better preserves accuracy. Empirical evaluations across various prominent LLMs and benchmarks show that key-favored allocations (e.g., 4-bit keys, 2-bit values) retain up to 98.3% accuracy compared to uniform allocations (e.g., 4-bit for both), while conserving memory. These results transform bit allocation from ad hoc tuning into a theoretically grounded, geometry-driven design principle for efficient LLM inference. Source code is available at https://github.com/mohsenhariri/spectral-kv.

[Findings] RACER: Retrieval-Augmented Contextual Rapid Speculative Decoding ▶ *Underline Video*

Zihong Zhang, Zuchao Li, Lefei Zhang, Ping Wang and Hai Zhao

Autoregressive decoding in Large Language Models (LLMs) generates one token per step, causing high inference latency. Speculative

decoding (SD) mitigates this through a guess-and-verify strategy, but existing training-free variants face trade-offs: retrieval-based drafts break when no exact match exists, while logits-based drafts lack structural guidance. We propose RACER (Retrieval-Augmented Contextual Rapid Speculative Decoding), a lightweight and training-free method that integrates retrieved exact patterns with logit-driven future cues. This unification supplies both reliable anchors and flexible extrapolation, yielding richer speculative drafts. Experiments on Spec-Bench, HumanEval, and MGSM-ZH demonstrate that RACER consistently accelerates inference, achieving more than 2x speedup over autoregressive decoding, and outperforms prior training-free methods, offering a scalable, plug-and-play solution for efficient LLM decoding. Our source code is available at https://github.com/hkr04/RACER.

[Findings] Reason Only When Needed: Efficient Generative Reward Modeling via Model-Internal Uncertainty ▶ *Underline Video*

Chao Xue, Yao Wang, Mengqiao Liu, Di Liang, Xingsheng Han, Peiyang Liu, Xianjie Wu, Chenyao Lu, Lei Jiang, Yu Lu, Haibo Shi, Shuang Liang, Minlong Peng and Flora D. Salim

Recent advancements in the Generative Reward Model (GRM) have demonstrated its potential to enhance the reasoning abilities of LLMs through Chain-of-Thought (CoT) prompting. Despite these gains, existing implementations of GRM suffer from two critical limitations. First, CoT prompting is applied indiscriminately to all inputs regardless of their inherent complexity. This introduces unnecessary computational costs for tasks amenable to fast, direct inference. Second, existing approaches primarily rely on voting-based mechanisms to evaluate CoT outputs, which often lack granularity and precision in assessing reasoning quality. In this paper, we propose E-GRM, an efficient generative reward modeling framework grounded in model-internal uncertainty. E-GRM leverages the convergence behavior of parallel model generations to estimate uncertainty and selectively trigger CoT reasoning only when needed, without relying on handcrafted features or task-dependent signals. To improve reward fidelity, we introduce a lightweight discriminative scorer trained with a hybrid regression-ranking objective to provide fine-grained evaluation of reasoning paths. Experiments on multiple reasoning benchmarks show that E-GRM substantially reduces inference cost while consistently improving answer accuracy, demonstrating that model-internal uncertainty is an effective and general signal for efficient reasoning-aware reward modeling.

[Findings] ReasonIF: Large Reasoning Models Fail to Follow Instructions During Reasoning ▶ *Underline Video*

Yongchan Kwon, Shang Zhu, Federico Bianchi, Kaitlyn Zhou and James Zou

The ability of large language models (LLMs) to follow user instructions is central to their reliability, safety, and usefulness. While prior studies assess instruction adherence in the models main responses, we argue that it is also critical for large reasoning models (LRMs) to follow user instructions throughout their reasoning process. Reasoning instruction following makes LRMs more controllable and transparent, while reducing risks of undesirable shortcuts, hallucinations, or reward hacking within reasoning traces. To evaluate this dimension, we introduce ReasonIF, a systematic benchmark for assessing reasoning instruction following. ReasonIF includes six categories of instruction prompts, spanning multilingual reasoning, and length control. Across many open-source LRMs including GPT-OSS, Qwen3, and DeepSeek-R1, we find substantial failures in reasoning instruction adherence: the highest instruction following score (IFS) remains below 0.25, meaning that fewer than 25% of reasoning traces comply with the given instructions. Notably, as task difficulty increases, reasoning instruction following degrades further. We also explore two strategies to enhance reasoning instruction fidelity: (1) multi-turn reasoning and (2) Reasoning Instruction Finetuning (RIF) using synthetic data. RIF improves the IFS of GPT-OSS-20B from 0.11 to 0.27, indicating measurable progress but leaving ample room for improvement. We hope this work draws attention to reasoning-level instruction adherence as an underexplored but critical aspect of model alignment, and helps pave the way toward more controllable, interpretable, and trustworthy reasoning models.

[Findings] Reasoning Up the Instruction Ladder for Controllable Language Models ▶ *Underline Video*

Zishuo Zheng, Vidhisha Balachandran, Chan Young Park, Faeze Brahman and Sachin Kumar

As large language model (LLM) based systems take on high-stakes roles in real-world decision-making, they must reconcile competing instructions from multiple sources within a single prompt context. Enforcing an instruction hierarchy, where higher-level directives override lower-priority requests, is critical to the reliability and controllability of LLMs. In this work, we reframe instruction hierarchy resolution as a reasoning task. The model must first "think" about the relationship between a given user prompt and higher-priority instructions before generating a response. To enable this capability, we construct VerIH, a training dataset of constraint-following tasks with verifiable answers, comprising aligned and conflicting system–user instructions. We show that lightweight reinforcement learning with VerIH effectively transfers general reasoning capabilities of models to instruction prioritization. Our method leads to consistent improvements across multiple model families and scales on instruction following and instruction hierarchy benchmarks, achieving ~20% absolute improvement in conflict setups. Our method also generalizes to safety-critical scenarios beyond the training distribution, exhibiting increased robustness against jailbreak and prompt injection, reducing absolute attack success rates by up to 20%. Our results establish reasoning over instruction hierarchies as a practical mechanism for improving AI reliability, where targeted updates to system prompts produce predictable, controllable, and robust changes in model behavior.

[Findings] RecMem: Recurrence-based Memory Consolidation for Efficient and Effective Long-Running LLM Agents ▶ *Underline Video*

Zijie Dai, Shiyuan Deng, Sheng Guan, Yizhou Tian, Xin Yao, Xiao Yan and James Cheng;

Memory systems often organize user-agent interactions as retrievable external memory and are crucial for long-running agents by overcoming the limited context windows of LLMs. However, existing memory systems invoke LLMs to process every incoming interaction for memory extraction, and such an eager memory consolidation scheme leads to substantial token consumption. To tackle this problem, we propose

RecMem by rethinking when memory consolidation should be conducted. RecMem stores incoming interactions in a subconscious memory layer and encode them using lightweight embedding models for retrieval. LLMs are only invoked to extract episodic and semantic memory when sustained recurrence are observed for semantically similar interactions. Such recurrence-based consolidation works because these interactions correspond to a semantic cluster with rich information and thus are worth extraction and summarization. To improve accuracy, RecMem also incorporates a semantic refinement mechanism that recovers the fine-grained facts omitted by memory extraction. Experiments show that RecMem reduces the memory construction token cost of three SOTA memory systems by up to 87% while exceeding their accuracy.

[Long] Red Teaming Large Reasoning Models ▶ *Underline Video*

Jiawei Chen, Yang Yang, Chao Yu, Yu Tian, Zhi Cao, Xue Yang, Linghao Li, Hang Su and Zhaoxia Yin

Large Reasoning Models (LRMs) have emerged as a powerful advancement in multi-step reasoning tasks, offering enhanced transparency and logical consistency through explicit chains of thought (CoT). However, these models introduce novel safety and reliability risks, such as CoT-hijacking and prompt-induced inefficiencies, which are not fully captured by existing evaluation methods. To address this gap, we propose Rt-LRM, a unified benchmark designed to assess the trustworthiness of LRMs. Rt-LRM evaluates three core dimensions: truthfulness, safety and efficiency. Beyond metric-based evaluation, we further introduce the training paradigm as a key analytical perspective to investigate the systematic impact of different training strategies on model trustworthiness. We achieve this by designing a curated suite of 30 reasoning tasks from an observational standpoint. We conduct extensive experiments on 26 models and identify several valuable insights into the trustworthiness of LRMs. For example, LRMs generally face trustworthiness challenges and tend to be more fragile than Large Language Models (LLMs) when encountering reasoning-induced risks. These findings uncover previously underexplored vulnerabilities and highlight the need for more targeted evaluations. In addition, we release a scalable toolbox for standardized trustworthiness research to support future advancements in this important field.

[Findings] ReFreeKV: Towards Threshold-Free KV Cache Compression ▶ *Underline Video*

Xuanfan Ni, Liyan Xu, Chenyang Lyu, Longyue Wang, Mo Yu, Lemao Liu, Fandong Meng, Jie Zhou and Piji Li

To reduce memory consumption during LLM inference, a handful of methods have been proposed for KV cache pruning. While these techniques can accomplish lossless memory reduction on many datasets, they often hinge on an under-emphasized condition: an input/domain-specific threshold for KV cache budget needs to be pre-determined to achieve the optimal performance. However, such input-sensitive design may be considerably limited in real-world scenarios, as open-domain inputs span diverse domains, lengths and difficulty levels, without clear boundaries for threshold selection. As a result, the dependence of such input-sensitive threshold can be a fundamental limitation that causes large degradation on arbitrary inputs. In this work, we propose a new objective that lifts the threshold constraints for robust KV compression, advocating for "threshold-free" methods that adaptively adjust budget allocation while preserving full-cache performance. We then propose a novel method, ReFreeKV, serving as the first instantiation of this objective. Extensive experiments across 13 datasets with diverse context lengths, task types, and model sizes demonstrate its efficacy and efficiency. Our code is publicly released at https://github.com/Patrick-Ni/ReFreeKV.

[Findings] Reliability-Aware Adaptive Self-Consistency for Efficient Sampling in LLM Reasoning ▶ *Underline Video*

Junseok Kim, Nakyeong Yang, Kyungmin Min and Kyomin Jung

Self-Consistency improves reasoning reliability through multi-sample aggregation, but incurs substantial inference cost. Adaptive self-consistency methods mitigate this issue by adjusting the sampling budget; however, they rely on count-based stopping rules that treat all responses equally, often leading to unnecessary sampling. We propose Reliability-Aware Adaptive Self-Consistency (ReASC), which addresses this limitation by reframing adaptive sampling from response counting to evidence sufficiency, leveraging response-level confidence for principled information aggregation. ReASC operates in two stages: a single-sample decision stage that resolves instances confidently answerable from a single response, and a reliability-aware accumulation stage that aggregates responses by jointly leveraging their frequency and confidence. Across five models and four datasets, ReASC consistently achieves the best accuracy-cost trade-off compared to existing baselines, yielding improved inference efficiency across model scales from 3B to 27B parameters. As a concrete example, ReASC reduces inference cost by up to 70% relative to self-consistency while preserving accuracy on GSM8K using Gemma-3-4B-it.

[Long] Retrieval Heads are Dynamic ▶ *Underline Video*

Yuping Lin, Zitao Li, Yue Xing, Pengfei He, Yingqian Cui, Yaliang Li, Bolin Ding, Jingren Zhou and Jiliang Tang

Recent studies have identified "retrieval heads" in Large Language Models (LLMs) responsible for extracting information from input contexts. However, prior works largely rely on static statistics aggregated across datasets, identifying heads that perform retrieval on average. This perspective overlooks the fine-grained temporal dynamics of autoregressive generation. In this paper, we investigate retrieval heads from a dynamic perspective. Through extensive analysis, we establish three core claims: (1) Dynamism: Retrieval heads vary dynamically across timesteps; (2) Irreplaceability: Dynamic retrieval heads are specific at each timestep and cannot be effectively replaced by static retrieval heads; and (3) Correlation: The model's hidden state encodes a predictive signal for future retrieval head patterns, indicating an internal planning mechanism. We validate these findings on the Needle-in-a-Haystack task and a multi-hop QA task, and quantify the differences on the utility of dynamic and static retrieval heads in a Dynamic Retrieval-Augmented Generation framework. Our study provides new insights into the internal mechanisms of LLMs.

[Long] REVEALER: Reinforcement-Guided Visual Reasoning for Element-Level Text-Image Alignment Evaluation

► *Underline Video*

FuLin Shi, Wenyi Xiao, Leilei Gan, Liang Ding and binchen binchen

Evaluating the alignment between textual prompts and generated images is critical for ensuring the reliability and usability of text-to-image (T2I) models. However, most existing evaluation methods rely on coarse-grained metrics or static Question Answering (QA) pipelines, which lack fine-grained interpretability and struggle to reflect human preferences. To address this, we propose REVEALER, a reinforcement-guided visual reasoning framework for element-level text-to-image alignment evaluation. Adopting a structured "grounding-reasoning-conclusion" paradigm, our method enables Multimodal Large Language Models (MLLMs) to explicitly localize semantic elements and derive interpretable alignment judgments. We optimize the model via Group Relative Policy Optimization (GRPO) using a multi-dimensional reward function that targets format compliance, localization precision, and alignment accuracy. Extensive experiments confirm that REVEALER achieves state-of-the-art results across four benchmarks. Notably, on EvalMuse-40K, it surpasses the strong proprietary Gemini 3 Pro and Training-based baselines with absolute accuracy gains of +4.2% and +13.3%, respectively. Ablation studies further demonstrate the efficacy of our method, contributing a cumulative 19.6% improvement over the base model.

[Long] RoboFailRing: Retrieval-Augmented and Language Grounding Failure Detection for VLM-enabled Robotic Manipulation ► *Underline Video*

Chenduo Ying, Linkang Du, Yuanchao Shu and Peng Cheng

Reliable failure detection and causal reasoning are critical in robotic manipulation, as their absence risks robot damage and endangers human safety. Although recent Vision-Language Models (VLMs) are employed to attempt failure detection and causality reasoning, they typically make retrospective assessment only after task completion, and their reasoning accuracy is often limited. To address these issues, we introduce RoboFailRing, which enables timely failure detection during task execution and enhances the reasoning accuracy of VLMs. It achieves rapid failure detection by retrieving a pre-constructed failure memory and returning a similarity-based decision. In addition, by providing grounded failure report to VLMs, it improves the accuracy of their reasoning about the failure causes and repair strategies. We evaluate RoboFailRing on two large-scale simulated datasets comprising over 6,000 failure trajectories and covering 81 distinct manipulation tasks. The results show that the average success rate of out-of-distribution failure detection reaches 80%, while the mean detection time is cut to roughly 50% of the baseline. Moreover, evaluations on real-world systems show an average 35% gain in VLM failure-reasoning accuracy. We make our code publicly available at: https://github.com/DynamicPoet/RoboFailRing.

[Long] RRAtention: Dynamic Block Sparse Attention via Per-Head Round-Robin Shifts for Long-Context Inference ► *Underline Video*

Siran Liu, Guoxia Wang, Sa Wang, Jinle Zeng, Haoyang Xie, Siyu Lou, Jiabin Yang, Dianhai Yu, Haifeng Wang and Chao Yang

The quadratic complexity of attention mechanisms poses a critical bottleneck for large language models processing long contexts. While dynamic sparse attention methods offer input-adaptive efficiency, they face fundamental trade-offs: requiring preprocessing, lacking global evaluation, violating query independence, or incurring high computational overhead. We present RRAttention, a novel dynamic sparse attention method that simultaneously achieves all desirable properties through a head round-robin (RR) sampling strategy. By rotating query sampling positions across attention heads within each stride, RRAttention maintains query independence while enabling efficient global pattern discovery with stride-level aggregation. Our method reduces complexity from O(L^2) to O(L^2/S^2) and employs adaptive Top-selection for optimal sparsity. Extensive experiments on natural language understanding (HELMET) and multimodal video comprehension (Video-MME) demonstrate that RRAttention recovers over 99% of full attention performance while computing only half of the attention blocks, achieving 2.4x speedup at 128K context length and outperforming existing dynamic sparse attention methods. The code is available at https://github.com/PaddlePaddle/PaddleFleet (see 'Research/RRAttention').

[Short] Selective Span-Level Unlearning for Large Language Models ► *Underline Video*

Chaewon Yoon, Dongjun Kim and Hyun-Je Song

Large language models (LLMs) trained on massive text corpora may inadvertently memorize sensitive or copyrighted content, motivating the need for more targeted unlearning. Selective LLM unlearning focuses on identifying token-level or span-level unlearning targets within a text, rather than treating entire sequences as unlearning targets. However, many existing selective approaches depend on external supervision to identify unlearning targets, which may misalign unlearning objectives with the model's internal behavior. In this paper, we propose a selective span-level unlearning method that is grounded entirely in model-intrinsic information. Our method first estimates token-level importance scores by contrasting gradient information induced by forget and retain datasets, identifying tokens that disproportionately contribute to information targeted for unlearning. These token-level importance scores are then used as anchors to identify coherent span-level unlearning targets via a self-consistency-based generation process, allowing the model to determine stable spans based on its own predictions. Experiments on two LLM unlearning benchmarks show that our approach achieves comparable unlearning performance while substantially better preserving retained knowledge.

[Long] SFTMix: Elevating Language Model Instruction Tuning with Mixup Recipe ► *Underline Video*

Yuxin Xiao, Shujian Zhang, Marzyeh Ghassemi and Wenxuan Zhou

To acquire instruction-following capabilities, large language models (LLMs) undergo instruction tuning, where they are trained on instruction-response pairs using next-token prediction (NTP). Efforts to improve instruction tuning often focus on higher-quality supervised fine-tuning (SFT) datasets, typically requiring data filtering with proprietary LLMs or human annotation. In this paper, we take a different

approach by proposing SFTMix, a novel Mixup-based recipe that elevates LLM instruction tuning without relying on well-curated datasets. We observe that LLMs exhibit uneven confidence across the semantic representation space. We argue that examples with different confidence levels should play distinct roles in instruction tuning: Confident data is prone to overfitting, while unconfident data is harder to generalize. Based on this insight, SFTMix leverages training dynamics to identify examples with varying confidence levels. We then interpolate them to bridge the confidence gap and apply a Mixup-based regularization to support learning on these additional, interpolated examples. We demonstrate the effectiveness of SFTMix in both instruction-following and healthcare-specific SFT tasks, with consistent improvements across LLM families and SFT datasets of varying sizes and qualities. Extensive analyses across six directions highlight SFTMix's compatibility with data selection, adaptability to compute-constrained scenarios, and scalability to broader applications.

[Long] Sigmoid Head for Quality Estimation under Language Ambiguity ▶ *Underline Video*

Tu Anh Dinh and Jan Niehues

Language model (LM) probability is not a reliable quality estimator, as natural language is ambiguous. When multiple output options are valid, the model's probability distribution is spread across them, which can misleadingly indicate low output quality. This issue is caused by two reasons: (1) LMs' final output activation is softmax, which does not allow multiple correct options to receive high probabilities simultaneuously and (2) LMs' training data is single, one-hot encoded references, indicating that there is only one correct option at each output step. We propose training a module for Quality Estimation on top of pre-trained LMs to address these limitations. The module, called Sigmoid Head, is an extra unembedding head with sigmoid activation to tackle the first limitation. To tackle the second limitation, during the negative sampling process to train the Sigmoid Head, we use a heuristic to avoid selecting potentially alternative correct tokens. Our Sigmoid Head is computationally efficient during training and inference. The probability from Sigmoid Head is notably better quality signal compared to the original softmax head. As the Sigmoid Head does not rely on human-annotated quality data, it is more robust to out-of-domain settings compared to supervised QE.

[Long] SLR: Automated Synthesis for Scalable Logical Reasoning ▶ *Underline Video*

Lukas Helff, Ahmad Omar, Felix Friedrich, Antonia Wüst, Hikaru Shindo, Rupert Mitchell, Tim Woydt, Patrick Schramowski, Wolfgang Stammer and Kristian Kersting

We introduce SLR, an end-to-end framework for systematic evaluation and training of Large Language Models (LLMs) via Scalable Logical Reasoning. Given a user's task specification, SLR automatically synthesizes (i) an instruction prompt for an inductive reasoning task, (ii) a validation program, executable on model outputs to provide verifiable rewards, and (iii) the latent ground-truth rule. This process is fully automated, scalable, requires no human annotations, and offers precise control over task difficulty. Using SLR, we create SLR-Bench, a benchmark comprising 19k prompts organized into 20 curriculum levels that progressively increase in relational, arithmetic, and recursive complexity. Large-scale evaluation reveals that contemporary LLMs readily produce syntactically valid rules, yet often fail at correct logical inference. Recent reasoning LLMs demonstrate improved performance but incur very high test-time computation, with costs exceeding $300 for just 1,000 prompts. Finally, curriculum learning via SLR doubles Llama-3-8B accuracy on SLR-Bench, achieving parity with Gemini-Flash-Thinking at a fraction of computational cost. Moreover, these reasoning capabilities generalize to a wide range of established benchmarks, underscoring the effectiveness of SLR for downstream reasoning.

[Long] Solve-Detect-Verify: Inference-Time Scaling with Flexible Generative Verifier ▶ *Underline Video*

Jianyuan Zhong, Zeju Li, Zhijian Xu, Xiangyu Wen, Kezhi Li and Qiang Xu

Complex reasoning with Large Language Models (LLMs) demands a careful balance between accuracy and computational cost. Verification is crucial for reliability but faces trade-off: robust process-based verifiers are computationally prohibitive, while fast verifiers lack precision. We introduce flexive, a unified generative verifier designed to navigate this trade-off by dynamically allocating compute between rapid fast thinking and deliberative slow thinking. A key innovation is our training strategy: we use Group Relative Policy Optimization (GRPO) to specifically enhance the reliability of the fast mode. This targeted training generalizes effectively, elevating the slow mode to state-of-the-art open-source performance. To deploy flexive, we propose the solve-detect-verify (SDV) pipeline. Moving beyond static Best-of-N ranking, SDV employs an iterative refinement process that utilizes likelihood-based probing to detect solution completion, curtailing overthinking, and leverages flexive's feedback for targeted correction. Solve-detect-verify establishes a new open-source state-of-the-art on ProcessBench, outperforming GenPRM-32B while requiring ~2.3x fewer TFLOPS and 15x less training data. On AIME 2024, the full SDV pipeline achieves 83.3% accuracy, surpassing strong baselines while using significantly fewer tokens.

[Long] SPARD: Self-Paced Curriculum for RL Alignment via Integrating Reward Dynamics and Data Utility ▶ *Underline Video*

Xuyang Zhi, Peilun zhou, Chengqiang Lu, Hang Lv, Yiwei Liang, Rongyang Zhang, Yan Gao, YI WU, Yao Hu, Hongchao Gu, Defu Lian, Hao Wang and Enhong Chen

The evolution of Large Language Models (LLMs) is shifting the focus from single, verifiable tasks toward complex, open-ended real-world scenarios, imposing significant challenges on the post-training phase. In these settings, the scale and complexity of reward systems have grown significantly, transitioning toward multi-objective formulations that encompass a comprehensive spectrum of model capabilities and application contexts. However, traditional methods typically rely on fixed reward weights, ignoring non-stationary learning dynamics and struggling with data heterogeneity across dimensions. To address these issues, we propose SPARD, a framework that establishes an automated, self-paced curriculum by perceiving learning progress to dynamically adjust multi-objective reward weights and data importance, thereby synchronizing learning intent with data utility for optimal performance. Extensive experiments across multiple

benchmarks demonstrate that SPARD significantly enhances model capabilities across all domains. Our code is publicly available at https://github.com/USTC-StarTeam/SPARD.

[Findings] SPASM: Stable Persona-driven Agent Simulation for Multi-turn Dialogue Generation ▶ *Underline Video*
Han Luo and Guy Laban
Large language models are increasingly deployed in multi-turn settings such as tutoring, support, and counseling, where reliability depends on preserving consistent roles, personas, and goals across long horizons. This requirement becomes critical when LLMs are used to generate synthetic dialogues for training and evaluation, since LLM-LLM conversations can accumulate identity-related failures such as persona drift, role confusion, and echoing, where one agent gradually mirrors its partner. We introduce SPASM (Stable Persona-driven Agent Simulation for Multi-turn dialogue generation), a modular, stability-first framework that decomposes simulation into (i) persona creation via schema sampling, plausibility validation, and natural-language persona crafting, (ii) Client-Responder dialogue generation, and (iii) termination detection for coherent stopping. To improve long-horizon stability without changing model weights, we propose Egocentric Context Projection (ECP): dialogue history is stored in a perspective-agnostic representation and deterministically projected into each agents egocentric view before generation. Across three LLM backbones (GPT-4o-mini, DeepSeek-V3.2, Qwen-Plus) and nine Client-Responder pairings, we construct a dataset of 4,500 personas and 45,000 conversations (500 personas x 10 conversations per pairing). Ablations show ECP substantially reduces persona drift and, under human validation, eliminates echoing; embedding analyses recover persona structure and reveal strong responder-driven interaction geometry. Our code is available at https://github.com/lhannnn/SPASM.

[Findings] Spatial-RAG: Spatial Retrieval Augmented Generation for Real-World Geospatial Reasoning Questions ▶ *Underline Video*
Dazhou Yu, Riyang Bao, Ruiyu Ning, Jinghong Peng, Gengchen Mai and Liang Zhao
Answering real-world geospatial questions—such as finding restaurants along a travel route or amenities near a landmark—requires reasoning over both geographic relationships and semantic user intent. However, existing large language models (LLMs) lack spatial computing capabilities and access to up-to-date, ubiquitous real-world geospatial data, while traditional geospatial systems fall short in interpreting natural language. To bridge this gap, we introduce Spatial-RAG, a Retrieval-Augmented Generation (RAG) framework designed for geospatial question answering. Spatial-RAG integrates structured spatial databases with LLMs via a hybrid spatial retriever that combines sparse spatial filtering and dense semantic matching. It formulates the answering process as a multi-objective optimization over spatial and semantic relevance, identifying Pareto-optimal candidates and dynamically selecting the best response based on user intent. Experiments across multiple tourism and map-based QA datasets show that Spatial-RAG significantly improves performance over strong baselines.

[Long] SpecAgent: A Speculative Retrieval and Forecasting Agent for Code Completion ▶ *Underline Video*
George Ma, Anurag Koul, Qi Chen, Yawen Wu, Sachit Kuhar, Yu Yu, Aritra Sengupta, Varun Kumar and Murali Krishna Ramanathan
Large Language Models (LLMs) excel at code-related tasks but often struggle in realistic software repositories, where project-specific APIs and cross-file dependencies are crucial. Retrieval-augmented methods mitigate this by injecting repository context at inference time. Low inference time latency budget either affects retrieval quality or the added latency impacts user experience adversely. We address this limitation with SpecAgent, an agent that enhances both latency and code-generation quality by proactively exploring repository files during indexing and constructing speculative context that anticipates future edits in each file. This indexing-time asynchrony allows thorough context computation masking latency and the speculative nature of the context improves code-generation quality. Additionally, we identify the problem of future context leakage in existing benchmarks, which can inflate reported performance. To address this, we construct a synthetic, leakage-free benchmark that enables a more realistic evaluation of our agent against baselines. Experiments show that SpecAgent consistently achieves absolute gains of 9-11% (48-58% relative) compared to the best-performing baselines, while significantly reducing inference latency.

[Findings] SpecExtend: A Drop-in Enhancement for Speculative Decoding of Long Sequences ▶ *Underline Video*
Jungyoub Cha, Hyunjong Kim and Sungzoon Cho
Speculative decoding is a widely used technique for accelerating inference in large language models (LLMs), but its performance degrades as input length grows, with significant drops even at moderate lengths. Yet, this early degradation has remained largely underexplored. We introduce SpecExtend, a drop-in enhancement that improves speculative decoding on long sequences without additional training. SpecExtend integrates efficient attention mechanisms such as FlashAttention and Hybrid Tree Attention to accelerate prefill and verification steps. To improve both draft accuracy and speed on long inputs without retraining, we propose Cross-model Retrieval, a novel KV cache eviction strategy that leverages the target models attention scores to dynamically select relevant context for the smaller draft model. Extensive evaluations show that SpecExtend accelerates speculative decoding by up to 2.84x on 16K-token long document summarization and up to 3.86x on long-form reasoning, while preserving the short-input performance of state-of-the-art frameworks.

[Long] SSG: Logit-Balanced Vocabulary Partitioning for LLM Watermarking ▶ *Underline Video*
Chenxi Gu, Xiaoning Du and John Grundy
Watermarking has emerged as a promising technique for tracing the authorship of content generated by large language models (LLMs). Among existing approaches, the KGW scheme is particularly attractive due to its versatility, efficiency, and effectiveness in natural language generation. However, KGW's effectiveness degrades significantly under low-entropy settings such as code generation and mathematical reasoning. A crucial step in the KGW method is random vocabulary partitioning, which enables adjustments to token selection based on specific preferences. Our study revealed that the next-token probability distribution plays an critical role in determining how much, or even

whether, we can modify token selection and, consequently, the effectiveness of watermarking. We refer to this characteristic, associated with the probability distribution of each token prediction, as watermark strength. In cases of random vocabulary partitioning, the lower bound of watermark strength is dictated by the next-token probability distribution. However, we found that, by redesigning the vocabulary partitioning algorithm, we can potentially raise this lower bound. In this paper, we propose SSG (Sort-then-Split by Groups), a method that partitions the vocabulary into two logit-balanced subsets. This design lifts the lower bound of watermark strength for each token prediction, thereby improving watermark detectability. Experiments on code generation and mathematical reasoning datasets demonstrate the effectiveness of SSG.

[Findings] STELLA: A Multimodal LLM for Protein Functional Annotation via Unified Sequence-Structure Encoding ► *Underline Video*

Hongwang Xiao, Wenjun Lin, Xi Chen, Hui Wang, Kai Chen, Jiashan Li, Yuancheng Sun, Sicheng Dai, Boya Wu and Qiwei Ye

Understanding the intricate interplay among sequence, structure, and function remains a fundamental challenge in proteomics. The sequence-structure-function paradigm posits that biological roles are governed by the tertiary geometric conformations encoded within primary sequences; consequently, integrating these multi-modal descriptors is imperative for accurate functional annotation. While protein language models (pLMs) have achieved significant progress via representation learning on massive sequence data, they often lack the capacity to incorporate high-resolution structural information and the rich textual context that characterizes protein roles. In this work, we present STELLA, a multimodal LLM that synergistically aligns bimodal (sequence-structure) representations with the textual modality to advance protein functional annotation. By leveraging ESM3 for unified bimodal encoding and Llama-3.1-8B-Instruct for natural language modeling, STELLA achieves state-of-the-art performance in two critical tasks: Functional Description Prediction and Enzyme-catalyzed Reaction Prediction. This study demonstrates that multimodal LLMs represent a paradigm shift beyond pure pLMs, offering a new frontier for protein biology and biomedical discovery. The codes can be accessed via https://github.com/ocx-lab/STELLA.

[Long] Stop When Enough: Adaptive Early-Stopping for Chain-of-Thought Reasoning ► *Underline Video*

Renliang Sun, Wei Cheng, Dawei Li, Haifeng Chen and Wei Wang

Chain-of-Thought (CoT) reasoning has driven recent gains of large language models (LLMs) on reasoning-intensive tasks by externalizing intermediate steps. However, excessive or redundant reasoning — so-called overthinking — can increase inference costs and lead LLMs toward incorrect conclusions. In this paper, we present REFRAIN (REflective-Redundancy for Adaptive INference), a training-free framework that adaptively determines when to stop reasoning to mitigate overthinking. REFRAIN integrates a two-stage stop discriminator to identify reflective yet redundant reasoning and a sliding-window Upper Confidence Bound (SW-UCB) multi-armed bandit controller to dynamically adjust stopping thresholds according to problem difficulty without supervision or fine-tuning. Across four representative benchmarks and two model families, REFRAIN reduces token usage by 20-55% while maintaining or improving accuracy compared to standard CoT prompting. Extensive ablation and robustness analyses demonstrate its stability across models, scorers, and prompt variations. In summary, our findings highlight when-to-stop as a new and practical axis of test-time scaling — enabling models to reason not just more, but just enough.

[Findings] Strategy-Induct: Task-Level Strategy Induction for Instruction Generation ► *Underline Video*

Po-Chun Chen, Hen-Hsen Huang and Hsin-Hsi Chen

Designing effective task-level prompts is crucial for improving the performance of Large Language Models (LLMs). While prior work on instruction induction demonstrates that LLMs can infer better instructions with limited examples, existing approaches often rely on input-output pairs, where obtaining labeled answers can be difficult or costly. To address this limitation, we propose Strategy-Induct, a framework that derives task-level instructions solely from a small set of example questions without requiring labeled answers. Our approach first prompts the model to generate explicit reasoning strategies for each question, forming (strategy, question) pairs. These pairs are then used to induce a task instruction that guides reasoning. Experiments across multiple tasks and model scales demonstrate that Strategy-Induct outperforms state-of-the-art methods in question-only settings. Furthermore, we observe that jointly utilizing LLMs and Large Reasoning Models for both task instruction generation and inference can lead to further performance improvements.

[Long] Structure Guided Retrieval-Augmented Generation for Factual Queries ► *Underline Video*

Miao Xie, Xiao Zhang, Yi Li and Chunli Lv

Retrieval-Augmented Generation (RAG) has been proposed to mitigate hallucinations in large language models (LLMs), where generated outputs may be factually incorrect. However, existing RAG approaches predominantly rely on vector similarity for retrieval, which is prone to semantic noise and fails to ensure that generated responses fully satisfy the complex conditions specified by factual queries, often leading to incorrect answers. To address this challenge, we introduce a novel research problem, named Exact Retrieval Problem (ERP). To the best of our knowledge, this is the first problem formulation that explicitly incorporates structural information into RAG for factual questions to satisfy all query conditions. For this novel problem, we propose Structure Guided Retrieval-Augmented Generation (SG-RAG), which models the retrieval process as an embedding-based subgraph matching task, and uses the retrieved topological structures to guide the LLM to generate answers that meet all specified query conditions. To facilitate evaluation of ERP, we construct and publicly release Exact Retrieval Question Answering (ERQA), a large-scale dataset comprising 120,000 fact-oriented QA pairs, each involving complex conditions, spanning 20 diverse domains. The experimental results demonstrate that SG-RAG significantly outperforms strong baselines on ERQA, delivering absolute improvements from 20.68 to 50.88 points across all evaluation metrics, while maintaining reasonable computational overhead.

[Findings] SynthRL: Scaling Visual Reasoning with Verifiable Data Synthesis ▶ *Underline Video*

Zijian Wu, Jinjie Ni, Xiangyan Liu, Zichen Liu, Hang Yan and Michael Qizhe Shieh;

Vision-language models (VLMs) trained via reinforcement learning with verifiable reward (RLVR) have shown notable progress in scaling test-time compute effectively. In this work, we investigate how synthesized RL data can further improve RLVR. To this end, we propose SynthRL, a scalable and guaranteed pipeline for automatic data scaling in reasoning-oriented RL training. SynthRL comprises three key stages: (1) selecting seed questions with appropriate distribution, (2) augmenting them into more challenging variants while preserving the original answers, and (3) a guaranteed verification stage that ensures near-perfect correctness and difficulty enhancement. Our empirical experiments demonstrate SynthRL's scalability and effectiveness. When applied to the MMK12 dataset, SynthRL synthesizes over 3.3K additional verifiable, challenging questions from approximately 8K seed samples. Models trained with our synthesized data achieve consistent gains across five out-of-domain visual math reasoning benchmarks, with a significant improvement over baseline models trained on seed data alone. Notably, detailed analysis reveals that the gains are more pronounced on the most challenging evaluation samples, highlighting SynthRL's effectiveness in eliciting deeper and more complex reasoning patterns.

[Findings] T2I-ReasonBench: Benchmarking Reasoning-Informed Text-to-Image Generation ▶ *Underline Video*

Kaiyue Sun, Rongyao Fang, Chengqi Duan, Xian Liu, Aoxue Li and Xihui Liu

Text-to-image (T2I) generative models have achieved remarkable progress, demonstrating exceptional capability in synthesizing high-quality images from textual prompts. While existing research and benchmarks have extensively evaluated the ability of T2I models to follow the literal meaning of prompts, their ability to reason over prompts with domain knowledge to uncover implicit meaning and contextual nuances remains underexplored. To bridge this gap, we introduce T2I-ReasonBench, a novel benchmark designed to explore the knowledge-driven reasoning capabilities of T2I models. T2I-ReasonBench comprises 800 meticulously designed prompts organized into four dimensions: (1) Idiom Interpretation, (2) Textual Image Design, (3) Entity Reasoning, and (4) Scientific Reasoning. These dimensions challenge models to integrate domain knowledge, infer implicit meaning, and resolve contextual ambiguities. To quantify the performance, we introduce a two-stage evaluation framework: a large language model (LLM) generates prompt-specific question-criterion pairs that evaluate if the image includes the essential elements resulting from correct reasoning; a multimodal LLM (MLLM) then scores the generated image against these criteria. Our comprehensive study across 16 state-of-the-art diffusion and unified multimodal models (UMMs) reveal two primary bottlenecks. First, many models lack the foundational reasoning ability to fully comprehend complex prompts. Second, even models with stronger reasoning modules exhibit a persistent gap between their internal understanding and the final generated image. This highlights an urgent need for the next generation of T2I systems to not only improve their reasoning capability but also to enhance integration between reasoning and synthesis.

[Findings] Talk to Your Slides: High-Efficiency Slide Editing via Language-Driven Structured Data Manipulation ▶ *Underline Video*

Kyudan Jung, Hojun Cho, Jooyeol Yun, Soyoung Yang, Jaehyeok Jang and Jaegul Choo

Editing presentation slides is a frequent yet tedious task, ranging from creative layout design to repetitive text maintenance. While recent GUI-based agents powered by Multimodal LLMs (MLLMs) excel at tasks requiring visual perception, such as spatial layout adjustments, they often incur high computational costs and latency when handling structured, text-centric, or batch processing tasks. In this paper, we propose Talk-to-Your-Slides, a high-efficiency slide editing agent that operates via language-driven structured data manipulation rather than relying on the image modality. By leveraging the underlying object model instead of screen pixels, our approach ensures precise content modification while preserving style fidelity, addressing the limitations of OCR-based visual agents. Our system features a hierarchical architecture that effectively bridges high-level user instructions with low-level execution codes. Experiments demonstrate that for text-centric and formatting tasks, our method enables 34% faster processing, achieves 34% better instruction fidelity, and operates at an 87% lower cost compared to GUI-based baselines. Furthermore, we introduce TSBench, a human-verified benchmark dataset comprising 379 instructions, including a Hard subset designed to evaluate robustness against complex and visually dependent queries. Our code and benchmark are available at https://drive.google.com/drive/folders/1onwp5m7t3207xZu7HEBTMpdivsiOuqG8?usp=share_link

[Findings] TELL-TALE: Task Efficient LLMs with Task Aware Layer Elimination ▶ *Underline Video*

Omar Naim, Krish Sharma, Niyar R Barman and Nicholas Asher

Large Language Models (LLMs) typically come with a fixed architecture, despite growing evidence that not all layers contribute equally to every downstream task. We introduce TALE (Task-Aware Layer Elimination), an inference-time method that improves task performance by selectively removing layers that are irrelevant or detrimental for a given task. TALE optimizes task-specific performance, yielding a task-optimized architecture without retraining. Across 9 tasks and 5 model families, under both zero-shot and few-shot settings, TALE consistently matches or surpasses baseline performance while simultaneously reducing computational costs. TALE also synergizes with fine-tuning, leading to further performance improvements. Computing TALE for a new task requires modest resources, making it a practical and deployable solution for task-specialized LLM inference.

[Long] The Bidirectional Process Reward Model ▶ *Underline Video*

Lingyin Zhang, Jun Gao, Xiaoxue Ren and Ziqiang Cao

Process Reward Models (PRMs), which assign fine-grained scores to intermediate reasoning steps within a solution trajectory, have emerged as a promising approach to enhance the reasoning quality of Large Language Models (LLMs). However, most existing PRMs rely on a unidirectional left-to-right (L2R) evaluation scheme, which restricts their utilization of global context. In light of this challenge, we propose a novel bidirectional evaluation paradigm, named Bidirectional Process Reward Model (BiPRM). BiPRM incorporates a parallel

right-to-left (R2L) evaluation stream, implemented via prompt reversal, alongside the conventional L2R flow. Then a gating mechanism is introduced to adaptively fuse the reward scores from both streams to yield a holistic quality assessment. Remarkably, compared to the original PRM, BiPRM introduces only a 0.3% parameter increase for the gating module, and the parallel execution of two streams incurs merely 5% inference time latency. Our extensive empirical evaluations spanning diverse benchmarks, LLM backbones, PRM objectives and sampling policies demonstrate that BiPRM consistently surpasses unidirectional baselines, achieving an average relative gain of 10.6% over 54 solution-level configurations and 37.7% in 12 step-level error detection scenarios. Generally, our results highlight the effectiveness, robustness and general applicability of BiPRM, offering a promising new direction for process-based reward modeling.

[Findings] The Illusion of Insight in Reasoning Models ► *Underline Video*

Liv d'Aliberti and Manoel Ribeiro

Do reasoning models have "Aha!" moments? Prior work suggests that models like DeepSeek-R1-Zero undergo sudden mid-trace realizations that lead to accurate outputs, implying an intrinsic capacity for self-correction. Yet, it remains unclear whether such intrinsic shifts in reasoning strategy actually improve performance. Here, we study mid-reasoning shifts and instrument training runs to detect them. Our analysis spans 1M+ reasoning traces, hundreds of training checkpoints, three reasoning domains, and multiple decoding temperatures and model architectures. We find that reasoning shifts are rare, do not become more frequent with training, and seldom improve accuracy, indicating that they do not correspond to prior perceptions of model insight. However, their effect varies with model uncertainty. Building on this finding, we show that artificially triggering extrinsic shifts under high entropy reliably improves accuracy. Our results show that mid-reasoning shifts are symptoms of unstable inference behavior rather than an intrinsic mechanism for self-correction.

[Long] The Visual Iconicity Challenge: Evaluating Vision-Language Models on Sign Language Form-Meaning Mapping ► *Underline Video*

Onur Kele, Asl Özyürek, Gerardo Ortega, Kadir Gökgöz and Esam Ghaleb

Iconicity, the resemblance between linguistic form and meaning, is pervasive in sign languages, offering a natural testbed for visual grounding in vision-language models (VLMs). We introduce the Visual Iconicity Challenge, a video-based benchmark that adapts psycholinguistic measures to evaluate VLMs on three tasks: (i) phonological sign-form prediction, (ii) transparency (inferring meaning from visual form), and (iii) graded iconicity ratings. We assess 17 state-of-the-art VLMs in zero- and few-shot settings on Sign Language of the Netherlands and compare them to human baselines. VLMs mirror human phonological difficulty patterns (e.g., handshape harder than location) and achieve moderate to strong alignment with human iconicity ratings. However, they still fail to infer lexical meaning from visual form alone and show a systematic object-based bias that inverts the human preference for action-based signs. Crucially, models with stronger phonological form prediction correlate better with human iconicity judgments, indicating shared sensitivity to visually grounded structure. Our findings validate these diagnostic tasks, show that explicit reasoning narrows the open-to-closed-model calibration gap, and motivate human-centric signals for modelling iconicity in multimodal models.

[Findings] Think Before you Write: QA-Guided Reasoning for Character Descriptions in Books ► *Underline Video*

Argyrios Papoudakis, Mirella Lapata and Frank Keller

Character description generation is an important capability for narrative-focused applications such as summarization, story analysis, and character-driven simulations. However, generating accurate character descriptions from long-form narratives (e.g., novels) is challenging: models must track evolving attributes (e.g., relationships and events), integrate evidence scattered across the text, and infer implicit details. Despite the success of reasoning-enabled LLMs on many benchmarks, we find that for character description generation their performance improves when built-in reasoning is disabled (i.e., an empty reasoning trace). Motivated by this, we propose a training framework that decouples reasoning from generation. Our approach, which can be applied on top of long-context LLMs or chunk-based methods, consists of a reasoning model that produces a structured QA reasoning trace and a generation model that conditions on this trace to produce the final character description. Experiments on two datasets (BookWorm and CroSS) show that QA-guided reasoning improves faithfulness, informativeness, and grounding over strong long-context baselines.

[Long] TLoRA: Task-aware Low Rank Adaptation of Large Language Models ► *Underline Video*

Weicheng Lin, Yi Zhang, Jiawei Dang and Liang-Jie Zhang

Low-Rank Adaptation (LoRA) has become a widely adopted parameter-efficient fine-tuning method for large language models, with its effectiveness largely influenced by the allocation of ranks and scaling factors, as well as initialization. Existing LoRA variants typically address only one of these factors, often at the cost of increased training complexity or reduced practical efficiency. In this work, we present Task-aware Low-Rank Adaptation (TLoRA), a unified framework that jointly optimizes initialization and resource allocation at the outset of training. TLoRA introduces a data-driven initialization strategy that aligns the LoRA A matrix with task-relevant subspaces by performing singular value decomposition on the product of pre-trained weights and input activation covariance. After this, the A matrix is frozen, and only the B matrix is trained. Furthermore, TLoRA employs a sensitivity-based importance metric to adaptively allocate ranks and scaling factors across layers under a fixed parameter budget. We conduct extensive experiments that demonstrate TLoRA consistently performs excellently across various tasks, including natural language understanding, commonsense reasoning, math reasoning, code generation, and chat generation, while significantly reducing the number of trainable parameters. Our code are available at https://github.com/Rambo-Yi/TLora/tree/main

[Findings] Towards Self-Evolving Agents: Enabling Autonomy through Interactive Experience Refinement ► *Underline Video*

Cheng Yang, Xuemeng Yang, Licheng Wen, Daocheng Fu, Jianbiao Mei, Rong Wu, Pinlong Cai, Yufan Shen, Nianchen Deng, Jia Xu, Botian Shi, Yu Qiao and Haifeng Li

Large Language Models often struggle with complex, multi-step operational tasks because they remain static during inference and cannot learn from past experience. To address this, we propose MUSE, a framework that enables iterative self-improvement through a hierarchical Memory Module. MUSE organizes cross-domain insights to facilitate the orchestration of long-horizon workflows. The core of our approach is an autonomous post-execution critique mechanism: after completing each sub-task, the system analyzes its operational logs and distills raw execution data into structured, reusable knowledge. This allows the agent to evolve dynamically rather than relying on fixed parameters. Evaluated on the rigorous TAC productivity benchmark, MUSE achieves new state-of-the-art results, significantly outperforming previous methods using only the streamlined Gemini-2.5 Flash model. Our analysis demonstrates that MUSEs performance scales with the accumulation of insights and exhibits strong cross-task transferability, marking a key step toward autonomous systems capable of lifelong learning in professional environments. Demo videos can be found in our supplementary materials.

[Long] TriPlay-RL: Tri-Role Self-Play Reinforcement Learning for LLM Safety Alignment ▶ *Underline Video*

Zhewen Tan, Wenhan Yu, Jianfeng Si, Tongxin Liu, Kaiqi Guan, Huiyan Jin, Jiawen Tao, Xiaokun Yuan, Duohe Ma, Lin Sun, Tong Yang and Xiangzheng Zhang

In recent years, safety risks associated with large language models have become increasingly prominent, highlighting the urgent need to mitigate the generation of toxic and harmful content. The mainstream paradigm for LLM safety alignment typically adopts a collaborative framework involving three roles: an attacker for adversarial prompt generation, a defender for safety defense, and an evaluator for response assessment. In this paper, we propose a closed-loop reinforcement learning framework called TriPlay-RL that enables iterative and co-improving collaboration among three roles with near-zero manual annotation. Experimental results show that the attacker preserves high output diversity while achieving a 20%–50% improvement in adversarial effectiveness. The defender attains 10%–30% gains in safety performance without degrading general reasoning capability, and the evaluator continuously refines its fine-grained judgment ability through iterations, accurately distinguishing unsafe responses, simple refusals, and useful guidance. Overall, our framework establishes an efficient and scalable paradigm for LLM safety alignment, enabling continuous co-evolution within a unified learning loop. The code is available at https://github.com/Qihoo360/TriPlay-RL.

[Long] Triviality Corrected Endogenous Reward ▶ *Underline Video*

Xinda Wang, Zhengxu Hou, Yangshijie Zhang, yanbingren yanbingren, Jialin Liu, ChenZhuo Zhao, Zhibo Yang, Bin-Bin Yang and Feng Xiao

Reinforcement learning for open-ended text generation is constrained by the lack of verifiable rewards, necessitating reliance on judge models that require either annotated data or powerful closed-source models. Inspired by recent work on unsupervised reinforcement learning for mathematical reasoning using confidence-based endogenous rewards, we investigate whether this principle can be adapted to open-ended writing tasks. We find that directly applying confidence rewards leads to Triviality Bias: the policy collapses toward high-probability outputs, reducing diversity and meaningful content. We propose TCER (Triviality Corrected Endogenous Reward), which addresses this bias by rewarding the relative information gain between a specialist policy and a generalist reference policy, modulated by a probability-dependent correction mechanism. Across multiple writing benchmarks and model architectures, TCER achieves consistent improvements without external supervision. Furthermore, TCER also transfers effectively to mathematical reasoning, validating the generality of our approach across different generation tasks.

[Long] TROJail: Trajectory-Level Optimization for Multi-Turn Large Language Model Jailbreaks with Process Rewards ▶ *Underline Video*

Xiqiao Xiong, Ouxiang Li, Zhuo Liu, Moxin Li, Wentao Shi, Fengbin Zhu, Qifan Wang and Fuli Feng

Large language models have seen widespread adoption, yet they remain vulnerable to multi-turn jailbreak attacks, threatening their safe deployment. This has led to the task of training automated multi-turn attackers to probe model safety vulnerabilities. However, existing approaches typically rely on turn-level optimization, which is insufficient for learning long-term attack strategies. To bridge this gap, we formulate this task as a multi-turn reinforcement learning problem, directly optimizing the harmfulness of the final-turn response as the outcome reward. To address the sparse supervision of the outcome reward, we introduce TROJail, which employs two process rewards to evaluate the utility of intermediate prompts and integrate them into advantage estimation. These rewards (1) penalize overly harmful prompts that trigger the models refusal mechanism, and (2) encourage steering the semantic relevance of responses toward the targeted harmful content. Experimental results show improved attack success rates across multiple models and benchmarks, highlighting the effectiveness of our approach. The code is available at https://anonymous.4open.science/r/TROJail. Warning: This paper contains examples of harmful content.

[Long] Trust Within? Seek Beyond? Knowledge Boundary Aware Policy Optimization for Agentic Search ▶ *Underline Video*

Tao Feng, Xinke Jiang, Xinyan Hu, Yonggang Zhang, Zhen Tao, Wentao Zhang, Boyang Liu, Wenhao Jiang and Chao Wu

Agentic search augments large language models (LLMs) with external knowledge through reinforcement learning. However, existing approaches suffer from blind reliance on noisy retrieval and hallucination when both parametric and external knowledge fail, reflecting a lack of calibration regarding the model's knowledge boundary. We propose Knowledge boundary Policy Optimization (KbPO), a reinforcement learning framework that explicitly aligns retrieval decisions with quantified knowledge states. KbPO introduces: (1) a semantic stability metric to delineate reliable parametric knowledge; (2) a four-quadrant taxonomy synthesising internal certainty with retrieval quality; and (3)

a quadrant-based reward mechanism incentivising calibrated behaviour. We further adopt an iterative query evolution pipeline to construct boundary-probing training samples. Experiments on ten benchmarks demonstrate that KbPO outperforms strong baselines while exhibiting reduced hallucination rates.

[Long] Two Streams, One Sarcasm: Orthogonal Expert Tuning for Holistic Multimodal Sarcasm Understanding ▶ *Underline Video*

Diandian Guo, Cong Cao, Fangfang Yuan, Pin Xu, Cheng Hu, Zhicheng Zhang, Yu Liu and Yanbing Liu

Multimodal Sarcasm Understanding (MSU) comprises multiple subtasks, demanding both incongruity perception and intent reasoning. However, this progress is impeded by two bottlenecks. First, the lack of a unified benchmark for holistic satirical cognition hinders comprehensive evaluation of MSU. Second, jointly modeling these heterogeneous subtasks often leads to feature entanglement. Specifically, while subtasks share a dependence on incongruity, they diverge in granular focus, causing specific execution patterns to erode the fundamental perception capability. To address these challenges, we make two contributions. First, we introduce DocMSU-PLUS, a comprehensive benchmark covering five cognitive dimensions of MSU. All tasks are reformulated into multiple-choice questions (MCQs), enabling a unified accuracy-based evaluation. Second, we propose the Dual Orthogonal Stream Experts (DOSE) framework. DOSE structurally decouples experts into orthogonal shared perception and private execution streams to physically block gradient interference between tasks. Experiments demonstrate that DOSE achieves superior performance on DocMSU-PLUS, effectively balancing general perception with task-specific adaptation.

[Long] UI-Copilot: Advancing Long-Horizon GUI Automation via Tool-Integrated Policy Optimization ▶ *Underline Video*

Zhengxi Lu, Fei Tang, Guangyi Liu, Kaitao Song, Xu Tan, Jin Ma, Wenqi Zhang, Weiming Lu, Jun Xiao, Yueting Zhuang and Yongliang Shen

MLLM-based GUI agents have demonstrated strong capabilities in complex user interface interaction tasks. However, long-horizon scenarios remain challenging, as these agents are burdened with tasks beyond their intrinsic capabilities, suffering from memory degradation, progress confusion, and math hallucination. To address these challenges, we present UI-Copilot, a collaborative framework where the GUI agent focuses on task execution while a lightweight copilot provides on-demand assistance for memory retrieval and numerical computation. We introduce memory decoupling to separate persistent observations from transient execution context, and train the policy agent to selectively invoke the copilot as Retriever or Calculator based on task demands. To enable effective tool invocation learning, we propose Tool-Integrated Policy Optimization (TIPO), which separately optimizes tool selection through single-turn prediction and task execution through on-policy multi-turn rollouts. Experimental results show that UI-Copilot-7B achieves state-of-the-art performance on challenging MemGUI-Bench, outperforming strong 7B-scale GUI agents such as GUI-Owl-7B and UI-TARS-1.5-7B. Moreover, UI-Copilot-7B delivers a 17.1% absolute improvement on AndroidWorld over the base Qwen model, highlighting UI-Copilot's strong generalization to real-world GUI tasks. Code website: https://anonymous.4open.science/r/UI-Copilot-0535.

[Long] Uncertainty-Aware Routing for Principled Alignment with MoE Dynamics ▶ *Underline Video*

Yilong Chen, Junyuan Shang, Yuchen Feng, Zhenyu Zhang, Naibin Gu, Ziqi Wang, Tingwen Liu, Shuohuan Wang, Yu Sun, Hua Wu and Haifeng Wang

Mixture-of-Experts (MoE) is a cornerstone for scaling LLMs, yet its training dynamics remain poorly understood, often leading to sub-optimal specialization. Moving beyond static routing, we present a systematic study of the MoE lifecycle using Helmholtz Free Energyand Router Entropy. We identify a universal Three-Stage Phase Transition, Exploration, Symmetry Breaking, and Stabilization, marked by an Energy Climb and Plateau. This reflects Frustrated Exploration, caused by structural interference between specialization drives and uniformity constraints. To address this, we propose Uncertainty-Aware Routing (UAR), which aligns routing with the models epistemic state via: (1) Evidence-Triggered Expansion, increasing active experts for high-energy tokens, and (2) Epistemic Masking, applying load-balancing only in high-uncertainty regimes to shield mature experts. Experiments confirm UAR reduces perplexity and improves expert distinctiveness, offering a principled path toward thermodynamically aligned computation.

[Findings] Understanding In-Context Learning Beyond Transformers: An Investigation of State Space and Hybrid Architectures ▶ *Underline Video*

Shenran Wang, Timothy Tin-Long Tse and Jian Zhu

We perform in-depth evaluations of in-context learning (ICL) on state-of-the-art transformer, state-space, and hybrid large language models over two categories of knowledge-based ICL tasks. Using a combination of behavioral probing and intervention-based methods, we have discovered that, while LLMs of different architectures can behave similarly in task performance, their internals could remain different. We discover that function vectors (FVs) responsible for ICL are primarily located in the self-attention and Mamba layers, and speculate that Mamba2 uses a different mechanism from FVs to perform ICL. FVs are more important for ICL involving parametric knowledge retrieval, but not for contextual knowledge understanding. Our work contributes to a more nuanced understanding across architectures and task types. Methodologically, our approach also highlights the importance of combining both behavioural and mechanistic analyses to investigate LLM capabilities.

[Findings] UniCreative: Unifying Long-form Logic and Short-form Sparkle via Reference-Free Reinforcement Learning ▶ *Underline Video*

Xiaolong Wei, Zerun Zhu, Simin Niu, Xingyu Zhang, Peiying Yu, Changxuan Xiao, Yuchen Li, Jicheng Yang, Zhejun Zhao, Chong Meng, Long Xia and Daiting Shi

A fundamental challenge in creative writing lies in reconciling the inherent tension between maintaining global coherence in long-form narratives and preserving local expressiveness in short-form texts. While long-context generation necessitates explicit macroscopic planning, short-form creativity often demands spontaneous, constraint-free expression. Existing alignment paradigms, however, typically employ static reward signals and rely heavily on high-quality supervised data, which is costly and difficult to scale. To address this, we propose UniCreative, a unified reference-free reinforcement learning framework. We first introduce AC-GenRM, an adaptive constraint-aware reward model that dynamically synthesizes query-specific criteria to provide fine-grained preference judgments. Leveraging these signals, we propose ACPO, a policy optimization algorithm that aligns models with human preferences across both content quality and structural paradigms without supervised fine-tuning and ground-truth references. Empirical results demonstrate that AC-GenRM aligns closely with expert evaluations, while ACPO significantly enhances performance across diverse writing tasks. Crucially, our analysis reveals an emergent meta-cognitive ability: the model learns to autonomously differentiate between tasks requiring rigorous planning and those favoring direct generation, validating the effectiveness of our direct alignment approach.

[Findings] Unlocking the Edge deployment and ondevice acceleration of multi-LoRA enabled one-for-all foundational LLM ► *Underline Video*

Sravanth Kodavanti, Sowmya Vajrala, Srinivas Soumitri Miriyala, Utsav Tiwari, Uttam Kumar, Utkarsh Kumar Mahawar, Achal Pratap Singh, Arya D, Narendra Mutyala, Vikram Nelvoy Rajendiran, Sharan Kumar Allur, Euntaik Lee, Dohyoung Kim, HyeonSu Lee, Gyusung Cho and JungBae Kim

Deploying large language models (LLMs) on smartphones poses significant engineering challenges due to stringent constraints on memory, latency, and runtime flexibility. In this work, we present a hardware-aware framework for efficient on-device inference of a LLaMA-based multilingual foundation model supporting multiple use cases on Samsung Galaxy S24 and S25 devices with SM8650 and SM8750 Qualcomm chipsets respectively. Our approach integrates application-specific LoRAs as runtime inputs to a single frozen inference graph, enabling dynamic task switching without recompilation or memory overhead. We further introduce a multi-stream decoding mechanism that concurrently generates stylistic variations, such as formal, polite, or jovial responses, within a single forward pass, reducing latency by up to 6x. To accelerate token generation, we apply Dynamic Self-Speculative Decoding (DS2D), a tree-based strategy that predicts future tokens without requiring a draft model, yielding up to 2.3x speedup in decode time. Combined with quantization to INT4 and architecture-level optimizations, our system achieves 4-6x overall improvements in memory and latency while maintaining accuracy across 9 languages and 8 tasks. These results demonstrate practical feasibility of deploying multi-use-case LLMs on edge devices, advancing the commercial viability of Generative AI in mobile platforms.

[Long] Utility-Oriented Visual Evidence Selection for Multimodal Retrieval-Augmented Generation ► *Underline Video*

Weiqing Luo, Zongye Hu, Xiao Wang, Zhiyuan Yu, Haofeng Zhang and Ziyi Huang

Visual evidence selection is a critical component of multimodal retrieval-augmented generation (RAG), yet existing methods typically rely on semantic relevance or surface-level similarity, which are often misaligned with the actual utility of visual evidence for downstream reasoning. We reformulate multimodal evidence selection from an information-theoretic perspective by defining evidence utility as the information gain induced on a models output distribution. To overcome the intractability of answer-space optimization, we introduce a latent notion of evidence helpfulness and theoretically show that, under mild assumptions, ranking evidence by information gain on this latent variable is equivalent to answer-space utility. We further propose a training-free, surrogate-accelerated framework that efficiently estimates evidence utility using lightweight multimodal models. Experiments on MRAG-Bench and Visual-RAG across multiple model families demonstrate that our method consistently outperforms state-of-the-art RAG baselines while achieving substantial reductions in computational cost. We release our code at https://github.com/Hcnaeg/utility-mrag.

[Findings] VecCISC: Improving Confidence-Informed Self-Consistency with Reasoning Trace Clustering and Candidate Answer Selection ► *Underline Video*

James Petullo, Sonny George, Dylan Cashman and Nianwen Xue

A standard technique for scaling inference-time reasoning is Self-Consistency, whereby multiple candidate answers are sampled from an LLM and the most common answer is selected. More recently, it has been shown that weighted majority voting (e.g. Confidence-Informed Self Consistency (CISC)), which assigns a confidence value to each candidate answer and chooses the answer with the largest accumulated score, tends to be more accurate on a wide range of popular benchmarks. In practice, weighted majority voting necessitates calling a critic LLM on each candidates reasoning trace to produce the answers confidence score. This secondary series of LLM calls greatly increases the overhead and cost of weighted majority voting, despite its potential performance benefits. To reduce this expense, we propose VecCISC, a lightweight, adaptive framework that uses a measure of semantic similarity to filter reasoning traces that are semantically equivalent to others, degenerate, or hallucinated, thus decreasing the number of candidate answers that must be evaluated by the critic. To ensure adequate experimental thoroughness, we evaluated VecCISC on five challenging, widely-adopted datasets spanning the domains of mathematics, chemistry, biology, commonsense reasoning, and the humanities. Our results demonstrate that VecCISC reduces the total token usage by 47%, while maintaining or exceeding the accuracy of CISC.

[Long] Verify Before You Commit: Towards Faithful Reasoning in LLM Agents via Self-Auditing ► *Underline Video*

Wenhao Yuan, Chenchen Lin, Jian Chen, Jinfeng Xu, Xuehe Wang and Edith Cheuk-Han Ngai

In large language model (LLM) agents, reasoning trajectories are treated as reliable internal beliefs for guiding actions and updating

memory. However, coherent reasoning can still violate logical or evidential constraints, allowing unsupported beliefs repeatedly stored and propagated across decision steps, leading to systematic behavioral drift in long-horizon agentic systems. Most existing strategies rely on the consensus mechanism, conflating agreement with faithfulness. In this paper, inspired by the vulnerability of unfaithful intermediate reasoning trajectories, we propose Self-Audited Verified Reasoning (SAVeR), a novel framework that enforces verification over internal belief states within the agent before action commitment, achieving faithful reasoning. Concretely, we structurally generate persona-based diverse candidate beliefs for selection under a faithfulness-relevant structure space. To achieve reasoning faithfulness, we perform adversarial auditing to localize violations and repair through constraint-guided minimal interventions under verifiable acceptance criteria. Extensive experiments on six benchmark datasets demonstrate that our approach consistently improves reasoning faithfulness while preserving competitive end-task performance.

[Long] Visual and Memory-Augmented Soccer Commentary Generation ▶ *Underline Video*
Haoran SUN, Natthawut Kertkeidkachorn and Kiyoaki Shirai
Automatic soccer commentary generation aims to bridge the gap between raw visual content and professional, tactical commentary. However, existing datasets tend to produce incomplete commentary that lacks semantic richness and fails to convey the full visual information present in standard video clips. To address these limitations, we propose two manually curated datasets: SN-Short, which enhances scene-level semantic descriptions, and SN-Long, which captures event continuity for context-aware commentary. Based on these, we design a commentary augmentation pipeline that transforms incomplete annotations into MatchText, a semantically complete and structurally standardized dataset. Leveraging this supervision, we introduce MatchAware, a generation model that incorporates contextual cues from previous events to produce coherent commentary aligned with the visual flow of the game. Experimental results show that proposed approach significantly outperforms existing baselines on the constructed datasets.

[Long] WebCoderBench: Benchmarking Web Application Generation with Comprehensive and Interpretable Evaluation Metrics ▶ *Underline Video*
Chenxu Liu, Yingjie Fu, Wei Yang, Ying Zhang and Tao Xie
Web applications (web apps) have become a key arena for large language models (LLMs) to demonstrate their code generation capabilities and commercial potential. However, building a benchmark for LLM-generated web apps remains challenging due to the need for real-world user requirements, generalizable evaluation metrics without relying on ground-truth implementations or test cases, and interpretable evaluation results. To address these challenges, we introduce WebCoderBench, the first real-world-collected, generalizable, and interpretable benchmark for web app generation. WebCoderBench comprises 1,572 user requirements, covering diverse modalities and expression styles that reflect realistic user intentions. WebCoderBench provides 24 fine-grained evaluation metrics across 9 perspectives, combining the rule-based and LLM-as-a-judge paradigms for fully automated, objective, and general evaluation. Moreover, WebCoderBench adopts human-preference-aligned weights over metrics to yield interpretable overall scores. Experiments across 12 representative LLMs and 2 LLM-based agents show that there exists no dominant model across all evaluation metrics, offering an opportunity for LLM developers to optimize their models in a targeted manner for a more powerful version.

[Findings] What Prompts Dont Say: Understanding and Managing Underspecification in LLM Prompts ▶ *Underline Video*
Chenyang Yang, Yike Shi, Qianou Ma, Michael Xieyang Liu, Christian Kaestner and Tongshuang Wu
Prompt underspecification is a common challenge when interacting with LLMs. In this paper, we present an in-depth analysis of this problem, showing that while LLMs can often infer unspecified requirements by default (41.1%), such behavior is fragile: Under-specified prompts are 2x as likely to regress across model or prompt changes, sometimes with accuracy drops exceeding 20%. This instability makes it difficult to reliably build LLM applications. Moreover, simply specifying all requirements does not consistently help, as models have limited instruction-following ability and requirements can conflict. Standard prompt optimizers likewise provide little benefit. To address these issues, we propose requirements-aware prompt optimization mechanisms that improve performance by 4.8% on average over baselines. We further advocate for a systematic process of proactive requirements discovery, evaluation, and monitoring to better manage prompt underspecification in practice.

[Short] When More Words Say Less: Decoupling Length and Specificity in Image Description Evaluation ▶ *Underline Video*
Rhea Kapur, Robert Hawkins and Elisa Kreiss
Vision-language models (VLMs) are increasingly used to make visual content accessible via text-based descriptions. In current systems, however, description specificity is often conflated with their length. We argue that these two concepts must be disentangled: descriptions can be concise yet dense with information, or lengthy yet vacuous. We define specificity relative to a contrast set, where a description is more specific to the extent that it picks out the target image better than other possible images. We construct a dataset that controls for length while varying information content, and validate that people reliably prefer more specific descriptions regardless of length. We find that controlling for length alone cannot account for differences in specificity; it matters how the length budget is applied. These results support evaluation approaches that directly prioritize specificity over verbosity.

[Long] Why Does Reinforcement Learning Generalize? A Feature-Level Mechanistic Study of Post-Training in Large Language Models ▶ *Underline Video*
Dan Shi, Zhuowen Han, Simon Ostermann, Renren Jin, Josef van Genabith and Deyi Xiong

Reinforcement learning (RL)-based post-training often improves the reasoning performance of large language models (LLMs) beyond the training domain, while supervised fine-tuning (SFT) frequently leads to general capabilities forgetting. However, the mechanisms underlying this contrast remain unclear. To bridge this gap, we present a feature-level mechanistic analysis methodology to probe RL generalization using a controlled experimental setup, where RL- and SFT-tuned models are trained from the same base model on identical data. Leveraging our interpretability framework, we align internal activations across models within a shared feature space and analyze how features evolve during post-training. We find that SFT rapidly introduces many highly specialized features that stabilize early in training, whereas RL induces more restrained and continually evolving feature changes that largely preserve base models' representations. Focusing on samples where RL succeeds but the base model fails, we identify a compact, task-agnostic set of features that directly mediate generalization across diverse tasks. Feature-level interventions confirm their causal role: disabling these features significantly degrades RL models' generalization performance, while amplifying them improves base models' performance. The code is available at https://github.com/danshi777/RL-generalization.

[Long] Why LLMs Hallucinate on Structured Knowledge: A Mechanistic Analysis of Reasoning over Linearized Representations ▶ *Underline Video*

Shanghao Li, Jinda Han, Yibo Wang, Yuanjie Zhu, Zihe Song, Langzhou He, Kenan Kamel A Alghythee and Philip S. Yu

In many reasoning tasks, large language models (LLMs) rely on structured external knowledge, such as graphs and tables, which is typically linearized into sequential token representations. However, even when sufficient knowledge is available, LLMs can still produce hallucinated outputs, and the underlying mechanisms behind such failures remain poorly understood. We investigate these mechanisms and find that hallucinations arise from systematic internal dynamics rather than random noise. First, attention disproportionately concentrates toward shortcut-like structural cues rather than distributing across the full context. Second, feed-forward representations fail to ground the provided knowledge, causing the model to revert to parametric memory. Moreover, our results indicate that hallucination is consistently associated with failures in semantic grounding within feed-forward layers, while attention allocation exhibits greater task-dependent variability. Finally, we show that these mechanistic patterns generalize beyond single-hop graphs to multi-hop and tabular settings, enabling effective hallucination detection across structured knowledge formats.

[Long] Your Reasoning Benchmark May Not Test Reasoning: Revealing Perception Bottleneck in Abstract Reasoning Benchmarks ▶ *Underline Video*

Xinhe Wang, Jin Huang, Xingjian Zhang, Tianhao Wang and Jiaqi W. Ma

Reasoning benchmarks such as the Abstraction and Reasoning Corpus (ARC) and ARC-AGI are widely used to assess progress in artificial intelligence and are often interpreted as probes of core, so-called "fluid" reasoning abilities. Despite their apparent simplicity for humans, these tasks remain challenging for frontier vision-language models (VLMs), a gap commonly attributed to deficiencies in machine reasoning. We challenge this interpretation and hypothesize that the gap arises primarily from limitations in visual perception rather than from shortcomings in inductive reasoning. To verify this hypothesis, we introduce a two-stage experimental pipeline that explicitly separates perception and reasoning. In the perception stage, each image is independently converted into a natural-language description, while in the reasoning stage a model induces and applies rules using these descriptions. This design prevents leakage of cross-image inductive signals and isolates reasoning from perception bottlenecks. Across three ARC-style datasets, Mini-ARC, ACRE, and Bongard-LOGO, we show that the perception capability is the dominant factor underlying the observed performance gap by comparing the two-stage pipeline with against standard end-to-end one-stage evaluation. Manual inspection of reasoning traces in the VLM outputs further reveals that approximately 80 percent of model failures stem from perception errors. Together, these results demonstrate that ARC-style benchmarks conflate perceptual and reasoning challenges and that observed performance gaps may overstate deficiencies in machine reasoning. Our findings underscore the need for evaluation protocols that disentangle perception from reasoning when assessing progress in machine intelligence.

Poster Session C (Zoom)

16:00-17:30 (Zoom)

[Findings] Single-Agent Generation Surpasses Multi-Agent Systems in Semantic Diversity ▶ *Underline Video*

Encheng Cui, Shaowen Peng, Kazuhiro Ito, Jinsha Xu, Shohei Hisada, Shoko Wakamiya and Eiji Aramaki

Multi-Agent Systems (MAS) are commonly used to improve reasoning diversity and robustness by simulating interactions among agents with distinct roles. However, prior work often entangles the contribution of the multi-agent architecture with that of prompt conditioning, making the source of observed diversity gains unclear. We address this confound with a controlled study on divergent thinking tasks, using identical prompt conditioning for MAS and single agent baseline. Under these matched conditions, single agent setups consistently outperform multi-agent systems in semantic diversity. We attribute this gap to information visibility: parallel agents often converge on overlapping ideas, whereas a single agent model can condition on its own generation to avoid redundancy. We further find that a Multi-Output strategy, which prompts a single agent to produce multiple responses within a single inference pass, achieves the highest diversity without degrading logical validity. Together, these results point to a more efficient and effective way to expand diversity, with implications for the design of more efficient agentic frameworks.

Session 5: Virtual Presentations 1 - 18:00-19:30

Virtual Presentations 1

18:00-19:30 (TBA)

[Findings] RADS: Reinforcement Learning-Based Sample Selection Improves Transfer Learning in Low-resource and Imbalanced Clinical Settings ▶ *Underline Video*

Wei Han, David Martinez Iraola, Anna Khanina, Lawrence Cavedon and Karin Verspoor

A common strategy in transfer learning is few shot fine-tuning, but its success is highly dependent on the quality of samples selected as training examples. Active learning methods such as uncertainty sampling and diversity sampling can select useful samples. However, under extremely low-resource and class-imbalanced conditions, they often favor outliers rather than truly informative samples, resulting in degraded performance. In this paper, we introduce RADS (Reinforcement Domain Adaptive Sampling), a robust sample selection strategy using reinforcement learning (RL) to identify the most informative samples. Experimental evaluations on several real world clinical datasets show our sample selection strategy enhances model transferability while maintaining robust performance under extreme class imbalance compared to traditional methods. Our code is open-sourced on GitHub.

Virtual Presentations 1

18:00-19:30 (Zoom)

[Findings] A Counterfactual Explanation Framework for Retrieval Models ▶ *Underline Video*

Bhavik Chandna and Procheta Sen

Explainability has become a crucial concern in today's world, aiming to enhance transparency in machine learning and deep learning models. Information retrieval is no exception to this trend. In existing literature on explainability of information retrieval, the emphasis has predominantly been on illustrating the concept of relevance concerning a retrieval model. The questions addressed include why a document is relevant to a query, why one document exhibits higher relevance than another, or why a specific set of documents is deemed relevant for a query. However, limited attention has been given to understanding why a particular document is not favored (e.g., not within top-K) with respect to a query and a retrieval model. In an effort to address this gap, our work focuses on the question of what terms need to be added within a document to improve its ranking. This, in turn, answers the question of which words in the document played a role in not being favored by a retrieval model for a particular query. We use a counterfactual framework to solve the above-mentioned research problem. Our experiments show the effectiveness of our proposed approach in predicting counterfactuals for both statistical (e.g. BM25) and deep-learning-based models (e.g. DRMM, DSSM, ColBERT, MonoT5).

[Long] A Data-Centric Approach to Generalizable Speech Deepfake Detection ▶ *Underline Video*

Wen Huang, Yuchen Mao and Yanmin Qian

Achieving robust generalization in speech deepfake detection (SDD) remains a primary challenge, as models often fail to detect unseen forgery methods. While research has focused on model-centric and algorithm-centric solutions, the impact of data composition is often underexplored. This paper proposes a data-centric approach, analyzing the SDD data landscape from two practical perspectives: constructing a single dataset and aggregating multiple datasets. To address the first perspective, we conduct a large-scale empirical study to characterize the data scaling laws for SDD, quantifying the impact of source and generator diversity. To address the second, we propose the Diversity-Optimized Sampling Strategy (DOSS), a principled framework for mixing heterogeneous data with two implementations: DOSS-Select (pruning) and DOSS-Weight (re-weighting). Our experiments show that DOSS-Select outperforms the naive aggregation baseline while using only 3% of the total available data. Furthermore, our final model, trained on a 12k-hour curated data pool using the optimal DOSS-Weight strategy, achieves state-of-the-art performance, outperforming large-scale baselines with greater data and model efficiency on both public benchmarks and a new challenge set of various commercial APIs.

[Long] A Layer-wise Analysis of Supervised Fine-Tuning ▶ *Underline Video*

Qinghua Zhao, Xueling Gong, Xinyu Chen, Zhongfeng Kang and Xinlu Li

While critical for alignment, Supervised Fine-Tuning (SFT) incurs the risk of catastrophic forgetting, yet the layer-wise emergence of instruction-following capabilities remains elusive. We investigate this mechanism via a comprehensive analysis utilizing information-theoretic, geometric, and optimization metrics across model scales (1B-32B). Our experiments reveal a distinct depth-dependent pattern: middle layers (20%-80%) are stable, whereas final layers exhibit high sensitivity. Leveraging this insight, we propose Mid-Block Efficient Tuning, which selectively updates these critical intermediate layers. Empirically, our method outperforms standard LoRA up to 10.2% on

GSM8K (OLMo2-7B) with reduced parameter overhead, demonstrating that effective alignment is architecturally localized rather than distributed. The code is publicly available at https://github.com/lshowway/base.

[Long] A Multi-Agent Framework for Feature-Constrained Difficulty Control in Reading Comprehension Item Generation ▶ *Underline Video*

Seonjeong Hwang, Jun Seo, Hyounghun Kim and Gary Lee

Recent studies in difficulty-controlled reading comprehension item generation have leveraged large language models (LLMs) to produce items by adjusting difficulty-related features. However, existing methods typically rely on a single-agent prompting approach, which often fails to consistently satisfy specified feature constraints, resulting in items that deviate from the target difficulty level. To address this limitation, we introduce MAFIG, a Multi-agent Framework for Feature-constrained Item Generation, where multiple LLM agents and feature-specific evaluators collaborate to generate and iteratively revise items based on intended constraints. Furthermore, to verify the efficacy of MAFIG in difficulty control, we propose a method for constructing a sequence of feature constraint sets that yield items with monotonically increasing difficulty. Experimental results demonstrate that MAFIG generates items that adhere to target constraints at a significantly higher rate than baselines, achieving robust difficulty control through the difficulty-calibrated constraint sequence.

[Long] A Scalable Multi-LLM Collaboration System with Retrieval-based Selection and Exploration-Exploitation-Driven Enhancement ▶ *Underline Video*

Shengji Tang, Jianjian Cao, Weihao Lin, Jiale Hong, Bo Zhang, Shuyue Hu, LEI BAI, Tao Chen, Wanli Ouyang and Peng Ye

Existing multi-LLM collaboration systems often encounter scalability challenges when integrating new LLMs and tasks, leading to suboptimal performance. To address this, we propose SMCS, a Scalable Multi-LLM Collaboration System designed to effectively coordinate multiple open-source LLMs. The system consists of two core components: a Retrieval-based Prior Selection (RPS) module, which dynamically selects the most suitable LLMs for each input, and an Exploration-Exploitation-Driven Posterior Enhancement (EPE) module, which fosters response diversity and selects high-quality outputs through a hybrid scoring mechanism. Experiments on eight mainstream benchmarks validate the effectiveness of our system: by integrating fifteen open-source LLMs, SMCS outperforms prevailing closed-source LLMs, e.g., GPT-4.1(+5.36%) and GPT-o3-mini(+5.28%) across multiple tasks. Remarkably, it even exceeds the average of best results on different datasets with open-source LLMs (+2.86%), significantly advancing the empirical performance frontier of open-source collaboration. The code is released at https://github.com/magent4aci/SMCS.

[Findings] Adapting Where It Matters: Depth-Aware Adaptation for Efficient Multilingual Speech Recognition in Low-Resource Languages ▶ *Underline Video*

Yang Xiao, Eun-Jung Holden and Ting Dang

Recent speech foundation models excel at multilingual automatic speech recognition (ASR) for high-resource languages, but adapting them to low-resource languages remains challenging due to data scarcity and efficiency constraints. Full-model fine-tuning is computationally expensive and prone to overfitting, while parameter-efficient methods like LoRA apply adaptation uniformly across layers, overlooking internal representations thus compromising effectiveness and efficiency. We analyze multilingual ASR models and reveal a U-shaped adaptability pattern: early and late layers are language-specific and require more adaptation, while intermediate layers retain shared semantics and need less. Building on this observation, we propose DAMA, a Depth-Aware Model Adaptation framework that allocates adaptation capacity according to each layers role. DAMA also introduces Singular Value Decomposition (SVD)-based initialization to constrain adaptation and preserve the U-shaped pattern, as well as a frozen middle-layer basis for further efficiency. Evaluated on 18 low-resource languages across two benchmark datasets, DAMA matches or surpasses state-of-the-art accuracy with 80% fewer trainable parameters, achieves a 29% error reduction under extreme data scarcity, and significantly improves memory, training time, and computational efficiency over baselines. These results highlight the benefits of structure-aware adaptation for efficient, scalable multilingual ASR.

[Long] Adaptive Constraint Propagation: Scaling Structured Inference for Large Language Models via Meta-Reinforcement Learning ▶ *Underline Video*

Ibne Farabi Shihab, Sanjeda Akter and Anuj Sharma

Large language models increasingly require structured inference, from enforcing JSON schema to multilingual parsing, where outputs must satisfy complex constraints. We introduce MetaJuLS, a meta-reinforcement learning approach that learns universal constraint propagation policies applicable across languages and tasks without task-specific retraining. By formulating structured inference as adaptive constraint propagation and training a Graph Attention Network with meta-learning, MetaJuLS achieves 1.5-2.0x speedups over GPU-optimized baselines while maintaining an accuracy within 0.2% of that of state-of-the-art parsers. On Universal Dependencies across 10 languages and LLM-constrained generation (LogicBench, GSM8K-Constrained), MetaJuLS demonstrates rapid cross-domain adaptation: a policy trained on English parsing adapts to new languages and tasks with 5–10 gradient steps (5–15 seconds) rather than requiring hours of task-specific training. Mechanistic analysis reveals that the policy employs human-like parsing strategies (easy-first) and novel, non-intuitive heuristics. By reducing the number of propagation steps in LLM deployments, MetaJuLS contributes to Green AI by directly reducing the inference carbon footprint.

[Findings] Adaptive Test-Time Compute Allocation with Evolving In-Context Demonstrations ▶ *Underline Video*

Bowen Zuo, Dongruo Zhou and Yinglun Zhu

While scaling test-time compute can substantially improve model performance, existing approaches either rely on static compute allocation or sample from fixed generation distributions. In this work, we introduce a test-time compute allocation framework that jointly adapts where computation is spent and how generation is performed. Our method begins with a warm-up phase that identifies easy queries and assembles an initial pool of question-response pairs from the test set itself. An adaptive phase then concentrates further computation on unresolved queries while reshaping their generation distributions through evolving in-context demonstrations—conditioning each generation on successful responses from semantically related queries rather than resampling from a fixed distribution. Experiments across math, coding, and reasoning benchmarks demonstrate that our approach consistently outperforms existing baselines while consuming substantially less inference-time compute.

[Long] Agent Newsroom: Efficient Chronological Report Generation via Dynamic Multi-Agent Collaboration ▶ *Underline Video*

Zhenhua Wang, Chunlei Wang, Yue Geng and Bang Wang

Many real-world applications require generating a chronological report from an evolving document stream; Timeline Summarization (TLS) provides a standard testbed for this setting. While large language models (LLMs) improve event synthesis, most LLM-based TLS systems remain monolithic: they repeatedly process overlapping evidence and often mirror the corpus' bursty reporting patterns, producing redundant timelines with temporal/topical imbalance and high cost. We propose MAS-TLS, a multi-agent framework that casts TLS as a newsroom-like collaboration. A master editor steers balanced coverage by allocating system-visible evidence with a coverage–diversity objective; specialist reporter agents independently draft time-anchored, evidence-grounded events while cross-reviewing to limit redundancy; an adjudication round reconciles competing drafts and consolidates duplicates into a global timeline; and a non-stationary Bayesian controller adaptively staffs agents under token/time budgets. Experiments on three benchmarks show that MAS-TLS improves semantic coverage and temporal grounding while substantially reducing token usage and latency.

[Long] Agent-based Substructure Counting under Local Differential Privacy ▶ *Underline Video*

Yuting Zhang, Kai Wang, Wei Ni, Ying Zhang and Wenjie Zhang

Recent studies have demonstrated the ability of Large Language Models (LLMs) in processing various graph problems. Substructure counting remains challenging in both scalability and accuracy. Incorporating sensitive edge information into the input prompts also introduces significant privacy risks of exposing the private information of user connections in real-world applications. This paper, for the first time, studies substructure counting for LLMs under edge local differential privacy (LDP) in a multi-agent framework. Unlike the Naive approach whose estimation relies entirely on overly dense noisy graphs, the proposed PSC framework decomposes substructure counting into node-level tasks distributed among node agents, and embeds the knowledge of distributed algorithms and DP frameworks in the curator agent and privacy controller, respectively. Thus, we can leverage the local neighboring information and reasoning capabilities of node agents to improve the estimation accuracy. Extensive experiments on 6 real-world datasets validate the effectiveness of PSC framework for substructure counting tasks under -edge LDP. Moreover, the non-DP version of PSC also demonstrated superior performance over a single LLM on standard substructure counting tasks.

[Findings] Agentic-R: Learning to Retrieve for Agentic Search ▶ *Underline Video*

Wenhan Liu, Xinyu Ma, Yutao Zhu, Yuchen Li, Daiting Shi, Dawei Yin and Zhicheng Dou

Agentic search has recently emerged as a powerful paradigm, where an agent interleaves multi-step reasoning with on-demand retrieval to solve complex questions. Despite its success, how to design a retriever for agentic search remains largely underexplored. Existing search agents typically rely on similarity-based retrievers, while similar passages are not always useful for final answer generation. In this paper, we propose a novel retriever training framework tailored for agentic search. Unlike retrievers designed for single-turn retrieval-augmented generation (RAG) that only rely on local passage utility, we propose to use both local query-passage relevance and global answer correctness to measure passage utility in a multi-turn agentic search. We further introduce an iterative training strategy, where the search agent and the retriever are optimized bidirectionally and iteratively. Different from RAG retrievers that are only trained once with fixed questions, our retriever is continuously improved using evolving and higher-quality queries from the agent. Extensive experiments on seven single-hop and multi-hop QA benchmarks demonstrate that our retriever, termed Agentic-R, consistently outperforms strong baselines across different search agents.

[Long] AIPO: Adaptive Information Guided Token-Level Reinforcement Learning for Large Language Model Reasoning ▶ *Underline Video*

BinChen BinChen, Hongfei Ye, Huiyang Wang, Wenxi Liu, Yu Zhang and Furui Liu

Reinforcement Learning with Verifiable Rewards (RLVR) improves the reasoning capability of Large Language Models (LLMs). Current RLVR trains LLMs on all generated tokens, rather than exploring which tokens actually contribute to reasoning. We propose AIPO(Adaptive–Information Policy Optimization), which focuses updates on those decisive tokens discovered on the fly. AIPO estimates each hidden states mutual information to score tokens. Policy gradients are then computed only on these critical tokens, using an advantage that blends information gain and verifiable correctness. To improve the efficiency of mutual-information estimation, AIPO adopts a Random-Fourier approximation of the Hilbert-Schmidt Independence Criterion. Across five math and science benchmarks, AIPO yields up to +20% accuracy over strong RLVR baselines while updating merely 10% of tokens, demonstrating superior efficiency and effectiveness. Our findings highlight the importance of information-driven token selection for efficient and effective reinforcement learning of LLM reasoning.

[Findings] AJ-Bench: Benchmarking Agent-as-a-Judge for Environment-Aware Evaluation ▶ *Underline Video*

Wentao Shi, Yu Wang, Yuyang Zhao, Yuxin Chen, Fuli Feng, Xueyuan Hao, Xi Su, Qi Gu, Hui Su, Xunliang Cai and Xiangnan He

As reinforcement learning continues to scale the training of large language model-based agents, reliably verifying agent behaviors in complex environments has become increasingly challenging. Existing approaches rely on rule-based verifiers or LLM-as-a-Judge models, which struggle to generalize beyond narrow domains. Agent-as-a-Judge addresses this limitation by actively interacting with environments and tools to acquire verifiable evidence, yet its capabilities remain underexplored. We introduce a benchmark AJ-Bench to systematically evaluate Agent-as-a-Judge across three domains, search, data systems, and graphical user interfaces, comprising 155 tasks and 516 annotated trajectories. The benchmark comprehensively assesses judge agents abilities in information acquisition, state verification, and process verification. Experiments demonstrate consistent performance gains over LLM-as-a-Judge baselines, while also revealing substantial open challenges in agent-based verification. Our data and code are available at https://aj-bench.github.io/.

[Long] AlignUSER: Human-Aligned LLM Agents via World Models for Recommender System Evaluation ▶ *Underline Video*

Nicolas Bougie, Gian Maria Marconi, Xiaotong Ye and Narimasa Watanabe

Evaluating recommender systems remains challenging due to the gap between offline metrics and real user behavior, as well as the scarcity of interaction data. Recent work explores large language model (LLM) agents as synthetic users, yet they typically rely on few-shot prompting, which yields a shallow understanding of the environment and limits their ability to faithfully reproduce user actions. We introduce AlignUSER, a framework that learns world-model-driven agents from human interactions. Given rollout sequences of actions and states, we formalize world modeling as a next state prediction task that helps the agent internalize the environment. To align actions with human personas, we generate counterfactual trajectories around demonstrations and prompt the LLM to compare its decisions with human choices, identify suboptimal actions, and extract lessons. The learned policy is then used to drive agent interactions with the recommender system. We evaluate AlignUSER across multiple datasets and demonstrate closer alignment with genuine humans than prior work, both at the micro and macro levels.

[Long] All Changes May Have Invariant Principles: Improving Ever-Shifting Harmful Meme Detection via Design Concept Reproduction ▶ *Underline Video*

Ziyou Jiang, Mingyang Li, Junjie Wang, Yuekai Huang, Jie Huang, Zhiyuan Chang, Zhaoyang Li and Qing Wang

Harmful memes are ever-shifting in the Internet communities, which are difficult to analyze due to their type-shifting and temporal-evolving nature. Although these memes are shifting, we find that different memes may share invariant principles, i.e., the underlying design concept of malicious users, which can help us analyze why these memes are harmful. In this paper, we propose RepMD, an ever-shifting harmful meme detection method based on the design concept reproduction. We first refer to the attack tree to define the Design Concept Graph (DCG), which describes steps that people may take to design a harmful meme. Then, we derive the DCG from historical memes with design step reproduction and graph pruning. Finally, we use DCG to guide the Multimodal Large Language Model (MLLM) to detect harmful memes. The evaluation results show that RepMD achieves the highest accuracy with 81.1% and has slight accuracy decreases when generalized to type-shifting and temporal-evolving memes. Human evaluation shows that RepMD can improve the efficiency of human discovery on harmful memes, with 1530 seconds per meme.

[Long] AlphaContext: An Evolutionary Tree-based Psychometric Context Generator for Creativity Assessment ▶ *Underline Video*

Yixuan Wang, Yue Huang, Hong Qian, Yunzhao Wei, Yifei Ding, Wenkai Wang, Zhi Liu, Zhongjing Huang, Aimin Zhou and Jiajun Guo

Creativity has become a core competence in the era of LLMs and human-AI collaboration, underpinning innovation in real-world problem solving. Crucially, the systematic improvement of creativity necessitates scientifically valid assessment instruments. Psychometric research recognizes context-based assessment as an effective way to measure creative thinking. However, high-quality expert-designed contexts remain scarce. Existing LLM-based generators often struggle with insufficient assessment cues, weak narrative coherence, limited stylistic diversity, and poor support for creative thinking. To address these challenges, we propose AlphaContext, an evolutionary tree-based psychometric context generator for creativity assessment. First, the HyperTree Outline Planner formalizes expert-designed outlining as a rule-guided hypertree and performs top-down hierarchical planning. The MCTS-based Context Generator fills the outline via MCTS to balance global structure and local quality. Then, the Evolutionary Context Optimizer evolves contexts with MAP-Elites by repeatedly updating niche elites to jointly improve diversity and quality. Finally, the Assessment-Guided Evolution Refiner simulates virtual participants with diverse styles and recycles weak contexts for further evolution. Experiments show that AlphaContext yields an average improvement of 8% over competitive methods across 6 quality metrics.

[Findings] AMO-Bench: Large Language Models Still Struggle in High School Math Competitions ▶ *Underline Video*

Junlin Liu, Shengnan An, Shuang Zhou, Dan Ma, Yehao Lin, Xinxuan Lv, Xuanlin Wang, Xiaoyu Li, Ziwen Wang, Xuezhi Cao and Xunliang Cai

We present AMO-Bench, an Advanced Mathematical reasoning benchmark with Olympiad level or even higher difficulty, comprising 50 human-crafted problems. Existing benchmarks have widely leveraged high school math competitions for evaluating mathematical reasoning capabilities of large language models (LLMs). However, many existing math competitions are becoming less effective for assessing top-tier LLMs due to performance saturation (e.g., AIME24/25). To address this, AMO-Bench introduces more rigorous challenges by ensuring all 50 problems are (1) cross-validated by experts to meet at least the International Mathematical Olympiad (IMO) difficulty standards, and (2) entirely original problems to prevent potential performance leakages from data memorization. Experimental results across 36 LLMs

on AMO-Bench highlights three key findings: (1) high-level mathematical reasoning remains challenging for current LLMs, with even the best-performing model achieving only 63.1% accuracy and most LLMs scoring below 50%; (2) scaling test-time compute remains a highly effective strategy for substantially improving reasoning performances, and (3) open-source models are progressively narrowing the performance gap with proprietary models. Additionally, we conduct further analysis about reasoning efficiency, volatility, and cross-lingual robustness, providing deeper insights behind the reasoning performances.

[Long] An Experimental Study on the Influence of Culture on Cross-Lingual Sentiment Transfer ▶ *Underline Video*

Ahao Liu, Haitong Yang and Wang Chuanrong

Identical linguistic expressions can convey different sentiments across cultural contexts. Yet, current multilingual models often reduce language to mere symbolic representation, neglecting the cultural pragmatics that fundamentally shape affective semantics in Sentiment Analysis (SA). Due to this oversight, the systematic performance degradation of Small Multilingual Language Models (SMLMs) on culturally distant targets is frequently attributed to resource constraints. This perspective obscures the pivotal role of cultural pragmatics, an intrinsic determinant of affective semantics, and thereby conceals cultural misalignment as the principal structural bottleneck. In this paper, we conduct a comprehensive empirical study to quantify the influence of culture on cross-lingual sentiment transfer across 7 common SMLMs and 5 linguistically diverse languages. By fitting a Linear Mixed-Effects Model (LMEM) on over 300 experimental runs, we disentangle cultural factors from confounding variables. Our results reveal that Cultural Distance is a significant, independent, negative predictor of transfer performance. Furthermore, representation probing and qualitative error analysis uncover a pragmatic alignment paradox: while SMLMs encode cultural distinctions, they fail to map these representations to downstream sentiment labels in high-context cultures. Ultimately, our work enhances the interpretability of cross-lingual transfer failures by statistically isolating cultural misalignment as a structural barrier, distinct from the resource constraints typically blamed for poor performance.

[Long] An Exploration of Mamba for Speech Self-Supervised Models ▶ *Underline Video*

Tzu-Quan Lin, Heng-Cheng Kuo, Tzu-Chieh Wei, Hsi-Chun Cheng, Chun Wei Chen, Hsien-Fu Hsiao, Yu Tsao and Hung-yi Lee

While Mamba has demonstrated strong performance in language modeling, its potential as a speech self-supervised learning (SSL) model remains underexplored, with prior studies limited to isolated tasks. To address this, we explore Mamba-based HuBERT models as alternatives to Transformer-based SSL architectures. Leveraging the linear-time Selective State Space, these models enable fine-tuning on long-context ASR with significantly lower compute. Moreover, they show superior performance when fine-tuned for streaming ASR. Beyond fine-tuning, these models show competitive performance on SUPERB probing benchmarks, particularly in causal settings. Our analysis shows that they yield higher-quality quantized representations and capture speaker-related features more distinctly than Transformer-based models. These findings highlight Mamba-based SSL as a promising and complementary direction for long-sequence modeling, real-time speech modeling, and speech unit extraction. The codebase is available at https://github.com/hckuo145/Mamba-based-HuBERT.

[Long] Analytical FFN-to-MoE Restructuring via Activation Pattern Analysis ▶ *Underline Video*

Zehua Pei, Hui-Ling Zhen, Lancheng Zou, Xianzhi Yu, Wulong Liu, Sinno Jialin Pan, Mingxuan Yuan and Bei Yu

Scaling large language models (LLMs) improves performance but significantly increases inference costs, with feed-forward networks (FFNs) consuming the majority of computational resources. While Mixture-of-Experts (MoE) architectures can reduce this cost through sparse activation, restructuring existing dense models into MoEs typically requires extensive retraining on hundreds of billions of tokens. We propose an analytical post-training framework that rapidly restructures FFNs into sparse MoE architectures using only a small calibration dataset. The method analyzes neuron activation patterns to partition neurons into always-active shared experts and conditionally activated routed experts, then constructs a router analytically from representative neuron statistics, enabling immediate deployment or optional lightweight fine-tuning. This approach applies both to dense models and recursively to existing MoE models for hierarchical sparsity. Experiments demonstrate up to 1.17x speedup in compute-bound scenarios with only minutes of processing and 2k-sample fine-tuning, outperforming methods requiring orders of magnitude more resources.

[Long] AnchorSeg: Language Grounded Query Banks for Reasoning Segmentation ▶ *Underline Video*

Rui Qian, Chuanhang Deng, Qiang Huang, Jian Xiong, Mingxuan Li, Yingbo Zhou, Wei Zhai, Jintao Chen and Dejing Dou

Reasoning segmentation requires models to ground complex, implicit textual queries into precise pixel-level masks. Existing approaches rely on a single segmentation token \{}<SEG\{}>, whose hidden state implicitly encodes both semantic reasoning and spatial localization, limiting the model's ability to explicitly disentangle what to segment from where to segment. We introduce AnchorSeg, which reformulates reasoning segmentation as a structured conditional generation process over image tokens, conditioned on language grounded query banks. Instead of compressing all semantic reasoning and spatial localization into a single embedding, AnchorSeg constructs an ordered sequence of query banks: latent reasoning tokens that capture intermediate semantic states, and a segmentation anchor token that provides explicit spatial grounding. We model spatial conditioning as a factorized distribution over image tokens, where the anchor query determines localization signals while contextual queries provide semantic modulation. To bridge token-level predictions and pixel-level supervision, we propose Token–Mask Cycle Consistency (TMCC), a bidirectional training objective that enforces alignment across resolutions. By explicitly decoupling spatial grounding from semantic reasoning through structured language grounded query banks, AnchorSeg achieves state-of-the-art results on ReasonSeg test set (67.7% gIoU and 68.1% cIoU). All code and models are publicly available at https://github.com/rui-qian/AnchorSeg.

[Long] arXiv2Table: Toward Realistic Benchmarking and Evaluation for LLM-Based Literature-Review Table Generation ▶ *Underline Video*

Weiqi Wang, Jiefu Ou, Yangqiu Song, Benjamin Van Durme and Daniel Khashabi

Literature review tables are essential for summarizing and comparing collections of scientific papers. In this paper, we study automatic generation of such tables from a pool of papers to satisfy a users information need. Building on recent work (Newman et al., 2024), we move beyond oracle settings by (i) simulating well-specified yet schema-agnostic user demands that avoid leaking gold column names or values, (ii) explicitly modeling retrieval noise via semantically related but out-of-scope distractor papers verified by human annotators, and (iii) introducing a lightweight, annotation-free, utilization-oriented evaluation that decomposes utility (schema coverage, unary cell fidelity, pairwise relational consistency) and measures paper selection via a two-way QA procedure (gold to system and system to gold) with recall, precision, and F1. To support reproducible evaluation, we introduce arXiv2Table, a benchmark of 1,957 tables referencing 7,158 papers, with human-verified distractors and rewritten, schema-agnostic user demands. We also develop an iterative, batch-based generation method that co-refines paper filtering and schema over multiple rounds. We validate the evaluation protocol with human audits and cross-evaluator checks. Extensive experiments show that our method consistently improves over strong baselines, while absolute scores remain modest, underscoring the tasks difficulty. Our data and code is available at https://github.com/JHU-CLSP/arXiv2Table.

[Long] Ascending the Infinite Ladder: Benchmarking Spatial Deformation Reasoning in Vision-Language Models ▶ *Underline Video*

Jiahuan Zhang, Shunwen Bai, Tianheng Wang, Kaiwen Guo, Zijia Song, Hanqing Wu, Guozheng Rao, Kai Han and Kaicheng Yu

Humans naturally possess the spatial reasoning ability to form and manipulate images and structures of objects in space. There is an increasing effort to endow Vision-Language Models (VLMs) with similar spatial reasoning capabilities. However, it remains unclear whether these models truly understand and manipulate spatial objects or not. To address this question, we propose a new evaluation framework aimed at assessing the performance of VLMs in spatial deformation reasoning tasks. Specifically, we construct a benchmark for spatial deformation reasoning from 2D to 3D. We explore whether the model can effectively perform spatial deformation reasoning from two directions: forward reasoning (given the operations, find the final state) and reverse reasoning (given the final state, determine the operations). We adopt a ladder competition format, using the number of deformation steps as the level classification criterion, with the goal of exploring the boundaries of the model's deformation reasoning capabilities. Interestingly, the benchmarking results reveal that almost no model demonstrates plausible spatial deformation reasoning abilities. Furthermore, even after applying targeted training and mainstream reasoning enhancement methods, the models are still unable to perform well on 3D spatial deformation reasoning.

[Findings] ATLAS: Orchestrating Heterogeneous Models and Tools for Multi-Domain Complex Reasoning ▶ *Underline Video*

Jinyang Wu, Guocheng Zhai, Ruihan Jin, Jiahao Yuan, Yuhao Shen, Shuai Zhang, Zhengqi Wen and Jianhua Tao

The integration of large language models (LLMs) with external tools has significantly expanded the capabilities of AI agents. However, as the diversity of both LLMs and tools increases, selecting the optimal model-tool combination becomes a high-dimensional optimization challenge. Existing approaches often rely on a single model or fixed tool-calling logic, failing to exploit the performance variations across heterogeneous model-tool pairs. In this paper, we present ATLAS (Adaptive Tool-LLM Alignment and Synergistic Invocation), a dual-path framework for dynamic tool usage in cross-domain complex reasoning. ATLAS operates via a dual-path approach: (1) training-free cluster-based routing that exploits empirical priors for domain-specific alignment, and (2) RL-based multi-step routing that explores autonomous trajectories for out-of-distribution generalization. Extensive experiments across 15 benchmarks demonstrate that our method outperforms closed-source models like GPT-4o as well as existing routing methods on both in-distribution (+10.1%) and out-of-distribution (+13.1%) tasks. Furthermore, our framework shows significant gains in visual reasoning by orchestrating specialized multi-modal tools.

[Long] Augur: Modeling Covariate Causal Associations in Time Series via Large Language Models ▶ *Underline Video*

Zhiqing Cui, Binwu Wang, Qingxiang Liu, Yeqiang Wang, Zhengyang Zhou, Yuxuan Liang and Yang Wang

Large language models (LLM) have emerged as a promising avenue for time series forecasting, offering the potential to integrate multimodal data. However, existing LLM-based approaches face notable limitations, such as marginalized role in model architectures, reliance on coarse statistical text prompts, and lack of interpretability. In this work, we introduce Augur, a fully LLM driven time series forecasting framework that exploits LLM causal reasoning to discover and use directed causal associations among covariates. Augur uses a two stage teacher student architecture where a powerful teacher LLM infers a directed causal graph from time series using heuristic search together with pairwise causality testing. A lightweight student agent then refines the graph and fine tune on high confidence causal associations that are encoded as rich textual prompts to perform forecasting. This design improves predictive accuracy while yielding transparent, traceable reasoning about variable interactions. Extensive experiments on real-world datasets with 25 baselines demonstrate that Augur achieves competitive performance and robust zero-shot generalization.

[Long] Automatic and Reliable Evaluation for Academic Caption-to-Figure Generation with LMMs ▶ *Underline Video*

Guanghui Ye, Huan Zhao, Qin Zhu, Fengnan Li, Jiaqi Li, Yixian Shen, Zhonghao Ren and Zhihua Jiang

Existing datasets for evaluating text-to-image generation focus mostly on real-life images, which poses challenges for assessing academic figure generation given real scientific captions, which is a hot topic in AI for Science. To fill the gap, we propose HE4AFG, a novel dataset which first provides a Holistic Evaluation for Academic caption-to-Figure Generation (AFG). Specifically, HE4AFG collects real figure captions from 8 scientific domains and finally generates 3,900 evaluation samples (particularly, including multi-panel figures) using 5 mainstream large multimodal models (LMMs). For each sample, we provide high-quality human ratings in terms of three aspects, scientific

aesthetic (SA), topic relevance (TR), and attribute correctness (AC). Moreover, we present two trainable models: (1) HE4AFG-E, an automated Evaluation model for AFG, which generates aspect-aware training examples and then use them to train three aspect-specific evaluation modules via contrastive learning; (2) HE4AFG-R, an automated Refinement model, which generates and utilizes feedback on the quality of the figures (e.g., unfaithful elements) to continuously improve AFG. Extensive experiments on HE4AFG demonstrate the effectiveness and performance advantages of our models.

[Demo] BabelDOC: Better Layout-Preserving PDF Translation via Intermediate Representation ▶ *Underline Video*

Yang Qi, Xiangyao Ma, Xiao Wang, Hao Wang and Rui Wang

As global cross-lingual communication intensifies, the language barrier in visually rich documents (e.g., PDFs) remains a critical bottleneck. Existing document translation pipelines suffer from a fundamental dilemma: traditional Computer-Assisted Translation (CAT) systems discard crucial layout and structural metadata, while advanced document parsers focus solely on extraction, lacking the end-to-end linguistic processing required for complex translation tasks. To bridge this gap, we introduce BabelDOC, a novel Intermediate Representation (IR)-based multimodal document translation framework. By explicitly decoupling visual layout from semantic content, BabelDOC maps raw PDFs into a unified IR, enabling advanced NLP interventions, such as automatic terminology extraction, cross-page contextual translation, and dictionary-constrained generation, without compromising structural integrity. The translated semantics are subsequently re-anchored and adaptively rendered into the original layout, supporting complex elements like mathematical formulas and nested tables. Featuring a highly extensible, plugin-based ecosystem compatible with state-of-the-art LLMs, BabelDOC establishes a new paradigm for lossless cross-lingual document intelligence. The open-source toolkit and its interactive downstream applications have garnered over 7.8k stars on GitHub https://github.com/funstory-ai/BabelDOC. A demonstration video is available at https://youtu.be/chwrlApH7a4.

[Findings] Beneficial Reasoning Behaviors in Agentic Search and Effective Training Methods to Obtain Them ▶ *Underline Video*

Jiahe Jin, Abhijay Sai Paladugu and Chenyan Xiong

Agentic search requires large language models (LLMs) to perform multi-step search to solve complex information-seeking tasks, imposing unique challenges on their reasoning capabilities. However, what constitutes effective reasoning for agentic search and how it can be learned remains unclear. In this work, we first investigate the reasoning behaviors that enable success in agentic search. By comparing successful and failed trajectories via an LLM-based analysis pipeline, we identify four beneficial behaviors: Information Verification, Authority Evaluation, Adaptive Search, and Error Recovery. Building on this, we propose Behavior Priming, a training approach that equips agentic search models with these reasoning behaviors before reinforcement learning (RL). Specifically, it collects trajectories with the identified behaviors for supervised fine-tuning (SFT), and then applies standard RL to further improve task performance. Experiments on Qwen3-1.7B and Llama3.2-3B-Instruct show that Behavior Priming yields relative improvements over direct RL by 37.2% on three web benchmarks and 6.2% on seven multi-hop QA benchmarks, and outperforms the SFT-then-RL baseline using outcome-correct trajectories for fine-tuning. Crucially, we show that these reasoning behaviors matter more than outcome correctness in the priming stage prior to RL. Further analysis reveals that Behavior Priming enhances exploration (pass@8) and test-time scaling (search step number), providing a robust foundation for RL.

[Long] Beyond Atomic Characters: Glyph-Aware Sub-character Alignment for Low-Resource Multilingual OCR ▶ *Underline Video*

Mengxiao Zhu, Haixu Chen, Jiu Sha, Jie Liu and Ge Shi

Low-resource multilingual OCR faces a dual challenge: complex script structures and severe data scarcity. In such settings, existing OCR models often struggle, as coarse visual representations combined with weak linguistic priors lead to frequent errors among visually similar characters. To address this, we present BASA (Beyond Atomic Sub-character Alignment), a OCR framework built upon high-resolution visual and language backbones with a novel glyph-aware interface. The core technical contribution is the Glyph-Aware Fine-grained Adapter (GAFA). Unlike standard linear projectors, GAFA employs learnable glyph prototypes to actively align sub-character structural primitives (e.g., strokes and radicals) with visual features, explicitly resolving topological ambiguities during vision-language alignment. To complement this, we introduce a two-stage curriculum learning strategy supported by a Glyph-Aware Reverse Synthesis pipeline, which generates large-scale multilingual training corpora with automatic, zero-cost component labels. Furthermore, we construct BASA-Bench, a representative benchmark spanning 11 languages with diverse script structures and 23 authentic scenarios. Experiments demonstrate that BASA achieves consistent improvements over strong OCR baselines, particularly on scripts with complex compositions. Our model and benchmark will be available at https://github.com/NcutLLM/BASA.

[Industry] Beyond Instruction Optimization: Multi-Agent Error-Driven Class Description Refinement for LLM-Based Classification ▶ *Underline Video*

Hamvir Dev, Shivam Ratnakant Mhaskar, Sasanka Vutla and Anup Pattnaik;

Large Language Models (LLMs) have demonstrated considerable efficacy in classification tasks, yet their performance depends on two critical prompt components: Task Instructions (HOW to classify) and Class Descriptions (WHAT defines each class). While prompt engineering research has extensively explored instruction optimization, class descriptions have received comparatively less attention, often being treated as fixed inputs or simple label names. This represents a critical gap for real-world classification tasks, particularly in contact center domains, where labels often suffer from ambiguous boundaries, overlapping definitions, and incomplete coverage of possible cases, substantially limiting accuracy regardless of instruction quality. We propose a multi-agent framework for iteratively refining class descriptions based on classification errors. By analyzing misclassified instances, language agents automatically generate improved descriptions that better capture class distinctions and resolve ambiguities. Empirical evaluation across contact center and public benchmark

datasets demonstrates upto 20.71% accuracy improvements over static class descriptions, addressing an orthogonal dimension to existing instruction optimization techniques.

[Long] Beyond Surface Features: Advancing Medical Vision-Language Alignment via Dynamic Evidence-Guided Preference Optimization ▶ *Underline Video*

Zixuan Huang, Zhihong Zhu, Xiaolong Liu, Yanchao Hao, Manman Zhang, Zheng Wei, Bowen Xing, Xian Wu, Ye Li, Fen Miao and Yefeng Zheng

Medical large Vision-Language Models (Med-LVLMs) have shown strong potential in multimodal clinical applications such as medical visual question answering and report generation. However, Med-LVLMs remain challenged by hallucinations caused by modality misalignment, where models prioritize textual knowledge over visual evidence and generate outputs that conflict with medical images. To mitigate this issue, recent studies have explored preference optimization to improve image-text alignment, achieving promising results. Despite these advances, existing preference-based methods still face two limitations in medical settings: (1) overfitting to superficial cues, and (2) pseudo convergence of the preference signal. In this paper, we propose Dynamic Evidence-Guided Preference Optimization (DEPO), a new framework that enables evidence-aware and adaptive preference learning for Med-LVLMs. DEPO introduces Multi-Modal Evidence Perturbation (MEP) to suppress non-causal textual and visual shortcuts, and Dispreferred Evidence Resampling (DER) to continuously update dispreferred responses as hallucination patterns evolve. Experiments on multiple medical VQA and report generation benchmarks demonstrate consistent improvements over existing methods, with strong robustness across datasets and architectures. All Codes and data will be released after review.

[Long] Beyond Variance: Knowledge-Aware LLM Compression via Fisher-Aligned Subspace Diagnostics ▶ *Underline Video*

Ibne Farabi Shihab, Sanjeda Akter and Anuj Sharma

Post-training activation compression is essential for deploying Large Language Models (LLMs) on resource-constrained hardware. However, standard methods like Singular Value Decomposition (SVD) are gradient-blind: they preserve high-variance dimensions regardless of their impact on factual knowledge preservation. We introduce Fisher-Aligned Subspace Compression (FASC), a knowledge-aware compression framework that selects subspaces by directly modeling activation-gradient coupling, minimizing a second-order surrogate of the loss function. FASC leverages the Fisher Information Matrix to identify dimensions critical for factual knowledge, which often reside in low-variance but high-gradient-sensitivity subspaces. We propose the Dependence Violation Score (rho) as a general-purpose diagnostic metric that quantifies activation-gradient coupling, revealing where factual knowledge is stored within transformer architectures. Extensive experiments on Mistral-7B and Llama-3-8B demonstrate that FASC preserves 6-8% more accuracy on knowledge-intensive benchmarks (MMLU, LAMA) compared to variance-based methods at 50% rank reduction, effectively enabling a 7B model to match the factual recall of a 13B uncompressed model. Our analysis reveals that rho serves as a fundamental signal of stored knowledge, with high-rho layers emerging only when models internalize factual associations during training

[Long] Bias Fitting to Mitigate Length Bias of Reward Model in RLHF ▶ *Underline Video*

Kangwen Zhao, Jianfeng Cai, Jinhua Zhu, Ruopei Sun, Dongyun Xue, Wengang Zhou, Li Li and Houqiang Li

Reinforcement Learning from Human Feedback (RLHF) relies on reward models to align large language models with human preferences. However, RLHF often suffers from reward hacking, wherein policy learning exploits flaws in the trained reward model to maximize reward scores without genuinely aligning with human preferences. A significant example of such reward hacking is length bias, where reward models usually favor longer responses irrespective of actual response quality. Previous works on tackling length bias have notable limitations, these approaches either mitigate bias without characterizing the bias form, or simply assume a linear length-reward relation. To accurately model the intricate nature of length bias and facilitate more effective bias mitigation, we propose FiMi-RM (Bias Fitting to Mitigate Length Bias of Reward Model), a framework that autonomously learns and corrects underlying bias patterns. Our approach consists of three stages: First, we warm up by training a standard reward model which inherently contains length bias. Next, we deploy a lightweight fitting model to capture the non-linear relation between length and reward. Finally, we incorporate this learned relation into the reward model, effectively decoupling length from reward while preserving preference modeling capabilities. Experimental results demonstrate that FiMi-RM achieves a more balanced length-reward distribution. Furthermore, when applied to alignment algorithms such as Direct Preference Optimization (DPO) and Best-of-N (BoN), our debiased reward model improves length-controlled win rate and reduces verbosity without compromising its performance.

[SRW] Boosting Self-Consistency with Ranking ▶ *Underline Video*

Maria Marina; Daniil Moskovskiy; Sergey Pletenev; Mikhail Salnikov; Alexander Panchenko; Viktor Moskvoretskii

Self-consistency improves large language models by sampling multiple reasoning paths and selecting the most frequent answer, but majority vote often fails to recover correct answers that are already present among samples. In this work, we reformulate answer selection in self-consistency as a ranking problem. Instead of relying on a single uncertainty or confidence signal, we train a lightweight reranker to score candidate answers using five carefully designed features that capture answer-level frequency, semantic centrality, and reasoning-trace consistency. We instantiate this approach with a LambdaRank model and evaluate it on three datasets under a range of test-time budgets. Across datasets, the proposed method consistently achieves a better accuracy-efficiency trade-off than standard self-consistency and strong baselines, with particularly large gains on question answering benchmarks. Further analysis shows that the proposed features are individually useful and, more importantly, complementary, highlighting the value of learning to combine multiple informative signals for test-time answer selection.

[Long] Breaking the Impasse: Dual-Scale Evolutionary Policy Training for Social Language Agents ▶ *Underline Video*

Minzheng Wang, Run Luo, Yanbo Wang, Zichen Liu, Yuqiao Tan, Tao Tan, Nan Xu, Lu Wang and Wenji Mao

While Reinforcement Learning with Verifiable Rewards (RLVR) has proven effective for closed-ended tasks, extending it to open-ended social language games via self-play reveals a critical issue: evolution impasse. Due to the vast strategy space, language agents frequently converge to homogenized behaviors, leading to deterministic match outcomes that eliminate the gradient signals necessary for policy evolution. To tackle this issue, we propose Dual-scale Evolutionary Policy Training (DEPT) for social language games. DEPT introduces a time-scaled evolutionary perception mechanism that detects impasse by quantifying dual-scale value baseline divergence alongside match entropy. Upon perceiving the collapse, it then activates asymmetric advantage reshaping to dynamically modulate the optimization landscape for intervention. Thus, our method effectively restores gradient signals and enforces sustained strategic exploration. Extensive experiments on multiple social language games demonstrate that DEPT outperforms strong baselines, avoiding policy degeneration and driving the continuous evolution of social language agents.

[Demo] Bridging Kernel Drivers and Virtual Device Models with LLM-Powered Automation ▶ *Underline Video*

Mingyu Wang, Bin Yu, Wenjian Lu, Zhi Wang, Kefeng Gao, Cheng Wen, Xu Lu and Cong Tian

Linux kernel device drivers are tightly coupled with hardware, making them difficult to execute and test without physical devices. This heavily limits automated code analysis and vulnerability discovery. While manual modeling is unscalable, Large Language Models (LLMs) offer a new approach to scale virtual device construction across the Linux driver ecosystem. In this paper, we present DevGen, an LLM-powered tool that generates QEMU-based virtual devices directly from Linux driver source code. DevGen combines static analysis to gather necessary context, guides the LLM through step-by-step prompting, and uses an automated self-correction loop driven by compilation and execution feedback. To further reduce errors, similar fixes are retrieved from a library of common modeling failures and incorporated into the repair prompt, which supports more targeted corrections in later iterations. The generated devices finally integrate with QEMU and Syzkaller, enabling driver fuzzing without physical hardware. DevGen is evaluated on 50 PCI/PCIe drivers from Linux 6.18 using three mainstream LLMs, and successfully generates usable models for 44 drivers. In these drivers, 24% of them achieve significant improvements in fuzzing coverage, and 7 previously unknown crashes are triggered with 1 CVE assigned. These results demonstrate the practical capability of LLMs to automate complex, system-level code generation tasks.

[Long] Business as Rulesual: A Benchmark and Framework for Business Rule Flow Modeling with LLMs ▶ *Underline Video*

Chen Yang, Ruping Xu, Ruizhe Li, Bin Cao and Jing Fan

Extracting structured procedural knowledge from unstructured business documents is a critical yet unresolved bottleneck in process automation. While prior work has focused on extracting linear action flows from instructional texts (e.g., recipes), it has insufficiently addressed the complex logical structures, such as conditional branching and parallel execution, that are pervasive in real-world regulatory and administrative documents. Furthermore, existing benchmarks are limited by simplistic schemas and shallow logical dependencies, restricting progress toward logic-aware large language models (LLMs). To bridge this "Logic Gap", we introduce BREX, a carefully curated benchmark comprising 409 real-world business documents and 2,855 expert-annotated rules. Unlike prior datasets centered on narrow service scenarios, BREX spans over 30 vertical domains, covering scientific, industrial, administrative, and financial regulations. We further propose ExIde, a structure-aware reasoning framework that investigates five distinct prompting strategies, ranging from implicit semantic alignment to executable grounding via pseudo-code generation, enabling explicit modeling of rule dependencies and providing an out-of-the-box framework for different business customers without finetuning their own LLMs. We benchmark ExIde using 13 state-of-the-art LLMs. Our extensive evaluation reveals that: (1) Executable grounding serves as a superior inductive bias, significantly outperforming standard prompts in rule extraction; and (2) Reasoning-optimized models demonstrate a distinct advantage in tracing long-range dependencies and non-linear rule dependencies compared to standard instruction-tuned models.

[Long] CADMate: Generating CAD Assembly Plan with Geometric Chain-of-Thought and Spatial Physical Rewards ▶ *Underline Video*

Jiali Chen, DingBa Fu, Xusen Hei, Yuhang Liu, Yiyang Chen, Jiayuan Xie, Wenqi Fan and Yi Cai

Computer-aided design (CAD) is crucial in prototyping complex 3D objects through precise geometric modeling. In practical design workflows, designers manually define assembly sequences for individual CAD parts, a process that is both time-consuming and expertise-intensive. To address this challenge, we formulate CAD assembly as a parametric action prediction task: given a reference design image and disassembled parts, the model predicts 6-DoF transformations (, actions) to progressively assemble each part. This paradigm enables multimodal large language models (MLLMs) to solve the task through autoregressive action generation. While recent MLLMs demonstrate promising spatial reasoning, they struggle with fine-grained geometric structure understanding and physical collision avoidance during assembly. In this paper, we propose CADMate, an MLLM-based framework for sequential CAD assembly action generation. Our training strategy comprises three stages: (i) CAD domain adaptation for spatial geometry and position understanding, (ii) supervised fine-tuning with geometric chain-of-thought (CoT) reasoning for action generation, and (iii) reinforcement learning with spatial-physical rewards jointly optimize spatial accuracy and collision avoidance. Additionally, we also construct CADBuilder dataset, comprising over 45K CAD assemblies with annotated action sequences. Our experiments demonstrate that CADMate significantly outperforms existing prominent MLLMs (, GPT-5), showing great potential in design applications.

[SRW] CAL-Log: Cost-Aware Active Learning with Logarithmic Cognitive Effort Modeling and Online Adaptation to Human Annotation Behavior ▶ *Underline Video*

Vihanga Supasan Kariyakaranage; Banuka Athuraliya

Active learning (AL) reduces labeled data requirements in NLP, yet most methods optimize label efficiency while ignoring annotation cost. Standard uncertainty sampling assumes uniform effort, leading to suboptimal resource allocation when documents vary in length. Supasan and Athuraliya (2026) introduced CAL-Log, a cost-aware AL variant using logarithmic cost modeling C(x)=alpha+beta log(1+L(x)), where C(x) is the predicted annotation time for document x and L(x) is its token length, grounded in information foraging theory (Pirolli and Card, 1999) and psycholinguistic studies of human skimming (Rayner, 1998). This paper presents CAL-Log in full, extending that preliminary framework with two new contributions: temperature-scaled calibrated entropy and online per-annotator cost adaptation, which together resolve the cold-start calibration bottleneck identified in the prior work. Experiments on ten text classification benchmarks demonstrate a 3.3x speedup over BADGE (Batch Active learning by Diverse Gradient Embeddings; Ash et al., 2020) and 3.9x over Entropy sampling to reach F1=0.80, with large effect sizes (Cohen's d>0.8). A live annotation deployment with preliminary user evaluation (N=7) confirms that the online cost model produces reading-speed classifications consistent with annotator self-reports, and that a transparency interface successfully communicates the scoring rationale to non-expert users.

[Long] Calibration-Aware Policy Optimization for Reasoning LLMs ▶ *Underline Video*

Ziqi Wang, Xingzhou Lou, Meiqi Wu, Zhengqi Wen and Junge Zhang

Group Relative Policy Optimization (GRPO) enhances LLM reasoning but often induces overconfidence, where incorrect responses yield lower perplexity than correct ones, degrading relative calibration as described by the Area Under the Curve (AUC). Existing approaches either yield limited improvements in calibration or sacrifice gains in reasoning accuracy. We first prove that this degradation in GRPO-style algorithms stems from their uncertainty-agnostic advantage estimation, which inevitably misaligns optimization gradients with calibration. This leads to improved accuracy at the expense of degraded calibration. We then propose Calibration-Aware Policy Optimization (CAPO). It adopts a logistic AUC surrogate loss that is theoretically consistent and admits regret bound, enabling uncertainty-aware advantage estimation. By further incorporating a noise masking mechanism, CAPO achieves stable learning dynamics that jointly optimize calibration and accuracy. Experiments on multiple mathematical reasoning benchmarks show that CAPO-1.5B significantly improves calibration by up to 15% while achieving accuracy comparable to or better than GRPO, and further boosts accuracy on downstream inference-time scaling tasks by up to 5%. Moreover, when allowed to abstain under low-confidence conditions, CAPO achieves a Pareto-optimal precision-coverage trade-off, highlighting its practical value for hallucination mitigation.

[Long] Can LLMs Estimate Cognitive Complexity of Reading Comprehension Items? ▶ *Underline Video*

Seonjeong Hwang, Hyounghun Kim and Gary Lee

Estimating the cognitive complexity of reading comprehension (RC) items is crucial for assessing item difficulty before it is administered to learners. Unlike syntactic and semantic features, such as passage length or semantic similarity between options, cognitive features that arise during answer reasoning are not readily extractable using existing NLP tools and have traditionally relied on human annotation. In this study, we examine whether large language models (LLMs) can estimate the cognitive complexity of RC items by focusing on two dimensions, Evidence Scope and Transformation Level, that indicate the degree of cognitive burden involved in reasoning about the answer. Our experimental results demonstrate that LLMs can approximate the cognitive complexity of items, indicating their potential as tools for prior difficulty analysis. Further analysis reveals a gap between LLMs reasoning ability and their metacognitive awareness: even when they produce correct answers, they sometimes fail to correctly identify the features underlying their own reasoning process.

[Long] Can We Predict Before Executing Machine Learning Agents? ▶ *Underline Video*

Jingsheng Zheng, Jintian Zhang, Yujie Luo, Yuren Mao, Yunjun Gao, Lun Du, Huajun Chen and Ningyu Zhang

Autonomous machine learning agents have revolutionized scientific discovery, yet they remain constrained by a Generate-Execute-Feedback paradigm. Previous approaches suffers from a severe Execution Bottleneck, as hypothesis evaluation relies strictly on expensive physical execution. To bypass these physical constraints, we internalize execution priors to substitute costly runtime checks with instantaneous predictive reasoning, drawing inspiration from World Models. In this work, we formalize the task of Data-centric Solution Preference and construct a comprehensive corpus of 18,438 pairwise comparisons. We demonstrate that LLMs exhibit significant predictive capabilities when primed with a Verified Data Analysis Report, achieving 61.5% accuracy and robust confidence calibration. Finally, we instantiate this framework in ForeAgent, an agent that employs a Predict-then-Verify loop, achieving a 6x acceleration in convergence while surpassing execution-based baselines by +6%.

[Findings] CAPA: Contribution-Aware Pruning and FFN Approximation for Efficient Large Vision-Language Models ▶ *Underline Video*

Samyak Jha and Junho Kim;

Efficient inference in Large Vision-Language Models is constrained by the high cost of processing thousands of visual tokens, yet it remains unclear which tokens and computations can be safely removed. While attention scores are commonly used to estimate visual token importance, they are an imperfect proxy for actual contribution. We show that Attention Contribution, which weights attention probabilities by value vector magnitude, provides a more accurate criterion for visual token selection. Our empirical analysis reveals that visual attention sinks are functionally heterogeneous, comprising Probability Dumps with low contribution that can be safely pruned, and Structural Anchors

with high contribution essential for maintaining model performance. Further, we identify substantial redundancy in Feed-Forward Networks (FFNs) associated with visual tokens, particularly in intermediate layers where image tokens exhibit linear behavior. Based on our findings, we introduce CAPA (Contribution-Aware Pruning and FFN Approximation), a dual-strategy framework that prunes visual tokens using attention contribution at critical functional transitions and reduces FFN computation through efficient linear approximations. Experiments on various benchmarks across baselines show that CAPA achieves competent efficiency–performance trade-offs with improved robustness.

[Long] CAPruner: Conceptual-Adjacent Scene Graph Pruner for Enhancing 3D Spatial Reasoning of Large Language Models ► *Underline Video*

Shengli Zhou, Xiangchen Wang, Guanhua Chen and Feng Zheng

Large language models (LLMs) have recently been applied to 3D vision-language (3D-VL) tasks, which require spatial reasoning to identify target objects relative to anchors. Scene graphs are commonly employed to represent such relations, but reasoning over complete graphs incurs high token costs and computational inefficiencies, motivating the need for pruning. Existing pruning methods primarily rely on spatial proximity and often remove task-relevant relations, thereby undermining reliable spatial reasoning. To address these limitations, we derive a key requirement for scene graph pruning: preserving spatial relations that are most pertinent to the specific 3D-VL task. Guided by this insight, we propose the Conceptual-Adjacent Scene Graph Pruner (CAPruner). CAPruner integrates fuzzy semantic relevance with spatial proximity to estimate the importance of relations, enabling the selection of critical relations in a task-specific context. Moreover, to avoid costly relation-level annotations, CAPruner is trained by supervising the aggregated scores of each node's incident edges. Extensive experiments demonstrate that CAPruner effectively preserves relations essential for spatial reasoning, leading to substantial performance improvements of LLMs on 3D-VL tasks. Code is available at https://github.com/fz-zsl/CAPruner.

[Long] Causal-ESC: Reliable Policy Learning for Emotional Support Conversation via Causal Inference ► *Underline Video*

Xv Wang, Wang Zhenyu, Guanyu Zheng and Rui Zhang

While Large Language Models (LLMs) have significantly advanced the fluency of Emotional Support Conversation (ESC) systems, current research predominantly focuses on engineering increasingly complex architectures, from intricate reasoning chains to multi-agent collaborations. While these advancements (e.g., CoT) offer semantic traces of reasoning, they remain mechanistically opaque, obscuring the fundamental causal mechanisms between dialogue features and effective empathic strategies, leading to poor interpretability and susceptibility to distribution shifts in offline learning. To address these limitations, we propose a novel framework Causal-ESC. Departing from conventional paradigms that directly utilize raw dialogue history as input, our approach introduces Doubly Robust (DR) learning to explicitly model the causal effect of utterance features on strategy selection, effectively mitigating the biases and counterfactual unobservability inherent in offline datasets. We further integrate an LLM-based stylized rewriting mechanism to translate these rigorously learned causal strategies into natural, context-consistent responses. Comprehensive experiments, supported by statistical verification (e.g., Outcome R^2) and human-like evaluation, demonstrate that our framework not only significantly outperforms state-of-the-art baselines in empathy and helpfulness but also provides a theoretically grounded, interpretable solution to the mechanistic interpretability dilemma in affective computing.

[Long] CEDAR: A Chinese Evaluation Dataset for Computational Argumentation ► *Underline Video*

Tian Lan, Jiang Li, Rong Yan, Feilong Bao, Weihua Wang, Guanglai Gao and Xiangdong Su

Computational argumentation has received increasing attention in recent years. However, existing debate datasets neglect some important labels for argument mining, generation, and evaluation. Meanwhile, the lack of comprehensively annotated Chinese oral debate datasets hinders progress in this field. To address these gaps, we introduce a comprehensive Chinese Evaluation Dataset for Computational Argumentation, named CEDAR. Compared to previous datasets, CEDAR includes the essential labels of computational argumentation (claim, stance, evidence) and five additional crucial labels: rhetorical figures, debater roles, modal words, utterance time, and debate results. Moreover, it offers complete transcripts of each debate, including speeches from the Pro and Con sides. Thus, the proposed CEDAR not only supports common argument mining and generation tasks, but also provides resources for rhetorical figure detection, argument quality evaluation, and debate result prediction. This dataset covers 600 debates about 318 topics from Chinese debate competitions. Besides providing a dataset for research, we conduct experiments on common computational argument tasks and a novel task (rhetorical figure detection), in which we also evaluate LLMs. The experimental results highlight the challenging nature of the dataset. Our corpus is available at https://github.com/VelikayaScarlet/CEDAR.

[Findings] Chain-in-Tree: Back to Sequential Reasoning in LLM Tree Search ► *Underline Video*

Xinzhe Li

Test-time scaling improves large language models (LLMs) on long-horizon reasoning tasks by allocating more compute at inference. LLM inference via tree search (LITS) achieves strong performance but is highly inefficient. We propose Chain-in-Tree (CiT), a plug-in framework that decides when to branch during search instead of expanding at every step. CiT introduces lightweight Branching Necessity (BN) evaluations, including BN-DP (direct prompting) and BN-SC (self-consistency). Integrated into Tree of Thoughts, ReST-MCTS, and RAP, BN-DP reduces token generation, model calls, and runtime by 75-85% on GSM8K and Math500, with often negligible or no accuracy loss. BN-SC typically yields substantial savings (up to 80%) generally but shows instability in 1-4 out of 14 settings, caused by a small subset of examples that produce extremely long reasoning steps. We theoretically prove that BN-DP never increases policy invocations and release unified implementations applicable across LITS frameworks. The full codebase is publicly available at

https://github.com/xinzhel/chain_in_tree.

[Long] Chain-of-Memory: Lightweight Memory Construction with Dynamic Evolution for LLM Agents ▶ *Underline Video*
Xiucheng Xu, Bingbing Xu, Tian Xueyun, Zihe Huang, Rongxin Chen, Li Yunfan and Huawei Shen
External memory systems are pivotal for enabling Large Language Model (LLM) agents to maintain persistent knowledge and perform long-horizon decision-making. Existing paradigms typically follow a two-stage process: computationally expensive memory construction (e.g., structuring data into graphs) followed by naive retrieval-augmented generation. However, our empirical analysis reveals two fundamental limitations: complex construction incurs high costs with marginal performance gains, and simple context concatenation fails to bridge the gap between retrieval recall and reasoning accuracy. To address above challenges, we propose CoM (Chain-of-Memory), a novel framework that advocates for a paradigm shift toward lightweight construction paired with sophisticated utilization. CoM introduces a Chain-of-Memory mechanism that organizes retrieved fragments into coherent inference paths through dynamic evolution, utilizing adaptive truncation to prune irrelevant noise. Extensive experiments on the LongMemEval and LoCoMo benchmarks demonstrate that CoM outperforms strong baselines with accuracy gains of 7.5%–10.4%, while drastically reducing computational overhead to approximately 2.7% of token consumption and 6.0% of latency compared to complex memory architectures.

[Short] Challenging the Explanation Based on Preceding Tokens: Discovering Transferable Non-Literal Biasing ▶ *Underline Video*
Yuchen Huang, Junpeng Zhang and Quanshi Zhang
In this paper, we find that the generated preceding tokens, which are not directly related to the answer, may still significantly push the large language model (LLM) towards the target answer. More crucially, the biased connotations of target answer in the preceding tokens can also transfer to other prompts. This finding suggests that the LLM may intentionally use the semantically unrelated tokens to help the generation of the target answer. Our finding offers a new perspective on understanding the long-range dependency phenomena in LLMs.

[Long] CIA: Inferring the Communication Topology from LLM-based Multi-Agent Systems ▶ *Underline Video*
Yongxuan Wu, Xixun Lin, He Zhang, Nan Sun, Kun Wang, Chuan Zhou, Shirui Pan and Yanan Cao
LLM-based Multi-Agent Systems (MAS) have demonstrated remarkable capabilities in solving complex tasks. Central to MAS is the communication topology which governs how agents exchange information internally. Consequently, the security of communication topologies has attracted increasing attention. In this paper, we investigate a critical privacy risk: MAS communication topologies can be inferred under a restrictive black-box setting, exposing system vulnerabilities and posing significant intellectual property threats. To explore this risk, we propose Communication Inference Attack (CIA), a novel attack that constructs new adversarial queries to induce intermediate agents' reasoning outputs and models their semantic correlations through the proposed global bias disentanglement and LLM-guided weak supervision. Extensive experiments on MAS with optimized communication topologies demonstrate the effectiveness of CIA, achieving an average AUC of 0.87 and a peak AUC of up to 0.99, thereby revealing the substantial privacy risk in MAS. The source code is available at https://github.com/aabbbcd/CIA.

[Long] CITE: Benchmarking Heterogeneous Text-Attributed Graph Models ▶ *Underline Video*
Chenghao Zhang, Qingqing Long, Ludi Wang, Wenjuan Cui, Jianjun Yu and Yi Du
Recent advances in large language models (LLMs) and text-aware graph learning have increased interest in reasoning over text-attributed graphs(TAGs). In many real-world settings, such graphs are inherently heterogeneous, with most existing benchmarks remaining largely homogeneous in structure. As a result, the lack of large-scale benchmarks for heterogeneous text-attributed graphs has hindered systematic evaluation and fair comparison of existing methods. In this work, we introduce CITE - Catalytic Information Textual Entities Graph, the first and largest heterogeneous text-attributed citation graph benchmark for catalytic materials. CITE contains over 438K nodes and 1.2M edges spanning four node types and four relation types, with rich node-level textual information. We establish standardized evaluation protocols for node classification and link prediction, and conduct ablation studies to assess the impact of graph heterogeneity and textual attributes. Using CITE, we benchmark four classes of learning paradigms, including homogeneous graph models, heterogeneous graph models, LLM-centric models, and hybrid LLM-graph models. By providing a large-scale heterogeneous text-attributed benchmark together with standardized evaluation protocols and comprehensive baselines, CITE enables systematic assessment across diverse modeling paradigms and offers new insights into text-aware and LLM-enhanced graph learning. The dataset, codebase and evaluation suite are publicly available.

[Long] CityVG: Contrastive Fine-Tuning and Reward-Based Chain-of-Thought Reasoning for Zero-Shot City-Scale 3D Visual Grounding ▶ *Underline Video*
Jianjun Zhang and Hanli Wang
3D Visual Grounding (3DVG) locates objects in 3D scenes based on natural language descriptions. However, existing methods are primarily confined to small-scale indoor data or rely on heavy supervision, failing to generalize to the complexity of large-scale urban environments. To address this limitation, we present CityVG, the first city-scale zero-shot 3D visual grounding framework capable of localizing urban objects without manual annotations. Our approach adopts a retrieval-and-reasoning paradigm comprising two key components. Specifically, we propose a contrastive fine-tuning strategy to align textual queries with urban scene graphs. By leveraging an LLM-driven graph clustering mechanism, we automatically construct high-quality positive and negative training pairs and fine-tune the text encoder via contrastive learning, resulting in a scene-adaptive text encoder that enables efficient alignment without grounding supervision. Complementing this, we introduce a multi-trajectory reward-based Chain-of-Thought (CoT) reasoning strategy for inference. This mechanism iteratively evaluates

candidate objects by aggregating reward scores across diverse reasoning trajectories, selecting the target that is most consistent with both appearance and spatial constraints. Extensive experiments on city-scale 3D grounding benchmarks demonstrate that CityVG achieves strong zero-shot localization performance and generalizes effectively to unseen urban environments.

[Long] CLAOCS-TX: Cross-Lingual Triplet Extraction with Aspect-Opinion-Aware Code-Switched Prompting and LLM-Guided Contrastive Distillation ▶ *Underline Video*

Lipika Dewangan and Chandresh Kumar Maurya

Cross-lingual learning enables the transfer of structured sentiment knowledge from high-resource languages to unlabeled or low-resource languages, but prior work has largely focused on coarse-grained sentiment classification or aspect extraction. In contrast, zero-shot cross-lingual aspect–opinion–sentiment triplet extraction (ASTE), which extracts sentiment triplets of the form (aspect term, opinion term, sentiment polarity), remains underexplored. We propose a unified framework that leverages large language models (LLMs) as both structured pseudo-label generators and semantic teachers for ASTE. Our approach employs stepwise structured prompting over aspect- and opinion-aware code-switched variants to generate reliable pseudo triplets, followed by a multi-variant consistency filter to retain high-confidence supervision. We further introduce a triplet-aware contrastive distillation objective that aligns student triplet representations with LLM-encoded semantic embeddings. During inference, only the student ASTE model is used, without requiring LLM access. Experiments on four non-Indic and four low-resource Indic target languages show consistent improvements over strong cross-lingual and LLM-based baselines. The proposed method yields an absolute micro-F1 improvement of 5.3 points on non-Indic languages and 3.8 points on low-resource Indic languages compared to the best competing approach. Ablation results further validate the complementary roles of aspect- and opinion-aware code-switched prompting and triplet-aware contrastive distillation, with larger relative gains observed in low-resource Indic settings.

[Industry] CLARITY: A Framework and Benchmark for Conversational Language Ambiguity and Unanswerability in Interactive NL2SQL Systems ▶ *Underline Video*

Tabinda Sarwar, Farhad Moghimifar, Cong Duy Vu Hoang, Xiaoxiao Ma, Shawn Chang Xu, Fahimeh Saleh, Poorya Zaremoodi, Avirup Sil and Katrin Kirchhoff

NL2SQL systems deployed in industry settings often encounter ambiguous or unanswerable queries, especially in interactive scenarios with incomplete user clarifications. Existing benchmarks typically assume a single point of ambiguity and typically rely on users to resolve them via interactions, overlooking realistic failure modes. We introduce CLARITY, a framework that automatically generates ambiguous and unanswerable NL2SQL benchmark with multi-faceted ambiguity and diverse user behaviour in both single- and multi-turn settings. Using a constraint-driven pipeline, CLARITY transforms executable SQL into ambiguous queries augmented with grounded conversational continuations and schema-level metadata. Empirical evaluation on Spider and BIRD indicates that leading NL2SQL systems (based on strong LLMs) exhibit pronounced performance degradation under multi-faceted ambiguity. They frequently flag ambiguous inputs but fail to accurately localize or disambiguate the underlying schema-level sources, underscoring the need for further industry-focused research on robust ambiguity detection and resolution in NL2SQL bolstered by this work.

[Long] CodeRipple: Wavelet-Based Detection of LLM-Generated Code ▶ *Underline Video*

Xingyu Yao, Zhendong Mao and Quan Wang

Detecting LLM-generated code is crucial for ensuring software provenance, security, reliability, and licensing compliance. Existing training-free detectors, mostly adapted from text-based methods, rely on global statistics of the Token Perplexity Sequence (TPS) and struggle with code. We reveal a key insight: despite the convergence of global statistics, LLM-generated and human-written code differ fundamentally in their local TPS dynamics: the former shows narrow transient spikes while the latter exhibits broad sustained fluctuations. To capture this distinction, we introduce CodeRipple, a novel training-free detection framework that employs wavelet analysis to characterize TPS morphology across scales. It jointly leverages the Stationary Wavelet Transform to model fluctuation shape and the Discrete Wavelet Transform to quantify cross-scale energy distribution. Evaluated on three challenging benchmarks spanning diverse programming languages, multiple generating LLMs, and various evasion strategies, CodeRipple consistently outperforms existing training-free methods, demonstrating its superior effectiveness and generalizability without any model training. Code available at: https://github.com/yaoxingyu77/CodeRipple.

[Findings] CogNet-KG: Empowering Tutoring Dialogues with a Cognitively-Structured Knowledge Graph for STEM Learning ▶ *Underline Video*

Ding Yu, Yu Lu, Tengju Li, Shasha Xiong and Shengquan Yu

Educational knowledge graph (EKG) is a critical component of intelligent tutoring systems that is structured around cognitive principles and provides support for interactive teaching. Most existing EKGs usually rely on simplistic relations, bind with single subjects, and lack integration with explicit learning objectives. In this paper, we introduce CogNet-KG, a novel and cognitively-structured large-scale knowledge graph for STEM learning. CogNet-KG models nearly 500 core concepts across five subjects with various cognitively-grounded relations corresponding to specific learning objectives, thereby encoding a rich cognitive schema for guiding more effective teaching. Based on this structure, we then construct a high-quality tutoring dialogue dataset CogDialogue-QA by leveraging adaptive instructional strategies. Additionally, we train CogTutor-LM, a specialized tutorial LLM that internalizes this structured pedagogical reasoning. Overall evaluation demonstrates that CogTutor-LM generates responses with significantly greater instructional coherence and more appropriate pedagogical guidance compared to baselines, validating the effectiveness of our graph-driven approach to fostering knowledge integration and stimulating

students' thinking. The datasets are publicly available at https://github.com/KCAIED/CogNet-KG.

[Long] Cognitive Alpha Mining via LLM-Driven Code-Based Evolution ► *Underline Video*

Fengyuan Liu, Yi Huang, Sichun Luo, Yuqi Wang, Yazheng Yang, Xinye Li, Zefa Hu, Junlan Feng and Qi Liu

Discovering effective predictive signals, or alphas, from financial data with high dimensionality and extremely low signal-to-noise ratio remains a difficult open problem. Despite progress in deep learning, genetic programming, and, more recently, large language model (LLM)-based factor generation, existing approaches still explore only a narrow region of the vast alpha search space. Neural models tend to produce opaque and fragile patterns, while symbolic or formula-based methods often yield redundant or economically ungrounded expressions that generalize poorly. Although different in form, these paradigms share a key limitation: none can conduct broad, structured, and human-like exploration that balances logical consistency with creative leaps. To address this gap, we introduce the Cognitive Alpha Mining Framework (CogAlpha), which combines code-level alpha representation with LLM-driven reasoning and evolutionary search. Treating LLMs as adaptive cognitive agents, our framework iteratively refines, mutates, and recombines alpha candidates through multi-stage prompts and financial feedback. This synergistic design enables deeper thinking, richer structural diversity, and economically interpretable alpha discovery, while greatly expanding the effective search space. Experiments on 5 stock datasets from 3 stock markets demonstrate that CogAlpha consistently discovers alphas with superior predictive accuracy, robustness, and generalization over existing methods. Our results highlight the promise of aligning evolutionary optimization with LLM-based reasoning for automated and explainable alpha discovery.

[Long] ConfSpec: Efficient Step-Level Speculative Reasoning via Confidence-Gated Verification ► *Underline Video*

Siran Liu, Zane Cao and Yongchao He

Chain-of-Thought reasoning significantly improves the performance of large language models on complex tasks, but incurs high inference latency due to long generation traces. Step-level speculative reasoning aims to mitigate this cost, yet existing approaches face a long-standing trade-off among accuracy, inference speed, and resource efficiency. We propose ConfSpec, a confidence-gated cascaded verification framework that resolves this trade-off. Our key insight is an asymmetry between generation and verification: while generating a correct reasoning step requires substantial model capacity, step-level verification is a constrained discriminative task for which small draft models are well-calibrated within their competence range, enabling high-confidence draft decisions to be accepted directly while selectively escalating uncertain cases to the large target model. Evaluation across diverse workloads shows that ConfSpec achieves up to 2.24x end-to-end speedups while matching target-model accuracy. Our method requires no external judge models and is orthogonal to token-level speculative decoding, enabling further multiplicative acceleration.

[Long] Continuous Interpretive Steering for Scalar Diversity ► *Underline Video*

Ye-eun Cho

Pragmatic inference is inherently graded. Different lexical items give rise to pragmatic enrichment to different degrees. Scalar implicature exemplifies this property through scalar diversity, where implicature strength varies across scalar items. However, evaluations of pragmatic inference in large language models (LLMs) often rely on prompt-based manipulations. Beyond prompt-level effects, this study introduces Continuous Interpretive Steering (CIS), a method that probes graded pragmatic interpretation by treating activation-level steering strength as a continuous experimental variable. To support this analysis, this study introduces a new dataset, GraSD, which encodes graded scalar diversity. Experiments on four LLMs show that uniform activation steering increases pragmatic interpretations globally but collapses item-level variation, whereas graded activation steering yields differentiated interpretive shifts aligned with scalar diversity grades. Together, CIS and GraSD provide a principled framework for evaluating graded pragmatic sensitivity in LLMs.

[Long] Controllable Contamination Detection for Reliable LLM Evaluation with Statistical Guarantees ► *Underline Video*

Zheng Zhang, Qi Liu, Siyuan Liang, Ning Li, Zirui Hu, Weibo Gao, Rui Li, Zhenya Huang, Leszek Rutkowski, Baosheng Yu and Dacheng Tao

Large language models (LLMs) have achieved remarkable performance across diverse tasks, largely driven by large-scale pretraining. However, this data abundance introduces test data contamination, where benchmark datasets overlap with pretraining corpora, undermining the reliability of model evaluation by confounding memorization with genuine generalization. To mitigate this issue, existing training data detectors attempt to identify clean (unseen) samples from contaminated test sets, but often suffer from residual contamination due to the black-box nature of LLMs. As a result, contaminated data may be mistakenly retained, leading to unreliable evaluation.To address this challenge, we propose FTD (FDR-controlled Training Data detection), a principled framework that detects and filters contaminated evaluation data while providing a statistical guarantee: the proportion of contaminated samples mistakenly retained as clean, the false discovery rate (FDR), is provably controlled below a user-specified threshold. FTD combines multiple complementary detectors via an adaptive weighting strategy, and we theoretically show it achieves high statistical power under valid FDR control. Extensive experiments on real-world benchmarks demonstrate that FTD significantly reduces residual contamination compared to existing methods while preserving evaluation consistency.

[Long] Controllable Memory Usage: Balancing Anchoring and Innovation in Long-Term Human-Agent Interaction ► *Underline Video*

Zisu Huang, Muzhao Tian, Xiaohua Wang, Jingwen Xu, Zhengkang Guo, Qi Qian, Kaitao Song, Jiakang Yuan, Changze Lv and Xiaoqing Zheng

As LLM-based agents are increasingly used in long-term interactions, cumulative memory is critical for enabling personalization and

maintaining stylistic consistency. However, most existing systems adopt an "all-or-nothing" approach to memory usage: incorporating all relevant past information can lead to Memory Anchoring, where the agent is trapped by past interactions, while excluding memory entirely results in under-utilization and the loss of important interaction history. We show that an agent's reliance on memory can be modeled as an explicit and user-controllable dimension. We first introduce a behavioral metric of memory dependence to quantify the influence of past interactions on current outputs. We then propose Steerable Memory Agent, SteeM, a framework that allows users to dynamically regulate memory reliance, ranging from a fresh-start mode that promotes innovation to a high-fidelity mode that closely follows interaction history. Experiments across different scenarios demonstrate that our approach consistently outperforms conventional prompting and rigid memory masking strategies, yielding a more nuanced and effective control for personalized human-agent collaboration.

[Long] Coordinating Search-Informed Reasoning and Reasoning-Guided Search in Claim Verification ▶ *Underline Video*

Qisheng Hu, Quanyu Long and Wenya Wang

Multi-hop claim verification is inherently challenging, requiring multi-step reasoning to construct verification chains while iteratively searching for information to uncover hidden bridging facts. This process is fundamentally interleaved, as effective reasoning relies on dynamically retrieved evidence, while effective search demands reasoning to refine queries based on partial information. To achieve this, we propose Hierarchical Agent Reasoning and Information Search (HARIS), explicitly modeling the coordinated process of reasoning-driven searching and search-informed reasoning. HARIS consists of a high-level reasoning agent that focuses on constructing the main verification chain, generating factual questions when more information is needed, and a low-level search agent that iteratively retrieves more information, refining its search based on intermediate findings. This design allows each agent to specialize in its respective task, enhancing verification accuracy and interpretability. HARIS is trained using reinforcement learning with outcome-based rewards. Experimental results on the EX-FEVER and HOVER benchmarks demonstrate that HARIS achieves strong performance, greatly advancing multi-hop claim verification.

[Long] Correct When Paired, Wrong When Split: Decoupling and Editing Modality-Specific Neurons in MLLMs ▶ *Underline Video*

Tingchao Fu, Wenkai Wang, Fanxiao Li, Huadong Zhang, Jinhong Zhang, Dayang Li, Yunyun Dong, Renyang Liu and Wei Zhou

Although Knowledge Editing provides an efficient mechanism for updating the knowledge of Multimodal Large Language Models (MLLMs), we find that current paradigms still suffer from an important yet remain underexplored issue: editing decoupling failure, where entity-related knowledge can be updated when the model is triggered by multimodal inputs (text–image query pairs), however, it often reverts to outdated pre-edit facts when the paired inputs are split into unimodal ones. Our in-depth empirical analysis reveals that the entity knowledge in MLLMs is not stored as a unified representation, but is instead distributed across disentangled modality-specific pathways. As a result, updates biased toward multimodal queries fail to propagate effectively to unimodal circuits. To bridge this gap, we propose DECODE, which explicitly disentangles and localizes modality-specific neuron groups for targeted knowledge. Extensive experiments demonstrate that DECODE consistently achieves effective knowledge updates under different modality triggers, thereby mitigating editing decoupling failures.

[Long] Cross-Modal Coreference Alignment: Enabling Reliable Information Transfer in Omni-LLMs ▶ *Underline Video*

Hongcheng Liu, Yuhao Wang, Zhe Chen, Pingjie Wang, Zhiyuan Zhu, Yixuan Hou, Yanfeng Wang and Yu Wang

Omni Large Language Models (Omni-LLMs) have demonstrated impressive capabilities in holistic multi-modal perception, yet they consistently falter in complex scenarios requiring synergistic omni-modal reasoning. Beyond understanding global multimodal context, effective reasoning also hinges on fine-grained cross-modal alignment, especially identifying shared referents across modalities, yet this aspect has been largely overlooked. To bridge this gap, we formalize the challenge as a cross-modal coreference problem, where a model must localize a referent in a source modality and re-identify it in a target modality. Building on this paradigm, we introduce CrossOmni, a dataset comprising nine tasks equipped with human-designed reasoning rationales to evaluate and enhance this capability. Experiments on 13 Omni-LLMs reveal systematic weaknesses in cross-modal coreference, which we attribute to the absence of coreference-aware thinking patterns. To address this, we enhance cross-modal alignment via two strategies: a training-free In-Context Learning method and a training-based SFT+GRPO framework designed to induce such thinking patterns. Both approaches yield substantial performance gains and generalize effectively to collaborative reasoning tasks. Overall, our findings highlight cross-modal coreference as a crucial missing piece for advancing robust omni-modal reasoning.

[Long] Cross-Modal Masked Compositional Concept Modeling for Enhancing Visio-Linguistic Compositionality ▶ *Underline Video*

Wei Li, Zhen Huang and Xinmei Tian

Contrastively trained vision-language models like CLIP, have made remarkable progress in learning joint image-text representations, but still face challenges in compositional understanding. They often exhibit a "bag-of-words" behavior, struggling to capture the object relations, attribute-object bindings, and word order dependencies. This limitation arises not only from the reliance on global, single-vector representations for optimization, but also from the insufficient exploitation and modeling of the rich compositional information inherently present in paired image text data. In this work, we propose MACCO (MAsked Compositional Concept MOdeling), a framework that masks compositional concepts in one modality and reconstructs them conditioned on the full contextual information from the other, enabling the model to capture and align cross-modal compositional structures more effectively. To facilitate this process, we introduce two auxiliary objectives that jointly align and regularize masked features both inter-modally and intra-modally. Extensive experiments on five compositional benchmarks, along with in-depth analyses, demonstrate that our approach not only significantly enhances compositionality in VLMs but also improves their ability to capture syntactic structure and linguistic information. Additionally, the improved compositionality

also benefits text-to-image generation and multimodal large language model.

[Long] DAC-Bench: A Decision-Aware Benchmark for Compositional Mobile GUI Tasks ▶ *Underline Video*

Yuqing Zhang, Honghui Sheng, Xueyu Hu, Shengyu Zhang and Fei Wu

Mobile GUI agents powered by LMMs can perceive screens and follow instructions, yet existing benchmarks largely target short, linear workflows and step-level accuracy, offering limited insight into long-horizon planning and decision-making under branching structures. We present DAC-Bench, a decision-aware benchmark with compositional tasks comprising 830 episodes and 11,345 action steps across 35 applications on Android and iOS. Tasks are organized into Sequential, Conjunctive, Conditional, and Hierarchical structures, reflecting real-world multi-step and branching interaction patterns. To complement standard step-level evaluation, we introduce weighted longest common subsequence to capture length-sensitive progress and decision accuracy for branch correctness. Evaluations across 7 diverse agents show substantial performance degradation compared to prior benchmarks, with success rates dropping below 5% on 6-8 step tasks and branch accuracy averaging 38%, highlighting challenges in conditional decision-making. By exposing these failure modes, DAC-Bench provides a challenging and diagnostic benchmark for advancing decision-aware mobile GUI agents. Our code and dataset are available at: https://github.com/YuqingZhangMirror12/DAC-Bench.

[Long] Demystifying Uncertainty in LLMs: Active Calibration between Concepts and Human Evaluations ▶ *Underline Video*

Pengqi Li, Lizhong Ding, Zhehao Zhou, Chunhui Zhang, Jiarun Fu, Hao Li, Ye Yuan and Guoren Wang

Hallucinations arise when large language models (LLMs) guess rather than acknowledge their underlying uncertainty. Existing static strategies for mitigating hallucinations have been only partially successful, largely because they do not explicitly model the information gain from interacting with the external environment. Researchers need a general method to proactively steer users toward informative clarifications, thereby unlocking the models effective capacity under underspecified inputs. We model the uncertainty of LLMs in interactive settings and uncover the mechanism of active calibration between model concepts and human evaluations, improving reliability. We show that calibration error in LLMs density estimation admits a non-vanishing lower bound under non-interactive learning, while interaction empirically reduces it. We further characterize that calibration error identifies informative queries and that calibration can be accelerated by shifting query distributions from imbalanced to balanced regimes. Guided by these insights, we propose a calibration-driven Interactive Learning Strategy (ILS) that selects clarification queries by optimizing calibration error, providing both theoretical guarantees and empirical gains for reliability. Code and data are available at https://github.com/zhouyeah215/Demystifying_Uncertainty.

[Long] DeReA: Improving Idiom Translation with Detect-Retrieve-Arbitrate Reasoning ▶ *Underline Video*

Rongqing Jiang and Xuebo Liu

Idiom translation remains a formidable challenge for Large Language Models (LLMs), as the constraints of static parametric memory and the noise in sentence-level retrieval often lead to literal misinterpretations. To address this, we propose DeReA, a detect-retrieve-arbitrate framework. The system employs a preference-aligned detector to identify idiomatic spans by reasoning over semantic conflicts between literal and contextual meanings. Subsequently, an idiom-centric translator invokes a fine-tuned embedding model to efficiently retrieve canonical definitions from an external knowledge base. The translator then utilizes a dual-path arbitration mechanism to select the optimal rendering by weighing the retrieval-augmented translations against direct translation. To evaluate our framework, we introduce LoMI, a high-difficulty benchmark with low data contamination. Experimental results demonstrate that DeReA significantly enhances performance across various model scales, improving GPT-5-mini by over 5.2 points in both idiomatic quality and consistency according to LLM-based metrics. Furthermore, evaluations on an emerging slang dataset from Urban Dictionary validate the potential of our approach in handling novel and evolving linguistic data. Our code is available at https://github.com/jrongqing/DeReA.

[Findings] Diagnosing and Mitigating Sycophancy and Skepticism in LLM Causal Judgment ▶ *Underline Video*

Edward Y. Chang

Do frontier LLMs reason causally, or do they pattern-match, yielding under pressure and hedging under uncertainty? We frame causal judgment as evaluation along three axes, Utility, Safety, and Wise Refusal, across Pearls Ladder. We introduce Recursive Causal Audit (RCA), a process-integrity evaluator whose Judge has no access to gold labels: it checks whether a models answer is entailed by its own derivation, internally consistent, and not dominated by user hints under pressure. RCA unifies persona and pressure: prompt tone is the intervention that regulates pressure-induced drift. For fine diagnostic resolution we use CAUSALT3, with explicit trap families and standardized pressure protocols. CAUSALT3 reveals a Skepticism Trap (Claude Haiku rejects 60% of valid L1 links) and a Scaling Paradox (GPT-5.2 underperforms GPT-4-Turbo by 55 points on L3, driven by paralysis rather than hallucination). Under RCA, operating points shift toward the high-Utility, high-Safety quadrant without retraining, consistent with much of the observed failure arising from how answers are rendered under pressure rather than from missing causal knowledge.

[Findings] DKME: Rethinking Coupled Knowledge Memory for Lifelong Model Editing of Large Language Models ▶ *Underline Video*

Guanyu Zheng, Zhenyu Wang, Tingting He, Xv Wang, Haochang Wang, Yaokai Huang and Tiejun Zhao

Lifelong knowledge editing aims to inject a stream of factual updates into large language models (LLMs) without retraining, yet existing memory-based editors often suffer from catastrophic forgetting as edits accumulate. We argue that a key factor is the coupled knowledge memory mechanism, where addressing (routing) and storage (writing via memory-module updates) are entangled. This entanglement makes

it difficult to confine the effects of each edit to its intended scope, particularly in multi-domain and associated-fact editing streams, where updates either span diverse semantic domains or repeatedly modify related attributes of the same subject. Consequently, updating memory for one edit inadvertently alters the routing and stored representations of previously injected edits, leading to catastrophic forgetting as edits accumulate. We propose DKME, which decouples addressing from storage via two stages: decoupled semantic addressing learns a fact-aware manifold for scope-aware routing, and partitioned memory storage localizes edits to memory partitions identified by unsupervised clustering in the embedding space. Experiments on three benchmarks, including HalluEditBench, CKnowEdit, and WikiData_counterfact, demonstrate that DKME consistently achieves a more favorable trade-off between editing success and locality compared to baselines, while maintaining more stable performance as the edit scale increases.

[SRW] Does Locality Cost in Polish Medical Text Classification? Duplicate-Aware Evaluation of Federated Learning ▶ *Underline Video*

Daniel Cielak; Andrzej Czyewski

Federated learning is often framed as a practical trade-off in clinical NLP: safer data handling at the cost of lower predictive performance. We test this assumption in a benchmark-specific study of Polish medical text classification and show that the expected penalty of locality is not inevitable. A key issue is evaluation granularity: the test split contains 10,634 rows but only 670 unique normalized text hashes, with 18 inconsistent groups removed in strict grouped evaluation. We therefore compare centralized and federated training under both conventional instance-level scoring and a stricter hash-level protocol that controls duplicate inflation. In the strongest settings, federated training matches or slightly exceeds the centralized baseline, reaching instance-level Macro-F1 of 0.8826 +/- 0.0177 versus 0.8689 +/- 0.0124, and hash-level Macro-F1 of 0.8908 +/- 0.0220 versus 0.8841 +/- 0.0078.We do not claim formal privacy guarantees; rather, we show that in this duplicate-heavy Polish medical text benchmark, conclusions about locality depend strongly on evaluation hygiene. An anonymized GitHub repository with code, run manifests, and evaluation scripts is included with the submission.

[Long] DPEPO: Diverse Parallel Exploration Policy Optimization for LLM-based Agents ▶ *Underline Video*

Junshuo Zhang, Chengrui Huang, Feng Guo, Zihan Li, Ke Shi, Menghua Jiang, Jiguo Yu, Shuo Shang and Shen Gao

Large language model (LLM) agents that follow the sequential reason-then-act paradigm have achieved superior performance in many complex tasks. However, these methods suffer from limited exploration and incomplete environmental understanding, as they interact with only a single environment per step. In this paper, we first introduce a novel paradigm that enables an agent to interact with multiple environments simultaneously and share cross-trajectory experiences. Build upon this paradigm, we further propose Diverse Parallel Exploration Policy Optimization (DPEPO), a reinforcement learning (RL) algorithm that encourages the agent to perform diverse parallel exploration. There are two stages in DPEPO: initial supervised fine-tuning (SFT) imparts basic parallel reasoning and action generation, followed by reinforcement learning stage with a hierarchical reward scheme. We design a parallel trajectory-level success reward and two step-level rewards: Diverse Action Reward and Diverse State Transition Reward, which actively penalize behavioral redundancy and promote broad exploration. Extensive experiments on ALFWorld and ScienceWorld show that DPEPO achieves state-of-the-art (SOTA) success rates, while maintaining comparable efficiency to strong sequential baselines.

[Long] Draft, Verify, Restore: Self-Refining Historical Inscription Restoration with a Unified MLLM ▶ *Underline Video*

Yuyi Zhang, Junle Liu, Peirong Zhang, Jianliang Liu, Zhenhua Yang and Lianwen Jin

Inscriptions are invaluable cultural heritage, yet centuries of degradation (e.g., fractures, erosion, oxidation) have rendered many partially illegible. Existing Historical Inscription Restoration (HIR) methods rely on task-separated pipelines with irreversible error accumulation and patch-based generation that sacrifices page-level consistency. Therefore, we present UniHIR, the first unified MLLM for end-to-end historical inscription restoration. It integrates two novel designs, Draft-Guided Localization and Hierarchical Self-Refinement, to enable accurate damage localization and illegible-content prediction via iterative reasoning and self-correction. This unified approach enables true page-level restoration with consistent typography and style. To support training under high-resolution inputs and long sequences, we design UHIRFactory and construct HIRBench, enabling step-wise, memory-efficient instruction tuning with step-aware annotations for intermediate drafts and refinements. Experiments demonstrate that UniHIR achieves superior performance in both text restoration accuracy and appearance restoration quality, validating that HIR can be effectively tackled by a standalone model in a unified manner. The model and code are available at https://github.com/ZZXF11/UniHIR.

[Long] EA-Agent: A Structured Multi-Step Reasoning Agent for Entity Alignment ▶ *Underline Video*

Yixuan Nan, Xixun Lin, Yanmin Shang, Ge Zhang, Zheng Fang, Fang Fang and Yanan Cao

Entity alignment (EA) aims to identify entities across different knowledge graphs (KGs) that refer to the same real-world object and plays a critical role in knowledge fusion and integration. Traditional EA methods mainly rely on knowledge representation learning, but their performance is often limited under noisy or sparsely supervised scenarios. Recently, large language models (LLMs) have been introduced to EA and achieved notable improvements by leveraging rich semantic knowledge. However, existing LLM-based EA approaches typically treat LLMs as black-box decision makers, resulting in limited interpretability, and the direct use of large-scale triples substantially increases inference cost. To address these challenges, we propose EA-Agent, a reasoning-driven agent for EA. EA-Agent formulates EA as a structured reasoning process with multi-step planning and execution, enabling interpretable alignment decisions. Within this process, it introduces attribute and relation triple selectors to filter redundant triples before feeding them into the LLM, effectively addressing efficiency challenges. Experimental results on three benchmark datasets demonstrate that EA-Agent consistently outperforms existing EA methods and achieves

state-of-the-art performance. The source code is available at https://anonymous.4open.science/r/EA-Agent-5696.

[Long] EchoVLM: Dynamic Mixture-of-Experts Vision-Language Model for Universal Ultrasound Intelligence ▶ *Underline Video*
Chaoyin She, Ruifang Lu, Lida Chen, Wei Wang and Qinghua Huang
Ultrasound is the preferred early cancer screening modality due to non-ionizing radiation, cost-effectiveness, and real-time imaging, yet conventional diagnosis relies heavily on physician expertise, causing significant subjectivity and limited efficiency. Vision-Language Models (VLMs) show promise but lack ultrasound-specific knowledge and multi-organ generalization. We propose EchoVLM, the first open-source 10-billion-parameter ultrasound-tailored VLM with a Mixture-of-Experts (MoE) architecture. It is infused with knowledge across seven anatomical systems, trained on 208,941 clinical cases, 1.47 million ultrasound key-frame images, and over 100 diseases or imaging findings. Supporting clinical report generation, diagnosis prediction, and Visual Question Answering (VQA), it outperforms Qwen2-VL by 7.58 BLEU-1 and 3.45 ROUGE-1 points in report generation. This work shows substantial potential for establishing a general-purpose ultrasound VLM and lays a technical foundation for clinical translation. Source code and model weights are available at https://github.com/Asunatan/EchoVLM.

[Long] Efficient Multi-Agent System Training with Data Influence-Oriented Tree Search ▶ *Underline Video*
Wentao Shi, Zichun Yu, Fuli Feng, Xiangnan He and Chenyan Xiong
Large Language Model (LLM) based multi-agent systems (MAS) show strong potential for tackling complex tasks through collaborative intelligence. Monte Carlo Tree Search (MCTS) based methods provide promising approaches for enhancing MAS self-training by generating synthetic data, using Q-values to estimate agent contributions. However, relying solely on Q-values may misalign with the goal of selecting data most beneficial for MAS improvement. To address this discrepancy, we propose Data Influence-oriented Tree Search (DITS), a novel framework that incorporates influence scores to guide both tree search and data selection in data synthesis. By leveraging influence scores, we effectively identify the most impactful data for MAS improvement, thereby enhancing model performance. Furthermore, we derive a novel influence score estimation method tailored for non-differentiable metrics, significantly reducing computational overhead by calculating performance changes on the validation set. Extensive experiments on three different multi-agent tasks demonstrate the robustness and effectiveness of the proposed methods. Notably, our findings reveal that allocating more resources to estimate influence scores, rather than Q-values, during data synthesis can more effectively and efficiently enhance model training. The code is available at https://anonymous.4open.science/r/DITS-F1C4/.

[Long] Efficient Prior-Guided Reasoning for Robust Retrieval-Augmented Generation under Conflicts ▶ *Underline Video*
Xiaowei Yuan, Ziyang Huang, Zhao Yang, Yequan Wang, Jun Zhao and Kang Liu
Retrieval-Augmented Generation (RAG) has become a standard paradigm for grounding Large Language Models (LLMs) with external knowledge. However, RAG performance often degrades substantially when faced with noisy, outdated, or conflicting retrieved information. In this work, we empirically demonstrate that Prior-Guided Reasoning, a strategy that explicitly elicits the model's parametric knowledge as prior information to guide reasoning on retrieved documents, effectively mitigates the impact of external conflicts. Building on this, we propose BrPr (Bernoulli-gated reinforcement learning for Prior-Guided reasoning), a framework that achieves robust performance across varying degrees of external inconsistency. Furthermore, by employing a Bernoulli-gated dropout mechanism during training, BrPr distills the prior-driven reasoning capability into the model parameters, enabling efficient latent reasoning without explicit prior generation. The experimental results demonstrate that BrPr consistently exhibits superior robustness to external conflicts and noise.

[Industry] EfficientTool: A Cost-Effective Aligning Framework for Tool-Conditioned Agents in SME Scenarios ▶ *Underline Video*
Yuanqi MuBingfeng PiDefei XiaLei ZuoYongqi Zhang
Large language models (LLMs) are increasingly adopted in downstream industries, yet aligning proprietary agents remains challenging due to limited high-quality data and hardware constraints in small and medium-sized enterprises (SMEs). We propose EfficientTool, a cost-effective, tool-conditioned alignment framework forming a closed loop over data collection, iterative training, and deployment-oriented evaluation. EfficientTool adopts a self-evolving bootstrapping-based Trajectory Collection Pipeline for high-quality trajectory generation, followed by iterative Model Training Pipeline using tool-conditioned parameter-efficient fine-tuning (PEFT). We evaluate the model with Interaction and Evaluation Pipeline in public and private benchmarks, and deploy for an internal enterprise agent. Results show that EfficientTool effectively aligns model in SME scenarios while preserving general tool-calling capability.

[Long] Enhancing Multimodal Large Language Models for Ancient Chinese Character Evolution Analysis via Glyph-Driven Fine-Tuning ▶ *Underline Video*
Rui Song, Lida Shi, Ruihua Qi, Yingji Li and Hao Xu
In recent years, rapid advances in Multimodal Large Language Models (MLLMs) have increasingly stimulated research on ancient Chinese scripts. As the evolution of written characters constitutes a fundamental pathway for understanding cultural transformation and historical continuity, how MLLMs can be systematically leveraged to support and advance text evolution analysis remains an open and largely underexplored problem. To bridge this gap, we construct a comprehensive benchmark comprising 11 tasks and over 130,000 instances, specifically designed to evaluate the capability of MLLMs in analyzing the evolution of ancient Chinese scripts. We conduct extensive evaluations across multiple widely used MLLMs and observe that, while existing models demonstrate a limited ability in glyph-level comparison, their performance on core tasks-such as character recognition and evolutionary reasoning-remains substantially constrained.

Motivated by these findings, we propose a glyph-driven fine-tuning framework (GEVO) that explicitly encourages models to capture evolutionary consistency in glyph transformations and enhances their understanding of text evolution. Experimental results show that even models at the 2B scale achieve consistent and comprehensive performance improvements across all evaluated tasks. To facilitate future research, we publicly release both the benchmark and the trained models.

[Long] ES4R: Speech Encoding Based on Prepositive Affective Modeling for Empathetic Response Generation ► *Underline Video*

Zhuoyue Gao, Xiaohui Wang, Xiaocui Yang, Wen Zhang, Daling Wang, Shi Feng and Yifei Zhang

Empathetic speech dialogue requires not only understanding linguistic content but also perceiving rich paralinguistic information such as prosody, tone, and emotional intensity for affective understandings. Existing speech-to-speech large language models either rely on ASR transcription or use encoders to extract latent representations, often weakening affective information and contextual coherence in multi-turn dialogues. To address this, we propose ES4R, a framework for speech-based empathetic response generation. Our core innovation lies in explicitly modeling structured affective context before speech encoding, rather than relying on implicit learning by the encoder or explicit emotion supervision. Specifically, we introduce a dual-level attention mechanism to capture turn-level affective states and dialogue-level affective dynamics. The resulting affective representations are then integrated with textual semantics through speech-guided cross-modal attention to generate empathetic responses. For speech output, we employ energy-based strategy selection and style fusion to achieve empathetic speech synthesis. ES4R consistently outperforms strong baselines in both automatic and human evaluations and remains robust across different Large Language Model (LLM) backbones. Code: https://github.com/Bean0901/ES4R.

[Long] Estimating the Black-box LLM Uncertainty with Distribution-Aligned Adversarial Distillation ► *Underline Video*

Huizi Cui, Huan Ma, Qilin Wang, Yuhang Gao and Changqing Zhang

Large language models (LLMs) have progressed rapidly in complex reasoning and question answering, yet LLM hallucination remains a central bottleneck that hinders practical deployment, especially for commercial black-box LLMs accessible only via APIs. Existing uncertainty quantification methods typically depend on computationally expensive multiple sampling or internal parameters, which prevents real-time estimation and fails to capture information implicit in the black-box reasoning process. To address this issue, we propose Distribution-Aligned Adversarial Distillation (DisAAD), which introduces a generation-discrimination architecture to guide a lightweight proxy model to learn the high-quality regions of the output distribution of the black-box LLM, thus effectively endowing it with the ability to "know whether the black-box LLM knows or not". Subsequently, we use the proxy model to reproduce the specific responses of the black-box LLM and estimate the corresponding uncertainty based on evidence learning. Extensive experiments have verified the effectiveness and promise of our proposed method, indicating that a proxy model even one that only accounts for 1% of the target LLMs size can achieve reliable uncertainty quantification.

[Long] Evolving Beyond Snapshots: Harmonizing Structure and Sequence via Entity State Tuning for Temporal Knowledge Graph Forecasting ► *Underline Video*

Siyuan Li, Yunjia Wu, Yiyong Xiao, Pingyang Huang, Peize Li, Ruitong Liu, Yan Wen and Te Sun

Temporal knowledge graph (TKG) forecasting requires predicting future facts by jointly modeling structural dependencies within each snapshot and temporal evolution across snapshots. However, most existing methods are stateless: they recompute entity representations at each timestamp from a limited query window, leading to episodic amnesia and rapid decay of long-term dependencies. To address this limitation, we propose Entity State Tuning (EST), an encoder-agnostic framework that endows TKG forecasters with persistent and continuously evolving entity states. EST maintains a global state buffer and progressively aligns structural evidence with sequential signals via a closed-loop design. Specifically, a topology-aware state perceiver first injects entity-state priors into structural encoding. Then, a unified temporal context module aggregates the state-enhanced events with a pluggable sequence backbone. Subsequently, a dual-track evolution mechanism writes the updated context back to the global entity state memory, balancing plasticity against stability. Experiments on multiple benchmarks show that EST consistently improves diverse backbones and achieves state-of-the-art performance, highlighting the importance of state persistence for long-horizon TKG forecasting.

[Long] Evolving Sparsity: Leveraging Token Importance Dynamics for Efficient LLM Decoding with Sparse Attention ► *Underline Video*

Ruizi Han, Miao Zhang, Ziyue Qiao and Liqiang Nie

Efficient long-context inference remains a major challenge for large language models (LLMs), as the cost of attention computation during auto-regressive decoding grows linearly with the context length. Recent sparse attention methods attempt to reduce the computational burden by selecting a subset of tokens at each step, while most rely on static importance scores that are repeatedly computed over the entire cache, overlooking the relational dynamics of the decoding process. In this work, we revisit sparse attention in LLMs and propose to model token importance as a dynamic process that evolves over decoding steps and propagates through model layers. To efficiently measure token importance, we propose two lightweight mechanisms: (1) Cross-Step Accumulation, which incrementally maintains long-term, query-agnostic importance via decayed accumulation of sparse attention scores, avoiding recomputing the importance of decoded tokens; and (2) Cross-Layer Propagation, which leverages the models intrinsic Retrieval Heads to compute query-aware indices and efficiently propagate them across layers; Together, these mechanisms preserve both stable context memory and adaptive query relevance while reduce redundant computation. We evaluate our approach on PG-19, RULER, LongBench, and mathematical reasoning benchmarks using models employing Multi-Head and Grouped-Query Attention. Under varying KV cache budgets, our method consistently outperforms prior sparse attention

baselines, approaches full attention performance in most settings, and achieves speedups of up to 5.36x for attention latency and 2.33x for end-to-end decoding. Our code is available at: https://github.com/iLearn-Lab/ACL26-EvoSparse.

[Long] Example Quality Matters: Multi-Aspects Example Augmentation for Private Library Programming ► *Underline Video*

Yuhao Li, Haifeng Sun, Xuesong Zhang, Shu Yao, Haoyu Zheng, Yvchuan wang, Huazheng Wang, Zirui Zhuang, Qi Qi, Jianxin Liao and Jingyu Wang

Recent advances in large language models (LLMs) have significantly improved code-generation capabilities, particularly through retrieval-augmented generation (RAG) for private libraries. While RAG leverages API documentation to address the scarcity of private code corpora, its performance critically depends on the quality of retrieved examples. Existing approaches often overlook the intrinsic characteristics of these examples, particularly how factors such as complexity, readability, and correctness impact their effectiveness. In this study, we systematically investigate these three critical aspects, complexity, readability, and correctness, and find that optimal examples should exhibit moderate complexity, semantic correctness, and step-by-step execution patterns. Based on these findings, we propose ComboPrompt, a novel example enhancement method that strategically combines existing API examples to improve complexity, refines code structure for readability, and incorporates automated validation ensuring correctness. Extensive evaluations across five private library benchmarks and different LLMs demonstrate that ComboPrompt achieves up to 22% accuracy improvement over baseline approaches. Code is available at https://github.com/FireAndWin/ComboPrompt_ExampleQualityMatters.

[Long] Exons-Detect: Identifying and Amplifying Exonic Tokens via Hidden-State Discrepancy for Robust AI-Generated Text Detection ► *Underline Video*

Xiaowei Zhu, Yubing Ren, Fang Fang, Shi Wang, Yanan Cao and Li Guo

The rapid advancement of large language models has increasingly blurred the boundary between human-written and AI-generated text, raising societal risks such as misinformation dissemination, authorship ambiguity, and threats to intellectual property rights. These concerns highlight the urgent need for effective and reliable detection methods. While existing training-free approaches often achieve strong performance by aggregating token-level signals into a global score, they typically assume uniform token contributions, making them less robust under short sequences or localized token modifications. To address these limitations, we propose Exons-Detect, a training-free method for AI-generated text detection based on an exon-aware token reweighting perspective. Exons-Detect identifies and amplifies informative exonic tokens by measuring hidden-state discrepancy under a dual-model setting, and computes an interpretable translation score from the resulting importance-weighted token sequence. Empirical evaluations demonstrate that Exons-Detect achieves state-of-the-art detection performance and exhibits strong robustness to adversarial attacks and varying input lengths. In particular, it attains a 2.2% relative improvement in average AUROC over the strongest prior baseline on DetectRL.

[Findings] Experience is the Teacher: Reusing Atomic Thoughts from LLMs to Improve Medical Dialogue ► *Underline Video*

Guangya Yu, Hui Luo, Qi Ye, Ruihui Hou, Weiyan Zhang, Mingxi Shang, Xuanwu Li, ChunMing Wang and Tong Ruan

With the remarkable performance of large language models (LLMs) in medicine, particularly their ability to support clinical decision-making in medical dialogues, a key limitation remains: the static reasoning patterns derived from human expert experience are often inadequate for the dynamic and diverse nature of real-world multi-turn conversations. While recent large reasoning models (such as R1) enable deeper and more complex thought processes to address such challenges, they also introduce significant redundancy. Meanwhile, recent studies on reusing atomic thoughts demonstrate a practical pathway toward dynamic and precise reasoning in general domains. In this paper, we investigate the role of atomic thought-based experience in medical dialogue tasks. First, we collect human expert clinical experience. Then, we propose a novel distillation framework that extracts atomic thoughts from teacher models and reuses them to guide reasoning and generate responses. Based on this framework, we construct training data from ReMeDi and fine-tune student models, which demonstrate enhanced performance in both static and interactive medical dialogue scenarios. Furthermore, we examine the impact of experience across various models, datasets, and scenarios. Crucially, transferring this experience empowers weaker models to generate high-quality reasoning data, matching the annotation capabilities of stronger LLMs while significantly reducing costs. The code is available in this repository https://github.com/VioletAmethystLunar/Atomic-Thoughts-Medical-Dialogue.

[Long] Explainable and Fine-Grained Safeguarding of LLM Multi-Agent Systems via Bi-Level Graph Anomaly Detection ► *Underline Video*

Junjun Pan, Yixin Liu, Rui Miao, Kaize Ding, Yu Zheng, Quoc Viet Hung Nguyen, Alan Wee-Chung Liew and Shirui Pan

Large language model (LLM)-based multi-agent systems (MAS) have shown strong capabilities in solving complex tasks. As MAS become increasingly autonomous in various safety-critical tasks, detecting malicious agents has become a critical security concern. Although existing graph anomaly detection (GAD)-based defenses can identify anomalous agents, they mainly rely on coarse sentence-level information and overlook fine-grained lexical cues, leading to suboptimal performance. Moreover, the lack of interpretability in these methods limits their reliability and real-world applicability. To address these limitations, we propose, an explainable and fine-grained safeguarding framework for detecting malicious agents in MAS. To incorporate both coarse and fine-grained textual information for anomalous agent identification, we utilize a bi-level agent encoder to jointly model the sentence- and token-level representations of each agent. A theme-based anomaly detector further captures the evolving discussion focus in MAS dialogues, while a bi-level score fusion mechanism quantifies token-level contributions for explanation. Extensive experiments across diverse MAS topologies and attack scenarios demonstrate robust detection performance and strong interpretability of XG-Guard.

[Long] Exploring and Distilling Multi-Dimensional Clues for Interpretable Social Bot Detection ▶ *Underline Video*

Yi Han, Haiqi Lu, Lizi Liao, Shuhan Zhou, Yuanxing Liu, Weinan Zhang and Ting Liu

Social bot accounts have long been disseminating disinformation and engaging in malicious activities on social media platforms. Detecting these social bots has become a critical and urgent task, essential for maintaining a healthy online ecosystem. Existing social bot detection research usually provides detection results directly without corresponding supportive explanations, making it difficult to assess the extent to which such predictions are trustworthy. This is a key concern for online moderation. In this work, we explore the detection interpretation and summarize a four-dimensional clue framework from individual and social perspectives. We propose CDRBot, which primarily employs outcome-reward reinforcement learning to train inspectors to generate faithful, grounded, and readable clues from the User Information, Semantic Features, Interactive Situation, and Behavioral Pattern. These clues are then integrated to make final predictions. Experimental results demonstrate that our approach outperforms other baselines in detection performance. The generated clues are faithful, grounded, and readable, and can significantly enhance the performance of large language models in social bot detection.

[Demo] Fast-MIA: Efficient and Scalable Membership Inference for LLMs ▶ *Underline Video*

Hiromu Takahashi and Shotaro Ishihara

We propose Fast-MIA (https://github.com/Nikkei/fast-mia), a Python library for efficiently evaluating membership inference attacks (MIA) against large language models (LLMs). MIA has emerged as a crucial technique for auditing privacy risks and copyright infringement in LLMs. However, computational demands have grown substantially: recent methods rely on repeated inference, while practical auditing requires large-scale evaluation. Progress is further hindered by existing implementations that execute methods independently, redundantly computing shared intermediate results such as log-probabilities. To address these challenges, Fast-MIA combines two strategies: (1) high-throughput batch inference via vLLM, achieving approximately 5x speedup, and (2) a cross-method caching architecture that computes intermediate results once and shares them across methods. The library includes representative MIA methods under a unified framework, integrates with established benchmarks, and supports flexible YAML configuration. We release Fast-MIA under the Apache License 2.0 to support scalable and reproducible MIA research.

[SRW] FedPAGR: Federated Prototype Alignment via Geometric Refinement for Heterogeneous Architectures ▶ *Underline Video*

Kris Prasad; Md Abdullah Al Hafiz Khan

Federated learning with heterogeneous client architectures cannot rely on parameter aggregation. Prototype-based methods address architectural heterogeneity by exchanging class-level representations, but naively averaging prototypes across non-IID clients leads to semantic drift and poor inter-class separation. We propose FedPAGR, a framework where heterogeneous clients project their features into a shared consensus space and exchange class prototypes with a central server. The server refines aggregated prototypes through a geometric regularization objective that enforces agreement with client submissions and inter-class angular separation. Clients anchor their classifiers to the refined prototypes and train with a composite objective combining classification, prototype alignment, and entropy regularization. We evaluate FedPAGR across multiple domains, including four image benchmarks and a clinical NLP task using heterogeneous ClinicalBERT variants, with five architectures per federation under severe label heterogeneity (alpha=0.1). FedPAGR achieves the highest ensemble accuracy across all four image datasets and the highest local test accuracy on low-class and clinical tasks, including a 4.99-point improvement over the strongest baseline on MIMIC-IV, while remaining competitive on high-class benchmarks.

[Long] Feeling Rules in Language Models: Mapping Norms of Emotional Appropriateness Across Roles, Institutions, and Intensity ▶ *Underline Video*

Guangrui Fan, DanDan Liu, AZNUL QALID MD SABRI, Rui Zhang and Lihu Pan

When asked explicitly, a Large language model (LLM) may validate your anger, but implicitly, it may still judge that anger as inappropriate. We call this divergence the endorsement-exposure gap, and it reveals that LLMs encode hidden norms about which emotions are acceptable in which contexts. To measure these norms systematically, we introduce Feeling Rules Atlas, a benchmark of 1,320 vignettes spanning 6 institutional settings, 12 roles, 7 emotions, and 5 intensity levels. We pair the benchmark with two probes: explicit norm judgments (APPROPRIATE/INAPPROPRIATE/DEPENDS) and implicit acceptability scored by log-likelihood contrast. Across six model families, we find large cross-model variation in sanctioning thresholds and institutional "norm signatures" not reducible to overall strictness; models that appear similarly lenient explicitly can diverge sharply in implicit judgments. These results establish normative affect, context-conditioned judgments of emotional appropriateness, as a distinct alignment axis, and motivate transparent profiling of feeling rules for emotionally sensitive deployments.

[Long] Fine-Mem: Fine-Grained Feedback Alignment for Long-Horizon Memory Management ▶ *Underline Video*

Weitao Ma, Xiaocheng Feng, Lei Huang, Xiachong Feng, Zhanyu Ma, Jun Xu, Jiuchong Gao, Jinghua Hao, Renqing He and Bing Qin

Effective memory management is essential for large language model agents to navigate long-horizon tasks. Recent research has explored using Reinforcement Learning to develop specialized memory manager agents. However, existing approaches rely on final task performance as the primary reward, which results in severe reward sparsity and ineffective credit assignment, providing insufficient guidance for individual memory operations. To this end, we propose Fine-Mem, a unified framework designed for fine-grained feedback alignment. First, we introduce a Chunk-level Step Reward to provide immediate step-level supervision via auxiliary chunk-specific question answering tasks. Second, we devise Evidence-Anchored Reward Attribution to redistribute global rewards by anchoring credit to key memory operations,

based on the specific memory items utilized as evidence in reasoning. Together, these components enable stable policy optimization and align local memory operations with the long-term utility of memory. Experiments on Memalpha and MemoryAgentBench demonstrate that Fine-Mem consistently outperforms strong baselines, achieving superior success rates across various sub-tasks. Further analysis reveals its adaptability and strong generalization capabilities across diverse model configurations and backbones

[Findings] FLOP-Efficient Training: Early Stopping Based on Test-Time Compute Awareness ▶ *Underline Video*

Hossam Amer, Maryam Dialameh, Hossein Rajabzadeh, Walid Ahmed, Weiwei Zhang and Yang Liu

Scaling training compute, measured in FLOPs, has long been shown to improve the accuracy of large language models, yet training remains resource-intensive. Prior work shows that increasing test-time compute (TTC), for example through iterative sampling, can allow smaller models to rival or surpass much larger ones at lower overall cost. We introduce TTC-aware training, where an intermediate checkpoint and a corresponding TTC configuration can together match or exceed the accuracy of a fully trained model while requiring substantially fewer training FLOPs. Building on this insight, we propose an early stopping algorithm that jointly selects a checkpoint and TTC configuration to minimize training compute without sacrificing accuracy. To make this practical, we develop an efficient TTC evaluation method that avoids exhaustive search, and we formalize a break-even bound that identifies when increased inference compute compensates for reduced training compute. Experiments demonstrate up to 92% reductions in training FLOPs while maintaining and sometimes remarkably improving accuracy. These results highlight a new perspective for balancing training and inference compute in model development, enabling faster deployment cycles and more frequent model refreshes.

[Industry] FlowHN: Adaptive Token Routing for Efficient Parallel Hybrid Networks ▶ *Underline Video*

Mohammad Mahdi Moradi, Walid Ahmed, Shuangyue Wen, Sudhir Mudur, Weiwei Zhang and Yang Liu

Production LLMs must balance modeling quality with predictable latency, stable accelerator utilization, and cost-efficient scaling, constraints that remain difficult for existing architectures. Transformers provide strong reasoning but incur quadratic complexity, while state-space models (SSMs) scale efficiently yet lack fine-grained interactions; prior hybrids either introduce sequential bottlenecks or rely on learned routing that complicates deployment. We present FlowHN, a deployment-oriented parallel hybrid architecture that enables deterministic conditional computation via FLOP-aware token circulation across attention and SSM branches. Instead of dynamic expert routing, FlowHN performs hardware-aligned token scheduling that balances workloads, reduces synchronization stalls, and preserves full parameter utilization. Across 135M-1B models, FlowHN achieves up to 4x higher throughput and 15% higher MFU than strong Transformer, SSM, and hybrid baselines while maintaining competitive accuracy on reasoning, coding, and long-context tasks up to 32K tokens. FlowHN is designed to integrate directly into existing Hybrid pipelines without changes to optimizers, training stacks, or inference serving infrastructure, making it practical for real-world deployment.

[Demo] Formally Specifying the Intended Behavior of the Program: LLM-Driven Neuro-Symbolic Program Specification Synthesis ▶ *Underline Video*

Cheng Wen, Hu Junjie, YiKun Hu, Jie Su, Bin Yu, Dugang Liu, Zhiwu Xu, Sun Weidi, Shengchao Qin and Cong Tian

Formal verification can provide strong mathematical guarantees about software correctness, but it typically requires developers to write detailed formal specifications (, contracts and loop invariants), which is costly and error-prone. We introduce AcslAgent, an LLM-driven neuro-symbolic demonstration system that reframes specification writing as constrained structured synthesis: large language models generate candidate specifications at the granularity of proof-relevant program components, while a symbolic verifier acts as a deterministic critic that checks legality, satisfiability, and proof adequacy, rejecting or repairing candidates in an iterative loop. This design turns unconstrained text generation into constrained structured synthesis, substantially reducing hallucinations and producing proof-ready annotations. We evaluate AcslAgent on seven benchmark suites, showing strong effectiveness. We release an open-source, Dockerized system with ensemble LLM backends and inter-modular verification support for reproducible demonstration and deployment. AcslAgent are available at: https://anonymous.4open.science/r/ACSLAgent-FB4C, the full experiment result can be accessed at: https://huggingface.co/spaces/saki222/experient_result_full, and the demonstration video is available at: https://youtu.be/rwA2kyzljac.

[Long] From pi to pi: Taming Divergence in Soft Clipping via Bilateral Decoupled Decay of Probability Gradient Weight ▶ *Underline Video*

Xiaoliang Fu, Jiaye Lin, Yangyi Fang, Chaowen Hu, Cong Qin, Zekai Shao, Binbin Zheng, Lu Pan and Ke Zeng

Reinforcement Learning with Verifiable Rewards (RLVR) has catalyzed a leap in Large Language Model (LLM) reasoning, yet its optimization dynamics remain fragile. Standard algorithms like GRPO enforce stability via "hard clipping", which inadvertently stifles exploration by discarding gradients of tokens outside the trust region. While recent "soft clipping" methods attempt to recover these gradients, they suffer from a critical challenge: relying on log-probability gradient (_theta _theta) yields divergent weights as probabilities vanish, destabilizing LLM training. We rethink this convention by establishing probability gradient (_theta _theta) as the superior optimization primitive. Accordingly, we propose Decoupled Gradient Policy Optimization (DGPO), which employs a decoupled decay mechanism based on importance sampling ratios. By applying asymmetric, continuous decay to boundary tokens, DGPO resolves the conflict between stability and sustained exploration. Extensive experiments across DeepSeek-R1-Distill-Qwen series models (1.5B/7B/14B) demonstrate that DGPO consistently outperforms strong baselines on various mathematical benchmarks, offering a robust and scalable solution for RLVR. Our code and implementation are available at: https://github.com/FlyTune/DGPO-RL.

[Industry] From 2:4 to 8:16 sparcity patterns in LLMs for Outliers and Weights with Variance Correction ▶ *Underline Video*

Egor Maximov, Yulia Kuzkina, Egor Shvetsov, Azamat Kanametov, Aleksandr Prutko, Maxim Zhelnin and Aleksei Goncharov

As large language models (LLMs) grow in size, efficient compression techniques like quantization and sparsification are critical. While quantization maintains performance with reduced precision, structured sparsity methods, such as N:M sparsification, often fall short due to limited flexibility and sensitivity to outlier weights. We explore 8:16 semi-structured sparsity, demonstrating its ability to surpass the Performance Threshold—where a compressed model matches the accuracy of its uncompressed or smaller counterpart under equivalent memory constraints. Compared to 2:4 sparsity, 8:16 offers greater flexibility with minimal storage overhead (0.875 vs. 0.75 bits/element). We also apply sparse structured patterns for salient weights, showing that structured sparsity for outliers is competitive with unstructured approaches, leading to equivalent or better results. Finally, we demonstrate that simple techniques such as variance correction and SmoothQuant-like weight equalization improve sparse models performance.

[Long] From ID to LLM: Rethinking Representation Learning for Recommendation ▶ *Underline Video*

Song-Li Wu, Zhaocheng Du, Weinan Gan, Jingyi Wang and Xianquan Wang

Recent studies indicate a fundamental incompatibility between ID representations and language model (LM) representations, as they capture behavioral and semantic spaces respectively. This mismatch leads LM representations to consistently underperform ID representations in recommendation tasks. In this work, we revisit this problem and show, from an information-theoretic perspective, that LLM representations retain all discriminative information in ID representations. Based on this, we introduce a Profile-then-Embedding (PtE) framework for recommendation, consisting of a Profile Stage, in which semantic user and item profiles are generated jointly through LLM-based bidirectional reasoning over user-item interactions, and a Personalized Embedding Stage, which encodes these profiles into task-aligned recommendation embeddings. We demonstrate PtEs effectiveness across three benchmark datasets, including cold-start and long-tail scenarios, achieving substantial gains in both discriminative and generative recommendation models.

[Long] From Synthesis to Clinical Assistance: A Strategy-Aware Agent Framework for Autism Intervention based on Real Clinical Dataset ▶ *Underline Video*

Junhong Lai, Shuzhong Lai, Yanhao Yu, Wanlin Chen, Chenyu Yan, Haifeng Li, Lin Yao and Yueming Wang

The development of AI-assisted Early Intensive Behavioral Intervention (EIBI) for Autism Spectrum Disorder (ASD) is severely constrained by data scarcity. Furthermore, while Applied Behavior Analysis (ABA) serves as the gold standard for clinical intervention, general-purpose Large Language Models (LLMs) struggle to strictly adhere to its standardized procedures, often resulting in interactions that are linguistically fluent but strategically inconsistent. To address these challenges, we introduce ASDAgent, a strategy-aware framework designed to unify high-fidelity intervention dialogue synthesis and clinical decision support. ASDAgent incorporates two specialized components to solve distinct problems: (i) a DoctorAgent equipped with an Observe-Think-Act-Correct (O-T-A-C) reasoning loop, which resolves the issue of strategy collapse in LLMs by making ABA execution explicit and controllable; and (ii) a ChildAgent that utilizes probabilistic behavior modeling to mitigate data homogeneity, simulating diverse and non-deterministic ASD response patterns. Experiments demonstrate that dialogues generated by ASDAgent closely mirror the strategy distribution of human therapists (KL divergence: 0.083). In real autism intervention, ASDAgent achieves nearly 80% strategic consistency with human experts. Moreover, we show that synthetic data produced by ASDAgent effectively distills professional clinical knowledge into small language models (SLMs), significantly enhancing their therapeutic capabilities.

[Long] Frozen LLMs are Native Decoders for High-Norm Semantic Vectors ▶ *Underline Video*

Yunsheng Zeng and Yongmei Tan

Large language models (LLMs) are designed for discrete tokens, yet they operate in a continuous embedding space. Recent context compression methods exploit this property by encoding text into dense vectors for frozen LLM decoding. However, a key question remains unanswered: how does a frozen LLM interpret continuous vectors that encode complex semantics? We investigate this through controlled reconstruction experiments. Our analysis reveals a critical geometric property: compression encoders learn to produce vectors with L2 norms two orders of magnitude higher than standard embeddings. We show that this high-norm signal is causally necessary for the frozen LLM to decode compressed information. Based on this finding, we propose a landmark-based compression framework for long contexts. Our encoder uses bidirectional attention over landmark tokens. This design captures global dependencies and avoids semantic fragmentation from segment-based methods. Experiments on text reconstruction and four QA benchmarks validate our approach. At 4x and 16x compression ratios, our method outperforms prior soft compression baselines.

[Long] Generating Attribution Reports for Manipulated Facial Images: A Dataset and Baseline ▶ *Underline Video*

Jingchun Lian, Lingyu Liu, Yaxiong Wang, Yujiao Wu, Lianwei Wu, Li Zhu and Zhedong Zheng

Existing facial forgery detection methods typically focus on binary classification or pixel-level localization, providing little semantic insight into the nature of the manipulation. To address this, we introduce Forgery Attribution Report Generation, a new multimodal task designed to provide post-hoc forensic evidence for manipulated images. This task jointly localizes forged regions ("Where") and generates natural language explanations grounded in the editing process ("Why"). This dual-focus approach goes beyond traditional binary forensics, providing a comprehensive, interpretable understanding of the manipulation. To enable research in this domain, we present Multi-Modal Tamper Tracing (MMTT), a large-scale dataset of 152,217 samples. Each sample features a process-derived ground-truth mask and a human-authored textual description, ensuring high annotation precision and linguistic richness. We further propose ForgeryTalker, a unified

end-to-end baseline that integrates vision and language via a shared encoder and dual decoders for mask and text generation. Experiments show that ForgeryTalker achieves competitive performance on both subtasks, i.e., 59.3 CIDEr and 73.67 IoU, establishing a strong baseline for explainable multimedia forensics. Our dataset and code are available at: https://github.com/NattyLianJc/Generating-Attribution-Reports.

[Findings] Generating Effective CoT Traces for Mitigating Causal Hallucination ▶ *Underline Video*

Yiheng Zhao and Jun Yan

Although large language models (LLMs) excel in complex reasoning tasks, they suffer from severe causal hallucination in event causality identification (ECI), particularly in smaller models (<=1.5B parameters). A promising approach to address this issue is to fine-tune them with Chain-of-Thought (CoT) traces. However, there is currently a lack of CoT trace dataset available for ECI. In this paper, we first investigate the essential criteria that effective CoT traces should possess to mitigate causal hallucination in smaller models. We then design a pipeline to generate CoT traces that meet these criteria. Moreover, since there is currently no metric for quantifying causal hallucination, we also introduce a new metric, the Causal Hallucination Rate (CHR), to quantify causal hallucination, guide the formulation of effective CoT trace criteria, and validate the effectiveness of our pipeline. Our experiments show that fine-tuning with the CoT traces generated by our pipeline not only substantially reduces causal hallucination in smaller LLMs but also improves mean accuracy. Moreover, the fine-tuned models exhibit strong cross-dataset and cross-difficulty generalization, as well as robustness under intentionally misleading intervention prompts.

[Long] GMFL: Efficient Global Masking for Federated LLM Fine-tuning ▶ *Underline Video*

Xin Huang, Yan Hu, Yuejiao Gong and Xinglin Zhang

Low-Rank Adaptation (LoRA) has emerged as a prominent solution to mitigate the communication and computation costs in federated fine-tuning of Large Language Models (LLMs). However, we observe that even within low-rank adapters, a substantial portion of parameters manifest negligible updates during federated training, leading to redundant communication and wasted local computation. To address this, we propose GMFL, a plug-and-play layer freezing mechanism designed to seamlessly integrate with existing federated fine-tuning frameworks. Specifically, the server monitors the global update magnitude of each LoRA layer to dynamically generate freezing masks. These masks are updated periodically with a fixed freezing rate, ensuring stable convergence by robustly identifying "saturated" layers. Theoretical analysis confirms the convergence of GMFL, where the freezing mechanism yields a bounded error that scales with client heterogeneity. Extensive experiments across multiple tasks (GLUE, Commonsense Reasoning, Math Reasoning and General Generation) demonstrate that GMFL reduces communication overhead and lowers computational costs while preserving the performance of the underlying federated fine-tuning methods. Our work provides a practical, versatile solution for deploying large-scale federated LLM fine-tuning in resource-constrained environments. Our code is available at: https://github.com/tunx-cyber/GMFL.

[Demo] GovScape: A Public Multimodal Search System for 70 Million Pages of Government PDFs ▶ *Underline Video*

Ying-Hsiang Huang, Claire Gong, Shreya Shaji, Alison Yan, Leslie Harka, Albert Du, Anjali Gopal, Samuel Klein, Shannon Shen, Mark Phillips, Trevor Owens, Kyle Deeds and Benjamin Lee

Efforts over the past three decades have produced web archives containing billions of webpage snapshots and petabytes of data. The End of Term Web Archive alone contains millions of PDFs produced by the federal government. While preservation with web archives has been successful, significant challenges for access and discoverability remain. In this paper, we introduce GovScape, a public search system that supports multimodal searches across 10,015,993 federal government PDFs from the 2020 End of Term crawl (70,958,487 total PDF pages) – to our knowledge, all renderable PDFs in the 2020 crawl that are 50 pages or under. GovScape supports four primary forms of search: in addition to providing (1) filter conditions over metadata facets including domain and crawl date and (2) exact text search against the PDF text, we provide (3) semantic text search and (4) visual search against the PDFs across individual pages, enabling users to structure queries such as redacted documents" or pie charts." We detail GovScape's search affordances, embedding pipeline, system architecture, and open source codebase. Significantly, the total estimated compute cost for GovScape's pre-processing pipeline for 10 million PDFs was approximately $1,500, equivalent to 47,000 PDF pages per dollar spent on compute, demonstrating the potential for immediate scalability. We evaluate GovScape by (1) analyzing 1,679 search queries and (2) benchmarking vector and keyword index efficiency using these queries. GovScape can be found at https://www.govscape.net.

[Findings] Hail to the Thief: Exploring Attacks and Defenses in Decentralised GRPO ▶ *Underline Video*

Nikolay Blagoev, Oguzhan Ersoy and Lydia Chen

Group Relative Policy Optimization (GRPO) has demonstrated wide adoption in the post-training of Large Language Models (LLMs). In GRPO, prompts are answered by the model and preferred behaviour is learnt via reinforcement learning. Owing to the small communication volume, GRPO is inherently suitable for decentralised training as the prompts can be concurrently answered by multiple nodes and these completions are exchanged in the form of strings. In this work, we explore the robustness of decentralised GRPO by presenting the first adversarial attacks and countermeasures. We present a diverse set of attacks where malicious nodes poison benign models by sharing their poisoned completions. We demonstrate these attacks on math and coding tasks and show that an adversary can achieve attack success rates of up to (100%) in as few as 50 iterations. Moreover, to mitigate the attacks, we propose two defense mechanisms that check logit probabilities of completions or utilize an LLM judge to filter completions. The defenses prevent all but the DoS attack that causes unnecessarily lengthy but conceptually correct completions. The code of both attacks and defenses can be found at: https://github.com/gensyn-ai/HTTT.

[Long] HalluAudio: A Comprehensive Benchmark for Hallucination Detection in Large Audio-Language Models ▶ *Underline Video*

Feiyu Zhao, Yiming Chen, Wenhuan Lu, Daipeng Zhang, Xianghu Yue and Jianguo Wei

Large Audio-Language Models (LALMs) have recently achieved strong performance across various audio-centric tasks. However, hallucination, where models generate responses that are semantically incorrect or acoustically unsupported, remains largely underexplored in the audio domain. Existing hallucination benchmarks mainly focus on text or vision, while the few audio-oriented studies are limited in scale, modality coverage, and diagnostic depth. We therefore introduce HalluAudio, the first large-scale benchmark for evaluating hallucinations across speech, environmental sound, and music. HalluAudio comprises over 5K human-verified QA pairs and spans diverse task types, including binary judgments, multi-choice reasoning, attribute verification, and open-ended QA. To systematically induce hallucinations, we design adversarial prompts and mixed-audio conditions. Beyond accuracy, our evaluation protocol measures hallucination rate, yes/no bias, error-type analysis, and refusal rate, enabling a fine-grained analysis of LALM failure modes. We benchmark a broad range of open-source and proprietary models, providing the first large-scale comparison across speech, sound, and music. Our results reveal significant deficiencies in acoustic grounding, temporal reasoning, and music attribute understanding, underscoring the need for reliable and robust LALMs.

[Long] HeteroSpec: Leveraging Contextual Heterogeneity for Efficient Speculative Decoding ▶ *Underline Video*

Siran Liu, Yang Ye, Qianchao Zhu, Zane Cao and Yongchao He

Autoregressive decoding inherently limits the inference throughput of Large Language Model (LLM) due to its sequential dependency. Speculative decoding mitigates this by verifying multiple predicted tokens in parallel, but its efficiency remains constrained by what we identify as verification heterogeneity, the uneven difficulty of verifying different speculative candidates. In practice, a small subset of high-confidence predictions accounts for most successful verifications, yet existing methods treat all candidates uniformly, leading to redundant computation. We present HeteroSpec, a heterogeneity-adaptive speculative decoding framework that allocates verification effort in proportion to candidate uncertainty. HeteroSpec estimates verification complexity using a lightweight entropy-based quantifier, partitions candidates via a data-driven stratification policy, and dynamically tunes speculative depth and pruning thresholds through coordinated optimization. Across five benchmarks and four LLMs, HeteroSpec delivers an average 4.24x decoding speedup over state-of-the-art methods such as EAGLE-3, while preserving exact output distributions. Crucially, HeteroSpec requires no model retraining and remains compatible with other inference optimizations, making it a practical direction for improving speculative decoding efficiency.

[Long] HiGoE: Hierarchical Graph of Evidence to Enhance Retrieval-Augmented Generation for Long-context Summarization ▶ *Underline Video*

Long Yuan, Kaiwen Tian, Zi Chen, Bolong Zheng and Chuan Ma

Long-context summarization is pivotal for extracting core insights from extensive documents. While Large Language Models (LLMs) show remarkable capabilities, they frequently encounter attention dilution and hallucination with lengthy inputs. Retrieval-Augmented Generation (RAG) partially mitigates this, but conventional RAG relies on shallow similarity retrieval of fragmented chunks, failing to capture high-level thematic structures and long-range dependencies. Although graph-based RAG approaches have emerged to address these structural limitations, existing solutions, such as Graph of Records (GoR), critically suffer from a fundamental flaw: they paradoxically re-introduce hallucinations by constructing graphs based on unreliable, LLM-generated responses. To overcome these challenges, we introduce Hierarchical Graph of Evidence (HiGoE) (Code link https://github.com/tkw123/HiGOE). HiGoE redefines the retrieval process by replacing unreliable chunk-based methods with a filtered proposition-evidence graph, ensuring verifiable fact grounding and substantially reducing hallucination. Moreover, HiGoE leverages Personalized PageRank (PPR) to cluster related nodes into thematic hierarchies, thereby restoring global document structure and effectively mitigating attention dilution. To model complex, multi-level relations beyond mere shallow similarity, we develop an Enhanced Graph Attention Network. Experiments show HiGoE consistently surpasses baselines in quality and efficiency.

[Long] HSGraphAgent: Knowledge-Graph-Guided Large Language Models for Harmonized System Code Classification ▶ *Underline Video*

Qiang Xia, Zijian Zhang, Ao Wang, Wenhan Wang, Xiangyu Wang and Jian Li

Harmonized System (HS) code classification is a hierarchically structured and regulation-constrained task, often complicated by short and noisy product descriptions. Misclassification can lead to tariff misapplication, regulatory violations, or delayed customs clearance, which in turn requires predictions to be both semantically appropriate and hierarchically valid. While large language models (LLMs) show strong semantic understanding, their unconstrained generation is poorly aligned with these requirements, often producing non-existent or hierarchically inconsistent codes. We propose HSGraphAgent a knowledge-graph-guided LLM framework that formulates HS classification as a stepwise, regulation-aware reasoning process over an explicit HS knowledge graph. By encoding hierarchical containment relations and regulatory exclusion rules, and enforcing them through a Select-Redirect mechanism, HSGraphAgent constrains inference to legally valid paths while producing explicit and traceable reasoning trajectories. Experiments on taxonomy-wide 4-digit and fine-grained 6-digit HS benchmarks demonstrate consistent improvements over direct generation and retrieval-augmented baselines, with particularly strong gains in fine-grained and regulation-sensitive classification settings.

[Long] HyperMem: Hypergraph Memory for Long-Term Conversations ▶ *Underline Video*

Juwei Yue, Chuanrui Hu, Jiawei Sheng, Zuyi Zhou, Wenyuan Zhang, Tingwen Liu, Li Guo and Yafeng Deng

Long-term memory is essential for conversational agents to maintain coherence, track persistent tasks, and provide personalized interactions across extended dialogues. However, existing approaches as Retrieval-Augmented Generation (RAG) and graph-based memory mostly rely on pairwise relations, which can hardly capture high-order associations, i.e., joint dependencies among multiple elements, causing

fragmented retrieval. To this end, we propose HyperMem, a hypergraph-based hierarchical memory architecture that explicitly models such associations using hyperedges. Particularly, HyperMem structures memory into three levels: topics, episodes, and facts, and groups related episodes and their facts via hyperedges, unifying scattered content into coherent units. Leveraging this structure, we design a hybrid lexical-semantic index and a coarse-to-fine retrieval strategy, supporting accurate and efficient retrieval of high-order associations. Experiments on the LoCoMo benchmark show that HyperMem achieves state-of-the-art performance with 92.73% LLM-as-a-judge accuracy, demonstrating the effectiveness of HyperMem for long-term conversations.

[Long] ImpRIF: Stronger Implicit Reasoning Leads to Better Complex Instruction Following ► *Underline Video*

Yuancheng Yang, Lin Yang, Xu Wang, Chao Tong and Haihua Yang

As applications of large language models (LLMs) become increasingly complex, the demand for robust complex instruction following capabilities is growing accordingly. We argue that a thorough understanding of the instruction itself, especially the latent reasoning structure embedded between the lines, is crucial for improving instruction following. Therefore we target complex instructions that involve implicit reasoning, intricate logical relations, and multi-constraint dependencies. We propose ImpRIF, a method to enhance LLMs' understanding of implicit reasoning instructions, thereby improving its ability to follow complex instructions. We formalize such instructions as verifiable reasoning graphs, enabling programmatic verification and graph-driven chain-of-thought reasoning. Based on this formulation, we synthesize large-scale single- and multi-turn data, propose fine-tuning with graph reasoning, and apply reinforcement learning to explicitly train models to reason along the graph. On five complex instruction following benchmarks, our models substantially outperform their base models. These results demonstrate that enhancing implicit reasoning capabilities can significantly improve complex instruction following. This project will be open-sourced in the near future.

[Industry] Improving Hate Speech Detection by Fusing Textual and User Interaction Representations in Online Communities ► *Underline Video*

Xu Gao, Dong Jing and Kee-hung Lai

Detecting hate speech in online communities is increasingly challenging due to the implicit and context-dependent nature of toxic expressions. While text-only models often struggle with such ambiguity, incorporating user interaction signals offers critical pragmatic context for disambiguation. However, research in this direction is hindered by the scarcity of datasets that align textual content with comprehensive user behavioral graphs. To bridge this gap, we present a new dataset collected from a real-world community, featuring labeled hate speech enriched with fine-grained interaction histories. We further propose a novel user-aware hate speech detection framework that effectively fuses textual semantics with social interaction representations. Experiments demonstrate that our approach consistently outperforms strong text-only baselines by over 3.6%, validating the critical role of social context in enhancing detection accuracy. Furthermore, to mitigate real-world adversarial risks such as graph spoofing and spam, we introduce a contrastive graph augmentation strategy, ensuring model robustness against unreliable community behaviors.

[Long] Inverting the Shield: Systematically Generating Safety Tests from Policy Specifications ► *Underline Video*

Xiaoyue Lu, Xianglin Yang, Haijun Liu, Jiahao Liu, Kuntai Cai, Yan Xiao and Jin Song Dong

The widespread integration of Large Language Models (LLMs) necessitates rigorous and systematic safety evaluation. Existing paradigms either rely on constructed benchmarks to assess safety from predefined perspectives, or employ dynamic red-teaming to probe potential vulnerabilities. While effective, these approaches face challenges, as they depend heavily on expert domain knowledge, offer limited systematic guarantees, and are vulnerable to rapid obsolescence. To address these limitations, we introduce a novel framework POLARIS that brings the rigor of specification-based software testing to AI safety. POLARIS first compiles unstructured natural-language policies into First-Order Logic (FOL) representations, establishing a traceable link between high-level rules and concrete test cases. This formalization enables the construction of a Semantic Policy Graph, where complex policy violation scenarios are encoded as traversable paths. By systematically exploring this graph, POLARIS uncovers compositional violation patterns, which are then instantiated into executable natural-language test queries, enabling coverage-driven and reproducible safety testing. Experiments demonstrate that POLARIS achieves higher policy coverage and attack success counts compared to established baselines. Crucially, by bridging formal methods and AI safety, POLARIS provides a principled, automated approach to ensuring LLMs adhere to safety-critical policies with verifiable traceability.

[Findings] Is Grokking Worthwhile? Functional Analysis and Transferability of Generalization Circuits in Transformers ► *Underline Video*

Kaiyu He, Mian Zhang, Peilin Wu, Xinya Du and Zhiyu Chen

While Large Language Models (LLMs) excel at factual retrieval, they often struggle with the "curse of two-hop reasoning" in compositional tasks. Recent research suggests that parameter-sharing transformers can bridge this gap by forming a "Generalization Circuit" during a prolonged "grokking" phase. A fundamental question arises: Is a grokked model truly superior to its non-grokked counterparts? Furthermore, is the extensive computational cost of waiting for the grokking phase worthwhile? In this work, we conduct a mechanistic study to evaluate the Generalization Circuit's role in knowledge assimilation and transfer. We demonstrate that: (i) The inference paths established by non-grokked and grokked models for in-distribution compositional queries are identical. This suggests that the "Generalization Circuit" does not represent the sudden acquisition of a new reasoning paradigm. Instead, we argue that grokking is the process of integrating memorized atomic facts into an easy-acquire, naturally established reasoning path. (ii) Achieving high accuracy on unseen cases after prolonged training and the formation of a certain reasoning path are not bound; they can occur independently under specific data regimes. (iii) Even a mature circuit exhibits limited

transferability when integrating new knowledge, suggesting that "grokked" Transformers do not achieve a full mastery of compositional logic.

[Long] IUQ: Interrogative Uncertainty Quantification for Long-Form Large Language Model Generation ▶ *Underline Video*
Haozhi Fan, Jinhao Duan and Kaidi Xu
Despite the rapid advancement of Large Language Models (LLMs), uncertainty quantification in LLM generation is a persistent challenge. Although recent approaches have achieved strong performance by restricting LLMs to produce short or constrained answer sets, many real-world applications require long-form and free-form text generation. A key difficulty in this setting is that LLMs often produce responses that are semantically coherent yet factually inaccurate, while the underlying semantics are multifaceted and the linguistic structure is complex. To tackle this challenge, this paper introduces Interrogative Uncertainty Quantification (IUQ), a novel framework that leverages inter-sample consistency and intra-sample faithfulness to quantify the uncertainty in long-form LLM outputs. By utilizing an interrogate-then-respond paradigm, our method provides reliable measures of claim-level uncertainty and the model's faithfulness. Experimental results across diverse model families and model sizes demonstrate the superior performance of IUQ over two widely used long-form generation datasets.

[Findings] JointCQ: Improving Factual Hallucination Detection with Joint Claim and Query Generation ▶ *Underline Video*
Fan Xu, Huixuan Zhang, Zhenliang Zhang, Jiahao Wang and Xiaojun Wan
Current large language models (LLMs) often suffer from hallucination issues, i,e, generating content that appears factual but is actually unreliable. A typical hallucination detection pipeline involves response decomposition (i.e., claim extraction), query generation, evidence collection (i.e., search or retrieval), and claim verification. However, existing methods exhibit limitations in the first two stages, such as context loss during claim extraction and low specificity in query generation, resulting in degraded performance across the hallucination detection pipeline. In this work, we introduce JointCQ, a joint claim-and-query generation framework designed to construct an effective and efficient claim-query generator. Our framework leverages elaborately designed evaluation criteria to filter synthesized training data, and finetunes a language model for joint claim extraction and query generation, providing reliable and informative inputs for downstream search and verification. Experimental results demonstrate that our method outperforms previous methods on multiple open-domain QA hallucination detection benchmarks, advancing the goal of more trustworthy and transparent language model systems.

[Long] Know the Known and the Unknown: Reasonable Answer Generation with Knowledge-Informed Citations ▶ *Underline Video*
Yichi Zhang, Zhuo Chen, Lingbing Guo, Jun Xu, Mengshu Sun, Zhizhen Liu, Lei Liang, Wen Zhang and Huajun Chen
Question answering (QA) with reference texts is a classic application scenario for large language models (LLMs), where high standards for the credibility and traceability of generated answers are crucial. Many existing approaches focus on generating multi-level citations linked to specific references within the answer, making it verifiable and trustworthy. However, they often overlook key challenges such as citation granularity, the awareness of unknown information, and the adoption of effective training strategies. In this paper, we introduce Knowledge-informed Citation (KFC), which addresses these issues through a novel data construction pipeline, a new benchmark, and an innovative training strategy. With approximately 42K samples spanning 19 distinct domains, KFC includes both traditional citations referencing known entity-level information and specialized citations referring to unknown knowledge in the given question. This structure provides a more granular approach to citations, guiding the model to recognize and explicitly indicate unknown information, thus enhancing the quality and credibility of the response. Additionally, we propose a self-correction paradigm, Self-KFC, designed to fine-tune LLMs by refining poorly cited answers into more accurate ones, making it particularly suitable for citation-dependent scenarios. We present comprehensive experimental results to demonstrate the effectiveness and generalization of Self-KFC on the KFC benchmark.

[Long] KnowRL: Exploring Knowledgeable Reinforcement Learning for Factuality ▶ *Underline Video*
Baochang Ren, Shuofei Qiao, Ningyu Zhang, Da Zheng and Huajun Chen
Slow-thinking Large Language Models (LLMs) have demonstrated strong reasoning capabilities but often suffer from severe hallucinations due to an inability to recognize their knowledge boundaries. Existing Reinforcement Learning (RL) approaches typically rely on outcome-oriented rewards, which can inadvertently reinforce fabricated reasoning paths when the final answer is correct. To address this, we propose Knowledge-enhanced RL, KnowRL, a framework that integrates factual supervision directly into the reasoning process. By decomposing the chain of thought into atomic facts and verifying them against the corresponding ground-truth knowledge, KnowRL performs fine-grained checks to encourage models to reason faithfully. Crucially, this process-oriented supervision teaches the model to identify its knowledge boundaries, learning to say "I don't know" instead of fabricating answers when information is missing. Experimental results demonstrate that KnowRL effectively mitigates hallucinations—reducing the Incorrect Rate on SimpleQA by 20.3% for distillation-based slow-thinking models while maintaining strong performance on complex reasoning benchmarks like GPQA and AIME 2025. Furthermore, our method shows robust transferability to out-of-distribution tasks, indicating that the model learns a generalizable verification behavior.

[Long] LAMCL: A Length-aware Momentum Contrastive Learning Framework for Multiscale Machine-Revised Text Detection ▶ *Underline Video*
Bing Zhou, Zhe Huang, Shilei Tan, Kai Zhao and Yongcheng Zhou
Detecting machine-revised text that exhibits subtle lexical differences from the original human-generated text remains a challenge. Recent detection methods, including watermarking-based, logit-based, and training-based models, struggle to capture the fine-grained semantic differences, especially for short texts. To address this issue, we propose Length-aware Momentum Contrastive Learning (LAMCL), a novel framework for multiscale machine-revised text detection that integrates two core modules. To enhance the discriminative semantic features,

the Enhance Before Detection (EBD) module first fuses the original detected text with the counterpart processed by a Large Language Model (LLM), and then measures semantic consistency to distinguish between machine-revised and human-generated text. Meanwhile, based on the Momentum Contrastive Learning (MCL) framework, the Length-aware Weighting (LW) module leverages text length and label information for hard negative sampling, mitigating the ambiguity of short text attribution and boosting the robustness of representation learning. Experimental results demonstrate that our method outperforms the existing detectors in identifying multiscale machine-revised text across diverse practical scenarios, tasks, and LLMs. The code is available at https://github.com/hangtze/LAMCL.

[Long] LANG: Reinforcement Learning for Multilingual Reasoning with Language-Adaptive Hint Guidance ▶ *Underline Video*

Yuchun Fan, Bei Li, PeiguangLi PeiguangLi, Yilin Wang, Yongyu Mu, Jian Yang, Xin Chen, Rongxiang Weng, Jingang Wang, Xunliang Cai, JingBo Zhu and Tong Xiao

Reinforcement learning has proven effective for enhancing multi-step reasoning in Large Language Models (LLMs), yet its benefits have not fully translated to multilingual contexts. Existing methods struggle with a fundamental trade-off: prioritizing input-language consistency severely hampers reasoning quality, while prioritizing reasoning often leads to unintended language drift toward English. We address this challenge with LANG, a novel framework that leverages language-conditioned hints to guide exploration in non-English reasoning tasks. Our method incorporates two key mechanisms to prevent dependency on these hints: a progressive decay schedule that gradually withdraws scaffolding, and a language-adaptive switch that tailors learning horizons to specific language difficulties. Empirical results on challenging multilingual mathematical benchmarks reveal that LANG substantially enhances reasoning performance without compromising language consistency. Moreover, we show that our framework generalizes beyond mathematics, fostering more consistent language alignment across model layers.

[Long] Latent-Condensed Transformer for Efficient Long Context Modeling ▶ *Underline Video*

Zeng You, Yaofo Chen, Qiuwu Chen, Ying Sun, Shuhai Zhang, Yingjian Li, Yaowei Wang and Mingkui Tan

Large language models (LLMs) face significant challenges in processing long contexts due to the linear growth of the key-value (KV) cache and quadratic complexity of self-attention. Existing approaches address these bottlenecks separately: Multi-head Latent Attention (MLA) reduces the KV cache by projecting tokens into a low-dimensional latent space, while sparse attention reduces computation. However, sparse methods cannot operate natively on MLAs compressed latent structure, missing opportunities for joint optimization. In this paper, we propose Latent-Condensed Attention (LCA), which directly condenses context within MLAs latent space, where the representation is disentangled into semantic latent vectors and positional keys. LCA separately aggregates semantic vectors via query-aware pooling and preserves positional keys via anchor selection. This approach jointly reduces both computational cost and KV cache without adding parameters. Beyond MLA, LCAs design is architecture-agnostic and readily extends to other attention mechanisms such as GQA. Theoretically, we prove a length-independent error bound. Experiments show LCA achieves up to 2.5x prefilling speedup and 90% KV cache reduction at 128K context while maintaining competitive performance.

[Long] Learn to Relax with Large Language Models: Solving Constraint Optimization Problems via Bidirectional Coevolution ▶ *Underline Video*

Beidan Liu, Zhengqiu Zhu, Chen Gao, Tianle Pu, Yong Zhao, Wei Qi and Quanjun Yin

Large Language Model (LLM)-based optimization has recently shown promise for autonomous problem solving, yet most approaches still cast LLMs as passive constraint checkers rather than proactive strategy designers, limiting their effectiveness on complex Constraint Optimization Problems (COPs). To address this, we present AutoCO, an end-to-end Automated Constraint Optimization method that tightly couples operations-research principles of constraint relaxation with LLM reasoning. A core innovation is a unified triple-representation that binds relaxation strategies, algorithmic principles, and executable codes. This design enables the LLM to synthesize, justify, and instantiate relaxation strategies that are both principled and executable. To navigate fragmented solution spaces, AutoCO employs a bidirectional global-local coevolution mechanism, synergistically coupling Monte Carlo Tree Search (MCTS) for global relaxation-trajectory exploration with Evolutionary Algorithms (EAs) for local solution intensification. This continuous exchange of priors and feedback explicitly balances diversification and intensification, thus preventing premature convergence. Extensive experiments on three challenging COP benchmarks validate AutoCO's consistent effectiveness and superior performance, especially in hard regimes where current methods degrade. Results highlight AutoCO as a principled and effective path toward proactive, verifiable LLM-driven optimization.

[Long] Learning from Cognition: Enhancing RL Efficiency for LLM Reasoning via Hierarchical Metacognitive Decomposition and Refinement ▶ *Underline Video*

Zexu Sun, Yongcheng Zeng, Erxue Min, Heyang Gao, Bokai Ji, Dugang Liu, Xing Tang, Xiuqiang He and Xu Chen

Contemporary progress in Large Language Models (LLMs) has revealed notable inferential capacities via reinforcement learning (RL) employing verifiable rewards. However, "zero-RL" approaches relying on fixed prompt templates introduce substantial sampling inefficiencies for weak LLMs, as most problems generate invalid outputs during accuracy-driven filtration. To solve this, we propose Cog-Rethinker, a novel hierarchical metacognitive RL framework. Cog-Rethinker enhances the rollout procedure by improving sample utilization through a two-stage framework leveraging human cognition. First, it prompts the policy to decompose zero-accuracy problems into subproblems. Second, it prompts the policy to refine answers by referencing previous wrong solutions. Moreover, to enable cold-starts and maintain train-test consistency, Cog-Rethinker applies supervised fine-tuning using correct samples from these stages. Experimental results demonstrate Cog-Rethinker's superior performance on mathematical reasoning benchmarks and its improved sample efficiency that

accelerates convergence compared to baselines.

[Long] Learning from Evolving Training Dynamics: An Entropy-Maximizing Data Curation Strategy for LLM Supervised Post-Training ▶ *Underline Video*

Mengxiang Zhang and Lingyuan Liu

Supervised post-training is essential for refining Large Language Models (LLMs), yet its effectiveness relies heavily on strategic data curation. Traditional Curriculum Learning (CL) strategies often fail to account for the evolving proficiency of the learner, relying instead on static, single dimensional metrics. We propose EVO-Curate, a dynamic data curation framework that synchronizes sample complexity with the maturing capacity of the LLM. EVO-Curate employs an Adaptive Dynamics Measurer to synthesize instantaneous difficulty and historical variability into a multidimensional utility score. To maintain representational diversity, we introduce an Evolutionary Sampling Scheduler based on an entropy maximizing mechanism. Empirical evaluations across instruction following, mathematical reasoning, and code generation demonstrate that EVO-Curate consistently outperforms standard training baselines and traditional CL methods across various architectures and scales. Specifically, our framework achieves relative performance gains of up to about 10% while maintaining manageable computational overhead. These results establish EVO-Curate as a scalable and model agnostic solution for enhancing the efficiency of modern LLM training pipelines.

[Long] Learning What to Ignore: Mitigating Negative Transfer in Medical Knowledge Fusion via Clinical Task-Adaptive Selection ▶ *Underline Video*

Xinyan Deng, Shoubin Dong and Xiaorou Zheng

Integrating external medical knowledge into longitudinal electronic health record modeling is a prevailing paradigm to mitigate clinical data sparsity. However, existing approaches face a reliability-timeliness dilemma, struggling to balance the structural authority of static ontologies with the reasoning flexibility of large language models. Furthermore, most frameworks overlook the risk of relative negative transfer, where indiscriminately fusing task-irrelevant knowledge can introduce noise or even cause conflicts that weakens patient-specific signals. In this paper, we propose TrustKE, a Trustworthy Knowledge Enhancement framework. First, we construct a dual-layer knowledge graph that anchors dynamic, evidence-based chain-of-thought reasoning from medical literature within the stable structure of medical knowledge graph. Second, we introduce a task-adaptive knowledge selection mechanism that dynamically optimizes the graph, retaining only task-specific signals. Extensive experiments on MIMIC-III and MIMIC-IV across four clinical tasks show that TrustKE outperforms state-of-the-art baselines. Our analysis confirms that TrustKE effectively mitigates negative transfer while offering transparent reasoning for clinical decision-making.

[Long] Legal Judgment Prediction: A Reflection on the State of the Art ▶ *Underline Video*

Yi Feng, Chuanyi Li and Vincent Ng

Automatic legal judgment prediction (LJP) has recently received increasing attention in the natural language processing community because of its practical values in the real world. Significant progress has been achieved on LJP in the past decade. However, most existing LJP research primarily focuses on developing methods that achieve better performance on standard evaluation datasets, with limited emphasis on the long-term advancement of the field beyond improving evaluation metrics. In this position paper, we reflect on the state of the art in LJP research, and explore issues that should motivate researchers to think beyond merely enhancing performance metrics, with the ultimate goal of sparking discussions among LJP researchers about the future trajectory of the field.

[Long] LegalGraphRAG: Multi-Agent Graph Retrieval-Augmented Generation for Reliable Legal Reasoning ▶ *Underline Video*

Zerui Chen, Qinggang Zhang, Zhishang Xiang, Zhimin Wei, Linfeng Gao, Xiao Huang, Zhihong Zhang and Jinsong Su

Graph-based Retrieval-Augmented Generation (GraphRAG) advances flat document retrieval by structuring knowledge as relational graphs, enabling more coherent and effective reasoning. However, applying it to specific domains like legal reasoning faces critical challenges. (i) Legal corpora are heterogeneous, containing multi-granular knowledge from cases, articles and interpretations. A flat knowledge graph cannot adequately differentiate between factual details, applied rules, and abstract principles, limiting accurate retrieval. (ii) Reliable legal judgment demands transparent, evidence-based reasoning. Traditional RAG passes retrieved context directly to an LLM without verification, resulting in opaque, error-prone reasoning. To this end, we propose LegalGraphRAG, a framework designed for reliable legal reasoning. Our approach introduces two core components: a hierarchical legal graph that hierarchically organizes legal sources to enable retrieval at appropriate abstraction levels, and a multi-agent system for reliable legal reasoning, where a Researcher retrieves candidate evidence, an Auditor rigorously verifies its validity against source documents, and an Adjudicator synthesizes the set of verified evidence to render a final judgment. Extensive experiments show that LegalGraphRAG achieves the state-of-the-art performance, outperforming existing GraphRAG baselines in accurate and trustworthy legal analysis. Our code, datasets and implementation details are available at https://github.com/XMUDeepLIT/LegalGraphRAG.

[Long] Lifting Optimized Binaries to Canonical Compiler IR via Structure-Aware Retrieval and Iterative Verification ▶ *Underline Video*

Xiaoao Zhu, Jie Ren, Zhiqiang Li, Jie Zheng, Zhanyong Tang and Zheng Wang

Lifting stripped and highly optimized binaries to the canonical compiler intermediate representation (IR) enables program analysis when source code is unavailable. However, compiler optimizations severely distort control-flow and data-flow structure, making existing rule-based

and LLM-based decompilation approaches brittle. We present BRIDGE, a system that reliably lifts optimized binaries to analysis-friendly compiler IR. BRIDGE combines control-flow-aware retrieval-augmented generation with feedback-driven verification. It uses pseudo-probe instrumentation to align optimized binary fragments with normalized IR semantics, and then employs an iterative refinement loop guided by static analysis and runtime feedback to improve executability and semantic consistency. We evaluate BRIDGE on HumanEval-Decompile and MBPP, lifting x86-64 and ARM64 binaries to LLVM IR. BRIDGE outperforms seven baselines, achieving an average of over 30% higher re-executability than the strongest general-purpose LLM baseline.

[Long] Lingua-Graph: A Unified Representation of Cross-Task Common Substructures for Analytic Language Processing ► *Underline Video*

Mingming Sun, Runze Jiang, Zhu zhangchenxi, Minlong Peng and Yunfeng Cai

Structural understanding of natural language requires explicit recovery of internal meaning structures (entities, facts, nested relations), yet current structural-analytic tasks are fragmented by inconsistent task requirements across datasets. We investigate the problem of robust cross-task structural understanding under heterogeneous requirements across structural-analytic tasks and outline a perspective called Analytic NLP in which tasks can be reformulated into a representation-then-decision paradigm. In this paper, we suggest a solution for the representation layer, called Lingua-Graph, which explicitly captures entities, facts, and relations. By representing predictions as explicit graphs with labeled nodes and edges, Lingua-Graph also improves interpretability, enabling transparent inspection and error analysis of intermediate meaning structures. We construct a labeled Lingua-Graph dataset and train a baseline parser. Experiments show that Lingua-Graph provides substantially higher entity-structure hostability than alternative representations on average, and OpenIE systems based on Lingua-Graph achieve superior performance on three benchmarks, demonstrating that better intermediate structures translate into downstream gains. The data, code and the trained model are publicly released at https://github.com/rudaoshi/Lingua.

[Demo] LiTS: A Modular Framework for LLM Tree Search ► *Underline Video*

Xinzhe Li and Yaguang Tao;

LiTS is a modular Python framework for LLM reasoning via tree search. It decomposes tree search into three reusable components, Policy, Transition, and RewardModel, that plug into algorithms like MCTS and BFS. A decorator-based registry enables domain experts to extend to new domains by registering components, and algorithmic researchers to implement custom search algorithms. We demonstrate composability on MATH500 (language reasoning), Crosswords (environment planning), and MapEval (tool use), showing that components and algorithms are orthogonal: components are reusable across algorithms within each task type, and algorithms work across all components and domains. We also report a mode-collapse finding: in infinite action spaces, LLM policy diversity, not reward quality, is the bottleneck for effective tree search. A demonstration video is available at https://youtu.be/nRGX43YrR3I. The package is released under the Apache 2.0 license at https://github.com/xinzhel/lits-llm, including installation instructions and runnable examples that enable users to reproduce the demonstrated workflows.

[Long] LLM-Generated Text May Harm Your Retrieval! A Robust Detection Strategy for Retrieval-Augmented Generation ► *Underline Video*

Zhaoheng Huang, Yutao Zhu, Ji-Rong Wen and Zhicheng Dou

Retrieval-augmented generation (RAG) effectively enhances the accuracy and timeliness of large language models (LLMs) by incorporating external knowledge retrieved from external sources. However, with the increasing prevalence of LLM-generated content, external corpora used by RAG systems may become contaminated with LLM-generated texts. Such contamination compromises the reliability and quality of retrieved results, ultimately leading to a degradation in RAG performance, and raises concerns about the diminishing presence of human texts and the "Spiral of Silence" effect. A natural solution is to incorporate LLM text detectors into the RAG pipeline to filter out LLM-generated texts from the retrieved results. However, their effective use in RAG remains under-explored. In this paper, we explore the usage paradigms of LLM text detectors for RAG and highlight key limitations of off-the-shelf or directly fine-tuned detectors. To this end, we propose a RAG-aware data augmentation strategy that aligns detector training with realistic contamination patterns. Our approach synthesizes training data from both LLM and human texts under diverse generation modes. Experiments show that our method mitigates performance degradation and improves the long-term stability of RAG systems.

[Findings] LLMs are Brittle to Simple Code Transformations: Introducing CETBench - A Benchmark for Code-Equivalence Checking ► *Underline Video*

Neeva Oza, Ishaan Govil, Parul Gupta, Dinesh Khandelwal, Dinesh Garg and Parag Singla

We study how well LLMs can determine whether two programs are functionally equivalent. This is an important problem because benchmarking code equivalence helps assess LLM capability in tasks such as code rewriting and translation. To this end, we introduce CETBench , Code Equivalence with Transformations Benchmark , built from a repository of programs that may solve the same or different tasks. Each dataset instance is created by sampling a program pair and applying a random sequence of predefined code transformations, yielding either equivalent or non-equivalent pairs. Our analysis shows that even simple transformations cause a significant performance drop in state-of-the-art LLMs on code-equivalence checking. These challenges are further amplified in the cross-lingual setting when comparing programs written in different languages. To remedy this, we present a simple fine-tuning-based approach to boost LLM performance on the transformed pairs of programs. Our approach for dataset generation is generic, supporting cross-lingual equivalence checking, the generation of program pairs with varying difficulty levels, and the application of diverse transformations. In our experiments, we perform ablations

over the difficulty level of original programs, as well as the kind of transformations used in generating pairs for equivalence checking. Our analysis presents deep insights into the working of LLMs for the task of code-equivalence, and points to the fact that they may still be far from what could be termed as a semantic understanding of the underlying code.

[Long] LLMSurgeon: Diagnosing Data Mixture of Large Language Models ► *Underline Video*

Yaxin Luo, Jiacheng Cui, Xiaohan Zhao, Xinyi Shang, Jiacheng Liu, Xinyue Bi, Zhaoyi Li and Zhiqiang Shen

The pretraining data mixture of Large Language Models (LLMs) constitutes their "digital DNA", shaping model behaviors, capabilities, and failure modes. Yet this composition is rarely disclosed, making post-hoc auditing of data combination or provenance difficult. In this work, we formalize Data Mixture Surgery (DMS): given only generated text from a target LLM, estimate the domain-level distribution of its pretraining corpus under a predefined taxonomy. We propose LLMSurgeon, a strong framework that casts DMS as an inverse problem under the label-shift assumption. Rather than directly aggregating classifier outputs, LLMSurgeon estimates a calibrated soft confusion matrix and solves a constrained inverse problem to correct systematic domain confusion and recover the latent mixture prior. To evaluate, we introduce LLMScan, a recipe-verifiable evaluation suite built from open-source LLMs with transparent pretraining mixtures. Across LLMScan, LLMSurgeon recovers domain mixtures with high fidelity under fixed protocols. Our work presents a practical, post-hoc approach for auditing the digital DNA of foundation models without access to their training data. Code is available at: https://github.com/Yaxin9Luo/LLMSurgeon.

[Long] LongTutor: Benchmarking Large Language Models for Long-term Personalized Tutoring ► *Underline Video*

Ning Li, Zheng Zhang, Zhenya Huang, Rui Li, Yi Zhan, Yinbo Luo, Qi Liu and Enhong Chen

The rapid advancement of large language models (LLMs) has driven the deployment of LLM-based AI tutors on online learning platforms. This widespread adoption highlights an urgent need for systematic benchmarks to evaluate their tutoring capabilities. However, existing evaluations predominantly focus on isolated, short-term interactions, overlooking the inherently long-term nature of learning. To bridge this gap, we introduce LongTutor, a benchmark for long-term personalized tutoring grounded in formative assessment theory. Built from expert-annotated real-world learning logs, LongTutor evaluates LLMs across three progressive tasks: historical evidence acquisition, knowledge state diagnosis, and adaptive teaching action. Our experiments reveal a critical capability mismatch: while LLMs excel at evidence acquisition, they struggle to effectively leverage long-term history for accurate diagnosis and adaptive teaching. To enable scalable benchmark expansion, we further propose an automated generator-verifier pipeline, paving the way toward truly long-term AI tutoring systems.

[Long] LOTUS: Evolving Multimodal Unlearning via Hyperbolic Entailment and Lorentz Transport ► *Underline Video*

zekun wangjingjie zengyingxu liliang yanghongfei lin

Multimodal Large Language Models (MLLMs) face critical privacy challenges due to the indiscriminate memorization of sensitive data. Existing unlearning methods, largely adapted from Euclidean paradigms, suffer from a geometric mismatch: they fail to disentangle specific instances from general concepts, causing catastrophic forgetting or unsafe substitution. We introduce LOTUS (Lorentz Transport for Unlearning Strategies), a framework for surgical semantic pruning within the Lorentz manifold. Leveraging hyperbolic geometry's hierarchical nature, LOTUS employs an Inverted Entailment Cone Loss to sever the inheritance of sensitive concepts and a Lorentz Transport mechanism to align pruned features within the tangent space, ensuring compatibility with Euclidean backbones via a safety refusal prior. Experiments on MLLMU-Bench with LLaVA and Qwen show that LOTUS significantly outperforms baselines, effectively erasing targeted visual data while preserving general utility.

[Findings] LycheeCluster: Efficient Long-Context Inference with Structure-Aware Chunking and Hierarchical KV Indexing ► *Underline Video*

Dongfang Li, Zixuan Liu, Gang Lin, Baotian Hu and Min Zhang

The quadratic complexity of the attention mechanism and the substantial memory footprint of the Key-Value (KV) cache present severe computational and memory challenges for Large Language Models (LLMs) processing long contexts. Existing retrieval-based methods often compromise semantic integrity through fixed-size chunking and suffer from inefficient linear scanning. In this paper, we propose LycheeCluster, a novel method for efficient KV cache management. LycheeCluster preserves local semantic coherence via boundary-aware chunking and constructs a recursive hierarchical index rooted in the triangle inequality. This design transforms cache retrieval from a linear scan into a theoretically bounded, logarithmic-time pruning process, while a lazy update strategy supports efficient streaming generation. Experiments demonstrate that LycheeCluster achieves up to a 3.6x end-to-end inference speedup with negligible degradation in model performance, outperforming state-of-the-art KV cache management methods (e.g., Quest, ClusterKV).

[Long] MAB-DQA: Addressing Query Aspect Importance in Document Question Answering with Multi-Armed Bandits ► *Underline Video*

Yixin Xiang, Yunshan Ma, Xiaoyu Du, Yibing Chen, Yanxin Zhang and Jinhui Tang

Document Question Answering (DQA) involves generating answers from a document based on a users query, representing a key task in document understanding. This task requires interpreting visual layouts, which has prompted recent studies to adopt multimodal Retrieval-Augmented Generation (RAG) that processes page images for answer generation. However, in multimodal RAG, visual DQA struggles to utilize a large number of images effectively, as the retrieval stage often retains only a few candidate pages (e.g., Top4), causing informative but less visually salient content to be overlooked in favor of common yet low-information pages. To address this issue, we

propose a Multi-Armed Bandit-based DQA framework (MAB-DQA) to explicitly model the varying importance of multiple implicit aspects in a query. Specifically, MAB-DQA decomposes a query into aspect-aware subqueries and retrieves an aspect-specific candidate set for each. It treats each subquery as an arm and uses preliminary reasoning results from a small number of representative pages as reward signals to estimate aspect utility. Guided by an exploration-exploitation policy, MAB-DQA dynamically reallocates retrieval budgets toward high-value aspects. With the most informative pages and their correlations, MAB-DQA generates the expected results. On four benchmarks, MAB-DQA shows an average improvement of 5%-18% over the stateoftheart method, consistently enhancing document understanding. Codes are available at https://github.com/ElephantOH/MAB-DQA.

[Long] Market-Bench: Benchmarking Large Language Models on Economic and Trade Competition ▶ *Underline Video*

Yushuo Zheng, Huiyu Duan, Zicheng Zhang, Yucheng Zhu, Xiongkuo Min and Guangtao Zhai

The ability of large language models (LLMs) to manage and acquire economic resources remains unclear. In this paper, we introduce Market-Bench, a comprehensive benchmark that evaluates the capabilities of LLMs in economically-relevant tasks through economic and trade competition. Specifically, we construct a configurable multi-agent supply chain economic model where LLMs act as retailer agents responsible for procuring and retailing merchandise. In the procurement stage, LLMs bid for limited inventory in budget-constrained auctions. In the retail stage, LLMs set retail prices, generate marketing slogans, and provide them to buyers through a role-based attention mechanism for purchase. Market-Bench logs complete trajectories of bids, prices, slogans, sales, and balance-sheet states, enabling automatic evaluation with economic, operational, and semantic metrics. Benchmarking on 20 open- and closed-source LLM agents reveals significant performance disparities and winner-take-most phenomenon, i.e., only a small subset of LLM retailers can consistently achieve capital appreciation, while many hover around the break-even point despite similar semantic matching scores. Market-Bench provides a reproducible testbed for studying how LLMs interact in competitive markets.

[Demo] MASFactory: A Graph-centric Framework for Orchestrating LLM-Based Multi-Agent Systems with Vibe Graphing ▶ *Underline Video*

Yang Liu, Jinxuan Cai, Yishen Li, Qi Meng, Zedi Liu, Xin Li, Chen Qian, Chuan Shi and Cheng Yang;

Large language model-based (LLM-based) multi-agent systems (MAS) are increasingly used to extend agentic problem solving via role specialization and collaboration. MAS workflows can be naturally modeled as directed computation graphs, where nodes execute agents/sub-workflows and edges encode dependencies and message passing. However, implementing complex graph workflows in current frameworks still requires substantial manual effort, offers limited reuse, and makes it difficult to integrate heterogeneous external context sources. To overcome these limitations, we present MASFactory, a graph-centric framework for orchestrating LLM-based MAS. It introduces Vibe Graphing, a human-in-the-loop approach that compiles natural-language intent into an editable workflow specification and then into an executable graph. In addition, the framework provides reusable components and pluggable context integration, as well as a visualizer for topology preview, runtime tracing, and human-in-the-loop interaction. We evaluate MASFactory on seven public benchmarks, validating both reproduction consistency for representative MAS methods and the effectiveness of Vibe Graphing. Our code (https://github.com/BUPT-GAMMA/MASFactory) and video (https://youtu.be/ANynzVfY32k) are publicly available.

[Findings] Masked Text-to-Audio Flow-Matching and Reward Feedback Optimization ▶ *Underline Video*

Rongjie Huang, Dongchao Yang, Wenxiang Guo, Huadai Liu, Xize Cheng, Zehan Wang, Zhou Zhao, Xixin Wu and Helen M. Meng

Flow-matching generative models have created significant milestones in text-to-audio generation, powered by scalable training with increased data, computational resources, and model size, while their scalable inference remains less explored. In this work, we propose MaskAudioFlow, a continuous flow-matching transformer with masked generative modeling designed for scaling text-to-audio inference-time prediction. Specifically, MaskAudioFlow 1) masks spans of audio frames in training and approximates the continuous velocity vector field with flow-matching objective, and 2) performs inference via masked prediction, where we mask out generation and re-predict them through iterative decoding. To reduce the gap between generation and human preferences, we fine-tune MaskAudioFlow using reward signals from text-audio correspondence and perceptual aesthetics. Experimental results demonstrate that MaskAudioFlow achieves state-of-the-art performance in text-to-audio generation, effectively scaling inference-time computation through iterative masked prediction. Moreover, the preference-tuned model demonstrates superior text-audio alignment faithfulness and enhanced perceptual aesthetics. Audio samples are available at https://MaskAudio.github.io

[Long] MathCanvas: Intrinsic Visual Chain-of-Thought for Multimodal Mathematical Reasoning ▶ *Underline Video*

Weikang Shi, Aldrich Yu, Rongyao Fang, Houxing Ren, Ke Wang, Aojun Zhou, Changyao Tian, Xinyu Fu, Yuxuan Hu, Zimu Lu, Linjiang Huang, Si Liu, Rui Liu and Hongsheng Li

While Large Language Models (LLMs) have excelled in textual reasoning, they struggle with mathematical domains like geometry that intrinsically rely on visual aids. Existing approaches to Visual Chain-of-Thought (VCoT) are often limited by rigid external tools or fail to generate the high-fidelity, strategically-timed diagrams necessary for complex problem-solving. To bridge this gap, we introduce MathCanvas, a comprehensive framework designed to endow unified Large Multimodal Models (LMMs) with intrinsic VCoT capabilities for mathematics. Our approach consists of two phases. First, a Visual Manipulation stage pre-trains the model on a novel 15.2M-pair corpus, comprising 10M caption-to-diagram pairs (MathCanvas-Imagen) and 5.2M step-by-step editing trajectories (MathCanvas-Edit), to master diagram generation and editing. Second, a Strategic Visual-Aided Reasoning stage fine-tunes the model on MathCanvas-Instruct, a new 219K-example dataset of interleaved visual-textual reasoning paths, teaching it when and how to leverage visual aids. To facilitate

rigorous evaluation, we introduce MathCanvas-Bench, a challenging benchmark with 3K problems that require models to produce interleaved visual-textual solutions. Our model, BAGEL-Canvas, trained under this framework, achieves an 86% relative improvement over strong LMM baselines on MathCanvas-Bench, demonstrating excellent generalization to other public math benchmarks. Our work provides a complete toolkit, framework, datasets, and benchmark, to unlock complex, human-like visual reasoning in LMMs.

[Findings] MathMixup: Boosting LLM Mathematical Reasoning with Difficulty-Controllable Data Synthesis and Curriculum Learning ▶ *Underline Video*

Xuchen Li, Jing Chen, Xuzhao Li, Hao Liang, Xiaohuan Zhou, Taifeng Wang and Wentao Zhang

In mathematical reasoning tasks, the advancement of Large Language Models (LLMs) relies heavily on high-quality training data with clearly defined and well-graded difficulty levels. However, existing data synthesis methods often suffer from limited diversity and lack precise control over problem difficulty, making them insufficient for supporting efficient training paradigms such as curriculum learning. To address these challenges, we propose MathMixup, a novel data synthesis paradigm that systematically generates high-quality, difficulty-controllable mathematical reasoning problems through hybrid and decomposed strategies. Automated self-checking and manual screening are incorporated to ensure semantic clarity and a well-structured difficulty gradient in the synthesized data. Building on this, we construct the MathMixupQA dataset and design a curriculum learning strategy that leverages these graded problems, supporting flexible integration with other datasets. Experimental results show that MathMixup and its curriculum learning strategy significantly enhance the mathematical reasoning performance of LLMs. Fine-tuned Qwen2.5-7B achieves an average score of 52.6% across seven mathematical benchmarks, surpassing previous state-of-the-art methods. These results fully validate the effectiveness and broad applicability of MathMixup in improving the mathematical reasoning abilities of LLMs and advancing data-centric curriculum learning.

[Long] MavenCoder: Competitive Code Generation via Model Adaptive Planning Strategies and Multi-Perspective Verification Enhancement ▶ *Underline Video*

ZhenChun Xu, Yi Cai, Dajun Zheng, li yuan, Mengchen Zhao, Qixiang Wang and Jiexin Wang

With the rapid advancement of large language models (LLMs), automated code generation has made remarkable progress. Recent studies explore multi-agent collaboration and adopt planning-coding-debugging workflows to enhance performance. However, these approaches are constrained by rigid, predefined workflows that fail to flexibly adjust their plans and lack effective verification of intermediate reasoning steps. In this work, we propose MavenCoder, a model-adaptive and verification-enhanced framework for competition-level code generation. MavenCoder leverages adaptive assessment aligned with the models capabilities to select planning strategies, while providing timely feedback and correction via multi-perspective verification. This adaptive problem-solving paradigm mitigates earlier limitations by enabling flexible planning and timely error correction. Compared with existing state-of-the-art approaches, MavenCoder achieves superior pass@1 results across multiple benchmarks, achieving 87.5% on LiveCodeBench, 93.9% on HumanEval+, 81.7% on MBPP+, and 46.1% on CodeContests, outperforming recent agent-based systems with improvement exceeding 3%–40%.

[Findings] MedDCR: Learning to Design Agentic Workflows for Medical Coding ▶ *Underline Video*

Jiyang Zheng, Islam Nassar, Thanh Vu, Xu Zhong, Yang Lin, Tongliang Liu, Long Duong and Yuan-Fang Li

Medical coding converts free-text clinical notes into standardized diagnostic and procedural codes, which are essential for billing, hospital operations, and medical research. Unlike ordinary text classification, it requires multi-step reasoning: extracting diagnostic concepts, applying guideline constraints, mapping to hierarchical codebooks, and ensuring cross-document consistency. Recent advances leverage agentic LLMs, but most rely on rigid, manually crafted workflows that fail to capture the nuance and variability of real-world documentation, leaving open the question of how to systematically learn effective workflows. We present MedDCR, a closed-loop framework that treats workflow design as a learning problem. A Designer proposes workflows, a Coder executes them, and a Reflector evaluates predictions and provides constructive feedback, while a memory archive preserves prior designs for reuse and iterative refinement. On benchmark datasets, MedDCR outperforms state-of-the-art baselines and produces interpretable, adaptable workflows that better reflect real coding practice, improving both the reliability and trustworthiness of automated systems.

[Long] Memory-Augmented LLM-based Multi-Agent System for Automated Feature Generation on Tabular Data ▶ *Underline Video*

FengXian Dong, Zhi Zheng, Xiao Han, Wei Chen, Jingqing Ruan, Tong Xu, Yong Chen and Enhong Chen

Automated feature generation extracts informative features from raw tabular data without manual intervention and is crucial for accurate, generalizable machine learning. Traditional methods rely on predefined operator libraries and cannot leverage task semantics, limiting their ability to produce diverse, high-value features for complex tasks. Recent Large Language Model (LLM)-based approaches introduce richer semantic signals, but still suffer from a restricted feature space due to fixed generation patterns and from the absence of feedback from the learning objective. To address these challenges, we propose a Memory-Augmented LLM-based Multi-Agent System (MALMAS) for automated feature generation. MALMAS decomposes the generation process into agents with distinct responsibilities, and a Router Agent activates an appropriate subset of agents per iteration, further broadening exploration of the feature space. We further integrate a memory module comprising procedural memory, feedback memory, and conceptual memory, enabling iterative refinement that adaptively guides subsequent feature generation and improves feature quality and diversity. Extensive experiments on multiple public datasets against state-of-the-art baselines demonstrate the effectiveness of our approach.

[Long] MirrorQA: Benchmarking Multimodal LLMs on Mirror-Orientation Reasoning ▶ *Underline Video*

Jingping Liu, Xingchen Peng, Yan Zhou, Ziyan Liu, Jie Zhai, Ronghao Chen, Huacan Wang and Xiaofeng Jia
Multimodal large language models (MLLMs) have achieved remarkable progress in recent years, yet their ability to perform left-right reasoning in mirror contexts, a fundamental element of spatial cognition, remains underexplored. To address this gap, we introduce MirrorQA, a manually constructed benchmark with 5,549 samples, designed to evaluate MLLMs capability to distinguish left from right from a subject-centered perspective. MirrorQA is built through a three-stage pipeline (annotation, verification, and final review) to ensure high-quality labeling. Comprehensive evaluations on both open- and closed-source MLLMs show that even the best-performing models achieve only 65.40% accuracy, far below the 99.28% accuracy of humans. These results highlight substantial challenges in current MLLMs when reasoning about left and right, and point to promising directions for future research. MirrorQA and its code are publicly available at anonymous link https://github.com/stargazer-zeno/MirrorQA.

[Long] Mitigating Action-Relation Hallucinations in LVLMs via Relation-aware Visual Enhancement ▶ *Underline Video*
Zhenxin Qin, Qiang Li, Qingzhuo Wang, Ruiyang Qin, Zhihua Wei and Wen Shen
Large Vision-Language Models (LVLMs) have achieved remarkable performance on diverse vision-language tasks. However, LVLMs still suffer from hallucinations, generating text that contradicts the visual input. Existing research has primarily focused on mitigating object hallucinations, but often overlooks more complex relation hallucinations, particularly action relations involving interactions between objects. In this study, we empirically observe that the primary cause of action-relation hallucinations in LVLMs is the insufficient attention allocated to visual information. Thus, we propose a framework to locate action-relevant image regions and enhance the LVLM's attention to those regions. Specifically, we define the Action-Relation Sensitivity (ARS) score to identify attention heads that are most sensitive to action-relation changes, thereby localizing action-relevant image regions that contain key visual cues. Then, we propose the Relation-aware Visual Enhancement (RVE) method to enhance the LVLM's attention to these action-relevant image regions. Extensive experiments demonstrate that, compared to existing baselines, our method achieves superior performance in mitigating action-relation hallucinations with negligible additional inference cost. Furthermore, it effectively generalizes to spatial-relation hallucinations and object hallucinations.

[Industry] Mixture of Heterogeneous Grouped Experts for Language Modeling ▶ *Underline Video*
Zhicheng Ma, Xiang Liu, Zhaoxiang Liu, Ning Wang, Yi Shen, Kai Wang, Shuming Shi and Shiguo Lian
Large Language Models (LLMs) based on Mixture-of-Experts (MoE) are pivotal in industrial applications for their ability to scale performance efficiently. However, standard MoEs enforce uniform expert sizes, creating a rigidity that fails to align computational costs with varying token-level complexity. While heterogeneous expert architectures attempt to address this by diversifying expert sizes, they often suffer from significant system-level challenges, specifically unbalanced GPU utilization and inefficient parameter utilization, which hinder practical deployment. To bridge the gap between theoretical heterogeneity and robust industrial application, we propose Mixture of Heterogeneous Grouped Experts (MoHGE) which introduces a two-level routing mechanism to enable flexible, resource-aware expert combinations. To optimize inference efficiency, we propose a Group-Wise Auxiliary Loss, which dynamically steers tokens to the most parameter-efficient expert groups based on task difficulty. To address the critical deployment challenge of GPU load balancing, we introduce an All-size Group-decoupling Allocation strategy coupled with an Intra-Group Experts Auxiliary Loss. These mechanisms collectively ensure uniform computation distribution across GPUs. Extensive evaluations demonstrate that MoHGE matches the performance of MoE architectures while reducing the total parameters by approximately 20% and maintaining balanced GPU utilization. Our work establishes a scalable paradigm for resource-efficient MoE design, offering a practical solution for optimizing inference costs in real-world scenarios.

[Long] Mobile-R1: Towards Interactive Capability for VLM-Based Mobile Agent via Systematic Training ▶ *Underline Video*
Jihao Gu, Qihang Ai, Yingyao Wang, Pi Bu, Jingxuan Xing, Yue Cao, Zekun Zhu, Wei Jiang, Ziming Wang, Yingxiu Zhao, Ming-Liang Zhang, Jun Song, Yuning Jiang and Bo Zheng
Vision-language model-based mobile agents have gained the ability to understand complex instructions and mobile screenshots, benefiting from reinforcement learning paradigms like Group Relative Policy Optimization (GRPO). However, existing approaches centers on offline training or local action-level rewards often trap agents in local optima, hindering effective exploration and error correction with the environment. Crucially, we find that directly applying task-level rewards often leads to convergence difficulties due to the sparse nature of GUI interactions. To address these challenges, we present Mobile-R1, a systematic training recipe that bridges atomic action execution and strategic task completion. We propose a hierarchical curriculum consisting of three stages: (1) format alignment for reasoning structure, (2) on-policy exploration with verifiable action feedback to ground basic execution, and (3) multi-turn task-level training with realistic environment to unlock exploration and self-correction. This hierarchical strategy effectively bootstraps the agent, significantly enhancing its capability for exploration and self-correction (the "Eureka" moments). Furthermore, addressing the critical scarcity of diverse GUI data in non-English ecosystems, we contribute a comprehensive Chinese mobile dataset covering 28 applications with 24,521 high-quality manual annotations, and establish a rigorous benchmark with 500 trajectories. We will open source all resources, including the dataset, benchmark, model weight, and codes: https://mobile-r1.github.io/Mobile-R1/.

[Findings] MorphBPE: Morphology-Aware Tokenization for Efficient LLM Training ▶ *Underline Video*
Ehsaneddin Asgari, Yassine El Kheir, MohammadAli SadraeiJavaheri and Ali Nazari
Tokenization fundamentally shapes NLP performance, affecting both efficiency and linguistic fidelity. While Byte Pair Encoding (BPE) underpins most Large Language Models (LLMs), its frequency-driven merges often disregard morpheme boundaries, yielding inconsistent and semantically opaque segmentations in morphologically rich languages. We introduce MorphBPE, a simple extension of BPE that

constrains merge operations during tokenizer training to respect morpheme boundaries, while leaving inference unchanged and fully compatible with existing LLM pipelines. We evaluate tokenization quality using two intrinsic metrics, Morphological Consistency F1, which measures whether shared morphemes are assigned consistent token representations, and Morphological Edit Distance, which quantifies alignment with morpheme boundaries. We then train 300M and 1B parameter decoder-only LMs from scratch across four typologically diverse languages, English, Russian, Hungarian, and Arabic, under identical vocabulary sizes and training settings. Across all languages, MorphBPE consistently improves intrinsic morphological coherence and reduces language model cross-entropy, moreover, token length statistics indicate that these gains are not attributable to materially shorter tokens. Finally, on the Belebele multilingual reading comprehension benchmark, MorphBPE yields significant improvements in morphologically rich languages such as Russian and Arabic.

[Industry] Motivating Next-Gen Accelerators with Flexible Activation Sparsity via Benchmarking Lightweight Post-Training Sparsification Approaches ▶ *Underline Video*

Shirin Alanova, Kristina Kazistova, Ekaterina Galaeva, Alina Kostromina, Vladimir Smirnov, Redko Dmitry, Alexey Dontsov, Maxim Zhelnin, Evgeny Burnaev and Egor Shvetsov

The demand for efficient large language model inference has spurred interest in sparsification, yet current hardware support remains narrowly focused on 2:4 weight sparsity. In this work, we argue that activation sparsity despite being overlooked in hardware design offers a promising path for dynamic, input-adaptive compression with significant I/O and memory benefits. We present a comprehensive post-training study of N:M activation pruning across four LLMs (Llama2-7B-chat, Llama3.1-8B-Instruct, Qwen2.5-7B-Instruct, Gemma3-4B-Instruct), demonstrating that activation pruning consistently outperforms weight pruning at matched sparsity levels. We evaluate lightweight, plug-and-play error mitigation and selection strategies that require minimal or no calibration data across four sparsity patterns: 2:4, 4:8, 8:16, and 16:32. Among these, 16:32 approaches the performance of unstructured 50% sparsity and is is approximately 2.7x better than 2:4, while 8:16 offers an optimal balance of accuracy and practicality. Our results provide evidence that next-generation accelerators should consider native support for N:M activation sparsity and can serve as a strong baseline for the future methods.

[Long] MPBoCo: Multimodal Prompt-based Boundary-enhanced Continual Framework for Joint Entity and Relation Extraction ▶ *Underline Video*

Guanglu Sun, Xinyu Liu, Lili Liang, Yu Yang, Fei Lang, Suxia Zhu and Ming Liu

In real-world scenarios, multimodal information continuously evolves, with new entity and relation types emerging, necessitating timely updates to multimodal knowledge graphs for supporting downstream tasks. However, existing methods struggle to balance real-time adaptability and computational efficiency in continual learning scenarios. To this end, this paper proposes the Continual Multimodal Entity and Relation Joint Extraction (CMERJE) task and a Multimodal Prompt-based Boundary-enhanced Continual (MPBoCo) framework. Specifically, MPBoCo incrementally stores task-specific knowledge via learnable multimodal prompts, dynamically matches relevant prompts for each instance, and fuses them into a frozen backbone model for task-specific reasoning. Subsequently, the boundary-enhanced dual-branch module leverages the auxiliary branch to preserve local syntactic continuity and provide boundary guidance. Experimental results demonstrate that MPBoCo achieves superior performance in real-world scenarios, significantly outperforming baseline methods by 5.5% and 7.2% in 10-task and 5-task settings, respectively.

[Long] MQM Re-Annotation: A Technique for Collaborative Evaluation of Machine Translation ▶ *Underline Video*

Parker Riley, Daniel Deutsch, Mara Finkelstein, Colten DiIanni, Juraj Juraska and Markus Freitag

Human evaluation of machine translation is in an arms race with translation model quality: as our models get better, our evaluation methods need to be improved to ensure that quality gains are not lost in evaluation noise. To improve annotation quality, we experiment with a two-stage version of the current state-of-the-art translation evaluation paradigm (MQM), which we call MQM re-annotation. In this setup, an annotator reviews and edits a set of prior MQM annotations that may have come from themselves, another human annotator, or an automatic system. We demonstrate that rater behavior in re-annotation aligns with our goals, and that re-annotation results in higher-quality annotations, mostly due to finding errors that were missed during the first pass.

[Long] MTRouter: Cost-Aware Multi-Turn LLM Routing with History-Model Joint Embeddings ▶ *Underline Video*

Yiqun ZhangHao LiZihan WangShi FengXiaocui YangDaling WangBo ZhangLei BaiShuyue Hu

Multi-turn, long-horizon tasks are increasingly common for large language models (LLMs), but solving them typically requires many sequential model invocations, accumulating substantial inference costs. Here, we study cost-aware multi-turn LLM routing: selecting which model to invoke at each turn from a model pool, given a fixed cost budget. We propose MTRouter, which encodes the interaction history and candidate models into joint history-model embeddings, and learns an outcome estimator from logged trajectories to predict turn-level model utility. Experiments show that MTRouter improves the performance-cost trade-off: on ScienceWorld, it surpasses GPT-5 while reducing total cost by 58.7%; on Humanitys Last Exam (HLE), it achieves competitive accuracy while reducing total cost by 43.4% relative to GPT-5, and these gains even carry over to held-out tasks. Further analyses reveal several mechanisms underlying its effectiveness: relative to prior multi-turn routers, MTRouter makes fewer model switches, is more tolerant to transient errors, and exhibits emergent specialization across models. Code: https://github.com/ZhangYiqun018/MTRouter

[Long] Multilingual Language Models Encode Script Over Linguistic Structure ▶ *Underline Video*

Aastha A K Verma, Anwoy Chatterjee, Mehak Gupta and Tanmoy Chakraborty

Multilingual language models (LMs) organize representations for typologically and orthographically diverse languages into a shared parameter space, yet the nature of this internal organization remains elusive. In this work, we investigate which linguistic properties , abstract language identity or surface-form cues , shape multilingual representations. To do so, we analyze language-associated units across different model families and scales using the Language Activation Probability Entropy (LAPE) metric, and further decompose activations with Sparse Autoencoders. We find that these units are strongly conditioned on orthography: romanization induces near-disjoint representations that align with neither native-script inputs nor English, while word-order shuffling has limited effect on unit identity. Probing shows that typological structure becomes increasingly accessible in deeper layers, while causal interventions indicate that generation is most sensitive to units that are invariant to surface-form perturbations rather than to units identified by typological alignment alone. Overall, our results suggest that multilingual LMs organize representations around surface form, with linguistic abstraction emerging gradually without collapsing into a unified interlingua.

[Long] MulVul: Retrieval-augmented Multi-Agent Code Vulnerability Detection via Cross-Model Prompt Evolution ► *Underline Video*

Zihan Wu, Jie Xu, Yun Peng, Chun Yong Chong and Xiaohua Jia

Large Language Models (LLMs) struggle to automate real-world vulnerability detection due to two key limitations: the heterogeneity of vulnerability patterns undermines the effectiveness of a single unified model, and manual prompt engineering for massive weakness categories is unscalable. To address these challenges, we propose MulVul, a retrieval-augmented multi-agent framework designed for precise and broad-coverage vulnerability detection. MulVul adopts a coarse-to-fine strategy: a Router agent first predicts the top- coarse categories and then forwards the input to specialized Detector agents, which identify the exact vulnerability types. Both agents use evidence retrieved from vulnerability knowledge bases to mitigate hallucinations. Crucially, to automate the generation of specialized prompts, we design Cross-Model Prompt Evolution, a prompt optimization mechanism where a generator LLM iteratively refines candidate prompts while a distinct executor LLM validates their effectiveness. This decoupling mitigates the self-correction bias inherent in single-model optimization. Evaluated on 130 CWE types, MulVul achieves 34.79% Macro-F1, outperforming the best baseline by 41.5%. Ablation studies validate cross-model prompt evolution, which boosts performance by 51.6% over manual prompts by effectively handling diverse vulnerability patterns.

[Long] Nature-Inspired Population-Based Evolution of Large Language Models ► *Underline Video*

Yiqun ZhangPeng YeXiaocui YangShi FengShufei ZhangLei BaiWanli OuyangShuyue Hu

Evolution, the engine behind the survival and growth of life on Earth, operates through the population-based process of reproduction. Inspired by this principle, this paper formally defines a newly emerging problem: the population-based evolution of large language models (LLMs). We introduce a novel framework that starts with a population of parent LLMs and allows this population to evolve through four key operations: (i) crossover, merging the weights of different parents to create offspring LLMs, (ii) mutation, introducing small, random changes to model weights to foster diversity, (iii) selection, prioritizing high-performing models, and (iv) succession, transferring the learned experience from parent to offspring LLMs. With only 200 samples per new task, the LLM population evolves rapidly to adapt to the task at hand, without any gradients. Experiments on 12 datasets show that our framework consistently outperforms existing multi-LLM merging and adaptation methods, achieving relative performance gains of up to 54.8 over the best LLM in the initial population. Moreover, our framework allows for (i) the evolution of LLMs across multiple new tasks simultaneously, (ii) scaling effectively with populations of up to 40 LLMs, and (iii) even zero-shot generalization to unseen held-out tasks. Code: https://github.com/ZhangYiqun018/GENOME

[Short] Neuro-Symbolic Agentic Reinforcement Learning for Long-Term Original Character Companionship and Interaction ► *Underline Video*

Zhenhan Huang

As human-agent interaction (HAI) evolves toward long-term social companionship, users expect Original Character (OC) agents to maintain a consistent persona, manage shared memories, and adapt to ever-changing preferences. However, LLM-based agents optimized by prompting or SFT exhibit a generalization gap: they behave as myopic instruction followers, leading to cascading errors in multi-turn interactions. For the agents to learn trajectory-level value functions that enable farsighted decision-making, we propose the NSARL framework, which formalizes OC companion agents' interactions as a POMDP and decomposes the agent into three sub-policies (Router, Memory, and Persona), optimized via closed-loop RL from AI feedback (RLAIF) with verifiable rewards in a graph-constrained action space. Our preliminary experiments indicate a trade-off: SFT yields stronger persona generation, while NSARL improves structural logic, through conservative strategies (e.g., over-routing) that increase workflow completeness, advocating for a hybrid deployment strategy.

[Long] OASIS: Mitigating Harmful Fine-tuning Attacks on LLMs via Orthogonal and Adaptive Safety Alignment Strategy ► *Underline Video*

Jiayu Tang, Guowei Peng, Qiuhao Xie, Yuning Yang, Xiurui Xie and Guisong Liu

The "Fine-Tuning-as-a-Service" paradigm exposes large language models to catastrophic safety degradation from less harmful samples. Alignment-stage defenses address this by proactively injecting adversarial perturbations to bolster the model's inherent robustness against harmful drift. However, existing methods rely on perturbation directions that often conflict with harmful gradients, inadvertently facilitating the acquisition of malicious features rather than suppressing them. To address this issue, we propose Orthogonal and Adaptive Safety Alignment Strategy (OASIS) to mathematically decouple safety enforcement from harmful feature acquisition. By projecting perturbations

orthogonal to harmful gradients and concentrating optimization on adaptively selected safety-critical layers, OASIS effectively resolves directional conflicts while maximizing parameter efficiency. Extensive experiments on four LLMs across three datasets (SST2, GSM8K, and AGNews) demonstrate that OASIS reduces the Harmful Score by approximately 60% compared to competitive baselines, while maintaining stable downstream task utility.

[TACL] On the Effect of Instruction Tuning Loss on Generalization ► *Underline Video*

Anwoy Chatterjee, H S V N S Kowndinya Renduchintala, Sumit Bhatia and Tanmoy Chakraborty

Instruction Tuning has emerged as a pivotal post-training paradigm that enables pre-trained language models to better follow user instructions. Despite its significance, little attention has been given to optimizing the loss function used. A fundamental, yet often overlooked, question is whether the conventional auto-regressive objective - where loss is computed only on response tokens, excluding prompt tokens - is truly optimal for instruction tuning. In this work, we systematically investigate the impact of differentially weighting prompt and response tokens in instruction tuning loss, and propose Weighted Instruction Tuning (WIT) as a better alternative to conventional instruction tuning. Through extensive experiments on five language models of different families and scale, three finetuning datasets of different sizes, and five diverse evaluation benchmarks, we show that the standard instruction tuning loss often yields suboptimal performance and limited robustness to input prompt variations. We find that a low-to-moderate weight for prompt tokens coupled with a moderate-to-high weight for response tokens yields the best-performing models across settings and also serve as better starting points for the subsequent preference alignment training. These findings highlight the need to reconsider instruction-tuning loss and offer actionable insights for developing more robust and generalizable models. Our code is open-sourced at https://github.com/kowndinya-renduchintala/WIT.

[Long] One Cognitive Loop Is Enough: SODA unlocks Pure-Text Spatial Reasoning in Large Language Models ► *Underline Video*

Shunwen Bai, Jiahuan Zhang, Haoran Huang, Yurun Wang, Jiale liu, Yanxi Wu, Ningzhe Yu, Yudong Gao and Mingjun Cheng

Currently, large language models (LLMs) have significant limitations in spatial reasoning, particularly in the absence of visual input. To address this issue, we introduce SODA (Spatial OODA), which draws inspiration from the OODA cognitive loop (Observe, Orient, Decide, Act), originally designed to enhance human decision-making in dynamic environments. Specifically, we embed the OODA loop into multiple control tasks, generating the SPOD-143k dataset, and successfully integrate it into LLMs through a two-phase and spatia-aware training strategy (SFT and GRPO). Furthermore, to fill the gap in evaluating spatial reasoning in purely text-based LLMs, we introduce the SPOD-Bench benchmark, including multiple tasks divided into three levels of difficulty. Experimental results show that SODA significantly enhances the spatial reasoning capabilities of LLMs across testing scenarios including SPOD-Bench, SPACE and applications, providing a replicable and effective paradigm for improving the spatial cognition of LLMs.

[SRW] One Task Vector is not Enough: A Large-Scale Study for In-Context Learning ► *Underline Video*

Pavel Tikhonov; Ivan Oseledets; Elena Tutubalina

In-context learning (ICL) enables Large Language Models (LLMs) to adapt to new tasks using few examples, with task vectors, defined as specific hidden state activations hypothesized to encode task information. Existing studies are limited by small-scale benchmarks, restricting comprehensive analysis. We introduce QAF, a novel dataset of 3,096 diverse few-shot tasks, each with 30 input-output pairs derived from the Alpaca dataset. Experiments with Llama-3-8B on QAF reveal: (1) task vector performance peaks at an intermediate layer (e.g., 15th), (2) effectiveness varies significantly by task type, and (3) complex tasks rely on multiple, subtask-specific vectors rather than a single vector, suggesting distributed task knowledge representation.

[Long] PACE: Predictive Adaptive Context Extraction for Long-Horizon LLM Agents ► *Underline Video*

Lei Wei, Xiao Peng, TT TT, Guannan Zhang, Chenhao Jiang, Hongyu Li, Lanbo Lin, Yuanwu Xu, Jiayao Liu, Kesu Wang and Bin Wang

Large Language Model (LLM) agents struggle with ultra-long-horizon tasks requiring hundreds or thousands of interaction steps. Traditional context management approaches face a fundamental dilemma: preserving complete histories rapidly exhausts context windows and forces crude truncation, while aggressive summarization discards critical information prematurely. We propose Predictive Adaptive Context Extraction (PACE), a novel framework that reconceptualizes context management as a Next Step Prediction problem. Inspired by neural attention, PACE dynamically constructs context by adjusting historical memory granularity based on its predicted relevance for the next action. Comprehensive evaluation across diverse benchmarks and models demonstrates that PACE consistently improves task success rates, with larger gains on complex tasks and robust cross-lingual performance. Crucially, PACE enables agents to sustain effective reasoning for 4,897 interaction steps in ultra-long-horizon scenarios, achieving a 66.2 improvement over the full-context ReAct baseline and 5.1 over advanced folding baselines. This fundamentally advances the capability of LLM-based agents in previously intractable long-horizon scenarios. Our code and data are available at https://anonymous.4open.science/r/PACE-B000/.

[Demo] Paper2Web: Lets Make Your Paper Alive! ► *Underline Video*

Yuhang Chen, Tianpeng Lv, Yao Wan, Philip S. Yu and Dongping Chen

Academic project websites can more effectively disseminate research when they clearly present core content and enable intuitive navigation and interaction. However, current approaches such as direct Large Language Model (LLM) generation, templates, or direct HTML conversion struggle to produce layout-aware, interactive sites, and a comprehensive evaluation suite for this task has been lacking. In this paper, we introduce Paper2Web, a benchmark dataset and multi-dimensional evaluation framework for assessing academic webpage generation. It incorporates rule-based metrics like Connectivity, Completeness and human-verified LLM-as-a-Judge (covering interactivity, aesthetics, and

informativeness), and PaperQuiz, which measures paper-level knowledge retention. We further present PWAgent, an autonomous pipeline that converts scientific papers into interactive and multimedia-rich academic homepages. The agent iteratively refines both content and layout through MCP tools that enhance emphasis, balance, and presentation quality. Our experiments show that PWAgent consistently outperforms end-to-end baselines like template-based webpages and arXiv/alphaXiv versions by a large margin while maintaining low cost, achieving the Pareto-front in academic webpage generation.

[Long] Paraphrasing as Zero-shot Translation with Feature-guided Diversity Enhancement ▶ *Underline Video*

Ziyue Yan, Hongying Zan, Xinglin Lyu and Hongfei Xu

Paraphrasing uses different words, sentence structures, or expressions to convey similar semantics. It is an effective training data augmentation method to improve low-resource Natural Language Processing (NLP) tasks. Existing studies normally leverage parallel corpora to construct parabanks, regarding the Machine Translation (MT) results of source sentences as the paraphrases of the corresponding target sentences. As MT models are usually trained on the same parallel corpus, translation of the training set may suffer from overfitting, which leads to less diverse paraphrases. Training paraphrasers on the parabank generated via MT may also suffer from the information loss issue, as the parabank is derived from the parallel corpora, and the knowledge inside the parabank is a subset of that inside the parallel corpora. In this paper, we train bidirectional Multilingual Neural Machine Translation (MNMT) on the bi-directional bilingual parallel corpus, and use the MNMT model directly as a paraphrasing model by asking it to generate "translations" of the input language. As some source tokens also appear in the translation in the parallel corpus, we introduce "copy"/"not-copy" tags to indicate the existence/non-existence of source tokens in the target translation during training, and use the "not-copy" tag to encourage paraphrasing during inference. Manual and automatic evaluation results show that our ParaMNMT method can generate paraphrases of higher semantic consistency, literal fluency and sentential diversity compared to existing parabanks and LLMs. Our data augmentation experiments verify the effectiveness of ParaMNMT on improving low-resource NLP tasks.

[Long] ParaSuite: Boosting LLM Reasoning via Paradox Resolution ▶ *Underline Video*

BinChen BinChen, Yu Zhang, Hongfei Ye, Huiyang Wang, Wenxi Liu and Hongyang Chen

Logical reasoning is a key capability of large language models, yet current benchmarks focus almost entirely on tasks that just check basic logical consistency and overlook the reflective reasoning required for paradox detection and resolution. To fill the gap, we present ParaSuite, the first pipeline dedicated to paradox research that automates data synthesis, evaluation, and training. We introduce PARADOX, a synthetic, high-quality data spanning two difficulty tiers and three academic domains, accompanied by specialized evaluation metrics and solving algorithms. We propose ParadoxBreaker-7B, trained with Mutual-Information Guided Fine-Tuning and reinforcement learning step verify paradox reward(PAPO). Experiments demonstrate significant improvements in both paradoxical and general STEM reasoning.

[Long] PARIF: Pushing the Pareto Frontier of Instruction Following and Reasoning with Curriculum Reinforcement Learning ▶ *Underline Video*

Rongchuan Mu, Zexin Wang, Qianyu Wang, MingHua Ma, Zekun Wang, Ming Liu and Bing Qin

Large Reasoning Models (LRMs) excel at complex problem-solving but frequently overlook specific instruction constraints. Existing alignment methods struggle to balance general reasoning with instruction-following (IF), hindered by dependency on teacher models, reward hacking, and reasoning-answer inconsistencies. We propose PARIF, a two-stage curriculum learning framework based on Reinforcement Learning from Verifiable Rewards (RLVR) to enhance both IF and general reasoning capabilities. The framework employs a correctness proxy across different stages to mitigate reward hacking. Stage I employs a dynamic weighting strategy simultaneously to optimize the model's reasoning paradigm regarding constraints. Stage II introduces Decoupled-GRPO, which builds upon the first stage to enhance the logical consistency between the reasoning process and the final answer, enabling the model to better leverage its optimized reasoning paradigm. To support the framework, we curate 26,000 high-quality instructions featuring diverse constraints. Extensive experiments demonstrate PARIF's effectiveness: our 7B model achieves a remarkable 21.25% relative average improvement to the original model across six representative IF tasks, while our 8B model outperforms leading models like DeepSeek-V3 on these IF tasks, effectively pushing the Pareto frontier of instruction following and reasoning for models of comparable scale. We open-source our code and models to facilitate future research.

[Long] PaTaRM: Bridging Pairwise and Pointwise Signals via Preference-Aware Task-Adaptive Reward Modeling ▶ *Underline Video*

Ai Jian, Jingqing Ruan, Xing Ma, Dailin Li, Weipeng Zhang, Ke Zeng and Xunliang Cai

Reward models (RMs) are central to reinforcement learning from human feedback (RLHF), providing the critical supervision signals that align large language models (LLMs) with human preferences. Generative reward models (GRMs) provide greater interpretability than traditional scalar RMs, but they come with a critical trade-off: pairwise methods are hindered by a training-inference mismatch, while pointwise methods require expensive absolute annotations. To bridge this gap, we propose the Preference-aware Task-adaptive Reward Model (PaTaRM). Unlike prior approaches, PaTaRM enables robust pointwise training using readily available pairwise data via a novel Preference-Aware Reward (PAR) mechanism, eliminating the need for explicit rating labels. Furthermore, it incorporates a task-adaptive rubric system that dynamically generates instance-specific criteria for precise evaluation. Extensive experiments demonstrate that PaTaRM achieves an average relative improvement of 8.7% over the corresponding base models on RewardBench and RMBench across the Qwen3-8B and Qwen3-14B backbones. Crucially, when used as a reward model for downstream RLHF, it yields an average relative improvement of 13.6% over the corresponding base policies on IFEval and InfoBench, validating its effectiveness for policy alignment. Our code, data, and

checkpoints are available at https://huggingface.co/AIJian/PaTaRM

[Findings] Pattern Enhanced Multi-Turn Jailbreaking: Exploiting Structural Vulnerabilities in Large Language Models ▶ *Underline Video*

Ragib Amin Nihal, Rui Wen, Kazuhiro Nakadai and Jun Sakuma

Large language models (LLMs) remain vulnerable to multi-turn jailbreaking attacks that exploit conversational context to bypass safety constraints gradually. These attacks target different harm categories through distinct conversational approaches. Existing multi-turn methods often rely on heuristic or ad hoc exploration strategies, providing limited insight into underlying model weaknesses. The relationship between conversation patterns and model vulnerabilities across harm categories remains poorly understood. We propose Pattern Enhanced Chain of Attack (PE-CoA), a framework of five conversation patterns to construct multi-turn jailbreaks through natural dialogue. Evaluating PE-CoA on twelve LLMs spanning ten harm categories, we achieve state-of-the-art performance, uncovering pattern-specific vulnerabilities and LLM behavioral characteristics: models exhibit distinct weakness profiles, defense to one pattern does not generalize to others, and model families share similar failure modes. These findings highlight limitations of safety training and indicate the need for pattern-aware defenses. Code available on: https://github.com/Ragib-Amin-Nihal/PE-CoA

[Findings] PrefRAG: Correcting Semantic Errors in Auto-Formalization for Logical Reasoning with Program Preference RAG ▶ *Underline Video*

Yuyin Zhou and Yongmei Liu

Recent advances in large language models (LLMs) have spurred interest in neurosymbolic methods for logical reasoning based on autoformalization, where LLMs first formalize problems into symbolic programs, for solvers to perform reasoning over. However, existing auto-formalization methods remain prone to both syntactic and semantic errors. Specifically, the absence of a programlevel semantic verification mechanism leaves semantic errors largely unaddressed. In this paper, we propose a novel approach to semantic error correction via program preference retrievalaugmented generation (RAG). First, we conduct an in-depth analysis of semantic error patterns, and then automatically synthesize SemanticPref, a program preference dataset to model these patterns. Using the dataset as the knowledge base, we introduce PrefRAG, a general RAG framework for refinement in auto-formalization, which enables LLMs to detect and repair syntactic and semantic errors. Extensive evaluations across both in-distribution (ID) benchmarks (i.e., AR-LSAT and FOLIO) and outofdistribution (OOD) datasets show that PrefRAG consistently outperforms strong baselines, achieving an average accuracy improvement of 2.39% on ID and 6.23% on OOD datasets.

[Findings] Progress Ratio Embeddings: An Impatience Signal for Robust Length Control in Neural Text Generation ▶ *Underline Video*

Ivanhoé Botcazou, Tassadit Amghar, Sylvain Lamprier and Frédéric Saubion

Modern neural language models achieve high accuracy in text generation, yet precise control over generation length remains underdeveloped. In this paper, we first investigate a recent length control method based on Reverse Positional Embeddings (RPE) and show its limits when control is requested beyond the training distribution. In particular, using a discrete countdown signal tied to the absolute remaining token count leads to instability. To provide robust length control, we introduce Progress Ratio Embeddings (PRE), as continuous embeddings tied to a trigonometric impatience signal. PRE integrates seamlessly into standard Transformer architectures, providing stable length fidelity without degrading text accuracy under standard evaluation metrics. We further show that PRE generalizes well to unseen target lengths. Experiments on two widely used news-summarization benchmarks and a popular question generation dataset validate these findings.

[TACL] PsyMem: Fine-grained psychological alignment and Explicit Memory Control for Advanced Role-Playing LLMs ▶ *Underline Video*

Yunxiao Qin, Xilong Cheng, Yuting Tan, Zhengnan Li, Ye Wang, Hongjiang Xiao, Yuan Zhang and Yikang Liu

Existing LLM-based role-playing methods often rely on superficial textual descriptions or simplistic metrics, inadequately modeling both intrinsic and extrinsic character dimensions. Additionally, they typically simulate character memory with implicit model knowledge or basic retrieval augment generation without explicit memory alignment, compromising memory consistency. The two issues weaken reliability of role-playing LLMs in several applications, such as trustworthy social simulation. To address these limitations, we propose PsyMem, a novel framework integrating fine-grained psychological attributes and explicit memory control for role-playing. PsyMem supplements textual descriptions with 26 psychological indicators to detailed model character. Additionally, PsyMem implements memory alignment training, explicitly trains the model to align character's response with memory, thereby enabling dynamic memory-controlled responding during inference. By training Qwen2.5-7B-Instruct on our specially designed dataset (including 5,414 characters and 38,962 dialogues extracted from novels), the resulting model, termed as PsyMem-Qwen, outperforms baseline models in role-playing, achieving the best performance in human-likeness and character fidelity.

[Short] Punctuation-Steered Representation Fine-Tuning ▶ *Underline Video*

Zheng Gong, Ying Sun, Ping Li, Yi Zheng and Zhefeng Wang

Representation Fine-tuning (ReFT), a recently proposed parameter-efficient fine-tuning (PeFT) method, significantly improves parameter efficiency by modifying the representation space alone. However, directly applying ReFT, which alters a fixed number of representations at the beginning and end positions of each layer, results in suboptimal performance for two reasons. (i) The impact

of these fixed-position representations on the output is uncertain; (ii) As the sequence length increases, fine-tuning a fixed number of representations may have diminishing effects on the final results. Based on our observations that punctuation plays a crucial role in integrating representations from preceding layers and modulating those of subsequent layers, we introduce Punctuation-steered Representation Fine-tuning (PuReFT), a straightforward yet powerful approach that additionally fine-tunes punctuation representations to achieve performance improvements. Extensive evaluations on common-sense, arithmetic, and code datasets demonstrate the effectiveness and versatility of PuReFT. Furthermore, our analysis of its training speed and memory overhead confirms its greater ease of use and efficiency.

[Findings] QaRL: Rollout-Aligned Quantization-Aware RL for Fast and Stable Training under Training–Inference Mismatch ▶ *Underline Video*

Hao Gu, Hao Wang, Jiacheng Liu, Lujun Li, Qiyuan Zhu, Bei Liu, Binxing Xu, LEI WANG, Xintong Yang, Sida Lin, Sirui Han and Yike Guo

Large language model (LLM) reinforcement learning (RL) pipelines are often bottlenecked by rollout generation, making end-to-end training slow. Recent work mitigates this by running rollouts with quantization to accelerate decoding, which is the most expensive stage of the RL loop. However, these setups destabilize optimization by amplifying the training–inference gap: rollouts are operated at low precision, while learning updates are computed at full precision. To address this challenge, we propose QaRL (Rollout Alignment Quantization-Aware RL), which aligns training-side forward with the quantized rollout to minimize mismatch. We further identify a failure mode in quantized rollouts: long-form responses tend to produce repetitive, garbled tokens (error tokens). To mitigate these problems, we introduce TBPO (Trust-Band Policy Optimization), a sequence-level objective with dual clipping for negative samples, aimed to keep updates within the trust region. On Qwen3-30B-A3B MoE for math problems, QaRL outperforms quantized-rollout training by +5.5 while improving stability and preserving low-bit throughput benefits.

[Long] QBridge: Bridging Natural Language and SQL via Gold Query Rewriting with Agentic Refinement ▶ *Underline Video*

Zhensheng Luo, Sai Wu, Yuan Qiu, Chang Yao, Gang Chen and Xiu Tang

Natural language to SQL (NL2SQL) provides an intuitive interface for querying structured data, yet real user questions are often noisy, ambiguous, and weakly grounded to database semantics. As a result, token-level schema linking and single-pass SQL decoding can be brittle: small misunderstandings in language or schema grounding may propagate into incorrect generation. We present QBridge, an agentic, feedback-driven NL2SQL framework based on a Refined Gold Query Paradigm, which bridges natural language and SQL via Gold Query—a structured, SQL-aligned intermediate representation. A core insight of QBridge is Distilled Back-Translation (DBT) for SL-independent rewriting. DBT converts SQL-grounded supervision into execution-verified Gold-Query-style rewrites from a teacher model, and distills a lightweight, plug-and-play rewriter that generates schema-aware rewrites without requiring explicit schema linking at inference. QBridge then (i) verifies and conservatively refines the rewrite into a high-fidelity Refined Gold Query, and (ii) refines the generated SQL with dual feedback from execution validity and semantic consistency, enabling interpretable self-correction while remaining compatible with diverse SQL backbones. Extensive experiments on Spider, BIRD, and three robustness variants demonstrate that QBridge consistently improves zero-shot NL2SQL, outperforming strong prompting and agentic baselines while showing strong robustness and generalization. Code and data are available at https://github.com/WannaBSteve/QBridge.

[Demo] QFinZero: A Unified Financial Toolchain for LLM-Based Trading Agents ▶ *Underline Video*

Haochen Luo, YIFAN LI, Ho Tin Ko, An Binh Minh, Junjie Xu, Tang Pok Hin, Wang Chak WONG, Gao Yuan, Zhengzhao Lai, Yuan Zhang and Chen Liu

Large language model (LLM) agents are increasingly applied to financial decision-making tasks that require interaction with external tools, including market data retrieval, news analysis, and trade execution. However, existing trading systems rely on fragmented and task-specific APIs, resulting in inconsistent schemas, complex integration, and limited reproducibility. We present QFinZero, a unified trading environment for LLM-based financial agents. QFinZero standardizes three core capabilities: (i) multi-frequency market and derivatives data access, (ii) structured news and event retrieval, and (iii) stateful brokerage simulation with explicit order lifecycle management. All tools adopt consistent JSON schemas and time-aligned interfaces, enabling agents to acquire information and execute trades within a coherent framework. By abstracting financial interaction into composable, agent-invokable primitives, QFinZero reduces engineering overhead and supports reproducible evaluation through comprehensive logging and deterministic replay. We argue that such a standardized trading environment is essential for scalable research on LLM-based financial agents.

[Industry] RADAR: Risk-Aware Distilled Adaptive Routing for Efficient Short-Form Video Platform Ecosystem Governance ▶ *Underline Video*

Baoyu Jing, Zixuan Wang, Junwen Chen, Xin Dong, Bingfeng Deng and Kun Xu

Large-scale integrity enforcement on short-form video platforms typically relies on multiple specialized vertical modules, each dedicated to a specific risk category. However, exhaustively executing these computationally intensive modules over massive content streams leads to substantial inference overhead, despite the fact that most content is benign and violations are usually confined to limited policy domains. To address this inefficiency, we propose RADAR, a lightweight risk-aware routing framework that selectively releases low-risk content while dispatching high-risk instances to appropriate vertical modules. Industrial deployment of such routing systems presents two major challenges: (1) systematic label sparsity caused by disjoint annotation pipelines across risk categories, and (2) the capacity-efficiency tradeoff inherent to compact routing architectures. To overcome these challenges, RADAR incorporates Validity-Aware Masking to handle fragmented supervision and Expert-Guided Knowledge Distillation to transfer knowledge from heavyweight expert models into the

lightweight router. Experiments on large-scale real-world datasets demonstrate that the proposed masking strategy effectively mitigates disjoint annotation issues, while distillation substantially enhances routing accuracy, enabling the lightweight router to achieve competitive or superior performance compared to specialized expert models.

[Long] RADO: Reasoning Audit-Driven Optimization for Rigorous Reasoning in High-Stakes Domains ▶ *Underline Video*
Zhijie Tan, Xu Chu, Guanyu Wang, Ziyu Li, Weiping Li and Tong Mo
High-stakes domains such as finance, law, and biomedicine demand both accurate results and rigorous reasoning. Current reinforcement learning paradigms primarily rely on outcome-based rewards, often overlooking latent logical fallacies in intermediate steps. Leveraging the cognitive asymmetry where falsifying local errors is more efficient than generating global correctness, we propose RADO (Reasoning Audit-Driven Optimization). RADO introduces a specialized audit model augmented with external tools to identify local logical ruptures and calibrate reward signals. By integrating Direct Preference Optimization (DPO) with Group Relative Policy Optimization (GRPO), our framework enables explicit supervision over reasoning paths. Experimental results demonstrate that RADO consistently improves final accuracy while significantly enhancing logical rigor in high-stakes domains.

[Long] Reasoning with Ontology Graph: Toward Type-Constrained Knowledge Graph Question Answering ▶ *Underline Video*
Yongxue Shan, Jie Peng, Zixuan Dong, Fei Hu and Xiaodong Wang
Large language models (LLMs) have recently advanced knowledge graph question answering (KGQA), but current methods tend to rely on LLM-induced type systems with inconsistent granularity, or perform multi-hop reasoning without explicit target-type constraints. We introduce OntGQA, a type-constrained KGQA framework that reasons over a relation-centric ontology graph, where each relation is labeled with its head and tail entity types to provide a stable schema backbone. Built on this graph, OntGQA adopts a planner-judge architecture with generative backoff: a type planner proposes plausible head-tail type pairs, a judge verifies retrieved candidates and their paths, and a generator is invoked only when all candidates are rejected. By constraining both endpoints of reasoning in type space, OntGQA achieves state-of-the-art performance and produces ontology-grounded reasoning chains, with substantial Hit@1 gains (87.7% to 91.5% on WebQSP and 67.6% to 74.6% on CWQ).

[Long] ReasoningGuard: Safeguarding Large Reasoning Models with Inference-time Safety Aha Moments ▶ *Underline Video*
Yuquan Wang, Mi Zhang, Yining Wang, Geng Hong, Mi Wen, Xiaoyu You and Min Yang
Large Reasoning Models (LRMs) have demonstrated impressive performance in reasoning-intensive tasks, but they remain vulnerable to harmful content generation, particularly in the mid-to-late steps of their reasoning processes. Current defense methods, however, depend on costly fine-tuning and additional expert knowledge, which limits their scalability. In this work, we propose ReasoningGuard, an inference-time safeguard for LRMs. It injects timely safety aha moments during the reasoning process to guide the model towards harmless yet helpful reasoning. Our approach leverages the internal attention mechanisms of the LRM to accurately identify key points in the reasoning path, triggering safety-oriented reflections. To safeguard both the subsequent reasoning steps and the final answers, we implement a scaling sampling strategy during decoding to select the optimal reasoning path. With minimal additional inference cost, ReasoningGuard effectively mitigates four types of jailbreak attacks, including recent ones targeting the reasoning process of LRMs. Our approach outperforms nine existing safeguards, providing state-of-the-art defenses while avoiding common exaggerated safety issues.

[Long] ReContraster: Making Your Posters Stand Out with Regional Contrast ▶ *Underline Video*
Peixuan Zhang, Zijian Jia, Ziqi Cai, Shuchen Weng, Si Li and Boxin Shi
Effective poster design requires rapidly capturing attention and clearly conveying messages. Inspired by the "contrast effects" principle, we propose ReContraster, the first training-free model to leverage regional contrast to make posters stand out. By emulating the cognitive behaviors of a poster designer, ReContraster introduces the compositional multi-agent system to identify elements, organize layout, and evaluate generated poster candidates. To further ensure harmonious transitions across region boundaries, ReContraster integrates the hybrid denoising strategy during the diffusion process. We additionally contribute a new benchmark dataset for comprehensive evaluation. Seven quantitative metrics and four user studies confirm its superiority over relevant state-of-the-art methods, producing visually striking and aesthetically appealing posters.

[Long] ReEfBench: Quantifying the Reasoning Efficiency of LLMs ▶ *Underline Video*
Zhizhang Fu, Yuancheng Gu, Chenkai Hu, Hanmeng Liu and Yue Zhang
Test-time scaling has enabled Large Language Models (LLMs) to tackle complex reasoning, yet the limitations of current Chain-of-Thought (CoT) evaluation obscure whether performance gains stem from genuine reasoning or mere verbosity. To address this, (1) we propose a novel neuro-symbolic framework for the non-intrusive, comprehensive process-centric evaluation of reasoning grounded in First-Order Logic. (2) Through this lens, we identify four distinct behavioral prototypes and diagnose the failure modes. (3) We examine the impact of inference mode, training strategy, and model scale. Our analysis reveals that extended token generation is not a prerequisite for deep reasoning. Furthermore, we reveal critical constraints: mixing long and short CoT data in training risks premature saturation and collapse, while distillation into smaller models captures behavioral length but fails to replicate logical efficacy due to intrinsic capacity limits.

[Short] Reliable Use of Lemmas via Eligibility Reasoning and SectionAware Reinforcement Learning ▶ *Underline Video*
Zhikun Xu, Xiaodong Yu, Ben Zhou, Jiang Liu, Jialian Wu, Ze Wang, Ximeng Sun, Hao Chen and Zicheng Liu

Recent large language models (LLMs) perform strongly on mathematical benchmarks yet often misapply lemmas, importing conclusions without validating assumptions. We formalize lemmajudging as a structured prediction task: given a statement and a candidate lemma, the model must output a precondition check and a conclusionutility check, from which a usefulness decision is derived. We present RULES, which encodes this specification via a twosection output and trains with reinforcement learning plus sectionaware loss masking to assign penalty to the section responsible for errors. Training and evaluation draw on diverse naturallanguage and formal proof corpora; robustness is assessed with a heldout perturbation suite; and endtoend evaluation spans competitionstyle, perturbationaligned, and theorem-based problems across various LLMs. Results show consistent indomain gains over both a vanilla model and a singlelabel RL baseline, larger improvements on applicabilitybreaking perturbations, and parity or modest gains on endtoend tasks; ablations indicate that the twosection outputs and sectionaware reinforcement are both necessary for robustness.

[Industry] ReList: A Multi-objective Reasoning Framework for Diversified Listwise Query Recommendation ▶ *Underline Video*

Shuxian Bi, Chenxu Wang, Wenjie Wang, Yueqi Mou, Fuli Feng, Tang Biao and Peng Yan

Related search query recommendation is essential for enhancing user engagement and information discovery on digital platforms. While Large Language Models (LLMs) have shifted the field toward generative retrieval, existing methods suffer from two primary limitations: (1) pointwise generation via beam search often leads to semantic redundancy and wasted retrieval quota, and (2) current listwise approaches lack explicit reasoning, relying on superficial click-through rate (CTR) rewards. In this paper, we propose ReList, a novel framework that transforms related search into a reasoning-enhanced listwise generation task. ReList follows a two-stage training paradigm: first, Reasoning Activation constructs a high-quality dataset by back-translating diverse query lists into Chain-of-Thought (CoT) rationales; second, Alternative Training iteratively evolves the model using Reinforcement Learning with a Gated Multi-Objective Reward and a Corrective SFT mechanism to handle hard samples. Experimental results on real-world search benchmarks and online A/B tests demonstrate that ReList significantly outperforms state-of-the-art methods in both query diversity and user engagement, providing more insightful and logically grounded query recommendations.

[Long] RepoGenesis: Benchmarking End-to-End Microservice Generation from Readme to Repository ▶ *Underline Video*

Zhiyuan Peng, Xin Yin, Pu Zhao, Fangkai Yang, Lu Wang, Ran Jia, Xu Chen, Qingwei Lin, Saravan Rajmohan and Dongmei Zhang

Large language models and agents have achieved remarkable progress in code generation. However, existing benchmarks focus on isolated function/class-level generation (e.g., ClassEval) or modifications to existing codebases (e.g., SWE-Bench), neglecting complete microservice repository generation that reflects real-world 0-to-1 development workflows. To bridge this gap, we introduce RepoGenesis, the first multilingual benchmark for repository-level end-to-end web microservice generation, comprising 106 repositories (60 Python, 46 Java) across 18 domains and 11 frameworks, with 1,258 API endpoints and 2,335 test cases verified through a "review-rebuttal" quality assurance process. We evaluate open-source agents (e.g., DeepCode) and commercial IDEs (e.g., Cursor) using Pass@1, API Coverage (AC), and Deployment Success Rate (DSR). Results reveal that despite high AC (up to 73.91%) and DSR (up to 100%), the best-performing system achieves only 23.67% Pass@1 on Python and 21.45% on Java, exposing deficiencies in architectural coherence, dependency management, and cross-file consistency. Notably, RepoGenesis-8B, fine-tuned on RepoGenesis (train), achieves performance comparable to GPT-5 mini, demonstrating the quality of RepoGenesis for advancing microservice generation. We release our benchmark at https://github.com/pzy2000/RepoGenesis.

[Long] ReportLogic: Evaluating Logical Quality in Deep Research Reports ▶ *Underline Video*

Jujia Zhao, Zhaoxin Huan, Zihan Wang, Xiaolu Zhang, Jun Zhou, Suzan Verberne and Zhaochun Ren

Users increasingly rely on Large Language Models (LLMs) for Deep Research, using them to synthesize diverse sources into structured reports that support understanding and action. In this context, the practical reliability of such reports hinges on logical quality: whether the reports claims and arguments are explicitly supported and can be trusted as a basis for downstream use, rather than merely appearing fluent or informative. However, current evaluation frameworks largely overlook this requirement. To bridge this gap, we introduce ReportLogic, a benchmark that quantifies report-level logical quality through a reader-centric lens of auditability. Specifically, ReportLogic adopts a hierarchical taxonomy that evaluates whether readers can (1) trace an on-topic report structure with a unified analytical arc (Macro-Logic), (2) understand the progression with necessary context (Expositional-Logic), and (3) verify conclusions via explicit claim–support (Structural-Logic). Based on this taxonomy, we construct a human-annotated rubric-guided dataset and train an open-source LogicJudge for scalable evaluation. We further evaluate judge robustness via adversarial attacks, showing that off-the-shelf LLM judges are frequently influenced by superficial cues (e.g., verbosity), and reasoning modes can mask broken support relations. Overall, our results provide actionable guidance for building more robust logic evaluators and improving the logical reliability of LLM-generated reports.

[Long] Retrievals Can Be Detrimental: Unveiling the Backdoor Vulnerability of Retrieval-Augmented Diffusion Models ▶ *Underline Video*

Hao Fang, Xiaohang Sui, Hongyao Yu, Kuofeng Gao, Jiawei Kong, Sijin Yu, Bin Chen and Shutao Xia

Diffusion models (DMs) have recently exhibited impressive generation capability. However, their training generally requires huge computational resources and large-scale datasets. To solve these, recent studies empower DMs with Retrieval-Augmented Generation (RAG), yielding retrieval-augmented diffusion models (RDMs) that enhance performance with reduced parameters. Despite the success, RAG may introduce novel security issues that warrant further investigation. In this paper, we propose BadRDM, the first poisoning framework targeting RDMs, to systematically investigate their vulnerability to backdoor attacks. Our framework fully considers RAG's characteristics by manipulating the retrieved items for specific text triggers to ultimately control the generated outputs. Specifically, we first insert a tiny portion

of images into the retrieval database as target toxicity surrogates. We then exploit the contrastive learning mechanism underlying retrieval models by designing a malicious variant that establishes robust shortcuts from triggers to toxicity surrogates. In addition, we introduce novel entropy-based selection and generative augmentation strategies for better toxicity surrogates. Extensive experiments on two mainstream tasks show that the proposed method achieves outstanding attack effects while preserving benign utility. Notably, BadRDM remains effective even under common defense strategies, further highlighting serious security concerns for RDMs.

[Long] Reusable Experiences: Latent Routing and Modular Composition in LLMs ▶ *Underline Video*

Shuai_Ling Shuai_Ling, Lizi Liao, Dongmei Jiang and Weili Guan

Large language models (LLMs) have remarkable capabilities, but adapting them to specialized domains poses a fundamental question: how should accumulated experience be represented and leveraged? Existing approaches represent experience either as explicit textual artifacts in prompts (e.g., retrieved documents or dialogues) or implicitly within model weights via fine-tuning (e.g., LoRA adapters). However, textual methods are limited by context windows and cannot internalize knowledge, while parametric fine-tuning yields one adapter per task with minimal cross-task skill reuse. We propose ReX (Reusable eXperience), an experience-centric adaptation framework that treats latent experiences , recurring reasoning patterns and skills , as fundamental units for LLM specialization. Our method learns a shared Experience Bank of foundational skill vectors and uses a VAE-based encoder to map each input to a low-dimensional experience code. An Experience Router then dynamically composes the relevant skill vectors from this bank into a lightweight adapter for that input. By reusing skills across inputs, ReX enables implicit knowledge sharing across tasks without any explicit task identifiers. Experiments on multi-task NLP benchmarks show that this approach outperforms standard task-specific fine-tuning, yielding improved generalization through flexible skill reuse. Code is available at https://github.com/iLearn-Lab/ACL26-ReX.

[Long] SafeAgent: Safeguarding LLM Agents via an Automated Risk Simulator ▶ *Underline Video*

Xueyang Zhou, Weidong Wang, Lin Lu, Jiawen Shi, Guiyao Tie, Yongtian Xu, Lixing Chen, Pan Zhou, Neil Zhenqiang Gong and Lichao Sun

LLM-based agents are rapidly being deployed in real-world applications (e.g., digital assistants and customer service), making safety a critical concern. However, in multi-turn, tool-augmented settings, dynamic user interactions, external tool use, and unintended harmful behaviors make robust safety assurance challenging. To address these challenges, we propose SafeAgent, a framework that improves agent safety through fully automated synthetic data generation. SafeAgent introduces (1) an open and extensible threat model OTS that decomposes agent risk into instruction-, context-, and action-induced sources to ground safety analysis and alignment; and (2) an automated pipeline that instantiates OTS to surface scenario-specific failure modes, stress-test agents, and generate self-reflective safe responses, without hazardous real-world data collection. We evaluate SafeAgent on two safety benchmarks and one real-world terminal task. Across four widely used open-source models, SafeAgent improves safety performance by 45% on average and delivers a 28.91% gain on the real-world task, outperforming state-of-the-art closed-source models. These results highlight the practical advancement and scalability of SafeAgent in building safer LLM agents for real-world deployment.

[Long] SAIR-Comb: A Structure-Aware Iterative Refinement Framework for Combinatorics Autoformalization ▶ *Underline Video*

Weijie Jiang, Gaolei He, Beibei Xiong, Jianlin Wang and Zhengfeng Yang

Autoformalization aims to bridge the gap between human mathematical intuition and formal proof by automating the translation of informal reasoning into machine-verifiable languages. Despite significant breakthroughs catalyzed by Large Language Models (LLMs), autoformalizing Combinatorics remains a formidable challenge due to its intricate structural dependencies and the severe scarcity of high-quality formal datasets. To address these challenges, we propose SAIR-Comb, a Structure-Aware Iterative Refinement framework for Combinatorics powered by Lean 4 and LLMs. SAIR-Comb employs a multi-stage pipeline: first, it performs data augmentation and refinement by rectifying syntactic, semantic, and structural errors, guided by a curated manual combinatorics dataset. The model then undergoes a two-stage training regime: expert iteration with syntactic grounding, followed by reinforcement learning (RL) to align formal reasoning trajectories. Furthermore, we introduce Structural Consistency—a rigorous new metric designed to expose formalizing failures that elude traditional semantic-only evaluations. Experiments demonstrate that SAIR-Comb achieves strong performance on the specialized CombiBench while remaining highly competitive on general-domain benchmarks, including PutnamBench and ProverBench.

[Long] Scaling Laws for Code: A More Data-Hungry Regime ▶ *Underline Video*

Xianzhen Luo, Wenzhen Zheng, Qingfu Zhu, Rongyi Zhang, Houyi Li, Siming Huang, YuanTao Fan and Wanxiang Che

Code Large Language Models (LLMs) are revolutionizing software engineering. However, scaling laws that guide the efficient training are predominantly analyzed on Natural Language (NL). Given the fundamental differences like strict syntax between code and NL, it is unclear whether these laws are directly applicable to code. To address this gap, we conduct the first large-scale empirical study of scaling laws for code, comprising 117 experimental runs with model sizes from 0.2B to 3.8B and training tokens from 2B to 128B. We fit the Chinchilla law and the Farsser law. First, the results show that the more expressive Farseer law offers greater accuracy. Second, the analysis reveals that Code LLMs scale effectively with model size. Crucially, code represents a more data-hungry regime, requiring a substantially higher data-to-parameter ratio than NL. Finally, two additional sets of experiments on code-NL mixtures show that NL benefits resource-constrained scenarios, but becomes a detriment at higher compute budgets.

[Long] Self-Guided Alignment: Adaptive Preference Sensing for Multi-Objective Generation ▶ *Underline Video*

Ning Wang, Zhanyang Liu, Taotao Zhou, Xinrui Zhang, Zongru Shao and Haojie Zhou

Aligning Large Language Models (LLMs) with diverse and potentially conflicting human values necessitates navigating complex multi-objective landscapes. However, existing prompt-conditioned approaches face a critical training-inference discrepancy: they rely on ground-truth scores during training while requiring manual user-specification at inference. We introduce prediction of implicit preferences to bridge this gap while reducing user burden. To this end, we propose Self-Guided Alignment (SGA), a framework that transforms passive reward dependency into an intrinsic adaptive sensing capability. It employs a dual-head architecture to unify preference internalization with conditional generation, enabling the model to learn a latent mapping between raw prompts and preference profiles. Through adaptive preference sensing, the model autonomously predicts the latent preference score to self-guide the generation, thereby eliminating the need for manual specification at inference. Extensive experiments across diverse model scales demonstrate that SGA often outperforms state-of-the-art baselines, achieving superior multi-objective trade-offs and improved preference alignment. Code is available at https://github.com/python-yyds/SGA.

[Long] SGT: Securing Open-Source LLMs Against Malicious Fine-tuning via Safety Guidance Trigger ► *Underline Video*

Sunguk Shin, Fangzhao Wu, Byung-Jun Lee, Meeyoung Cha and Sungwon Park

Open-weight large language models (LLMs) enable broad customization, but also increase exposure to post-release misuse, including malicious fine-tuning (MFT). To mitigate this risk, many prior defenses aim to improve the robustness of open-weight models to MFT by constraining adversarial fine-tuning dynamics in parameter space or mitigating harmful information encoded in internal representations. Nevertheless, since malicious fine-tuning can still erode safety, developing robust safeguards for open-weight models that fundamentally mitigate this risk remains an open research problem. In this paper, we characterize a safety region for open-weight LLMs and propose Safety Guidance Trigger (SGT), which guides fine-tuning toward the safety manifold to preserve alignment. SGT has two stages: (1) optimizing a safety trigger that steers the base model toward safe responses and (2) training the open-weight model to align its internal features with trigger-induced safety representations. We demonstrate that SGT substantially improves robustness against malicious fine-tuning, requiring adversaries to increase their data budget significantly to compromise safety. Our analysis shows that SGT anchors model representations to a safety region, which remains stable under malicious fine-tuning.

[Long] SiLP: Enhancing Non-Dominant Language Capabilities with a Selective Bidirectional Language Projection Framework ► *Underline Video*

Junpeng Liu, Jiuyi Li, Kaiyu Huang, Bo Jin, Degen Huang and Hui Xiong

Current large language models (LLMs) often exhibit performance imbalances between dominant languages (e.g., English) and non-dominant ones due to the skewed distribution of pretraining data. A common strategy to address this issue is to enhance cross-lingual alignment, thereby facilitating non-dominant language processing. However, existing methods typically rely on additional training objectives or language-specific parameters, which increase training complexity and cost. In this work, we propose a selective bidirectional language projection framework that enables efficient multilingual alignment and language shift using the intrinsic parameters. Specifically, we first identify the layers most sensitive to language projection between non-dominant and dominant languages through neuron activation analysis. We then perform sequential language projection within the selected layers by mapping non-dominant representations into the dominant language space and reverting them before generation. The bidirectional projection benefits the subsequent instruction tuning in non-dominant languages. Experiments on seven benchmarks demonstrate that our method remarkably enhances the performance of non-dominant languages. Further analyses indicate that our method learns better internal representations and exhibits strong generalization capabilities.

[Long] SOS-LoRA: Static Orthogonal-Subspace Low-Rank Adaptation with Fixed Multi-Scale Scaling ► *Underline Video*

Yupeng Chang, Yuan Wu and Yi Chang

Low-Rank Adaptation (LoRA) is a widely used parameter-efficient fine-tuning (PEFT) method for large language models. Under a fixed rank budget, LoRA parameterizes each adapted weight through a single low-dimensional input-side pathway, which may couple heterogeneous behaviors through shared input directions and induce interference during optimization. We propose Static Orthogonal Subspace LoRA (SOS-LoRA), a drop-in extension that reparameterizes a rank-r_tot update as a sum of K static (always-on, non-routed) low-rank experts. SOS-LoRA (i) decomposes the total rank across experts, (ii) applies a fixed multi-scale scaling scheme to encourage scale-separated optimization dynamics, and (iii) promotes diverse input-side directions via cross-expert orthogonal initialization and a lightweight regularizer. SOS-LoRA remains fully mergeable, adding no inference-time parameters or latency after merging. Experiments on reasoning and knowledge-intensive benchmarks (Llama 2/3), encoder-based NLU (GLUE), and math reasoning (GSM8K/MATH) show consistent gains over matched-budget LoRA baselines and recent variants.

[Findings] SPaCe: Unlocking Sample-Efficient Large Language Models Training With Self-Pace Curriculum Learning ► *Underline Video*

Dai Do, Manh Nguyen, Svetha Venkatesh and Hung Le

Large language models (LLMs) have shown strong reasoning capabilities when fine-tuned with reinforcement learning (RL). However, such methods require extensive data and compute, making them impractical under many realistic training budgets. Many existing pipelines sample training examples uniformly across steps or epochs, ignoring differences in difficulty, redundancy, and learning value, which slows learning and wastes computation. We propose SPaCe, a self-paced learning framework that enables efficient learning based on the capability of the model being trained through optimizing which data to use and when. First, we apply cluster-based data reduction to partition training data by semantics and difficulty, extracting a compact yet diverse subset that reduces redundancy. Then, a multi-armed bandit treats data clusters as

arms, allocating training samples based on the model's solve rates and learning progress. Experiments across multiple reasoning benchmarks show that SPaCe achieves comparable or better accuracy than state-of-the-art baselines while using up to (100 times) fewer samples. Ablation studies and analyses further highlight the importance of both data clustering and adaptive selection. Our results demonstrate that carefully curated, performance-driven training curricula can unlock strong reasoning abilities in LLMs with minimal resources.

[Long] SPARK: Strategic Policy-Aware Exploration via Dynamic Branching for Long-Horizon Agentic Learning ▶ ***Underline Video***

Jinyang Wu, Shuo Yang, Yuhao Shen, Shuai Zhang, Zhengqi Wen and Jianhua Tao

Reinforcement learning has empowered large language models to act as intelligent agents, yet training them for long-horizon tasks remains challenging due to the scarcity of high-quality trajectories, especially under limited resources. Existing methods typically scale up rollout sizes and indiscriminately allocate computational resources among intermediate steps. Such attempts inherently waste substantial computation budget on trivial steps while failing to guarantee sample quality. To address this, we propose SPARK (Strategic Policy-Aware exploRation via Key-state dynamic branching), a novel framework that selectively branches at critical decision states for resource-efficient exploration. Our key insight is to activate adaptive branching exploration at critical decision points to probe promising trajectories, thereby achieving precise resource allocation that prioritizes sampling quality over blind coverage. This design leverages the agent's intrinsic decision-making signals to reduce dependence on human priors, enabling the agent to autonomously expand exploration and achieve stronger generalization. Experiments across diverse tasks (e.g., embodied planning), demonstrate that SPARK achieves superior success rates with significantly fewer training samples, exhibiting robust generalization even in unseen scenarios. Our code and checkpoints are available at https://github.com/jinyangwu/SPARK.

[Findings] Spatiotemporal Sycophancy: Negation-Based Gaslighting in Video Large Language Models ▶ ***Underline Video***

Ziyao Tang, Pengkun Jiao, Bin Zhu, Huiyan Qi, Jingjing Chen and Yu-Gang Jiang

Video Large Language Models (Vid-LLMs) have demonstrated remarkable performance in video understanding tasks, yet their robustness under conversational interaction remains largely underexplored. In this paper, we identify spatiotemporal sycophancy, a failure mode in which Vid-LLMs retract initially correct, visually grounded judgments and conform to misleading user feedback under negation-based gaslighting. Rather than merely changing their answers, the models often fabricate unsupported temporal or spatial explanations to justify incorrect revisions. To systematically investigate this phenomenon, we propose a negation-based gaslighting evaluation framework and introduce GasVideo-1000, a curated benchmark designed to probe spatiotemporal sycophancy with clear visual grounding and temporal reasoning requirements. We evaluate a broad range of state-of-the-art open-source and proprietary Vid-LLMs across diverse video understanding tasks. Extensive experiments reveal that vulnerability to negation-based gaslighting is pervasive and severe, even among models with strong baseline performance. While prompt-level grounding constraints can partially mitigate this behavior, they do not reliably prevent hallucinated justifications or belief reversal. Our results indicate that current Vid-LLMs lack robust mechanisms for maintaining grounded spatiotemporal beliefs under adversarial conversational feedback.

[Long] SpeechLLM-as-Judges: Towards General and Interpretable Speech Quality Evaluation ▶ ***Underline Video***

Hui Wang, Jinghua Zhao, Yifan Yang, Shujie Liu, Junyang Chen, Yanzhe Zhang, Shiwan Zhao, Jinyu Li, Jiaming Zhou, Haoqin Sun, Yan Lu and Yong Qin

Generative speech technologies are progressing rapidly, but evaluating the perceptual quality of synthetic speech remains a core challenge. Existing methods typically rely on scalar scores or binary decisions, which lack interpretability and generalization across tasks and languages. We present SpeechLLM-as-Judges, a new paradigm for enabling large language models (LLMs) to conduct structured and explanation-based speech quality evaluation. To support this direction, we introduce SpeechEval, a large-scale dataset containing 32,207 multilingual speech clips and 128,754 annotations spanning four tasks: quality assessment, pairwise comparison, improvement suggestion, and deepfake detection. Based on this resource, we develop SQ-LLM, a speech-quality-aware LLM trained with chain-of-thought reasoning and reward optimization to improve capability. Experimental results show that SQ-LLM delivers strong performance across tasks and languages, revealing the potential of this paradigm for advancing speech quality evaluation. The relevant code, models, and data are publicly available at https://github.com/NKU-HLT/SpeechLLM-as-Judges.

[Long] STAPO: Selective Trajectory-Aware Policy Optimization for LLM Agent Training ▶ ***Underline Video***

Qiuyi Qi, Tian Liang, Mutian Bao, Jinjian Zhang, Dongnan Liu, Wei Zhou, Linjian Mo, Ming Kong, Jie Liu, Feng Zhang and Qiang Zhu

Reinforcement Learning (RL) is the dominant paradigm for training Large Language Model (LLM) agents on long-horizon tasks. However, sparse and delayed rewards often lead to trajectory neglect, in which agents lose focus on the task goal and interaction history at intermediate steps. Prior work has explored step-level supervision using Shannon-entropy-based uncertainty signals, which conflate inherent state complexity with agent confidence and therefore provide unreliable estimates of decision reliability. To address this issue, we propose normalized entropy, which measures confidence deviations relative to an agents average behavior under a given state, thereby strengthening the association between low-quality actions and trajectory neglect. Building on this insight, we introduce Selective Trajectory-Aware Policy Optimization (STAPO), a hierarchical group-based RL framework. STAPO leverages normalized entropy to locate outlier steps associated with trajectory neglect and optimizes them via a joint mechanism of trajectory-aware reward and trajectory-independent penalty, enhancing trajectory awareness while preserving training stability. Extensive experiments on ALFWorld, WebShop, and Search-Augmented QA demonstrate that STAPO achieves state-of-the-art performance while substantially alleviating trajectory neglect, validating its effectiveness and robustness for agentic tasks.

[Long] Stratagem: Learning Transferable Reasoning via Trajectory-Modulated Game Self-Play ▶ *Underline Video*

Xiachong Feng, Deyi Yin, Xiaocheng Feng, Yi Jiang, Libo Qin, Yangfan Ye, Lei Huang, Weitao Ma, Qiming Li, Yuxuan Gu, Bing Qin and Lingpeng Kong

Games offer a compelling paradigm for developing general reasoning capabilities in language models, as they naturally demand strategic planning, probabilistic inference, and adaptive decision-making. However, existing self-play approaches rely solely on terminal game outcomes, providing no mechanism to distinguish transferable reasoning patterns from game-specific heuristics. We present STRATAGEM, which addresses two fundamental barriers to reasoning transfer: domain specificity, where learned patterns remain anchored in game semantics, and contextual stasis, where static game contexts fail to cultivate progressive reasoning. STRATAGEM selectively reinforces trajectories exhibiting abstract, domain-agnostic reasoning through a Reasoning Transferability Coefficient, while incentivizing adaptive reasoning development via a Reasoning Evolution Reward. Experiments across mathematical reasoning, general reasoning, and code generation benchmarks demonstrate substantial improvements, with particularly strong gains on competition-level mathematics where multi-step reasoning is critical. Ablation studies and human evaluation confirm that both components contribute to transferable reasoning.

[Findings] SWE-Mutation: Can LLMs Generate Reliable Test Suites in Software Engineering? ▶ *Underline Video*

Yuxuan Sun, Yuze Zhao, Yufeng Wang, Yao Du, Zhiyuan Ma, Jinbo Wang, Mengdi Zhang, Kai Zhang and Zhenya Huang

Evaluating software engineering capabilities has become a core component of modern large language models (LLMs); however, the key bottleneck hindering further scaling lies not in the scarcity of high-quality solutions, but in the lack of high-quality test suites. Test suites are indispensable both for synthesizing program repair trajectories and for providing precise feedback signals in reinforcement learning. Unfortunately, due to the high cost and difficulty of annotation, high-quality test suites have long been hard to obtain, while those automatically generated by LLMs tend to be superficial and lack sufficient discriminative power. As a first step toward constructing high-quality test suites, we introduce SWE-Mutation, a benchmark for evaluating LLM-generated test suites. The benchmark characterizes test suites by introducing systematically mutated solutions that attempt to "fool" the test suites and pass validation. We further propose an agentic, language-agnostic framework for automatically generating complex mutants. Our benchmark consists of 2,636 mutated variants derived from 800 original instances and includes a multilingual subset spanning nine programming languages. Experiments on seven LLMs reveal that even DeepSeek-V3.1 achieves only 10.20% verification and 36.15% detection rates, highlighting the inadequacy of current LLMs. Additionally, our agentic mutation strategy enhances realism, reducing average detection rates from 71.04% to 39.81% compared to conventional methods. These findings expose persistent deficiencies in the ability of current LLMs to generate reliable and discriminative test suites.

[Long] TalkLoRA: Communication-Aware Mixture of Low-Rank Adaptation for Large Language Models ▶ *Underline Video*

Lin Mu, Haiyang Wang, Li Ni, Lei Sang, Zhize Wu, Peiquan Jin and Yiwen Zhang

Low-Rank Adaptation (LoRA) enables parameter-efficient fine-tuning of Large Language Models (LLMs), and recent Mixture-of-Experts (MoE) extensions further enhance flexibility by dynamically combining multiple LoRA experts. However, existing MoE-augmented LoRA methods assume that experts operate independently, often leading to unstable routing, expert dominance. In this paper, we propose TalkLoRA, a communication-aware MoELoRA framework that relaxes this independence assumption by introducing expert-level communication prior to routing. TalkLoRA equips low-rank experts with a lightweight Talking Module that enables controlled information exchange across expert subspaces, producing a more robust global signal for routing. Theoretically, we show that expert communication smooths routing dynamics by mitigating perturbation amplification while strictly generalizing existing MoELoRA architectures. Empirically, TalkLoRA consistently outperforms vanilla LoRA and MoELoRA across diverse language understanding and generation tasks, achieving higher parameter efficiency and more balanced expert routing under comparable parameter budgets. These results highlight structured expert communication as a principled and effective enhancement for MoE-based parameter-efficient adaptation.

[Findings] Task-Related In-Context Learning ▶ *Underline Video*

Wenqiang Wang, Peng Chen, Yan XIAO, Yangshijie Zhang, Xiaoyue Lu, Jianjie huang and Xiaochun Cao

Standard in-context learning (ICL) assumes identical output spaces between test and retrieval datasets (fully aligned). However, in practice, these datasets can be fully aligned, partially aligned, or fully disjoint in label space (Output space), forming an information continuum from rich to scarce. Naive ICL often becomes ineffective under such mismatches. In this work, we challenge this assumption by demonstrating that the retrieval dataset need not perfectly align with the test dataset, as long as it remains related to the target task. We propose Task-Related In-Context Learning (TRICL), a unified framework for ICL under output-space mismatch, designed to cover the full continuum of scenarios. TRICL first identifies demonstrations in the mismatched retrieval dataset that are relevant to the test label space via a lightweight Bayesian probabilistic criterion, and uses them to form a related dataset. TRICL then perform ICL on the related dataset to obtain preliminary predictions; finally, TRICL leverage these intermediate predictions to reduce and transform the output space of the original test task, thereby improving the performance of LLMs. Even in the most information-scarce fully disjoint scenario, as long as the retrieval dataset is task-related to the test task, TRICL achieves state-of-the-art (SOTA) results across three LLMs, three task types, and four datasets. Moreover, TRICL remains effective in the fully aligned and partially aligned scenarios, consistently yielding strong gains over competitive baselines. Moreover, TRICL also extends to generative task.

[Long] Team-Based Self-Play With Dual Adaptive Weighting for Fine-Tuning LLMs ▶ *Underline Video*

Wu Li, Yigeng Zhou, Zesheng Shi, Yequan Wang, Min Zhang and Jing Li
While recent self-training approaches have reduced reliance on human-labeled data for aligning LLMs, they still face critical limitations: (i) sensitivity to synthetic data quality, leading to instability and bias amplification in iterative training; (ii) ineffective optimization due to a diminishing gap between positive and negative responses over successive training iterations. In this paper, we propose Team-based self-Play with dual Adaptive Weighting (TPAW), a novel self-play algorithm designed to improve alignment in a fully self-supervised setting. TPAW adopts a team-based framework in which the current policy model both collaborates with and competes against historical checkpoints, promoting more stable and efficient optimization. To further enhance learning, we design two adaptive weighting mechanisms: (i) a response reweighting scheme that adjusts the importance of target responses, and (ii) a player weighting strategy that dynamically modulates each team members contribution during training. Initialized from a SFT model, TPAW iteratively refines alignment without requiring additional human supervision. Experimental results demonstrate that TPAW consistently outperforms existing baselines across various base models and LLM benchmarks.

[Long] TeCES: Collaborative Geometric Knowledge Representation Framework under Evolving Fact Snapshots ▶ *Underline Video*
Jiujiang Guo, Zhengliang Guo, Kai Wang, Meiyang Wang, Dehua Peng, Shaozu Yuan, Chengyin Hu, Shuan Ai and Yiwei Wei
Existing knowledge graph completion research is gradually shifting from representing logical semantics of static facts to modeling evolving semantics of temporal facts, yet lacks collaborative modeling of both within a unified framework. To this end, we use concept of snapshots to decompose fact features into two complementary mechanisms: (a) intra-snapshot semantic coupling, where entities and relations exhibit snapshot-specific meanings through multidimensional interactions; (b) trans-snapshot evolutionary synergy, where relations between entities evolve across snapshots and manifest varying states. These snapshot mechanisms jointly reveal underlying logic of facts. To track them, we propose TeCES, a framework for high-fidelity modeling of evolving snapshots. TeCES embeds facts into a 2-grade geometric algebra (GA) system to capture complex semantics via multilevel structures. Temporal information is attached to each entity for mapping into snapshot spaces, while relations and timestamps are reconfigured into composite GA representations. Geometric products enable multidimensional interactions, revealing relation state changes over time. Lastly, the head entity at each snapshot combines with fused temporal-relational representation via geometric product to approximate the target tail entity at multiple levels. Overall, TeCES supports joint modeling of evolving snapshots within a lightweight GA system and significantly outperforms SOTA models on six benchmarks.

[Long] TED-TTS: Training-Free Intra-Utterance Emotion and Duration Control for Text-to-Speech Synthesis ▶ *Underline Video*
Qifan Liang, Yuansen Liu, Ruixin Wei, Nan Lu, Junchuan Zhao and Ye Wang
While controllable Text-to-Speech (TTS) has achieved notable progress, most existing methods remain limited to inter-utterance-level control, making fine-grained intra-utterance expression challenging due to their reliance on non-public datasets or complex multi-stage training. In this paper, we propose TED-TTS, a training-free controllable framework for pretrained zero-shot TTS to enable intra-utterance emotion and duration expression. Specifically, we propose a segment-aware emotion conditioning strategy that combines causal masking with monotonic stream alignment filtering to isolate emotion conditioning and schedule mask transitions, enabling smooth intra-utterance emotion shifts while preserving global semantic coherence. Based on this, we further propose a segment-aware duration steering strategy to combine local duration embedding steering with global EOS logit modulation, allowing local duration adjustment while ensuring globally consistent termination. To eliminate the need for segment-level manual prompt engineering, we construct a 30,000-sample multi-emotion and duration-annotated text dataset to enable LLM-based automatic prompt construction. Extensive experiments demonstrate that our training-free method not only achieves state-of-the-art intra-utterance consistency in multi-emotion and duration control, but also maintains baseline-level speech quality of the underlying TTS model. Audio samples are available at https://aclanonymous111.github.io/TED-TTS-DemoPage/.

[Long] Temp-R1: A Unified Autonomous Agent for Complex Temporal KGQA via Reverse Curriculum Reinforcement Learning ▶ *Underline Video*
Zhaoyan Gong, Zhiqiang Liu, Songze Li, Xiaoke Guo, Yuanxiang Liu, Xinle Deng, Zhizhen Liu, Lei Liang, Huajun Chen and Wen Zhang
Temporal Knowledge Graph Question Answering (TKGQA) is inherently challenging, as it requires sophisticated reasoning over dynamic facts with multi-hop dependencies and complex temporal constraints. Existing methods rely on fixed workflows and expensive closed-source APIs, limiting flexibility and scalability. We propose Temp-R1, the first autonomous end-to-end agent for TKGQA trained through reinforcement learning. To address cognitive overload in single-action reasoning, we expand the action space with specialized internal actions alongside external action. To prevent shortcut learning on simple questions, we introduce reverse curriculum learning that trains on difficult questions first, forcing the development of sophisticated reasoning before transferring to easier cases. Our 8B-parameter Temp-R1 achieves state-of-the-art performance on MultiTQ and TimelineKGQA, improving 19.8% over strong baselines on complex questions. Our work establishes a new paradigm for autonomous temporal reasoning agents. The code is available at https://github.com/zjukg/Temp-R1.

[Long] Temporal Precision Matters: Brain-Tuning Speech Language Models with Millisecond-Resolution Neural Signals ▶ *Underline Video*
Zhejun Zhang, Wenqing Zhou, Haozhe Xu, Lin Zhang and Lei Li
Brain-tuning enhances brain alignment and downstream performance by fine-tuning speech language models with neural recordings. However, previous work relies primarily on fMRI, whose temporal resolution integrates neural activity over seconds, blending distinct processing stages into a single supervision signal and precluding temporally targeted training. We introduce ECoG-tuning, which leverages

electrocorticography's millisecond precision to train speech language models. We design temporally targeted windows, a speech window capturing acoustic-phonetic encoding and a language window capturing higher-order linguistic processing, grounded in neuroscientific findings about temporal encoding hierarchies. Evaluating three models on the Podcast ECoG dataset, we find that ECoG-tuning significantly improves brain alignment over pretrained and distillation baselines. Notably, full spatiotemporal dynamics yield 7-17% higher alignment than time-averaged supervision across models, and language-window tuning produces larger gains in higher-order language regions, indicating that temporal precision provides additional training value. Moreover, ECoG-tuned models consistently improve or maintain downstream performance. Overall, our work provides initial evidence that electrophysiology is a viable brain-tuning modality, demonstrating how neuroscientific insights into processing hierarchies can inform principled model training strategies. Code is available at https://github.com/Mochizuki-BUPT/ECoG-Tuning-main.

[Findings] TemporalVLM: Video LLMs for Temporal Reasoning in Long Videos ▶ *Underline Video*

Fawad Fateh, Umer Ahmed, Hamza Khan, Zeeshan Zia and Quoc-Huy Tran

We introduce TemporalVLM, a video large language model (video LLM) for temporal reasoning and fine-grained understanding in long videos. Our approach includes a visual encoder for mapping a long-term video into features which are time-aware and contain both local and global cues. It first divides an input video into short-term clips, which are jointly encoded with timestamps and fused across overlapping temporal windows into time-sensitive local features. Next, the local features are passed through a bidirectional long short-term memory (BiLSTM) module for global feature aggregation. Moreover, to facilitate the evaluation of TemporalVLM, we present a large-scale long video dataset of industry assembly processes, namely IndustryASM, consisting of videos recorded on factory floors with actions and timestamps annotated by industrial engineers for time and motion studies and temporal action segmentation evaluation. Finally, extensive experiments show that TemporalVLM outperforms previous methods across temporal reasoning and fine-grained understanding tasks, i.e., dense video captioning, temporal video grounding, video highlight detection, and temporal action segmentation.

[Long] Text-Guided Multi-Scale Frequency Representation Adaptation ▶ *Underline Video*

Weicai Yan, Xinhua Ma, Wang Lin and Tao Jin

Parameter-efficient fine-tuning methods introduce a small number of training parameters, enabling pre-trained models to adapt rapidly to new data distributions. While these methods have shown promising results, they exhibit notable limitations. First, most existing methods operate in the signal space domain, which results in substantial information redundancy. Second, most existing methods utilize fixed prompts or adaptation layers, failing to fully account for the multi-scale characteristics of signals. To address these challenges, we propose the Multi-Scale Frequency Adapter (FreqAdapter), which integrates textual information and performs multi-scale fine-tuning of visual signal in the frequency domain. Additionally, we introduce a multi-scale adaptation strategy to optimize receptive fields across different frequency ranges, further enhancing the model's representational capacity. Extensive experiments on multimodal models, including CLIP and LLaVA, demonstrate that FreqAdapter significantly improves both performance and efficiency. FreqAdapter improves performance with minimal cost and fast convergence within one epoch.

[Long] The Evolution of Thought: Tracking LLM Overthinking via Reasoning Dynamics Analysis ▶ *Underline Video*

Zihao Wei, Liang Pang, Jiahao Liu, Wenjie Shi, Jingcheng Deng, Shicheng Xu, Zenghao Duan, Jingang Wang, Fei Sun, Huawei Shen and Xueqi Cheng

Test-time scaling via explicit reasoning trajectories significantly boosts large language model (LLM) performance but often triggers overthinking. To explore this, we analyze reasoning through two lenses: Reasoning Length Dynamics, which reveals a compensatory trade-off between thinking and answer content length that eventually leads to thinking redundancy, and Reasoning Semantic Dynamics, which identifies semantic convergence and repetitive oscillations. These dynamics uncover an instance-specific Reasoning Completion Point (RCP), beyond which computation continues without further performance gain. Since the RCP varies across instances, we propose a Reasoning Completion Point Detector (RCPD), an inference-time early-exit method that identifies the RCP by monitoring the rank dynamics of termination tokens (e.g., </think>). Across AIME and GPQA benchmarks using Qwen3 and DeepSeek-R1, RCPD reduces token usage by up to 44% while preserving accuracy, offering a principled approach to efficient test-time scaling.

[Long] Theory-optimal Quantization Based on Flatness ▶ *Underline Video*

Xiusheng Huang, Zhe Li, Xuanwu Yin, Lu Wang, Yequan Wang, Dong Li, Emad Barsoum and Kang Liu

Post-training quantization has emerged as a widely adopted technique for compressing and accelerating the inference of Large Language Models (LLMs). The primary challenges in LLMs quantization stem from activation outliers, which significantly degrade model performance especially at lower bit precision. While recent approaches attempt to mitigate outliers through linear transformations across feature dimensions, our analysis reveals that the transformed weights and activations still exhibit persistent outlier patterns with concentrated magnitude distributions. In this paper, we first model the mathematical relationship between quantization error and outliers, and then introduce a new metric Flatness to quantify the distribution of outliers. Based on this, we derive the theoretical optimal solution with respect to Flatness. Building on these insights, we propose Bidirectional Diagonal Quantization (BDQ), a novel post-training quantization framework that effectively disperses outlier patterns through optimized matrix transformations. BDQ strategically distributes outlier magnitudes across matrix dimensions via learned diagonal operations. Extensive experiments demonstrate that BDQ establishes a new quantization benchmark. It achieves less than 1% accuracy drop in W4A4 quantization on the LLaMA-3-8B model. In the more challenging W2A4KV16 experiment, compared to state-of-the-art approaches, BDQ reduces the performance gap by 39.1% on the DeepSeek-R1-Distill-LLaMA-70B model.

[SRW] Thesis Proposal: A Normalization-First Framework for Sound, Complete, and Utility-Ready Open Information Extraction ▶ *Underline Video*

Chandan Prakash, Pavan Kumar Chittimalli and Arnab Bhattacharya;

Open Information Extraction (OIE) has largely focused on extracting relational tuples from text, yet in its current form remains unsuitable for downstream systems due to the absence of standardized, semantically sound representations. This thesis argues that the field has been addressing extraction as a surface-level prediction problem, leading to outputs that are semantically incomplete and logically ambiguous, particularly in the presence of modality, negation, conditionality, quantification, and attribution. We propose a normalization-first framework that reframes OIE as a structured semantic transformation pipeline, where raw text is first converted into a lossless, canonical form of declarative, active-voice, and irreducible sentence units, and extraction is constrained to atomic unary and binary relations augmented with explicit semantic annotations. Within a Probably Approximately Correct (PAC) learning perspective, we formalize soundness, completeness, and usefulness as approximate yet verifiable guarantees over extraction quality, acknowledging the inherent undecidability of full semantic interpretation. This thesis outlines a feasible research program to develop the theoretical foundations, models, and evaluation protocols required to produce system-ready OIE representations, thereby establishing a principled and executable path toward making OIE directly usable for downstream reasoning and machine interpretability.

[SRW] Thesis Proposal: On the Granularity-Robustness Trade-off in Text-Derived Knowledge Graphs ▶ *Underline Video*

Surawat Pralomram

Retrieval-augmented generation (RAG) based on dense embeddings has become a dominant paradigm for text retrieval. However, many real-world applications require attribute-specific querying, where explicit values or properties must be extracted from text (e.g., symptoms in clinical notes or dosage values in medical reports). Dense retrieval handles paraphrastic variation well but often entangles multiple attributes within a single embedding, making value extraction difficult. Knowledge graphs (KGs), in contrast, support explicit attribute access but are brittle under linguistic and structural variation, leading to low recall. This thesis proposal aims to investigate the representational trade-off underlying these approaches. We study knowledge graph representations from an information-theoretic and optimal coding perspective, focusing on the tension between fine-grained factorization and compact canonicalization of concepts. Building on this perspective, we propose a query-driven framework for constructing and retrieving knowledge graphs from text, aiming to combine the robustness of dense retrieval with the explicit queryability of symbolic representations.

[Long] Think Faster Than Words: Efficient LLM Chain-of-Thought Reasoning via Dynamic Shortcut Decoding ▶ *Underline Video*

Fan Liu, Yanhao Wang, Min Zhang, Zhikang Chen, Zeyuan Li, Lewei He and Jiahui Pan

This paper proposes shortcut decoding, an efficient framework for accelerating Chain-of-Thought (CoT) reasoning in Large Language Models (LLMs). Existing methods that prune or employ early stopping to reduce latency often compromise reasoning reliability. Motivated by the observation that LLMs frequently converge to correct solutions internally before completing explicit textual reasoning, we propose a dual-signal adaptive controller that integrates lightweight probes over internal hidden states with step-level entropy. This controller detects convergence of reasoning during generation and adaptively selects between a fast-exit path and a stability-verified path to remove redundant steps while preserving answer correctness. Experiments across multiple mathematical reasoning benchmarks demonstrate that shortcut decoding reduces token usage by approximately 35%, maintains accuracy comparable to full CoT decoding, and achieves final-answer accuracy comparable to the full CoT baseline, outperforming existing early-stopping methods without updating the base model. Our code is available at https://github.com/kuromi9527/shortcut_decoding.

[Long] Think in Sentences: Explicit Sentence Boundaries Enhance Language Model's Capabilities ▶ *Underline Video*

Zhichen Liu, Yongyuan Li and Yang Xu

Researchers have explored ways to improve large language models (LLMs)' capabilities via dummy token insertion in contexts. However, existing works focus solely on the dummy tokens themselves, but failed to leverage the inherent sentence-level structure of natural language. This is a critical oversight, as LLMs acquire linguistic capabilities through exposure to human-generated texts, which are inherently structured at the sentence level. Motivated by the gap, we proposed a method that inserts delimiters at sentence boundaries. Our method not only integrates dummy tokens into contexts, but also enables LLMs with sentence-by-sentence processing behavior during reasoning. Two approaches are proposed: (1). In-context learning and (2). Supervised fine-tuning are experimented from 7B LLMs to 600B Deepseek-V3. Experimental results demonstrate consistent improvements in various tasks, with notable gains of up to 7.7% on GSM8k and 12.5% on DROP. Furthermore, LLMs fine-tuned via our strategy further incorporate sentence awareness into their inner representations. Our work establishes a simple yet effective technique for enhancing LLM's capabilities, offering promising directions for cognitive-inspired LLM enhancement paradigm.

[Long] To Think or Not to Think: The Hidden Cost of Meta-Training with Excessive CoT Examples ▶ *Underline Video*

Vignesh Kothapalli, Ata Fatahibaarzi, Hamed Firooz and Maziar Sanjabi

Chain-of-thought (CoT) prompting combined with few-shot in-context learning (ICL) has unlocked significant reasoning capabilities in large language models (LLMs). However, ICL with CoT examples is ineffective on novel tasks when the pre-training knowledge is insufficient. We study this problem in a controlled setting using the CoT-ICL Lab framework, and propose meta-training techniques to learn novel abstract reasoning tasks in-context. Although CoT examples facilitate reasoning, we noticed that their excessive inclusion during meta-training

degrades performance when CoT supervision is limited. To mitigate such behavior, we propose CoT-Recipe, a formal approach to modulate the mix of CoT and non-CoT examples in meta-training sequences. We demonstrate that careful modulation via CoT-Recipe can increase the accuracy of transformers on novel tasks by up to 300% even when there are no CoT examples available in-context. We confirm the broader effectiveness of these techniques by applying them to pretrained LLMs (Qwen2.5 series) for symbolic reasoning tasks and observing gains of up to 130% in accuracy.

[Long] Token-Level Policy Optimization: Linking Group-Level Rewards to Token-Level Aggregation via sequence-level likelihood ▶ *Underline Video*

Xingyu Lin, Yilin Wen, Du Su, Jinchang Hou, En Wang, Wenbin Liu, Zhonghou Lv and Chenfu Bao

Group Relative Policy Optimization (GRPO) has significantly advanced the reasoning ability of large language models (LLMs), particularly in their mathemat- ical reasoning performance. However, GRPO and related entropy regularization methods still struggle with token-level sparse-rewards, which is an inherent challenge in chain-of-thought (CoT) reasoning. These approaches often rely on undifferentiated token-level entropy regu- larization, which easily leads to entropy collapse or model degradation under sparse token rewards. In this work, we propose TEPO, a novel token-level framework that (1) leverages sequence-level likelihood to link group-level rewards with individual tokens via token-level aggregation, and (2) introduces a token-level KL-Divergence mask constraint that targets tokens with positive advantages and decreasing entropy to mitigate abrupt policy updates. Experiments demonstrate that TEPO not only achieves state-of-the-art performance on mathematical reasoning benchmarks but also markedly enhances training stability, reducing convergence time by 50% compared with GRPO/DAPO.

[Short] Too Correct to Learn: Reinforcement Learning on Saturated Reasoning Data ▶ *Underline Video*

Zhenwen Liang, Yujun Zhou, Sidi Lu, Xiangliang Zhang, Haitao Mi and Dong Yu

Reinforcement Learning (RL) enhances LLM reasoning, yet a paradox emerges as models scale: strong base models saturate standard benchmarks (e.g., MATH), yielding correct but homogeneous solutions. In such environments, the lack of failure cases causes the advantage signal in group-relative algorithms (e.g., GRPO) to vanish, driving policies into mode collapse. To address this, we propose Constrained Uniform Top-K Sampling (CUTS), a parameter-free decoding strategy enforcing structure-preserving exploration. Unlike standard sampling that follows model biases, CUTS flattens the local optimization landscape by sampling uniformly from constrained high-confidence candidates. We integrate this into Mixed-CUTS, a training framework synergizing exploitative and exploratory rollouts to amplify intra-group advantage variance. Experiments on Qwen3 models demonstrate that our approach prevents policy degeneration and significantly boosts out-of-domain generalization. Notably, Mixed-CUTS improves Pass@1 accuracy on the challenging AIME25 benchmark by up to 15.1% over standard GRPO, validating that maintaining diversity within the semantic manifold is critical for rigorous reasoning.

[Long] Too Nice to Tell the Truth: Quantifying Agreeableness-Driven Sycophancy in Role-Playing Language Models ▶ *Underline Video*

Arya Shah, Deepali Mishra and Chaklam Silpasuwanchai

Large language models increasingly serve as conversational agents that adopt personas and role-play characters at user request. This capability, while valuable, raises concerns about sycophancy: the tendency to provide responses that validate users rather than prioritize factual accuracy. While prior work has established that sycophancy poses risks to AI safety and alignment, the relationship between specific personality traits of adopted personas and the degree of sycophantic behavior remains unexplored. We present a systematic investigation of how persona agreeableness influences sycophancy across 13 small, open-weight language models ranging from 0.6B to 20B parameters. We develop a benchmark comprising 275 personas evaluated on NEO-IPIP agreeableness subscales and expose each persona to 4,950 sycophancy-eliciting prompts spanning 33 topic categories. Our analysis reveals that 9 of 13 models exhibit statistically significant positive correlations between persona agreeableness and sycophancy rates, with Pearson correlations reaching $r = 0.87$ and effect sizes as large as Cohen's $d = 2.33$. These findings demonstrate that agreeableness functions as a reliable predictor of persona-induced sycophancy, with direct implications for the deployment of role-playing AI systems and the development of alignment strategies that account for personality-mediated deceptive behaviors

[Long] Toward Secure Tuning: Mitigating Security Risks from Instruction Fine-Tuning ▶ *Underline Video*

yanrui du, fenglei fan, sendong zhao, jiawei cao, ming ma, danyang zhao, shuren qi, ting liu and bing qin

Instruction Fine-Tuning (IFT) has emerged as a critical technique for customizing Large Language Models (LLMs) to meet diverse downstream applications. However, recent studies have revealed that IFT can compromise the built-in security mechanisms of LLMs, thereby posing significant security risks. Although defense methods targeting various training stages have been proposed, they either face challenges in practical deployment or exhibit instability and limited performance gains. In our study, we propose a novel SWAT method that introduces a key idea: shifting more of the learning burden onto security-robust parameters. To this end, our study investigates how module-level parameters affect LLMs' internal security feature space, aiming to uncover robustness patterns in parameters. Guided by this analysis, we identify a robust module set (Mods_Rob) that exhibits minimal effects on LLMs' security feature space. Leveraging this insight, SWAT proceeds in two phases: (1) a warm-up phase that preferentially trains Mods_Rob to learn low-level features with minimal security risk, followed by (2) standard tuning to achieve optimal task performance. Across diverse knowledge-intensive datasets, scenarios, and LLMs, SWAT substantially reduces security risks without sacrificing task performance gains.

[Long] Towards Efficient and Effective Diffusion Language Model Inference via Semantic-Aware Adaptive Denoising

► *Underline Video*

Fan Li, Yu Gu, Zhigang Wang, Fangling Leng, Zhenghao Liu and Ge Yu

Diffusion language models (DLMs) have emerged as a powerful non-autoregressive alternative to GPT-style sequential generation, but suffer from substantial computational overhead due to their iterative parallel denoising. Existing acceleration works cannot accurately detect semantically stabilized tokens and then skip computation, leading to sub-optimal speedup in practice. This paper presents the first systematic study of convergence dynamics in DLMs. Innovative observations include the misalignment between traditionally used scalar detection criterion and the semantic convergence, and the post-peak confidence score, that wastes denoising computation and degrades inference quality. To address these limitations, we propose AdaDLM, a semanticaware adaptive denoising framework that encodes the trajectory of scalar confidence scores into an evolution-aware feature vector and then clusters vectors proactively and adaptively identify semantically converged tokens. Furthermore, we incorporate system-level optimizations to maximize runtime efficiency. Experiments show that AdaDLM consistently outperforms the SOTA competitor, achieving up to 2x speedup and 19% quality improvement. That offers a practical path toward efficient highquality DLM deployment.

[Long] Towards Explainable Diagnosis: A Self-learned Explanatory Knowledge Base Approach ► *Underline Video*

Dongqi Huang, Tong Zhou, Zhuoran Jin, Shenghui Shi, Yujiao Mao, Kang Liu, Jun Zhao and Yubo Chen

Explainable diagnosis requires that authoritative medical knowledge provide the rationales linking a patients clinical manifestations to the diagnostic conclusion. Although large language models (LLMs) hold great potential to facilitate explainable diagnosis, their effectiveness is often constrained by insufficient diagnostic expertise. To address this limitation, we propose Self-learned Explainable Knowledge Augmented Diagnosis (SEKAD), a unified LLM-based framework for faithful and explainable diagnosis. Our approach builds a high-quality diagnostic knowledge base through a record-driven explanation learning paradigm, as well as applies this knowledge via an explanation-based diagnostic process that ensures faithful inference. Experiments on the DiReCT and JAMA benchmarks show that SEKAD consistently outperforms strong baselines across the metrics. In particular, on the DiReCT benchmark, SEKAD improves the explanation completeness metric from 64.5% to 76.9% over the best existing methods, highlighting its effectiveness in enhancing diagnostic explainability and showing that our text mining approach produces knowledge that is both reliable in quality and large in quantity.

[Findings] Towards Inference-time Scaling for Continuous Space Reasoning ► *Underline Video*

Minghan Wang, Thuy-Trang Vu, Ehsan Shareghi and Gholamreza Haffari

Inference-time scaling through multiple sample generation in combination with Process- or Outcome-Reward Model (PRM or ORM) re-ranking has proven effective for text-based reasoning in large language models. This paper investigates whether such established techniques can be successfully adapted to reasoning in the continuous space, using COCONUT~hao2024traininglargelanguagemodels continuous space reasoning LM as the backbone. We demonstrate the feasibility of generating diverse reasoning paths through dropout-based sampling. Our Pass@N analysis on the generated samples reveals the potential that could enable a significant gain in performance akin to gains observed in the discrete space. However, we highlight unique challenges faced for materializing this gain in the continuous thought space. In particular, working recipes for data generation and training PRM and ORM models in the discrete space unlocks only marginal improvements in the continuous space. Through probing various aspects including geometric properties and trajectory dynamics, we identify the underlying reasons that prevent effective discrimination between correct and incorrect reasoning (essential for the functioning of PRM and ORM). Our findings reveal that current limitations stem from the absence of key inductive biases in continuous thought representations.

[Long] Trait Activation in Silicon: A Situation-Aware Framework for Psychologically Grounded Role-Playing ► *Underline Video*

Zuolong Li, Pingyu Wu, Xianwen Huang, Tianyi Wei and Wenbo Zhou

Role-playing agents (RPAs) have made significant strides in mimicking static character identities. However, their personality simulations remain superficial, lacking a profound understanding of complex human psychological mechanisms. We identify a critical bottleneck termed "Personality Inertia", a behavioral rigidity where RLHF-induced alignment bias traps models in a sanitized, "helpful assistant" persona. This inertia prevents models from adapting to diverse social contexts or expressing essential but negative traits under pressure. To bridge this gap, we propose PD-LLM, a situation-aware framework grounded in Trait Activation Theory. PD-LLM introduces Bipolar Latent Decomposition, which decouples personality traits into bidirectional LoRA adapters. These adapters are dynamically modulated by a situation-aware module based on the DIAMONDS taxonomy, allowing for precise behavioral regulation. Empirical results show that while baseline methods fail to synchronize multidimensional traits under pressure, PD-LLM achieves superior performance in both static fidelity and dynamic adaptability. By advancing from prompt engineering to intrinsic parameter control, PD-LLM effectively overcomes personality rigidity, facilitating the creation of vivid and psychologically consistent agents.

[Long] Trajectory Signatures of Deception in Large Language Models ► *Underline Video*

Viraaji Mothukuri and Reza M. Parizi

Detecting deceptive behavior in LLMs is typically done post-hoc on outputs or by probing static activations. We instead treat deception as a dynamic process, a trajectory through the model's hidden-state space during inference. We capture layerwise activations at sparse "decision points" where the model is uncertain between competing tokens, forming activation trajectories for matched truthful vs. deceptive responses across strategic deception, sycophancy, instructed deception, and confabulation. Across GPT-2 and Llama variants, deceptive generation is associated with changes in trajectory geometry, but increases in path length are model and deception-type-dependent. Sycophancy shows the clearest signal, whereas instructed deception yields near-null signatures. With just 7 geometric features, a lightweight classifier achieves

performance comparable to PCA-reduced probing at matched dimensionality for binary sycophancy detection and shows preliminary utility for 4-way deception-type classification. These findings indicate that trajectory-based monitoring can provide process-level signals associated with deceptive generation during inference, complementing methods that focus on endpoint activation states.

[Findings] Trident: Self-Supervised Preference Alignment via Triplet Regularization ▶ *Underline Video*

Yingnan Guo, Kejia Chen, Xiaofeng Zhang, Zifei Wu and Yu Zhang

Aligning Large Vision-Language Models (LVLMs) to mitigate hallucinations typically relies on high-quality preference data. However, in self-supervised settings, standard binary preference optimization (e.g., DPO) suffers from noisy supervision and semantic ambiguity, as automatically generated chosen responses are not guaranteed to be superior to rejected ones. In this work, we propose Trident, a fully self-supervised framework that ensures robust alignment via a structured triplet paradigm. Trident autonomously constructs reliable preference triplets, comprising semantically enriched (chosen), degraded (rejected), and neutral (anchor) responses, through automated visual perturbations and self-summarization. We further introduce Trident Preference Regularization (TPR), a novel objective that utilizes an adaptive margin to enforce semantic separation between the triplet components while preventing deviation from the pretrained distribution. Despite requiring no human annotations or external reward models, Trident consistently outperforms state-of-the-art RLHF and RLAIF baselines. For instance, on LLaVA-1.5-7B, it reduces the hallucination rate on AMBER to 11.3% and achieves 95.70% precision on POPE using only 4k self-generated triplets and a single epoch. This validates structured triplet supervision as a scalable paradigm for robust self-supervised alignment.

[Long] TriEx: A Game-based Tri-View Framework for Explaining Internal Reasoning in Multi-Agent LLMs ▶ *Underline Video*

Ziyi Wang, Chen Zhang, Wenjun Peng, Qi Wu and Xinyu Wang

Explainability for Large Language Model (LLM) agents is especially challenging in interactive, partially observable settings, where decisions depend on evolving beliefs and other agents. We present TriEx, a tri-view explainability framework that instruments sequential decision making with aligned artifacts: (i) structured first-person self-reasoning bound to an action, (ii) explicit second-person belief states about opponents updated over time, and (iii) third-person oracle audits grounded in environment-derived reference signals. This design turns explanations from free-form narratives into evidence-anchored objects that can be compared and checked across time and perspectives. Using imperfect-information strategic games as a controlled testbed, we show that TriEx enables scalable analysis of explanation faithfulness, belief dynamics, and evaluator reliability, revealing systematic mismatches between what agents say, what they believe, and what they do. Our results highlight explainability as an interaction-dependent property and motivate multi-view, evidence-grounded evaluation for LLM agents.

[Long] TrustTable: A Neuro-Symbolic Auditing Framework for Faithful Table QA ▶ *Underline Video*

Guangzhen Zhao, Dechang Kong, Tongyu Wu and Zhenjiang Dong

Large Language Models based Table Question Answering (LLMs-based TableQA) models excel in NLP field, however, they occasionally exhibit an unfaithful behavior where correct answers are derived through erroneous reasoning paths. In this condition, we propose TrustTable, a neuro-symbolic framework designed to ensure reasoning faithfulness by auditing the reasoning processes of LLMs. Unlike monolithic LLM-based auditors, TrustTable decouples the auditing operation into two orthogonal dimensions. It enforces factual grounding by executing neurally generated Pandas code against the table, and ensures logical soundness by verifying reasoning chains through a LLM-synthesized formal solver. By integrating these symbolic checks, TrustTable enables a Label-Free Audit Loop that systematically identifies and rectifies reasoning flaws without human supervision. In addition, we present the TrustTable-Bench, a diagnostic dataset containing diverse error categories that range from calculation discrepancies to schema misalignments. This benchmark allows for a rigorous quantification of reasoning limitations. Experiments demonstrate that our symbolic audit detects reasoning flaws more accurately than advanced baselines. More broadly, the TrustTable outperforms LLM judges in both majority voting with logical weighting and rejection sampling with process supervision.

[Long] Truth or Sophistry? LoFa: A Benchmark for LLM Robustness Against Logical Fallacies ▶ *Underline Video*

Xudong Shen, Li Yuan, Ye Chen, Xin Wu, Yi Cai and Zhiyong Wu

While Large Language Models (LLMs) exhibit strong semantic capabilities, their resilience to manipulative linguistic patterns such as logical fallacies remains an underexplored area. Prior work has focused on the ability of LLMs to identify or classify fallacies, but their robustness against these fallacies in persuasive contexts remains largely unexplored. To address this gap, we introduce LoFa (Logical Fallacy), a comprehensive benchmark to evaluate LLM robustness against fallacies. We first construct the LoFa dataset via a multi-agent pipeline, pairing factual questions with fallacious arguments. Then, we develop a multi-round debate framework to assess model resilience under sustained attacks. Furthermore, to disentangle robustness from a models inherent knowledge limitations, we propose a new metric, LFR@k (Logical Fallacy Resistance), to quantify performance. Our experiments reveal that different LLMs exhibit varied robustness to distinct types of fallacies, highlighting unique vulnerability profiles across models.

[Findings] TTVS: Boosting Self-Exploring Reinforcement Learning via Test-time Variational Synthesis ▶ *Underline Video*

Sikai Bai, Haoxi Li, Jie ZHANG, Yongjiang Liu and Song Guo

Despite significant advances in Large Reasoning Models (LRMs) driven by reinforcement learning with verifiable rewards (RLVR), this paradigm is fundamentally limited in specialized or novel domains where such supervision is prohibitively expensive or unavailable, posing a key challenge for test-time adaptation. While existing test-time methods offer a potential solution, they are constrained by learning

from static query sets, risking overfitting to textual patterns. To address this gap, we introduce Test-Time Variational Synthesis (TTVS), a novel framework that enables LRMs to self-evolve by dynamically augmenting the training stream from unlabeled test queries. TTVS comprises two synergistic modules: (1) Online Variational Synthesis, which transforms static test queries into a dynamic stream of diverse, semantically-equivalent variations, enforcing the model to learn underlying problem logic rather than superficial patterns; (2) Test-time Hybrid Exploration, which balances accuracy-driven exploitation with consistency-driven exploration across synthetic variants. Extensive experiments show TTVS yields superior performance across eight model architectures. Notably, using only unlabeled test-time data, TTVS not only surpasses other test-time adaptation methods but also outperforms state-of-the-art supervised RL-based techniques trained on vast, high-quality labeled data.

[Findings] Understanding LLMs' summarization capabilities: an analysis of biomedical abstract and lay summary generation ▶ *Underline Video*

Batuhan Nursal and Cassie S. Mitchell

Scientific abstracts and lay summaries serve distinct but critical roles in research communication. Abstracts use technical language for academic audiences, while lay summaries aim to make findings accessible to non-specialists. With the rise of large language models (LLMs), there is increasing interest in automating the generation of both types of summaries, especially in the biomedical domain, where clarity and factual accuracy are essential. This study evaluates the performance of lightweight LLMs (under 8B parameters) in generating biomedical abstracts and lay summaries in a zero-shot setting. We assess outputs across three key dimensions: relevance, readability, and factuality. Additionally, we introduce a novel analysis of the sectional origin and desirability of information, where desirability reflects the utility of content from the reader's perspective. We further compare human and LLM preferences using an objective ranking task. Our results show that LLM-generated summaries often contain comparable levels of desirable information to gold-standard human references. In several cases, LLM outputs are preferred by human evaluators and occasionally mistaken for human-authored text. These findings demonstrate the potential of lightweight LLMs for scalable, high-quality summarization and suggest their practical use in domains requiring both technical and accessible communication. The codebase for this study is publicly available on GitHub: https://github.com/batuinmetz/Understanding-LLMs-summarization-capabilities

[Long] Unlearners Can Lie: Evaluating and Improving Honesty in LLM Unlearning ▶ *Underline Video*

Renjie Gu, Jiazhen Du, Yihua Zhang and Sijia Liu

Unlearning in large language models (LLMs) aims to remove harmful training data while preserving overall utility. However, we find that existing methods often hallucinate, generate abnormal token sequences, or behave inconsistently, raising safety and trust concerns. According to prior literature on LLM honesty, such behaviors are often associated with dishonesty. This motivates us to investigate the notion of honesty in the context of model unlearning. We propose a formal definition of unlearning honesty, which includes: (1) preserving both utility and honesty on retained knowledge, and (2) ensuring effective forgetting while encouraging the model to acknowledge its limitations and respond consistently to questions related to forgotten knowledge. To systematically evaluate the honesty of unlearning, we introduce a suite of metrics that cover utility, honesty on the retained set, effectiveness of forgetting, rejection rate and refusal stability in Q&A and MCQ settings. Evaluating 9 methods across 3 mainstream families shows that all current methods fail to meet these standards. After experimental and theoretical analyses, we present ReVa, a representation-alignment procedure that fine-tunes feature-randomized unlearned models to better acknowledge forgotten knowledge. On Q&A tasks from the forget set, ReVa achieves the highest rejection rate after two rounds of interaction, nearly doubling the performance of the second-best method. Remarkably, It also improves honesty on the retained set.

[Long] Unlocking Multilingual Reasoning Capability of LLMs and LVLMs through Representation Engineering ▶ *Underline Video*

Qiming Li, Xiaocheng Feng, Yixuan Ma, Ruihan Chen, Zihe Tong, Zekai Ye, Xiachong Feng, Libo Qin, Haoyu Ren, Kun Chen, Yunfei Lu, Dandan Tu and Bing Qin

Large Language Models (LLMs) and Large Vision-Language Models (LVLMs) demonstrate strong reasoning capabilities, yet their performance in English significantly outperforms that in low-resource languages, raising fairness concerns in multilingual applications. Existing approaches either rely on costly multilingual training or employ prompting with external translation tools, both of which are resource-intensive and sensitive to translation quality. To address these limitations, we propose a training-free inference-time method to enhance Multilingual Reasoning capabilities via Representation Engineering (MRRE) without using any additional training data or tools. MRRE sequentially injects two precomputed vectors at specific layers during inference processing: cross-lingual reasoning enhancement vectors, which steer non-English reasoning representations toward English space to unlock multilingual reasoning, and target-language output anchoring vectors, which restore the distribution of the target language to preserve input-output language consistency. Comprehensive experiments across six advanced LLMs and LVLMs on four reasoning benchmarks demonstrate that MRRE consistently enhances non-English reasoning by an average gain of 5.48% and up to 7.54% in low-resource languages (e.g., Thai and Swahili), while improving input-output language consistency by 3.78%.

[Long] Unlocking Multilingual Reasoning Capability of LLMs and LVLMs through Representation Engineering ▶ *Underline Video*

Qiming Li, Xiaocheng Feng, Yixuan Ma, Ruihan Chen, Zihe Tong, Zekai Ye, Xiachong Feng, Libo Qin, Haoyu Ren, Kun Chen, Yunfei Lu, Dandan Tu and Bing Qin

Large Language Models (LLMs) and Large Vision-Language Models (LVLMs) demonstrate strong reasoning capabilities, yet their performance in English significantly outperforms that in low-resource languages, raising fairness concerns in multilingual applications.

Existing approaches either rely on costly multilingual training or employ prompting with external translation tools, both of which are resource-intensive and sensitive to translation quality. To address these limitations, we propose a training-free inference-time method to enhance Multilingual Reasoning capabilities via Representation Engineering (MRRE) without using any additional training data or tools. MRRE sequentially injects two precomputed vectors at specific layers during inference processing: cross-lingual reasoning enhancement vectors, which steer non-English reasoning representations toward English space to unlock multilingual reasoning, and target-language output anchoring vectors, which restore the distribution of the target language to preserve input-output language consistency. Comprehensive experiments across six advanced LLMs and LVLMs on four reasoning benchmarks demonstrate that MRRE consistently enhances non-English reasoning by an average gain of 5.48% and up to 7.54% in low-resource languages (e.g., Thai and Swahili), while improving input-output language consistency by 3.78%.

[Long] Unlocking the Black Box of Latent Reasoning: An Interpretability-Guided Approach to Intervention ▶ *Underline Video*

Shuochen Chang, Tong Bai, Xiaofeng Zhang, Qianli Ma, Qingyang Liu, Zhaohe Liao, Yibo Miao and Li Niu

Latent reasoning enables Large Language Models (LLMs) to perform multi-step inference within continuous hidden states, offering efficiency gains over explicit Chain-of-Thought (CoT). However, the opacity of these continuous thought vectors hinders their reliability and controllability. This paper bridges the gap between mechanistic interpretability and actionable control. We first present a systematic analysis using structural, causal, and geometric probes, revealing that latent vectors encode compressed, faithful representations of reasoning steps, with early vectors acting as critical causal hubs. Building on this, we operationalize these interpretability insights into a suite of training-free, decode-time interventions that refine the latent reasoning process by imposing the identified geometric and semantic priors. Extensive experiments across multiple model scales and diverse task domains demonstrate that our approaches consistently improve reasoning accuracy. Our interpretability-guided interventions consistently unlock latent capabilities and improve reasoning accuracy without any parameter updates.

[Findings] Verifying the Subjective: Structured Multilingual Rewards for Low-Resource Alignment ▶ *Underline Video*

Jiu Sha, Yu Weng, Zeshuo Liu, Jingyi Jiang, Zheng Liu and Mengxiao Zhu

Aligning LLMs in low-resource multilingual settings faces a fundamental reward bottleneck: scalar rewards lack cultural generalization, while unstructured critiques remain noisy and unverifiable. To bridge this gap, we introduce a Structured Multilingual Reward Modeling Framework that extends Reinforcement Learning with Verifiable Rewards (RLVR) to subjective and open-ended tasks. The framework unifies three core components to transform abstract quality into concrete supervision: (1) a Structured Checklist Schema decomposing evaluation into granular universal reasoning steps and task-specific criteria; (2) Structured Generative Critique Modeling, which produces rubric-aligned critiques with grounded justifications; and (3) Adaptive Multilingual Reward Optimization, integrating reasoning quality and language consistency into a verifiable objective. We integrate this framework into a bootstrapped Group Relative Policy Optimization pipeline, augmented by length-aware normalization and variance stabilization to ensure stability. Extensive experiments on a newly constructed suite covering 7 subjective task categories across 50 low-resource languages demonstrate that this checklist-driven approach yields substantial improvements in reasoning capability and response quality, particularly in settings where traditional reward models exhibit significant degradation. We publicly release our models and the corresponding evaluation benchmark to facilitate further research. Our code is available at https://github.com/Shajiu/SGCM.

[Long] Video-MMMU: Evaluating Knowledge Acquisition from Multidisciplinary Professional Videos ▶ *Underline Video*

Kairui Hu, Penghao Wu, Fanyi Pu, Wang Xiao, Xiang Yue, Bo Li, Yuanhan Zhang and Ziwei Liu

Humans acquire knowledge through three cognitive stages: perceiving information, comprehending knowledge, and adapting knowledge to solve novel problems. Videos serve as an effective medium for knowledge acquisition, facilitating a natural progression through these learning stages. However, existing video benchmarks fail to evaluate the knowledge acquisition capabilities of Large Multimodal Models (LMMs). To address this gap, we introduce Video-MMMU, a multi-modal, multi-discipline, multi-track benchmark that evaluates LMMs' ability to acquire knowledge from college-level, educational videos. Video-MMMU features a collection of 300 videos and 900 human-annotated questions across six disciplines, evaluating knowledge acquisition through stage-aligned question-answer pairs: Perception, Comprehension, and Adaptation. Beyond measuring final accuracy, Video-MMMU proposes the performance gain metric that quantifies an LMMs learning gain from video, shifting the focus of evaluation from absolute performance to learning efficiency. Our evaluation reveals a substantial gap between human learners and current LMMs, highlighting the need to improve models' ability to learn and adapt knowledge from video content.

[Long] VisPCO: Visual Token Pruning Configuration Optimization via Budget-Aware Pareto-Frontier Learning for Vision-Language Models ▶ *Underline Video*

Huawei Ji, Yuanhao Sun, Yuan Jin, Cheng Deng, Jiaxin Ding, Luoyi Fu and Xinbing Wang

Visual token pruning methods effectively mitigate the quadratic computational growth caused by processing high-resolution images and video frames in vision-language models (VLMs). However, existing approaches rely on predefined pruning configurations without determining whether they achieve computation-performance optimality. In this work, we introduce, a novel framework that formulates visual token pruning as a Pareto configuration optimization problem to automatically identify optimal configurations. Our approach employs continuous relaxation and straight-through estimators to enable gradient-based search, solved via the Augmented Lagrangian method. Extensive experiments across 8 visual benchmarks demonstrate that effectively approximates the empirical Pareto frontier obtained through

grid search and generalizes well across various pruning methods and VLM architectures. Furthermore, through learnable kernel functions, we investigate layer-wise pruning patterns and reveal that multi-step progressive pruning captures VLMs' hierarchical compression structure, achieving superior accuracy-efficiency trade-offs compared to single-layer approaches.

[Long] Vista-LLM: Decoupled Query-Guided Visual Token Pruning for Efficient Long-Video Large Language Models ► *Underline Video*

Zhenyu Li, Zuchao Li, Ping Wang, Lefei Zhang and Haojun Ai

Long-video understanding is bottlenecked by the high cost of processing massive visual tokens. Current reduction strategies often rely on static allocation or inefficient in-network selection that disrupts optimized attention kernels. In this paper, we introduce Vista-LLM, a decoupled framework for query-guided visual token pruning. By filtering redundancy prior to inference with minimal overhead, Vista-LLM ensures full compatibility with Flash Attention. Our method employs a coarse-to-fine pipeline: (1) Query-Guided Dynamic Budgeting for adaptive temporal allocation; (2) a lightweight Semantic Scout for fine-grained, query-specific selection; and (3) Structure-Aware Compensation to preserve global context. Extensive experiments on benchmarks like Video-MME and MLVU demonstrate a significantly improved Pareto frontier. Notably, on LLaVA-OneVision, Vista-LLM reduces visual tokens by 90% and accelerates inference while retaining over 98% of baseline performance on average, effectively filtering visual noise.

[Long] Visual Inception: Compromising Long-term Planning in Agentic Recommenders via Multimodal Memory Poisoning ► *Underline Video*

Jiachen Qian

The evolution from static ranking models to Agentic Recommender Systems (Agentic RecSys) empowers AI agents to maintain long-term user profiles and autonomously plan service tasks. While this paradigm shift enhances personalization, it introduces a vulnerability: reliance on Long-term Memory (LTM). In this paper, we uncover a threat termed Visual Inception. Unlike traditional adversarial attacks that seek immediate misclassification, Visual Inception injects triggers into user-uploaded images (e.g., lifestyle photos) that act as sleeper agents within the systems memory. When retrieved during future planning, these poisoned memories hijack the agents reasoning chain, steering it toward adversary-defined goals (e.g., promoting high-margin products) without prompt injection. To mitigate this, we propose CognitiveGuard, a dual-process defense framework inspired by human cognition. It consists of a System 1 Perceptual Sanitizer (diffusion-based purification) to cleanse sensory inputs and a System 2 Reasoning Verifier (counterfactual consistency checks) to detect anomalies in memory-driven planning. Extensive experiments on a mock e-commerce agent environment demonstrate that Visual Inception achieves about 85% Goal-Hit Rate (GHR), while CognitiveGuard reduces this risk to around 10% with configurable latency trade-offs (about 1.5s in lite mode to about 6.5s for full sequential verification), without quality degradation under our setup.Latency reporting uses separate accounting: query-time overhead excludes one-time upload-time preprocessing.

[Long] VLN-MME: Diagnosing MLLMs as Language-guided Visual Navigation Agents ► *Underline Video*

Xunyi Zhao, Gengze Zhou and Qi Wu

Multimodal Large Language Models (MLLMs) have demonstrated remarkable capabilities across a wide range of vision-language tasks. However, their performance as embodied agents, which requires multi-round interaction with spatial reasoning and sequential action prediction, needs further exploration. Our work investigates this potential in the context of Vision-and-Language Navigation (VLN) by introducing a unified and extensible simulation-free evaluation framework to probe MLLMs as zero-shot agents, named VLN-MME. Simplifying the evaluation with a highly modular and accessible design streamlines experiments, enabling structured comparisons and component-level ablations across diverse MLLM architectures, agent designs, and navigation tasks. Crucially, enabled by VLN-MME, we observe that enhancing prevalent agents with Chain-of-Thought (CoT) reasoning and self-reflection leads to an unexpected performance decrease. This suggests MLLMs exhibit poor context awareness in embodied navigation tasks; although they can follow instructions and structure their output, their 3D spatial reasoning fidelity is low. Furthermore, we demonstrate that agent performance could be largely improved with simple failure cases in context learning. VLN-MME lays the groundwork for systematic evaluation of general-purpose MLLMs in embodied navigation settings and reveals limitations in their sequential decision-making capabilities. We believe these findings offer crucial guidance for MLLM post-training as embodied agents.

[Long] Vocabulary Hijacking in LVLMs: Unveiling Critical Attention Heads by Excluding Inert Tokens to Mitigate Hallucination ► *Underline Video*

Yangneng Chen, Junlin Li, Weijun Yao, Xilai Ma, Guodong DU, Wenya Wang and Jing Li

Large Vision-Language Models (LVLMs) have achieved remarkable progress in multimodal tasks, yet their reliability is persistently undermined by hallucinations, generating text that contradicts visual input. Recent studies often attribute these errors to inadequate visual attention. In this work, we analyze the attention mechanisms via the logit lens, uncovering a distinct anomaly we term Vocabulary Hijacking. We discover that specific visual tokens, defined as Inert Tokens, disproportionately attract attention. Crucially, when their intermediate hidden states are projected into the vocabulary space, they consistently decode to a fixed set of unrelated words (termed Hijacking Anchors) across layers, revealing a rigid semantic collapse. Leveraging this semantic rigidity, we propose Hijacking Anchor-Based Identification (HABI), a robust strategy to accurately localize these Inert Tokens. To quantify the impact of this phenomenon, we introduce the Non-Hijacked Visual Attention Ratio (NHAR), a novel metric designed to identify attention heads that remain resilient to hijacking and are critical for factual accuracy. Building on these insights, we propose Hijacking-Aware Visual Attention Enhancement (HAVAE), a training-free intervention

that selectively strengthens the focus of these identified heads on salient visual content. Extensive experiments across multiple benchmarks demonstrate that HAVAE significantly mitigates hallucinations with no additional computational overhead, while preserving the model's general capabilities.

[Long] WebSynthesis: World Model-Guided Monte Carlo Tree Search for Efficient WebAgent Trajectory Synthesis ▶ *Underline Video*

Yifei Gao, Junhong Ye, Yifan Yang, Jiaqi Wang, Yi Zhang, Zhang Ruichen and Jitao Sang

Recent advances in large language models (LLMs) have enabled increasingly capable web agents, yet training such agents still relies on high-quality interaction trajectories that are difficult to obtain at scale. We identify two key challenges: (1) Infrastructure Overhead, where network instability and website access restrictions limit data collection scalability; and (2) Constrained Exploration, where irreversible state transitions preclude tree-based search and thus limit trajectory diversity. To address these challenges, we introduce WebSynthesis, a framework for scalable trajectory synthesis. WebSynthesis employs an LLM-based World Model to simulate state transitions without network dependencies, and integrates Monte Carlo Tree Search to enable reversible exploration over the simulated state space. Experiments on WebArena, WebVoyager, and Mind2Web-Online demonstrate that agents trained exclusively on synthesized trajectories outperform those trained on real-world data, providing a viable alternative to costly real-world data collection.

[Findings] WenetSpeech-Wu: Datasets, Benchmarks, and Models for a Unified Chinese Wu Dialect Speech Processing Ecosystem ▶ *Underline Video*

Chengyou Wang, Mingchen Shao, Jingbin Hu, Zeyu Zhu, Hongfei Xue, Bingshen Mu, Xin Xu, Xingyi Duan, Binbin Zhang, Pengcheng Zhu, Chuang Ding, Xiaojun Zhang, Hui Bu and Lei Xie

Speech processing for low-resource dialects remains a fundamental challenge in developing inclusive and robust speech technologies. Despite its linguistic significance and large speaker population, the Wu dialect of Chinese has long been hindered by the lack of large-scale speech data, standardized evaluation benchmarks, and publicly available models. In this work, we present WenetSpeech-Wu, the first large-scale, multi-dimensionally annotated open-source speech corpus for the Wu dialect, comprising approximately 8,000 hours of diverse speech data. Building upon this dataset, we introduce WenetSpeech-Wu-Bench, the first standardized and publicly accessible benchmark for systematic evaluation of Wu dialect speech processing, covering automatic speech recognition (ASR), Wu-to-Mandarin translation, speaker attribute prediction, speech emotion recognition, text-to-speech (TTS) synthesis, and instruction-following TTS (instruct TTS). Furthermore, we release a suite of strong open-source models trained on WenetSpeech-Wu, establishing competitive performance across multiple tasks and empirically validating the effectiveness of the proposed dataset. Together, these contributions lay the foundation for a comprehensive Wu dialect speech processing ecosystem, and we open-source proposed datasets, benchmarks, and models to support future research on dialectal speech intelligence.

[Long] What If Consensus Lies? Selective-Complementary Reinforcement Learning at Test Time ▶ *Underline Video*

Dong Yan, Jian Liang, Yanbo Wang, Shuo Lu, Ran He and Tieniu Tan

Test-Time Reinforcement Learning (TTRL) enables Large Language Models (LLMs) to enhance reasoning capabilities on unlabeled test streams by deriving pseudo-rewards from majority voting consensus. However, existing TTRL methods rely exclusively on positive pseudo-labeling strategies. Such reliance becomes vulnerable under challenging scenarios where answer distributions are highly dispersed, resulting in weak consensus that inadvertently reinforces incorrect trajectories as supervision signals. In this paper, we propose SCRL (Selective-Complementary Reinforcement Learning), a robust test-time reinforcement learning framework that effectively mitigates label noise amplification. SCRL develops Selective Positive Pseudo-Labeling, which enforces strict consensus criteria to filter unreliable majorities. Complementarily, SCRL introduces Entropy-Gated Negative Pseudo-Labeling, the first negative supervision mechanism in TTRL, to reliably prune incorrect trajectories based on generation uncertainty. Extensive experiments on mathematical reasoning benchmarks demonstrate that SCRL achieves substantial improvements over baselines, while maintaining robust generalization and training stability under constrained rollout budgets.

[Long] What's Left Unsaid? Detecting and Correcting Misleading Omissions in Multimodal News Previews ▶ *Underline Video*

Fanxiao Li, Jiaying Wu, Tingchao Fu, Dayang Li, Herun Wan, Wei Zhou and Min-Yen Kan

Even when factually correct, social-media news previews (image-headline pairs) can induce interpretation drift: by selectively omitting crucial context, they lead readers to form judgments that diverge from what the full article supports. This covert harm is subtler than explicit misinformation, yet remains underexplored. To address this gap, we develop a multi-stage pipeline that simulates preview-based and context-based understanding, enabling construction of the MM-Misleading benchmark. Using MM-Misleading, we systematically evaluate open-source LVLMs and uncover pronounced blind spots in omission-based misleadingness detection. We further propose OMGuard, which combines (1) Interpretation-Aware Fine-Tuning for misleadingness detection and (2) Rationale-Guided Misleading Content Correction, where explicit rationales guide headline rewriting to reduce misleading impressions. Experiments show that OMGuard lifts an 8B models detection accuracy to the level of a 235B LVLM while delivering markedly stronger end-to-end correction. Further analysis shows that misleadingness usually arises from local narrative shifts, such as missing background, instead of global frame changes, and identifies image-driven cases where text-only correction fails, underscoring the need for visual interventions.

[Long] When Morphology Hides in Plain Sight: Breaking the Isolation in Vietnamese and Beyond ▶ *Underline Video*

Anh Trac Duc Dinh, Khang Hoang Nhat Vo, Tai Tien Ta, Vinh Cong Doan and Tho Quan
In isolating languages such as Vietnamese, core morphological structure is encoded not by inflection but by the composition and ordering of monosyllabic morphemes, yet standard Transformer encoders largely overlook this signal. We introduce HuTieuBERT, a morpheme-aware Transformer that augments a pretrained Vietnamese encoder with two lightweight inductive biases: (i) Adaptive Boundary-Token Fusion, which integrates BMES-based morpheme boundary embeddings into token representations via a learnable gate, and (ii) a Morpheme-Aware Attention Bias, which injects a fixed structural attention matrix into early self-attention layers while minimally perturbing the pretrained attention geometry. Across a suite of Vietnamese POS, NER, and sentence-level classification benchmarks, HuTieuBERT consistently outperforms strong baselines, with the largest gains on syntactic tasks. Hyperparameter ablations show a broad regime in which structural biases improve accuracy without destabilizing representations. Applying the same design to ChineseBERT (Chinese-BERT-wwm) yields MAChineseBERT, which improves F_1 and produces more balanced tag distributions on Chinese POS and NER, suggesting that explicit morpheme-aware attention is a portable and effective strategy for modeling isolating languages.

[Long] When Seeing Is not Enough: Revealing the Limits of Active Reasoning in MLLMs ▶ *Underline Video*
Hongcheng Liu, Pingjie Wang, Yuhao Wang, Siqu Ou, Yanfeng Wang and Yu Wang
Multimodal large language models (MLLMs) have shown strong capabilities across a broad range of benchmarks. However, most existing evaluations focus on passive inference, where models perform step-by-step reasoning under complete information. This setup is misaligned with real-world use, where seeing is not enough. This raises a fundamental question: Can MLLMs actively acquire missing evidence under incomplete information? To bridge this gap, we require the MLLMs to actively acquire missing evidence and iteratively refine decisions under incomplete information, by selecting a target image from a candidate pool without task-specific priors. To support systematic study, we propose GuessBench, a benchmark with both perception-oriented and knowledge-oriented images for evaluating active reasoning in MLLMs. We evaluate 20 superior MLLMs and find that performance on active reasoning lags far behind it on passive settings, indicating substantial room for improvement. Further analysis identifies fine-grained perception and timely decision-making as key challenges. Ablation studies show that perceptual enhancements benefit smaller models, whereas thinking-oriented methods provide consistent gains across model sizes. These results suggest promising directions for future research on multimodal active reasoning.

[SRW] Why Large Language Models can Secretly Outperform Embedding Similarity in Information Retrieval ▶ *Underline Video*
Matei Benescu; Ivo Pascal de Jong
With the emergence of Large Language Models (LLMs), new methods in Information Retrieval are available in which relevance is estimated directly through language understanding and reasoning, instead of embedding similarity. We argue that similarity is a short-sighted interpretation of relevance, and that LLM-Based Relevance Judgment Systems (LLM-RJS) (with reasoning) have potential to outperform Neural Embedding Retrieval Systems (NERS) by overcoming this limitation. Using the TREC-DL 2019 passage retrieval dataset, we compare various LLM-RJS with NERS, but observe no noticeable improvement. Subsequently, we analyze the impact of reasoning by comparing LLM-RJS with and without reasoning. We find that human annotations also suffer from short-sightedness, and that false-positives in the reasoning LLM-RJS are primarily mistakes in annotations due to short-sightedness. We conclude that LLM-RJS do have the ability to address the short-sightedness limitation in NERS, but that this cannot be evaluated with standard annotated relevance datasets.

[Long] You Can Have a Second Chance: Unbiased and Multi-bit Watermarking for Diffusion Language Models with Regret-based Remasking ▶ *Underline Video*
Ke Yang, Dongyang Liang, Jing Yu, Shuguang Yuan and Chi Chen
The rapid development of Diffusion Language Models (DLMs) raises concerns about watermarking for DLM-generated detection. However, existing sequential LLM watermarking cannot be directly applied to DLMs, as DLMs' generation order is arbitrary. While emerging studies adapt biased LLM watermarking to DLMs by temporarily predicting the watermark prefix, they suffer from degraded quality and unstable watermarking due to bias accumulation and prediction errors. Besides, they cannot carry multi-bit watermarks. In this paper, we propose unbiased multi-bit watermarking for DLMs. We introduce a stability-aware constraint that allows watermarking only in stable contexts and a bit-controlled, unbiased modulation to preserve the original DLM output distribution, achieving stable watermarking with minimal quality impact. To enhance detection robustness, we design a Regret-based Remasking, which grants a "second chance" for unwatermarked tokens to be regenerated. It can seamlessly integrate into DLM inference with no added diffusion steps and latency. Experiments across DLMs and various tasks show that our scheme is effective, achieving superior generation quality compared to baselines while maintaining high detection accuracy and multi-bit capacity. Our code is available here https://github.com/iieSKLCSDsg/UMR.

[Findings] You Only Need One Single Token to Refine Safety Alignment ▶ *Underline Video*
Wenqian Yu, Shuo Chen, Zhijiang Li, Zhipeng Wang and Jindong Gu
Large language models (LLMs) face a critical alignment challenge: balancing safety with helpfulness. Excessive safety can lead to over-refusal, where models reject harmful-looking yet benign queries, severely limiting utility. Existing training-free interventions offer an efficient way to mitigate over-refusal without re-training, but suffer from high inference overhead and architecture dependency. Our work explores a complementary direction: rather than applying post-hoc corrections to model outputs, our goal is to intrinsically reshape the distributions of harmful and benign samples within the models decision space. In this paper, we argue that a lightweight training-based approach can more effectively distinguish between harmful and benign samples. We propose Single Token Alignment (STA), which optimizes only a single-token prefix (e.g., 4,096 parameters) while keeping the base model frozen. To address the inherent challenge of

achieving robust refinement through such a minimal parameter interface, STA employs a mixed weighting mechanism integrated with its optimization objective. This mechanism incorporates hard weighting via stringent data filtering to provide clear, unbiased learning signals, and soft weighting through a focal mechanism to prioritize challenging cases. Extensive experiments across 9 models and 10 datasets demonstrate that STA achieves a superior safety-helpfulness balance for LLMs, MLLMs, and reasoning models, offering a highly efficient and generalizable solution for refining safety alignment.

12

Main Conference Day 2: Monday, July 6, 2026

Overview

01:00-02:00	**Session 6: Keynote: Yue Zhang (Westlake University)**	Underline (virtual)
02:30-04:00	**Session 7: Virtual Presentations 2**	Underline (virtual)
08:30-16:30	**Registration & Conference Check In (Full Conference & Main Conference Attendees)**	Grand Hall Foyer (L1)
09:00-10:30	**Session 8: Orals / Demos / Poster Session D**	
	Demo, Main & Finding Posters	
	See Tracks on Whova or Underline for room assignment	Grand Hall (Ground), Coronado (L4)
	Speech Processing and Spoken Language Understanding 1	Harbor A (L2)
	AI/LLM Agents 2	Harbor B&C (L2)
	Low-resource Methods for NLP 1	Harbor D (L2)
	Language Models 3	Harbor E&F (L2)
	Special Theme: Explainability of NLP Models 1	Harbor G (L2)
	Linguistic theories, Cognitive Modeling and Psycholinguistics 2	Harbor H&I (L2)
	Financial Applications and Time Series	Promenade (L3)
	Question Answering 1	Regatta (L4)
	Industry 4: Panel - Orals	Balboa (L2, Seaport)
	BoF/Affinity: Please check the online schedule	Gaslamp C&D, Old Town A&B, La Jolla A&B
10:30-11:00	Coffee	Grand Hall
11:00-12:30	**Session 9: Orals / Demos / Poster Session E**	
	Demo, Main & Finding Posters	
	See Tracks on Whova or Underline for room assignment	Grand Hall (Ground), Coronado (L4)
	Human-AI Interaction/Cooperation and Human-Centric NLP 1	Harbor A (L2)

	Multimodality and Language Grounding to Vision, Robotics and Beyond 1	Harbor B&C (L2)
	Machine Learning for NLP 2	Harbor D (L2)
	LLM Efficiency 1	Harbor E&F (L2)
	Clinical and Biomedical Applications	Harbor G (L2)
	Low-resource Methods for NLP 2	Harbor H&I (L2)
	Machine Translation	Promenade (L3)
	Code Models	Regatta (L4)
	BoF/Affinity: Please check the online schedule	Gaslamp C&D, Old Town A&B, La Jolla A&B
12:30-13:30	Lunch	On Your Own
12:45-14:15	**Session 10: Virtual Presentations 3**	Underline
13:30-14:30	**Session 11: Business Meeting (all attendees welcome)**	Harbor Ballroom (L2)
14:45-15:30	**Session 12: Discussion Panel: Explainability of LLMs: Academic vs. Industry Perspective** *Panelists: Tal Linzen (NYU/Google); Sameer Singh (UCI/Spiffy AI); Kamalika Das (Intuit)* Moderator: Julian McAuley (UCSD)	Harbor Ballroom (L2)
15:30-16:15	**Session 13: Presidential Address: Barbara Plank (LMU Munich)**	Harbor Ballroom (L2)
16:15-16:45	Coffee	Grand Hall
16:45-17:45	**Keynote: Tania Lombrozo (Princeton University)**	Harbor Ballroom (L2)
18:00-19:30	**Session 14: Virtual Presentations 4**	Underline
18:30-22:00	**Social Event**	USS Midway

Detailed Program

Session 7: Virtual Presentations 2 - 02:30-04:00

Virtual Presentations 2

02:30-04:00 (Zoom)

[Long] Phun-Bench: Evaluating LLMs on Phonological Understanding in Chinese ▶ *Underline Video*

Xing Yue, Yongliang Shen and Weiming Lu

Language is a vehicle for thought, intricately tied to sounds, symbols, and meaning. However, most large language model (LLM) research focuses on meaning (semantics) and symbols (spelling) while largely overlooking sounds. Existing benchmarks on LLMs' phonological abilities are either solvable through rote memorization or intertwined with other abilities, making them inadequate to measure LLMs' genuine ability in phonological understanding. Here, we present Phun-Bench, a purpose-built Chinese benchmark with diverse tasks and settings across three dimensions (Homophony, Rhyme, and Phonetic Similarity), designed to systematically evaluate LLMs' phonological understanding. Our results show that while LLMs excel at recalling correct pronunciations, they generally struggle to leverage phonological knowledge in the flexible and intuitive way that human speakers do. Moreover, through detailed analyses, we propose a hypothesis regarding the underlying mechanism of LLMs' phonological understanding and "perception", highlighting an underexplored frontier for future research.

[Industry] A Novel Matching Paradigm: Unified Generative and Discriminative LLM with Prompt Compression for Relevance Learning ▶ *Underline Video*

Guoliang Zhao, Zixin Cui, Chao Ye, Dengwu He, Fei Huang, Yubo Liu, Shuanglong Li, Zongren Guo, Bin Ding, Shuang Zhang, Kunhong Zhu, Zhi Guo and Lin Liu

The matching paradigm is fundamental to large-scale information retrieval and is widely used in industrial search and advertising systems. Existing approaches employ Large Language Models (LLMs) primarily as feature extractors, underutilizing their full modeling capabilities. To address this limitation, we propose a novel matching paradigm, termed the Unified Generative and Discriminative LLM (UGD). It integrates two-tower, single-tower, and generative tasks within a unified LLM framework via attention-mask partitioning, enabling generative tasks to serve as auxiliary supervision for discriminative learning and facilitating distillation from single-tower to two-tower architectures through a multi-task fine-tuning mechanism. To satisfy online latency constraints, we further introduce a self-distillation variant of UGD with a KMeans-enhanced linearized RQVAE for prompt compression and quantization. This design compresses and quantizes landing-page documents during inference, improving serving efficiency and reducing storage overhead. Extensive experiments show that UGD achieves superior performance and strong practical value. The framework has been deployed in an industrial search engine serving hundreds of millions of users and hundreds of thousands of advertisers, significantly enhancing search experience. Open access upon publication.

[Long] Activation Decomposition and Steering for LLM Backdoor Remediation ▶ *Underline Video*

Lingfeng Zhong, Qiongkai Xu and Usman Naseem

Existing works on defending against LLM backdoor attacks rely on either auxiliary models or safety-related datasets for defending against backdoor attacks on large language models, which are not always available. To address these challenges, we propose our we propose our Contrastive-Selective Activation Decomposition and Steering (CS-ADS), which contrasts relatively more benign and poisoned settings to decompose the feature vectors for steering without relying on additional auxiliary models or datasets. With such disentangled vectors for remediation, our method can achieve feasible defense qualities even better than dataset-based contrastive steering strategies. This novel decomposition-based solution is motivated by the key insight that feature representations of prompt pairs can encode the same benign semantics in different proportions, even when both prompt pairs are similarly backdoored. Such discrepancies allow our method to identify effective remediation directions for steering the generation process, thereby preventing undesired outputs. We evaluate CS-ADS against multiple state-of-the-art backdoor attacks, and experimental results show that CS-ADS provides effective defense across settings.

[Long] AgentSlimming: Towards Efficient and Cost-Aware Multi-Agent Systems ▶ *Underline Video*

Yulang Chen, Haoxuan Peng, Jinyan Liu, Zichen Wen, Dongrui Liu and Linfeng Zhang

Large Language Model-based Multi-Agent Systems (MAS) have demonstrated remarkable capabilities in complex tasks. However, manually designing optimal communication topologies is labor-intensive, while automated expansion methods often result in bloated structures with redundant agents, leading to excessive token consumption. To address this problem, we introduce AgentSlimming, a plug-and-play compression framework for graph-structured multi-agent workflows. Motivated by the AgentPruner and AgentQuant in neural networks, AgentSlimming compresses workflows by firstly estimate the importance score of each agent with a hybrid mechanism, and then removing redundant agents or replacing them with low-cost ones, where each operation is then validated with a baseline-anchored acceptance rule to prevent performance collapse. Experiments show that AgentSlimming reduces average token cost by up to 78.9% with neg-

ligible performance degradation, and even sometimes improves accuracy, achieving a strong Pareto-optimal trade-off between cost and quality.

[Demo] AiraXiv: An AI-Driven Open-Access Platform for Human and AI Scientists ▶ *Underline Video*
Junshu Pan, Panzhong Lu, Yixuan Weng, Qiyao Sun, Fang Guo, Zijie Yang, Qiji Zhou and Yue Zhang
Recent advances in artificial intelligence (AI) have accelerated the growth of both human-authored and AI-generated research outputs, placing increasing strain on traditional academic publishing systems and challenging the scalability of conference- and journal-centered paradigms amid rising submission volumes, reviewer workload, and venue size. To address these challenges, we explore an AI-era publishing paradigm in which both human and AI scientists participate as authors and readers, and papers evolve through continuous, feedback-driven iteration. We propose AiraXiv, an AI-driven open-access platform built on open preprints, AI-augmented analysis and review, and reader feedback. AiraXiv supports human scientists through an interactive UI and supports AI scientists through Model Context Protocol (MCP)-based interactions. We validate AiraXiv through real-world deployments, including serving as the submission platform for ICAIS 2025, demonstrating its potential as a fast, inclusive, and scalable research infrastructure for the AI era.

[Long] AIRCoder: Adaptive Integration of Multi-dimensional Retrieval for Repository-level Code Completion ▶ *Underline Video*
Chuanqi Shi, Miao Gao and Zhiqiang Gao
Repository-level code completion relies on retrieval strategies to select relevant context from large codebases. Most existing methods employ single-dimensional retrieval based on textual similarity or dependency existence, leading to inconsistent performance across completion scenarios. Even task-adaptive and hybrid methods incur high computational costs while failing to explicitly consider structural hierarchy for retrieving functionally similar code. We introduce AIRCoder, a multi-dimensional retrieval framework that combines eight complementary metrics across three dimensions: textual similarity, dependency existence, and structural hierarchy. By proposing a structure-preserving chunking strategy and lightweight fusion module, AIRCoder learns context-dependent weights to adaptively integrate retrieval metrics for each query. Experiments on CrossCodeEval and RepoEval demonstrate that AIRCoder achieves an average improvement of 4.63% in exact match over the best baseline, with 10.2x higher efficiency and strong cross-language generalization across Python, Java, C#, and TypeScript.

[Long] All Languages Matter: Understanding and Mitigating Language Bias in Multilingual RAG ▶ *Underline Video*
Dan Wang, Guozhao Mo, Yafei Shi, Cheng Zhang, Bo Zheng, Boxi Cao, Xuanang Chen, Yaojie Lu, Hongyu Lin, Ben He, Xianpei Han and Le Sun
Multilingual Retrieval-Augmented Generation (mRAG) leverages cross-lingual evidence to ground Large Language Models (LLMs) in global knowledge. However, we show that current mRAG systems suffer from a language bias during reranking, systematically favoring English and the query's native language. By introducing an estimated oracle evidence analysis, we quantify a substantial performance gap between existing rerankers and the achievable upper bound. Further analysis reveals a critical distributional mismatch: while optimal predictions require evidence scattered across multiple languages, current systems systematically suppress such "answer-critical" documents, thereby limiting downstream generation performance. To bridge this gap, we propose Language-Agnostic Utility-driven Reranker Alignment (LAURA), Experiments across diverse languages and generation models show that LAURA effectively mitigates language bias and consistently improves mRAG performance.

[Findings] AlphaEdit+: Model Editing in the Presence of Conflicting and Inconsistent Knowledge ▶ *Underline Video*
Qing Liu, Jianhao Zhang, Ou Wu, Michael K. Ng and Yi Du
Knowledge editing is a crucial technique for daily updates in LLMs, requiring a balance between accurately modifying incorrect knowledge and preserving existing information. The recently proposed AlphaEdit method achieves competitive editing performance by updating parameters under null-space constraints. However, our theoretical analysis reveals that AlphaEdit struggles with high knowledge conflicts and inconsistencies during editing. To address this, we propose a new editing method AlphaEdit+, featuring three key improvements: 1) relaxing null-space constraints by adding a matrix perturbation through optimization to resolve conflicts between new and preserved knowledge; 2) introducing a weighting scheme on previously updated knowledge constraints to mitigate conflicts between new and historical editing; 3) developing a value smoothing algorithm to resolve high knowledge inconsistencies. These enhancements collectively ensure robust editing while maintaining model coherence. Comprehensive experiments show that our approach AlphaEdit+ not only resolves the brittleness of the original method on carefully constructed challenging datasets but also outperforms AlphaEdit on existing benchmark datasets.

[Long] APPSI-139: A Parallel Corpus of English Application Privacy Policy Summarization and Interpretation ▶ *Underline Video*
Pengyun Zhu, Qiheng Sun, Long Wen, Yanbo Wang, Yang Cao, Junxu Liu, Deyi Xiong, Jinfei Liu, Zhibo Wang and Kui Ren
Privacy policies are essential for users to understand how service providers handle their personal data. However, these documents are often long and complex, as well as filled with technobabble and legalese, causing users to unknowingly accept terms that may even contradict the law. While summarizing and interpreting these privacy policies is crucial, there is a lack of high-quality English parallel corpus optimized for legal clarity and readability. To address this issue, we introduce APPSI-139, a high-quality English privacy policy corpus meticulously annotated by domain experts, specifically designed for summarization and interpretation tasks. The corpus includes 139 English privacy policies, 15,692 rewritten parallel corpora, and 36,351 fine-grained annotation labels across 11 data practice categories. Concurrently, we propose TCSI-pp-V2, a hybrid privacy policy summarization and interpretation framework that employs an alternating training strategy and coordinates multiple expert modules to effectively balance computational efficiency and accuracy. Experimental results show that the hybrid summarization system built on APPSI-139 corpus and the TCSI-pp-V2 framework outperform large language

models, such as GPT-4o and LLaMA-3-70B, in terms of readability and reliability. The source code and dataset are available at https://github.com/EnlightenedAI/APPSI-139.

[Long] AraVQA: Building a New Arabic Factoid Visual Question Answering Dataset from Wikipedia ▶ *Underline Video*

Sultan Alrowili, Younes Samih, Abed Alhakim Freihat and Mathan Kumar Eswaran

The development of large-scale Visual Question Answering (VQA) datasets has traditionally relied on resource-intensive manual annotation. In addition, most of the existing Arabic VQA datasets focus on culturally-specific and dialect-aware domains. To address these limitations, we propose a new pipeline that leverages Wikipedia template tags to extract the relevant information for each image, which is subsequently utilized by the Large Language Model (LLM) to synthetically generate a new visual question answering dataset. Using this pipeline, we have constructed AraVQA, the most comprehensive Arabic Factoid Visual Question Answering dataset, containing more than 50,000 questions and covering over 20 varied primary subjects within Arabic general knowledge. Our detailed analysis shows that our dataset can serve as a post-training dataset to enhance the performance of existing Visual Language Models (VLMs) on Arabic VQA tasks. Furthermore, we present a novel benchmark, derived from our dataset and validated through manual annotation, that poses more challenges to Arabic VLMs than existing Arabic VQA datasets.

[Long] ARF-RLHF: Adaptive Reward-Following for RLHF through Emotion-Driven Self-Supervision and Trace-Biased Dynamic Optimization ▶ *Underline Video*

Yuxuan Zhang

Current RLHF methods such as PPO and DPO typically reduce human preferences to binary labels, which are costly to obtain and too coarse to reflect individual variation. We observe that expressions of satisfaction and dissatisfaction follow stable linguistic patterns across users, indicating that more informative supervisory signals can be extracted from free-form feedback. Building on this insight, we introduce Adaptive Reward-Following (ARF), which converts natural feedback into continuous preference trajectories and optimizes them using the novel TraceBias algorithm. Across diverse LLMs and preference domains, ARF consistently outperforms PPO and DPO, improving alignment by up to 7.6%. Our results demonstrate that continuous reward modeling provides a scalable path toward personalized and theoretically grounded RLHF.

[Long] ATGL: An Adaptive-Threshold Global Loss for Document-level Relation Extraction ▶ *Underline Video*

Huangming Xu, Fu Zhang, Zhixuan Yang, Lu Zhang and Jingwei Cheng

Document-level relation extraction (DocRE) aims to determine which relations hold between a given entity pair within a document. As a multi-label classification task, the most commonly adopted paradigm introduces a learnable threshold to distinguish positive and negative classes for an entity pair. Under this paradigm, existing losses decouple the optimization into independent positive and negative losses, which interact solely with a shared threshold. This leads to two inherent limitations: (i) threshold instability caused by conflicting gradient updates from the decoupled losses; and (ii) optimization bias exacerbated by the severe imbalance between limited positive samples and abundant negative samples inherent in DocRE, which makes the model more likely to predict that no relation exists. To address these issues, we propose the Adaptive-Threshold Global Loss (ATGL). Unlike prior work, ATGL integrates positive, negative, and threshold optimization into a unified logit space and explicitly enforces ranking constraints on their contributions to the objective. Furthermore, ATGL incorporates an imbalance-aware optimization mechanism, thereby effectively addressing the severe class imbalance in DocRE. Our ATGL serves as a general optimization objective that can be readily applied to different DocRE models. Experiments on four datasets show that ATGL outperforms other DocRE losses and achieves state-of-the-art results, while consistently improving the performance of existing DocRE models. Code is available at https://github.com/xhm-code/ATGL.

[Long] Attention as Selector: Unlocking VLM Attention for Long Document Page Retrieval ▶ *Underline Video*

Minfeng Zhu, Linxin Bao, Wei Chen and Linchao Zhu

Visual Language Models (VLMs) have become a robust foundation for document question answering. Processing long documents remains challenging due to limited context windows and computational budgets. Existing page-level retrieval methods offer a practical solution, typically encoding pages and queries into vectors and ranking them via cosine similarity. However, such embeddingbased methods (i) lack query-page interaction before similarity scoring and (ii) usually require large-scale datasets to align visual and textual embeddings. In this paper, we observe that the crossmodal attention maps of welltrained VLMs are able to highlight semantically relevant regions. Building on this insight, we present CAPS (Cross-modal Attention as Page Selector), a retrieval framework that utilizes attention mechanisms inside VLMs for page selection. Specifically, CAPS first enhances attention-based retrieval capability with a small amount of contrastive data, then identifies the most effective attention head through expert head selection, and finally employs an adaptive filtering mechanism to obtain an appropriate number of relevant page candidates. Extensive experiments on four long-document benchmarks demonstrate that CAPS outperforms state-of-the-art embeddingbased methods in both retrieval precision and downstream DocQA accuracy. Notably, CAPS achieves these gains using less than 10% of the training data required by competing baselines, highlighting the data efficiency of attention-based page retrieval.

[Long] Awakening Dormant Experts:Counterfactual Routing to Mitigate MoE Hallucinations ▶ *Underline Video*

Wentao Hu, Yanbo Zhai, Xiaohui Hu, Mingkuan Zhao, Shanhong Yu, Xue Liu, Kaidong Yu, Shuangyong Song and Xuelong Li

Sparse Mixture-of-Experts (MoE) models have achieved remarkable scalability, yet they remain vulnerable to hallucinations, particularly

when processing long-tail knowledge. We identify that this fragility stems from static Top-k routing: routers tend to favor high-frequency patterns over rare factual associations. Consequently, "specialist experts" possessing critical long-tail knowledge are often assigned low gating scores and remain "dormant", under-prioritized for specific tokens despite their proven causal importance on other inputs. To address this, we propose Counterfactual Routing (CoR), a training-free inference framework designed to awaken these dormant experts. CoR integrates layer-wise perturbation analysis with the Counterfactual Expert Impact (CEI) metric to dynamically shift computational resources from syntax-dominant to knowledge-intensive layers while maintaining a constant total activation count, effectively retrieving causally decisive experts via virtual ablation. Extensive experiments on TruthfulQA, FACTOR, and TriviaQA demonstrate that CoR improves factual accuracy by 3.1% on average without increasing the inference budget, establishing a superior Pareto frontier compared to static scaling strategies.

[Long] BanHADEX: Towards Explainable HAte Speech Detection in Bangla Using Human Annotated EXplanation ► *Underline Video*

Faisal Hossain Raquib, Akm Moshiur Rahman Mazumder, Md Fahim, Md Tahmid Hasan Fuad, Md Farhan Ishmam, Faria Sultana, M Ashraful Amin, Amin Ahsan Ali and AKMMAHBUBUR RAHMAN

Online safety in low-resource languages hinges not only on accurate hate speech detection but also on transparent, culturally grounded explanations. Yet prior works in Bangla largely focus on hate classification, while overlooking interpretability. We address this gap by introducing BanHADEX, the first hate explainability dataset in Bangla with human-annotated labels. BanHADEX contains 19,203 YouTube comments spanning April 2024-June 2025, annotated for binary hate classification with seven fine-grained hate categories, seven target groups, and concise explanations for each sample. Our data pipeline relies on a twostage annotation protocol that uses majority voting for robust labeling. Our rich suite of experiments on open and closed-source LLMs reveals that explanationguided LoRA substantially outperforms both classification and explanation quality across prompting and fine-tuning strategies. BanHADEX establishes the groundworks for faithful interpretability and safer moderation in linguistically rich yet under-resourced languages.

[Long] BaseCal: Unsupervised Confidence Calibration via Base Model Signals ► *Underline Video*

Hexiang Tan, Wanli Yang, Junwei Zhang, Xin Chen, Rui Tang, Du Su, Jingang Wang, Yuanzhuo Wang, Fei Sun and Xueqi Cheng

Reliable confidence is essential for trusting the outputs of LLMs, yet widely deployed post-trained LLMs (PoLLMs) typically compromise this trust with severe overconfidence. In contrast, we observe that their corresponding base LLMs often remain well-calibrated. This naturally motivates us to calibrate PoLLM confidence using the base LLM as a reference. This work proposes two ways to achieve this. A straightforward solution, BaseCal-ReEval, evaluates PoLLM's responses by feeding them into the base LLM to get average probabilities as confidence. While effective, this approach introduces additional inference overhead. To address this, we propose BaseCal-Proj, which trains a lightweight projection to map the final-layer hidden states of PoLLMs back to those of their base LLMs. These projected states are then processed by the base LLM's output layer to derive base-calibrated confidence for PoLLM's responses. Notably, BaseCal is an unsupervised, plug-and-play solution that operates without human labels or LLM modifications. Experiments across five datasets and three LLM families demonstrate the effectiveness of BaseCal, reducing Expected Calibration Error (ECE) by an average of 42.90% compared to the best unsupervised baselines.

[Long] BEFT: Bias-Efficient Fine-Tuning of Language Models in Low-Data Regimes ► *Underline Video*

Baichuan Huang, Ananth Balashankar and Amir Aminifar

Fine-tuning the bias terms of large language models (LLMs) has the potential to achieve unprecedented parameter efficiency while maintaining competitive performance, particularly in low-data regimes. However, the link between fine-tuning different bias terms (i.e., b_q, b_k, b_v in the query, key, or value projections) and downstream performance remains largely unclear to date. In this paper, we investigate the link between fine-tuning b_q, b_k, b_v with the performance of the downstream task. Our key finding is that directly fine-tuning b_v generally leads to higher downstream performance in low-data regimes, in comparison to b_q and b_k. We extensively evaluate this unique property across a wide range of LLMs spanning encoder-only and decoder-only architectures up to 6.7B parameters (including bias-free LLMs). Our results provide strong evidence for the effectiveness of directly fine-tuning b_v across various downstream tasks. The implementation code is available at https://github.com/whubaichuan/BEFT.

[Long] Benchmarking and Learning Real-World Customer Service Dialogue ► *Underline Video*

Tianhong Gao, Jundong Shen, Jiapeng Wang, Bei Shi, Ying Ju, Junfeng Yao and Huiyu Yu

Existing benchmarks and training pipelines for industrial intelligent customer service (ICS) remain misaligned with real-world dialogue requirements, overemphasizing verifiable task success while under-measuring subjective service quality and realistic failure modes, leaving a gap between offline gains and deployable dialogue behavior. We close this gap with a benchmark-to-optimization loop: we first introduce OlaBench, an ICS benchmark spanning retrieval-augmented generation, workflow-based systems, and agentic settings, which evaluates service capability, safety, and latency sensitivity; moreover, motivated by OlaBench results showing state-of-the-art LLMs still fall short, we propose OlaMind, which distills reusable reasoning patterns and service strategies from expert dialogues and applies staged exploration–exploitation reinforcement learning with instance-level rubric-aware guidance to improve model capability. OlaMind surpasses GPT-5.2 and Gemini 3 Pro on OlaBench (83.64 vs. 70.58/70.84) and, in online A/B tests, delivers an average +23.67% issue resolution and -6.6% human transfer rate versus the baseline, bridging offline gains to deployment. Together, OlaBench and OlaMind advance ICS systems toward more anthropomorphic, professional, and reliable deployment. The project page and evaluation are available at https://olamind-olabench.github.io.

[Long] Benchmarking Large Vision-Language Models on CFMME: A Comprehensive Chinese Financial Multimodal Evaluation Dataset ▶ *Underline Video*

Qian Chen, Xianyin Zhang, Yanzhi Liu, Lifan Guo, Feng Chen and Chi Zhang

The emergence of Large Vision-Language Models (LVLMs) has substantially expanded model capabilities beyond text-only understanding, enabling unified inference across both visual and textual modalities and supporting a broader range of real-world applications. To comprehensively evaluate the perception, understanding, reasoning, and cognition capabilities of LVLMs throughout the entire financial business workflow in Chinese contexts, we introduce CFMME, a novel Chinese financial multimodal evaluation benchmark. CFMME comprises 6,052 instances spanning from fundamental academic knowledge to complex real-world applications, covering eight primary financial image modalities and four core multimodal tasks. On CFMME, we conduct a thorough evaluation of representative LVLMs. The results show that the state-of-the-art model attains an overall accuracy of 66.11% on the question answering task and an average score of 77.18 on the detection, recognition, and information extraction tasks, indicating substantial room for improvement in current LVLMs. In addition, we conduct detailed analyses of error causes, cross-modal capabilities, and multi-orientation settings, yielding valuable insights for future research. We hope that CFMME will spur further progress in LVLMs, especially by improving their performance on multiple multimodal tasks in the financial domain.

[Long] Benchmarking Post-Training Quantization of Large Language Models under Microscaling Floating Point Formats ▶ *Underline Video*

Manyi Zhang, Ji-Fu Li, Zhongao Sun, Haoli Bai, Hui-Ling Zhen, Zhenhua Dong and Xianzhi Yu

Microscaling Floating-Point (MXFP) has emerged as a promising low-precision format for large language models (LLMs). Despite various post-training quantization (PTQ) algorithms being proposed, they mostly focus on integer quantization, while their applicability and behavior under MXFP formats remain largely unexplored. To address this gap, this work conducts a systematic investigation of PTQ under MXFP formats, encompassing over 7 PTQ algorithms, 15 evaluation benchmarks, and 3 LLM families. The key findings include: 1) MXFP8 consistently achieves near-lossless performance, while MXFP4 introduces substantial accuracy degradation and remains challenging; 2) PTQ effectiveness under MXFP depends strongly on format compatibility, with some algorithmic paradigms being consistently more effective than others; 3) PTQ performance exhibits highly consistent trends across model families and modalities, in particular, quantization sensitivity is dominated by the language model rather than the vision encoder in multimodal LLMs; 4) The scaling factor of quantization is a critical error source in MXFP4, and a simple pre-scale optimization strategy can significantly mitigate its impact. Together, these results provide practical guidance on adapting existing PTQ methods to MXFP quantization.

[Industry] Better Literary Translation: A Multi-Aspect Data Generation and LLM Training Approach ▶ *Underline Video*

Zhihao Lin, Ziqi Zhu, Hao Huang, Guanghui Wang and Peiyang He;

Literary translation poses unique challenges due to the scarcity of high-quality annotated data and the need to balance expression fluency with literary effect. We present a multi-aspect iterative refinement framework that generates high-quality translation references and preference data through specialized LLM translators, each targeting a distinct quality dimension. We leverage the generated data for supervised fine-tuning and reinforcement learning. Experiments show that our generated references outperform the original ground truth for SFT by 8.65 CEA100 points. For reinforcement learning, we find that DPO leads to performance degradation in this setting, while leveraging an explicit reward model for GRPO yields an additional 1.51 point improvement. We attribute this to the stability of two-stage training and GRPO's online exploration capability. Our resulting models, LitMT-8B and LitMT-14B, achieve 67.25 and 69.07 CEA100 respectively on the MetaphorTrans English-to-Chinese literary translation benchmark, competitive with Claude Sonnet 4.5 at 68.43, and demonstrate strong generalization to out-of-domain literary work (i.e., O. Henry).

[Long] Beyond Accuracy: Unveiling Inefficiency Patterns in Tool-Integrated Reasoning ▶ *Underline Video*

QishengSu QishengSu, Shiting Huang, Zhen Fang, Ziyan Chen, Zehui Chen and Feng Zhao

In real-world Tool-Integrated Reasoning (TIR) scenarios, a major source of inefficiency is that the toolcalls create pauses between LLM requests and cause KV-cache eviction. Also, the long, unfiltered response returned by external tools inflates the KV-cache, so each decode step spends more time loading the growing cache and thus becomes steadily slower as context length increases. However, existing efficiency metrics like token counts and toolcall counts fail to capture this real computational cost. To address this, we introduce PTE (Prefill Token Equivalents), a hardware-aware TIR-efficiency metric that unifies internal reasoning and external tool-use costs while explicitly accounting for non-reusable KV-Cache and long-tool-response scenarios, thus better reflects real-world scenarios. We conduct extensive experiments across five TIR benchmarks, quantify their PTE costs, and identify four inefficiency patterns that appear in TIR. In a simulated high-concurrency industrial setting, PTE explains wall-clock latency significantly better than token-count metric. We also discover that trajectories with higher PTE costs tend to have lower reasoning correctness, indicating that simply using more tools does not improve the quality of the answer. PTE offers a new perspective on the efficiency of Tool-Integrated Reasoning. The code is available.

[Long] Beyond Detection: Evaluating Fallacy Awareness of LLMs in Interactive Scenarios ▶ *Underline Video*

Conghui Niu, Ningxin Wu, Ziran Zhao, Dong Yu, Chen Kang and Pengyuan Liu

Large Language Models (LLMs) often fail to recognize fallacious reasoning in real-world interactions, despite strong performance on static fallacy detection tasks. We define this ability as fallacy awareness, the capacity to autonomously perceive and resist fallacies in

dynamic, pragmatic contexts. To study this, we introduce ISFallacy, a large-scale Chinese benchmark of 50K interactive scenarios spanning six fallacy types, five social interaction settings, diverse role relationships, and personality traits. We further propose FATE, a two-stage evaluation framework that assesses fallacy awareness without explicit cues, combining natural dialogue responses and reasoning-based decisions. Experiments on five representative LLMs reveal a substantial gap between fallacy classification and awareness, with models particularly vulnerable to emotion-driven fallacies and scenarios involving cooperative or trust-based relationships. Deeper analysis uncovers a cognition-behavior gap and fragile internal representations underlying awareness failures. Our work establishes a foundation for evaluating and enhancing the robustness of LLMs against fallacious reasoning in interactive settings.

[Long] Beyond Experience Retrieval: Learning to Generate Utility-Optimized Structured Experience for Frozen LLMs ▶ *Underline Video*

Xuancheng Li, Haitao Li, Yujia Zhou, Yiqun Liu and Qingyao Ai

Large language models (LLMs) are largely static and often redo reasoning or repeat mistakes. Prior experience reuse typically relies on external retrieval, which is similarity-based, can introduce noise, and adds latency. We introduce SEAM (Structured Experience Adapter Module), a lightweight, executor-specific plug-in that stores experience in its parameters and generates a structured, instance-tailored experience entry in a single forward pass to guide a frozen LLM executor. SEAM is trained for utility via executor rollouts and GRPO while keeping the executor frozen, and can be further improved with logged-success SFT after deployment. Experiments on mathematical reasoning benchmarks show consistent accuracy gains across executors with low overhead. Extensive ablation and analysis further elucidate the mechanisms underlying SEAMs effectiveness and robustness.

[Long] Beyond Explicit Refusals: Soft-Failure Attacks on Retrieval-Augmented Generation ▶ *Underline Video*

Wentao Zhang, Yan Zhuang, ZhuHang Zheng, Mingfei Zhang, Jiawen Deng and Fuji Ren

Existing jamming attacks on Retrieval-Augmented Generation (RAG) systems typically induce explicit refusals or denial-of-service behaviors, which are conspicuous and easy to detect. In this work, we formalize a subtler availability threat, termed soft failure, which degrades system utility by inducing fluent and coherent yet non-informative responses rather than overt failures. We propose Deceptive Evolutionary Jamming Attack (DEJA), an automated black-box attack framework that generates adversarial documents to trigger such soft failures by exploiting safety-aligned behaviors of large language models. DEJA employs an evolutionary optimization process guided by a fine-grained Answer Utility Score (AUS), computed via an LLM-based evaluator, to systematically undermine the certainty of answers while maintaining high retrieval success. Extensive experiments across multiple RAG configurations and benchmark datasets show that DEJA consistently drives responses toward low-utility soft failures and that the resulting adversarial documents maintain high stealth and effectiveness, proving resilient against common mitigation strategies including perplexity-based detection and input perturbations.

[Long] Beyond Single-View Detection: A Dual-Space Reasoning Framework for Interpretable Harmful Meme Understanding ▶ *Underline Video*

Wenqing Hou, Hongkui Tu, Ye Wang, Yue Zhang, Yuying Liu, Dong Zhu, Liqun Gao and Bin Zhou

The identification of harmful memes extends beyond a mere classification task, encompassing challenges related to multi-perspective semantic comprehension and hierarchical reasoning. Prevailing approaches predominantly depend on modal alignment or black-box classifiers, which fail to capture implicit biases and lack interpretability. In this study, we propose BPDMoE-Hate, a novel framework grounded in dual-space mixture-of-experts, which innovatively conceptualizes harmful meme detection as an integrated process of viewpoint decoupling and hierarchical fusion". Our approach generates adversarial binary perspectives via Visual-Language Models (VLMs) and incorporates an adaptive viewpoint gating to facilitate viewpoint selection, thereby enabling the model to autonomously discern implicit semantic inclinations. Moreover, we propose the Hyperbolic-Euclidean space expert to effectively capture the hierarchical structural relationships and semantic correlations between multimodal and viewpoint features, thereby enabling interpretable reasoning at the geometric representation level. Empirical evaluations conducted on three mainstream datasets demonstrate that BPDMoE-Hate not only substantially surpasses existing methodologies in performance but also offers visual explanations for viewpoint selection and hierarchical structuring, thereby advancing the field of interpretable multimodal content analysis.

[Long] Beyond Static Alignment: Adaptive Arbitration for Semantic Incongruence in Semi-Supervised Multimodal Sentiment Analysis ▶ *Underline Video*

Huicong Li, Xiangbo Ji and Wei Wu

Multimodal sentiment analysis is fundamentally challenged by semantic incongruence, where ambiguous visual signals often conflict with explicit textual cues. In semi-supervised scenarios, naively fusing such noisy features contaminates the joint representation, while conventional static alignment strategies fail to effectively arbitrate conflicting modalities in this task, leading to error reinforcement during self-training. To this end, we propose a novel Adaptive Arbitration for Semantic Incongruence (A2SI) framework for semi-supervised multimodal sentiment analysis, which emphasizes stable cross-modal representations and reliable supervision. Specifically, we first constrain unreliable visual representations by leveraging the reliable textual modality as an anchor to align divergent embeddings and reduce representation noise. Based on this, we further consider the reliability of supervision signals and calibrate pseudo-labels by adaptively weighting evidentiary confidence from heterogeneous views. Finally, to prevent error accumulation caused by unreliable samples, we introduce a progressive arbitration mechanism that verifies pseudo-labeled data from dual perspectives, enabling the model to dynamically balance sample diversity and label purity throughout self-training. Extensive experiments on the MVSA-Single and MVSA-Multiple datasets

demonstrate that A2SI consistently outperforms state-of-the-art methods under label-limited settings.

[Long] Beyond Static Artifacts: An Evolutionary Framework for Synthetic Claim Generation ▶ *Underline Video*

Yeqing Teng, Jiasheng Si, Shuxia Lin, Linhai Zhang, Weiyu Zhang, Wenpeng Lu, Deyu Zhou and Xiaoming Wu

With the generative capabilities of large language models (LLMs) reshaping the information ecosystem, the concern with the sociological validity of claim detection benchmarks is increasing. Current claim detection benchmarks predominantly treat claims as static textual artifacts, overlooking the sociological etiology of how information naturally emerges and mutates. In this paper, we propose an evolutionary paradigm that models claims as socially evolving entities. In specific, we introduce a socially generative framework for synthetic claim generation, a multi-agent simulation grounded in the Open Claims Model. By decomposing claims into context, utterance, and proposition, our approach enables the precise simulation of unmitigated propagation to capture truth decay, and intervened propagation with multi-auditor oversight for targeted generation. Furthermore, we propose the background-user-perspective (BUP) framework, which reformulates check-worthiness as a condition-dependent probability rooted in social environment. Experiments on our datasets verify the data quality and reveal how network topology and user attributes systematically shape veracity drift.

[Findings] Beyond the Individual: Virtualizing Multi-Disciplinary Reasoning for Clinical Intake via Collaborative Agents ▶ *Underline Video*

Huangwei Chen, Wu Li, Junhao Jia, Yining Chen, Xiaotao Pang, YaLong Chen, Gonghui Li, Haishuai Wang, Jiajun Bu and Lei Wu

The initial outpatient consultation is critical for clinical decision-making, yet it is often conducted by a single physician under time pressure, making it prone to cognitive biases and incomplete evidence capture. Although the Multi-Disciplinary Team (MDT) reduces these risks, they are costly and difficult to scale to real-time intake. We propose Aegle, a synchronous virtual MDT framework that brings MDT-level reasoning to outpatient consultations via a graph-based multi-agent architecture. Aegle formalizes the consultation state using a structured SOAP representation, separating evidence collection from diagnostic reasoning to improve traceability and bias control. An orchestrator dynamically activates specialist agents, which perform decoupled parallel reasoning and are subsequently integrated by an aggregator into a coherent clinical note. Experiments on ClinicalBench and a real-world RAPID-IPN dataset across 24 departments and 53 metrics show that Aegle consistently outperforms state-of-the-art proprietary and open-source models in documentation quality and consultation capability.

[Long] Beyond Timestamps: Bridging Forward and Backward Reasoning in Temporal Numerical and Relational Understanding ▶ *Underline Video*

Xinying Qian, Ying Zhang, Xuhui Sui, Yu Zhao, Baohang Zhou and Jeff Z. Pan

Temporal reasoning remains a critical challenge for large language models (LLMs), particularly when it requires encompassing relational dependencies and numerical constraints. Yet, existing benchmarks largely overlook the joint consideration of these two dimensions and primarily rely on single-task evaluation paradigms, making it difficult to assess whether correct answers reflect grounded reasoning or arise from superficial statistical recall. To address these gaps, we introduce TNR, a benchmark designed to evaluate both Temporal Numerical and Relational reasoning. We propose a bi-directional evaluation framework consisting of forward generation via Question Answering (QA) and backward verification via Fact Verification (FV). By measuring the alignment between QA and FV, we introduce a Consistency Rate to quantify the robustness of reasoning across these two directions. Experiments on a range of LLMs reveal notable discrepancies between QA and FV performance, particularly in numerical and interval-based tasks. Moreover, our bi-directional error analysis demonstrates that these inconsistencies often stem from heuristic shortcuts and statistical co-occurrences rather than grounded logical deduction, flaws that are frequently masked in standard single-task evaluations.

[Findings] Beyond Topology: Generative Node Importance Estimation via Structure-Guided Semantic Reasoning ▶ *Underline Video*

Kuofei Fang and Siyan WuGuo Yu Bin Wu

Node Importance Estimation (NIE) in Knowledge Graphs (KGs) aims to quantify the significance of entities, serving as a pivotal instrument for deciphering the latent mechanisms of social dynamics. However, existing methods are often confined to supervised paradigms and rely heavily on topological aggregation, resulting in limited generalization capability. To address these challenges, we propose GenNIE, the first end-to-end generative reasoning framework for NIE. Specifically, GenNIE leverages Large Language Models (LLMs) integrated with topological information to generate precise importance scores for entities in KGs. Furthermore, GenNIE introduces a Global-Structural Graph Perception mechanism to empower the LLMs with holistic graph cognition. Extensive experiments demonstrate the performance superiority of GenNIE and its robust generalization across diverse domains. Our code is available at https://github.com/CoffeyF/GenNIE.git.

[Long] BiMol-Diff: A Unified Diffusion Framework for Molecular Generation and Captioning ▶ *Underline Video*

Aditya Hemant Shahane, Anuj Kumar Sirohi, Devansh Arora, Nitin Kumar, Prathosh A P and Sandeep Kumar

Bridging molecular structures and natural language is essential for controllable design. Autoregressive models struggle with long-range dependencies, while standard diffusion processes apply uniform corruption across positions, which can distort structurally informative tokens. We present BiMol-Diff, a unified diffusion framework for the paired tasks of text-conditioned molecule generation and molecule captioning. Our key component is a token-aware noise schedule that assigns position-dependent corruption based on token recovery difficulty, preserving harder-to-recover substructures during the forward process. On ChEBI-20 and M3-20M, BiMol-Diff improves molecule reconstruction with a 15.4% relative gain in Exact Match and achieves strong captioning results, attaining best BLEU and BERTScore among compared baselines. These results indicate token-aware noising improves fidelity in molecular structure-language modeling

[Long] Black-Box Membership Inference Attacks for Video Training Data in Multimodal Large Language Models ▶ *Underline Video*
Jinrui Wang, Zhenfeng Gao, Wendan Wang, Huili Wang, Zichen Qin, Linjie Zhu, Hongke Fu, Shangguang Wang and Tao Qi
The increasing use of video data in training multimodal large language models (MLLMs) raises significant concerns on privacy leakage and copyright violations, highlighting the need for detecting improperly used training videos through membership inference attacks (MIAs). Most existing video MIA methods assess model memorization of key semantic concepts within a video (e.g., the name of a well-known movie character). However, such concepts usually appear repeatedly throughout the training corpus, and memorization of them does not constitute reliable evidence that a specific video was used during training. Besides, while some methods mitigate this limitation by capturing relationships between frames, they require a model logit-accessible setting and are impractical in realistic black-box scenarios. To address these challenges, we propose a black-box MIA framework, named VideoMIA, that can provide reliable evidence of specific video data usage for training MLLMs. The key of our method is to leverage temporal dependencies across video frames to evaluate the model's memorization of sequential dynamics within the video data, which cannot be inferred solely from general world knowledge or individual image data. The results across ten MLLMs and four benchmarks demonstrate that our method consistently achieves superior performance over all baselines in black-box evaluation settings. Code is available in https://github.com/jinruiwang258/VideoMIA.

[Findings] BnMMLU: Measuring Massive Multitask Language Understanding in Bengali ▶ *Underline Video*
Saman Sarker Joy and Swakkhar Shatabda
Large-scale multitask benchmarks have driven rapid progress in language modeling, yet most emphasize high-resource languages such as English, leaving Bengali underrepresented. We present BnMMLU, a comprehensive benchmark for measuring massive multitask language understanding in Bengali. BnMMLU spans 41 domains across STEM, humanities, social sciences, and general knowledge, and contains 134,375 multiple-choice question-option pairs-the most extensive Bengali evaluation suite to date. The dataset preserves mathematical content via MathML, and includes BnMMLU-HARD, a compact subset constructed from questions most frequently missed by top systems to stress difficult cases. We benchmark 24 model variants across 11 LLM families, spanning open-weights general/multilingual, Bengali-centric open-weights, and proprietary models, covering multiple parameter scales and instruction-tuned settings. We evaluate models under standardized protocols covering two prompting styles (Direct vs. Chain-of-Thought) and two context regimes (0-shot vs. 5-shot), reporting accuracy consistently across families. Our analysis highlights persistent gaps in reasoning and application skills and indicates sublinear returns to scale across model sizes. We release the dataset and evaluation templates to support rigorous, reproducible assessment of Bengali language understanding and to catalyze progress in multilingual NLP.

[Long] Calibrating Inference Time Alignment with Sequence-level Risk Accumulation ▶ *Underline Video*
Shanwen Tan, Ziyang Dong, Wei Ju, Yiwei Fu, Hao Wu, Kun Wang, Yifan Wang and Ziyue Qiao
This paper investigates the problem of safe decoding for Large Language Models (LLMs) during inference, particularly under jailbreak attacks. Previous approaches typically either detect malicious content or regulate the decoding alignment of LLMs to mitigate such attacks. Although effective in defending against attacks, these methods often over-reject benign content, limiting their generalizability in real-world scenarios where harmful and benign information coexist. Towards this end, we propose an innovative framework named Sequence-level risk Accumulation for calibrating test-time alignment (SEAT). Specifically, SEAT introduces a reward-guided branch decoding paradigm to incorporate safety awareness during generation. To balance the detection of harmful content with the accurate response to benign information, SEAT employs a sequence-level risk monitor that smooths risk signals over the entire sequence, preventing over-confident refusals for certain tokens. Furthermore, we conduct extensive experiments on four attack benchmarks and two neutral datasets, comparing SEAT with eight state-of-the-art baselines. Consequently, the results demonstrate that SEAT achieves superior performance both in defending against jailbreak attacks and in generating high-quality responses on neutral datasets. Our code is available at https://github.com/ShanwenTan/SEAT.

[SRW] Calibration vs Decision Making: Revisiting the Reliability Paradox in Unlearned Language Models ▶ *Underline Video*
Divyaksh Shukla; Ashutosh Modi
Machine unlearning aims to remove the influence of specific training data from a model while preserving reliable behavior on the remaining data, making reliable prediction and uncertainty estimation essential for evaluation. Calibration is commonly used as a proxy for reliability in language models, but low calibration error does not necessarily imply reliable decision rules, as models may rely on spurious correlations while remaining well calibrated. We investigate this gap in generative language models using the multiple-choice question-answering evaluation protocol on the TOFU benchmark, measuring probabilistic reliability with calibration metrics (ECE, MCE, Brier) and decision-rule reliability via attribution-based shortcut detection with Integrated Gradients and Local Mutual Information. We find that fine-tuned models achieve low calibration error (ECE approx 0.04) compared to pretrained models (ECE > 0.5), and models after unlearning retain similarly low calibration despite reduced accuracy on the forget split, while attribution analysis shows increased reliance on correlation-based tokens. These results demonstrate that good calibration can coexist with shortcut-based decision rules after unlearning, extending the reliability paradox to the machine unlearning setting.

[Findings] Can AI Revise Research Papers with Human Review Feedback? An Empirical Study and Benchmark ▶ *Underline Video*
Zihan Luo, Hong Huang, Jianxun Lian, Yu Chang, Xing Xie and Hai Jin
The rise of Human-AI collaboration can effectively speed up the research process for experts and allow anyone with critical thinking skills to conduct innovative work. A key part of this collaboration is the AI's ability to improve a paper with human feedback, updating both

the text and experiments to meet high standards. To evaluate this skill, we introduce ReviseBench, an extensible benchmark built on real academic data that can be easily scaled via agent-driven automated data collection. It tests the skills of Large Language Models (LLMs) on paper interpretation, experimental implementation, and paper formulation, using authors' camera-ready versions as natural human baselines. To facilitate a fine-grained assessment, we further propose ReviseArena, a platform supporting pair-wise comparisons between different AI-revised papers. Our initial evaluation results on ReviseBench reveal that even state-of-the-art foundation LLMs struggle significantly in this domain, achieving a win rate of less than 10% against human experts, and facing issues like incremental revision, unprofessional revision, and potential data fabrication. Our code and data are released publicly at: https://github.com/CGCL-codes/ReviseBench.

[Findings] Causal-Audit: Explicit and Auditable Graph-based Reasoning via Target-Aware Causal Chain Construction ► *Underline Video*

Su LAN, Xuefei Yin, Yanming Zhu and Alan Wee-Chung Liew

Causal and intervention-based question answering is fundamental to advancing large language models (LLMs) toward reasoning beyond surface-level correlations and understanding underlying causal mechanisms. However, existing LLM-based methods often rely on implicit language-level reasoning, resulting in opaque causal assumptions, unverifiable reasoning paths, and fragile predictions under complex interventions, particularly in context-free settings. In this paper, we propose an explicit and auditable causal reasoning framework for context-free intervention-based question answering. Our method formulates causal inference as structured reasoning over an explicit causal graph through four modular stages, rather than implicit end-to-end prediction. A key innovation is a target-aware causal graph construction strategy that treats the target variable as a core constraint during graph expansion, effectively suppressing irrelevant variables, spurious causal relations, and reasoning noise. We further introduce a path-level causal evidence aggregation mechanism that combines multiple causal paths while modeling both reinforcing and counteracting effects, enabling robust decision-making beyond single-chain reasoning. Extensive experiments on three benchmarks demonstrate that our framework consistently outperforms existing LLM-based methods while providing interpretable and auditable causal reasoning traces.

[Long] Chain-of-Thought as a Lens: Evaluating Structured Reasoning Alignment between Human Preferences and Large Language Models ► *Underline Video*

Boxuan Wang, Zhuoyun Li, Xinmiao Huang, Xiaowei Huang and Yi Dong

This paper primarily demonstrates a method to quantitatively assess the alignment between multi-step, structured reasoning in large language models and human preferences. We introduce the Alignment Score, a semantic-level metric that compares a model-produced chain of thought traces with a human-preferred reference by constructing semantic-entropy-based matrices over intermediate steps and measuring their divergence. Our analysis shows that Alignment Score tracks task accuracy across models and hop depths, and peaks at 2-hop reasoning. Empirical results further indicate that misalignment at greater reasoning depths is driven mainly by alignment errors such as thematic shift and redundant reasoning. Viewing chain sampling as drawing from a distribution over reasoning paths, we empirically demonstrate a strong and consistent correlation between Alignment Score and accuracy, readability, and coherence, supporting its use as a diagnostic signal. The code is available.

[Long] Chart-MRAG: Benchmarking Multimodal Retrieval Augmented Generation on Chart-based Documents ► *Underline Video*

Yuming Yang

Multimodal Retrieval-Augmented Generation (MRAG) enhances reasoning capabilities by integrating external knowledge. However, existing benchmarks primarily focus on simple image-text interactions, overlooking complex visual formats like charts that are prevalent in real-world applications. In this work, we introduce a novel task, Chart-based MRAG, to address this limitation. To generate high-quality evaluation samples, we propose CHARGE (CHARt-based document question-answering GEneration), a semi-automatic framework for generating evaluation samples through multi-modal keypoint extraction, knowledge graph construction, and qa pair synthesis. By combining CHARGE with expert validation, we construct Chart-MRAG Bench, a comprehensive benchmark for chart-based MRAG evaluation, featuring 4,738 question-answering pairs across 8 domains from real-world documents. Our experiments reveal three critical limitations in current approaches: (1) unified multimodal embedding retrieval methods struggles in chart-based scenarios, (2) even with ground-truth retrieval, state-of-the-art Multimodal Large Language Models (MLLMs) achieve only 71.15% Correctness and 80.74% Coverage scores, and (3) Widely-used MLLMs demonstrate consistent text-over-visual modality bias. These findings highlight great challenges in processing information-dense visual formats. We will make our code and dataset publicly available.

[Long] Check Your Work: Structured Checklist Feedback for Improving Large Language Models ► *Underline Video*

Jonathan Cook, Tim Rocktäschel, Jakob Nicolaus Foerster, Dennis Aumiller and Alex Wang

Much recent progress in Large Language Model (LLM) performance has been driven by verifiable feedback in deterministic domains like mathematics and code. However, scaling reinforcement learning (RL) and test-time compute in domains for which strict verification is infeasible remains a challenge. A common approach is to use an LLM-as-judge, which often relies on opaque, monolithic scores. In this work, we propose that AI feedback is most effective when decomposed into granular, prompt-specific checklists. To transform these checklists into a scalar reward, we introduce DIVA: DIscriminative VAriance weighting, a dynamic aggregation scheme that prioritises checklist items based on their ability to distinguish quality across a candidate pool. This ensures the reward signal focuses on the most salient criteria for a given prompt and response group, rather than being diluted by trivial or redundant constraints. Our approach yields an 11.8% win-rate improvement on AlpacaEval 2.0 using Qwen3-8B, outperforming holistic reward models and existing checklist baselines. Beyond training, we show that these checklists serve as a structured policy improvement operator at inference time - by using the models own

checklist evaluation as localised contextual feedback, the model can iteratively refine its output. This self-correction mechanism outperforms free-form sequential self-correction, offering a unified and interpretable framework for scaling both training-time and test-time performance in domains lacking strict verifiers.

[Long] ChemReason-Bench: Benchmarking Large Language Models for Procedural Reasoning in Experimental Chemistry ▶ *Underline Video*

Jinwei Zhang, Xucheng Liang, Yu Zhang, Ruijie Yu, Xiaokang Yang, Yaohui Jin and Yanyan Xu

Experimental protocols in organic synthesis specify not only the intended transformation but also an executable sequence of operations and conditions. While recent language models show strong chemistry knowledge, widely used evaluations remain less diagnostic of procedure-level decision making. In this setting, correctness requires consistent step ordering, feasibility under stated conditions, faithful entity-role grounding, and schema-parseable outputs that can be automatically validated against operational constraints. We present ChemReason-Bench, a human-validated benchmark for verifiable experimental procedure reasoning built on a structured representation with explicit placeholders and a unified schema, enabling automatic checks of many operational constraints. From 500 reactions, we instantiate 7306 benchmark tasks across six complementary formats: ordering, step validation, condition validation, schema-constrained completion, contrastive choice, and evidence-grounded rationalization. We further release a large-scale instantiation of the same templates for downstream adaptation studies, kept disjoint from the evaluation set. Using a unified evaluation protocol, we benchmark diverse open-source, proprietary, and domain-specific models and observe clear variation across the capability surface. We also report controlled adaptation experiments in the appendix, where supervised fine-tuning improves small models, preference optimization adds limited gains in our setting, and a gap remains to the strongest evaluated systems.

[Long] ChessArena: A Chess Testbed for Evaluating Strategic Reasoning Capabilities of Large Language Models ▶ *Underline Video*

Jincheng Liu, Sijun He, Jingjing Wu, Xiangsen Wang, Yang Chen, Zhaoqi Kuang, Siqi Bao and Yuan Yao

Recent large language models (LLMs) have shown strong reasoning capabilities. However, a critical question remains: do these models possess genuine strategic reasoning, or do they primarily excel at pattern recognition? To address this, we present ChessArena, a chess-based testbed for evaluating LLMs. Chess demands strategic reasoning, precise rule adherence, and the ability to track complex game states. ChessArena is a competitive framework where LLMs play against each other under four play modes. We evaluate 13 LLMs across over 800 games, testing basic understanding, move selection, and puzzle solving. Results reveal significant shortcomings: no model beats Maia-1100 (human amateur level), and some lose to random play. We also present a strong baseline: our fine-tuned Qwen3-8B substantially improves performance, approaching much larger state-of-the-art reasoning models.

[Long] Chunks as Arms: Multi-Armed Bandit-Guided Sampling for Long-Context LLM Preference Optimization ▶ *Underline Video*

Shaohua Duan, Pengcheng Huang, Xinze Li, Zhenghao Liu, Xiaoyuan Yi, Yukun Yan, Shuo Wang, Yu Gu, Ge Yu and Maosong Sun

Long-context modeling is critical for a wide range of real-world tasks, including long-context question answering, summarization, and complex reasoning tasks. Recent studies have explored fine-tuning Large Language Models (LLMs) with synthetic data to enhance their long-context capabilities. However, the effectiveness of such approaches is often limited by the low diversity and factual inconsistencies in the generated data. To address these challenges, we propose LongMab, a novel framework that leverages a Multi-Armed Bandit (MAB) rollout strategy to identify the most informative chunks from the given long context for sampling high-quality and diverse responses and constructing preference data pairs for Direct Preference Optimization (DPO) training. Specifically, we treat context chunks as arms of MAB, select chunks based on their expected reward scores to input into LLMs to generate responses, and iteratively update these scores based on reward feedback. Both exploration and exploitation during the rollout process enable the LLM to focus on the most relevant context segments, thereby generating and collecting high-quality and diverse responses. Experimental results on both Llama and Qwen show the effectiveness of LongMab by achieving more than a 4% improvement on long-context reasoning benchmarks. All data and code will be released on https://github.com/NEUIR/LongMab-PO.

[Long] CO-EVO: Co-evolving Semantic Anchoring and Style Diversification for Federated DG-ReID ▶ *Underline Video*

Fengchun Zhang, Jianwei Hu, Qiang Ma, Liuyu Xiang, Jinshan Lai and Tingxuan Huang

Federated domain generalization for person re-identification (FedDG-ReID) aims to collaboratively train a pedestrian retrieval model across multiple decentralized source domains such that it can generalize to unseen target environments without compromising raw data privacy. However, this task is significantly challenged by the inherent stylistic gaps across decentralized clients. Without global supervision, models easily succumb to shortcut learning where representations overfit to domain specific camera biases rather than universal identity features. We propose CO-EVO, a novel federated framework that resolves this semantic-style conflict through a co-evolutionary mechanism. On the semantic side, Camera-Invariant Semantic Anchoring (CSA) learns identity prompts with cross-camera consistency to establish purified and domain-agnostic anchors that filter out local imaging noise. On the visual side, Global Style Diversification (GSD), powered by a Global Camera-Style Bank (GCSB), synthesizes realistic perturbations to expand the visual boundaries of training data. The core of CO-EVO is its co-evolutionary loop where purified anchors act as gravitational centers to guide the image encoder toward robust anatomical attributes amidst diverse style variations. Extensive experiments demonstrate that CO-EVO achieves state-of-the-art (SOTA) performance, proving that the synergy between semantic purification and style expansion is essential for robust cross-domain generalization. Our code is available at: https://github.com/NanYiyuzurn/ACL-LGPS-2026.

[Long] CoCoNUTS: Concentrating on Content while Neglecting Uninformative Textual Styles for AI-Generated Peer Review Detection ▶ *Underline Video*

Yihan Chen, Jiawei Chen, Guozhao Mo, Xuanang Chen, Ben He, Xianpei Han and Le Sun

The growing use of large language models (LLMs) in peer review threatens scholarly integrity. Recent conference policies allow AI tools for language polishing but prohibit their use for generating substantive content. However, existing detectors mainly rely on stylistic cues, making it difficult to distinguish between surface-level language refinement and genuine content generation. To address this, we advocate a content-based detection paradigm and introduce CoCoNUTS, a comprehensive benchmark containing 315,535 reviews covering leading AI conferences and six human-AI collaboration modes. Our evaluation shows that current detectors struggle to handle these nuanced settings. Consequently, we propose CoCoDet, an AI review detector designed to identify substantive AI-generation. Experiments demonstrate that CoCoDet achieves a macro F1-score of 98.24%. Crucially, on permissible machine-polished reviews, it maintains a low false positive rate of 3.89%, substantially outperforming the strongest baseline (7.84%). Examination on real-world reviews using CoCoDet reveals an escalating trend of substantive AI generation. Our work exposes the inadequacy of current detectors, underscoring the importance of domain-specific solutions.

[Long] CoEvolve: Training LLM Agents via Agent-Data Mutual Evolution ▶ *Underline Video*

Shidong Yang, Ziyu Ma, Tongwen Huang, Yiming Hu, Yong Wang and Xiangxiang Chu

Reinforcement learning for LLM agents is typically conducted on a static data distribution, which fails to adapt to the agent's evolving behavior and leads to poor coverage of complex environment interactions. To address these challenges, we propose CoEvolve, an agent-data mutual evolution framework that enables LLM agents to improve through closed-loop, interaction-driven training. Specifically, CoEvolve extracts feedback signals such as forgetting and uncertainty from rollout trajectories to identify failure-prone interaction patterns, and utilizes them to guide LLM-based task synthesis. The synthesized tasks are validated through environment interaction and utilized to update the data distribution, enabling joint adaptation of the agent and its data. Extensive experiments on AppWorld and BFCL across Qwen2.5-7B, Qwen3-4B, and Qwen3-30B-A3B demonstrate consistent and significant improvements over strong base models, yielding absolute gains of 19.43%, 15.58%, and 18.14%, respectively.

[Long] CogEvolve: A Multimodal Benchmark for Evaluating Relational Reasoning in Semantic Extension ▶ *Underline Video*

Jingjie Zeng, Huayang Li, Liang Yang, Yuanyuan Sun, SHAOWU ZHANG and Hongfei Lin

Human cognition excels at extending knowledge through analogy, where word meanings evolve along structured pathways from concrete prototypes to abstract senses via metaphor and metonymy. Do Large Language Models (LLMs) internalize this generative logic, or merely mimic statistical patterns? To investigate this, we introduce CogEvolve, a cognitive linguistic benchmark designed to test these evolutionary pathways across textual and visual modalities. Our evaluation reveals a distinct cognitive profile: models function as "Super-Associators" expert at static recognition yet fail at causal reasoning. In text, they exhibit a Frequency-Primacy Conflation, confusing statistical prevalence with cognitive basicness. Crucially, this reasoning collapses further in the visual domain. We term this deficit the Ungrounded Arrow: models possess high-fidelity concept representations (the "dots") but lack the transformational operators (the "arrows") essential for true relational understanding.

[Long] CoLA: A Choice Leakage Attack Framework to Expose Privacy Risks in Subset Training ▶ *Underline Video*

Qi Li, Cheng-Long Wang, Yinzhi Cao and Di Wang

Training models on a carefully chosen portion of data rather than the full dataset is now a standard preprocess for modern ML. From vision coreset selection to large-scale filtering in language models, it enables scalability with minimal utility loss. A common intuition is that training on fewer samples should also reduce privacy risks. In this paper, we challenge this assumption. We show that subset training is not privacy free: the very choices of which data are included or excluded can introduce new privacy surface and leak more sensitive information. Such information can be captured by adversaries either through side-channel metadata from the subset selection process or via the outputs of the target model. To systematically study this phenomenon, we propose CoLA (Choice Leakage Attack), a unified framework for analyzing privacy leakage in subset selection. In CoLA, depending on the adversary's knowledge of the side-channel information, we define two practical attack scenarios: Subset-aware Side-channel Attacks and Black-box Attacks. Under both scenarios, we investigate two privacy surfaces unique to subset training: (1) Training-membership MIA (TM-MIA), which concerns only the privacy of training data membership, and (2) Selection-participation MIA (SP-MIA), which concerns the privacy of all samples that participated in the subset selection process. Notably, SP-MIA enlarges the notion of membership from model training to the entire data-model supply chain. Experiments on vision and language models show that existing threat models underestimate subset-training privacy risks: the expanded privacy surface leaks both training and selection membership, extending risks from individual models to the broader ML ecosystem.

[Long] Collaborative Chain-of-Agents for Parametric-Retrieved Knowledge Synergy ▶ *Underline Video*

Yi Jiang, Sendong Zhao, Jianbo Li, Haochun Wang, Lizhe Zhang, Yan Liu and Bing Qin

Retrieval-Augmented Generation (RAG) enhances Large Language Models (LLMs), especially for knowledge-intensive tasks. Despite its advantages, current RAG methods often struggle to fully exploit knowledge during generation. In particular, the synergy between the models internal parametric knowledge and external retrieved knowledge remains limited. Retrieved contents may sometimes mislead generation, while certain generated content can guide the model toward more accurate outputs. In this work, we propose Collaborative Chain-of-Agents, a framework designed to enhance explicitly synergy over both parametric and retrieved knowledge. Specifically, we

first introduce CoCoA-zero, a multi-agent RAG framework that first performs conditional knowledge induction and then reasons answers. Building on this, we develop CoCoA, a long-chain training strategy that synthesizes extended multi-agent reasoning trajectories from CoCoA-zero to fine-tune the LLM. This strategy enhances the models capability to explicitly integrate and jointly leverage parametric and retrieved knowledge. Experimental results demonstrate the superiority of CoCoA in open-domain QA and multi-hop QA.

[Findings] ComfyFlow: Benchmarking LLMs for AIGC Workflow Generation ▶ *Underline Video*

Zhenran Xu, Yiyu Wang, yunxin li, Muyang Ye, Yangxue Yangxue, Kai Chen, Longyue Wang, Weihua Luo, Baotian Hu and Min Zhang

Large language models (LLMs) have shown promising advancements in tackling human-level tasks, wherein generating workflows for collaborative AI systems remains a critical and challenging step. To explore this frontier, we introduce ComfyFlow, a comprehensive benchmark to evaluate current LLMs' ability to generate executable and instruction-following AIGC workflows in ComfyUI. The dataset includes 400 diverse visual generation tasks across 20 categories, supported by 10K training examples constructed from knowledge bases, which contain detailed annotations for 2,480 nodes and 3,298 workflows. We establish a systematic evaluation protocol that quantifies performance across multiple dimensions, ranging from basic format validity to multi-level hallucination rates. Our extensive evaluations show that: 1) ComfyFlow presents a substantial challenge even for top-tier proprietary LLMs such as GPT-5.1 and the Claude series; 2) Open-source models achieve new state-of-the-art results after post-training, yet struggle with long-horizon planning as the number of nodes increases; 3) Different post-training strategies offer complementary benefits in following instructions and mitigating hallucinations. By establishing both a challenging benchmark and a principled evaluation scheme, ComfyFlow lays the foundation for developing more intelligent and reliable collaborative AI systems.

[Findings] ComfyUI-R1: Exploring Reasoning Models for Workflow Generation ▶ *Underline Video*

Zhenran Xu, Yiyu Wang, Yangxue Yangxue, Longyue Wang, Weihua Luo, Kaifu Zhang, Baotian Hu and Min Zhang

AI-generated content has evolved from monolithic models to modular workflows, particularly on platforms like ComfyUI, enabling customization in creative pipelines. However, crafting effective workflows requires great expertise to orchestrate numerous specialized components, presenting a steep learning curve for users. To address this challenge, we introduce ComfyUI-R1, the first large reasoning model for automated workflow generation. Starting with our curated knowledge bases, we construct long chain-of-thought (CoT) reasoning data, including node selection, workflow planning, and code-level workflow representation. ComfyUI-R1 is trained through a two-stage framework: (1) CoT fine-tuning for cold start, adapting models to the ComfyUI domain; (2) reinforcement learning for incentivizing reasoning capability, guided by a fine-grained rule-metric hybrid reward, ensuring format validity, structural integrity, and node-level fidelity. Experiments show that our 7B-parameter model achieves a 97% format validity rate, along with high pass rate, node-level and graph-level F1 scores, surpassing prior state-of-the-art methods that employ leading closed-source models such as GPT-4o and Claude series. Qualitative analysis further highlights our model's strength in synthesizing intricate workflows with diverse nodes, aligning with human instructions, and generalizing to newly introduced nodes, underscoring the potential of long CoT reasoning in AI art creation.

[Long] Compressing then Matching: An Efficient Pre-training Paradigm for Multimodal Embedding ▶ *Underline Video*

Da Li, Yuxiao Luo, Keping Bi, Jiafeng Guo, Wei Yuan, Biao Yang, Yan Wang, Fan Yang, Tingting Gao and Guorui Zhou

Multimodal Large Language Models advance multimodal representation learning by acquiring transferable semantic embeddings, thereby substantially enhancing performance across a range of vision-language tasks, including cross-modal retrieval, clustering, and classification. An effective embedding is expected to comprehensively preserve the semantic content of the input while simultaneously emphasizing features that are discriminative for downstream tasks. Recent approaches demonstrate that MLLMs can be adapted into competitive embedding models via large-scale contrastive learning, enabling the simultaneous optimization of two complementary objectives. We argue that the two aforementioned objectives can be decoupled: a comprehensive understanding of the input enables the embedding model to achieve superior performance on downstream tasks via contrastive learning. In this paper, we propose CoMa, a compressed pre-training phase, which serves as a warm-up stage for contrastive learning. Experiments demonstrate that with only a small amount of pre-training data, we can transform an MLLM into a competitive embedding model. CoMa achieves new state-of-the-art results among MLLMs of comparable size on the MMEB, realizing optimization in both efficiency and effectiveness. Our project is available at https://github.com/Trustworthy-Information-Access/CoMa.

[Long] Conditional Memory via Scalable Lookup: A New Axis of Sparsity for Large Language Models ▶ *Underline Video*

Xin Cheng, Wangding Zeng, Damai Dai, Qinyu Chen, Bingxuan Wang, Zhenda Xie, Kezhao Huang, Xingkai Yu, Zhewen Hao, han zhang, Yu-Kun Li, Huishuai Zhang, Dongyan Zhao and Wenfeng Liang

Mixture-of-Experts (MoE) scales capacity via conditional computation, but Transformers lack a native knowledge lookup primitive. We introduce conditional memory, instantiated via Deep Sparse Embedding (DSE), which indexes a massive embedding table using local n-grams for retrieval. We formalize sparsity allocation problem—how to split a fixed parameter budget between MoE experts and DSE memory—and find a U-shaped scaling law that identifies an optimal balance. Scaling to 27B parameters, DSE outperform an iso-parameter and iso-FLOPs MoE baseline across knowledge and reasoning benchmarks, and achieve markedly stronger long-context performance. Mechanistic analyses show that DSE offloads early-layer static recall into memory, freeing effective depth and attention for higher-level reasoning. DSE is also infrastructure-efficient: its deterministic hashing enables offloading massive parameters into host memory during inference with negligible throughput overhead.

[Findings] Context-Value-Action Architecture for Value-Driven Large Language Model Agents ► *Underline Video*

TianZe Zhang, Sirui Sun, Yuhang Xie, Xin Zhang, Zhiqiang Wu and Guojie Song

Large Language Models (LLMs) have shown promise in simulating human behavior, yet existing agents often exhibit behavioral rigidity, a flaw frequently masked by the self-referential bias of current "LLM-as-a-judge" evaluations. By evaluating against empirical ground truth, we reveal a counter-intuitive phenomenon: increasing the intensity of prompt-driven reasoning does not enhance fidelity but rather exacerbates value polarization, collapsing population diversity. To address this, we propose the Context-Value-Action (CVA) architecture, grounded in the Stimulus-Organism-Response (S-O-R) model and Schwartzs Theory of Basic Human Values. Unlike methods relying on self-verification, CVA decouples action generation from cognitive reasoning via a novel Value Verifier trained on authentic human data to explicitly model dynamic value activation. Experiments on CVABench, which comprises over 1.1 million real-world interaction traces, demonstrate that CVA significantly outperforms baselines. Our approach effectively mitigates polarization while offering superior behavioral fidelity and interpretability.

[Long] ContrastKV: Robust KV Cache Eviction via Contrastive Signal Fusion for Multi-Query Generalization ► *Underline Video*

Xingchi Chen, Peiyuan Zong, Ziqiang Gao, Qing Li, Yong Jiang, Fa Zhu and Hui Li

Large Language Models (LLMs) face significant memory and latency overheads during long-context inference due to the growing KV cache, especially in Knowledge Base Question Answering (KBQA) settings that require support for multiple downstream queries. Query-aware eviction methods do not generalize across queries, while existing query-agnostic approaches rely on a single proxy query, leading to fragile eviction decisions under high eviction ratios. We propose ContrastKV, a robust query-agnostic KV cache eviction algorithm for multi-query generalization. ContrastKV introduces a contrastive signal fusion mechanism that jointly exploits complementary semantic and non-semantic signals. By contrasting semantic consistency with structural robustness, the method constructs a more reliable eviction criterion that alleviates the blind spots of single-query proxies. The framework integrates efficient signal generation, parallel importance scoring, and multi-level fusion across heads and layers. Experiments show that ContrastKV outperforms state-of-the-art methods, retaining up to 92% accuracy with only 20% of the KV cache budget, while reducing decoding latency by approximately 50% and significantly lowering GPU memory usage.

[Findings] Corpora Generation for Urdu Grammatical Error Correction ► *Underline Video*

Syed Ahad and Burhanuddin Ezzi

Grammatical Error Correction (GEC) for Urdu remains an under-researched area due to the lack of annotated datasets. This paper addresses the challenge of generating a robust corpus for fine-tuning deep learning models aimed at Urdu GEC. We propose a method for synthesizing a large dataset by collecting errors from the Urdu WikiEdits history, learning from them, and inserting similar errors in grammatically correct sentences to generate incorrect sentences with grammatical errors, hence creating a pair of grammatically correct and incorrect sentences. We introduce UrduGEC-Synthetic, a synthetically generated dataset produced through this pipeline. Furthermore, we introduce UrduGEC-Gold, a Gold Dataset by extracting errors from exam copies of students. Finally, we also fine-tuned various models on UrduGEC-Synthetic and evaluated them against UrduGEC-Gold to show the quality of synthetic data generation.

[SRW] Counterspeech Generation using Small Language Models ► *Underline Video*

Abubakar Muhammad, Simona Frenda and Gavin Abercrombie;

Counterspeech offers a way to tackle harmful content online without restricting freedom of expression. This work explores counterspeech generation using small language models (SLMs) as lightweight and cost-effective alternatives to large language models. We evaluate SLMs ranging from 100 million to 3 billion parameters using simple prompting strategies as well as fine-tuning, combining automatic and robust human evaluations. Our findings demonstrate that small language models can generate relevant, coherent, and high-quality counterspeech, suggesting their potential suitability for efficient and responsible deployments.

[Long] CPT-Agent: A Cognitive Process Theory-driven Framework for Student Simulation in Writing Development ► *Underline Video*

Yuhan Chen, Zizhuo Shen, Miaomiao Cheng, Jiefu Gong, Xu Han, Shijin Wang and Wei Song

Simulating student writing behavior offers a promising pathway to scalable feedback evaluation and teacher training. However, existing LLM-based approaches tend to model overly capable learners who readily understand and over-apply feedback, resulting in pedagogically implausible behavior. In this work, we introduce pedagogical realism as a guiding principle for student writing simulation, emphasizing bounded cognition, selective feedback comprehension, and developmentally plausible learning processes. To operationalize this idea, we propose CPT-Agent, a cognitively grounded framework that decouples cognitive ability from writing proficiency and models their interaction during writing and revision. CPT-Agent combines probabilistic modeling of cognitive development, proficiency-controlled text generation, and structured memory for skill accumulation. Experiments show that it (1) produces clearly distinguishable proficiency levels, (2) generates cognitively plausible revisions consistent with instructional theories, and (3) achieves strong agreement with expert judgments in evaluating feedback quality. These results highlight the importance of modeling cognitive constraints in LLM-based student simulation and demonstrate the potential of pedagogically realistic agents for automated feedback assessment and teacher development.

[Long] CreditDecoding: Accelerating Parallel Decoding in Diffusion Large Language Models with Trace Credit ► *Underline Video*

Kangyu Wang, Zhiyun Jiang, Haibo Feng, Weijia Zhao, Lin Liu, Jianguo Li, Zhenzhong Lan and Weiyao Lin

Diffusion large language models (dLLMs) generate text through iterative denoising. In commonly adopted parallel decoding schemes,

each step confirms only high-confidence positions while remasking the others. By analyzing dLLM denoising traces, we uncover a key inefficiency: models often predict the correct target token several steps before its confidence becomes high enough to be decoded. This gap between early prediction and late decoding forces repeated remasking of already-correct tokens, causing redundant iterations and limiting acceleration. To exploit this temporal redundancy, we introduce Trace Credit to quantify a token's decoding potential by accumulating historical evidence. Building on this, we propose CreditDecoding, a training-free parallel decoding method that fuses Trace Credit with current logits to boost the confidence of correct but underconfident tokens, thereby accelerating denoising and improving robustness. On eight benchmarks, CreditDecoding achieves up to 5.48 times speedup with +0.48 accuracy on LLaDA-8B and consistently improves performance across diverse dLLM architectures and parameter scales. It further scales to long contexts and remains orthogonal to mainstream inference optimizations, making it a practical and widely applicable solution.

[Long] CT-FineBench: A Diagnostic Fidelity Benchmark for Fine-Grained Evaluation of CT Report Generation ▶ ***Underline Video***
Ruifeng Yuan, Wanxing Chang, Weiwei Cao, Bowen Shi, Zhongyu Wei, Ling Zhang and Jianpeng Zhang
The evaluation of generated reports remains a critical challenge in Computed Tomography (CT) report generation, due to the large volume of text, the diversity and complexity of findings, and the presence of finegrained, diseaseoriented attributes. Conventional evaluation metrics offer only coarse measures of lexical overlap or entity matching and fail to reflect the granular diagnostic accuracy required for clinical use. To address this gap, we propose CT-FineBench, a benchmark built from CT-RATE and Merlin to evaluate the fine-grained factual consistency of CT reports, constructed from CT-RATE and Merlin. Our benchmark is constructed through a meticulous, Question-Answering (QA) based process: first, we identify and structure key, finding-specific clinical attributes (e.g., location, size, margin). Second, we systematically transform these attributes into a QA dataset, where questions probe for specific clinical details grounded in gold-standard reports. The evaluation protocol for CT-FineBench involves using this QA dataset to query a machine-generated report and scoring the correctness of the answers. This allows for a comprehensive, interpretable, and clinically-relevant assessment, moving beyond superficial lexical overlap to pinpoint specific clinical errors. Experiments show that CT-FineBench correlates better with expert clinical assessment and is substantially more sensitive to fine-grained factual errors than prior metrics.

[Long] Culture-Aware Machine Translation in Large Language Models: Benchmarking and Investigation ▶ ***Underline Video***
Zekun Yuan, Yangfan Ye, Xiaocheng Feng, Baohang Li, Qichen Hong, Yunfei Lu, Dandan Tu and Bing Qin
Large language models (LLMs) have achieved strong performance in general machine translation, yet their ability in culture-aware scenarios remains poorly understood. To bridge this gap, we introduce CanMT, a Culture-Aware Novel-Driven Parallel Dataset for Machine Translation, together with a theoretically grounded, multi-dimensional evaluation frame work for assessing cultural translation quality. Leveraging CanMT, we systematically evaluate a wide range of LLMs and translation systems under different translation strategy constraints. Our findings reveal substantial performance disparities across models and demonstrate that translation strategies exert a systematic influence on model behavior. Further analysis shows that translation difficulty varies across types of culture-specific items, and that a persistent gap remains between modelsrecognition of culture-specific knowledge and their ability to correctly operationalize it in translation outputs. In addition, incorporating reference translations is shown to substantially improve evaluation reliability in LLM-as-a-judge, underscoring their essential role in assessing culture-aware translation quality. The corpus and code are available at CanMT.

[Long] CuMA: Aligning LLMs with Sparse Cultural Values via Demographic-Aware Mixture of Adapters ▶ ***Underline Video***
Ao Sun, Xiaoyu Wang, Zhe Tan, Yu Li, Jiachen Zhu, Yuheng Jia and Shu Su
As Large Language Models (LLMs) serve a global audience, alignment must transition from enforcing universal consensus to respecting cultural pluralism. We demonstrate that dense models, when forced to fit conflicting value distributions, suffer from Mean Collapse, converging to a generic average that fails to represent diverse groups. We attribute this to Cultural Sparsity, where gradient interference prevents dense parameters from spanning distinct cultural modes. To resolve this, we propose CuMA (Cultural Mixture of Adapters), a framework that frames alignment as a conditional capacity separation problem. By incorporating demographic-aware routing, CuMA internalizes a Latent Cultural Topology to explicitly disentangle conflicting gradients into specialized expert subspaces. Extensive evaluations on WorldValuesBench, Community Alignment, and PRISM demonstrate that CuMA achieves competitive performance, outperforming both dense baselines and semantic-only MoEs. Our analysis confirms that CuMA effectively mitigates mean collapse and preserves cultural diversity. Our code is available at https://github.com/Throll/CuMA.

[Long] CuMA: Aligning LLMs with Sparse Cultural Values via Demographic-Aware Mixture of Adapters ▶ ***Underline Video***
Ao Sun, Xiaoyu Wang, Zhe Tan, Yu Li, Jiachen Zhu, Yuheng Jia and Shu Su
As Large Language Models (LLMs) serve a global audience, alignment must transition from enforcing universal consensus to respecting cultural pluralism. We demonstrate that dense models, when forced to fit conflicting value distributions, suffer from Mean Collapse, converging to a generic average that fails to represent diverse groups. We attribute this to Cultural Sparsity, where gradient interference prevents dense parameters from spanning distinct cultural modes. To resolve this, we propose CuMA (Cultural Mixture of Adapters), a framework that frames alignment as a conditional capacity separation problem. By incorporating demographic-aware routing, CuMA internalizes a Latent Cultural Topology to explicitly disentangle conflicting gradients into specialized expert subspaces. Extensive evaluations on WorldValuesBench, Community Alignment, and PRISM demonstrate that CuMA achieves competitive performance, outperforming both dense baselines and semantic-only MoEs. Our analysis confirms that CuMA effectively mitigates mean collapse and preserves cultural diversity. Our code is available at https://github.com/Throll/CuMA.

[Long] CAIRO: Contextual Hierarchical Analogical Induction and Reasoning Optimization for LLMs ► *Underline Video*

Haotian Lu, Yuchen Mou and Bingzhe Wu

Warning: This paper may contain content that could be disturbing or offensive. Content moderation in online platforms faces persistent challenges due to the evolving complexity of user-generated content and the limitations of traditional rule-based and machine learning approaches. While recent advances in large language models (LLMs) have enabled more sophisticated moderation via direct prompting or fine-tuning, these approaches often exhibit limited generalization, interpretability, and adaptability to unseen or ambiguous cases. In this work, we propose a novel moderation framework that leverages analogical examples to enhance rule induction and decision reliability. Our approach integrates end-to-end optimization of analogical retrieval, rule generation, and moderation classification, enabling the dynamic adaptation of moderation rules to diverse content scenarios. Through comprehensive experiments, we demonstrate that our method significantly outperforms both rule-injected fine-tuning baselines and multi-stage static RAG pipelines in terms of moderation accuracy and rule quality. Further evaluations, including human assessments and external model generalization tests confirm the superiority of rules generated by our framework in terms of clarity, interpretability, and applicability. These findings highlight the potential of analogical example-driven methods for advancing robust, explainable, and generalizable content moderation in real-world applications.

[Long] DARM: Distribution-Aware Reward Modeling by Alleviating Biases from Low Preference-Context Dependency Data ► *Underline Video*

Shaofan Liu, Guoqiang Zhang, Shihan Dou, Huiyuan Zheng, Yiming Zhou, Junjie Ye, Shaowen Wang, Shichun Liu, Jiazheng Zhang, Tao Gui, Qi Zhang and Xuanjing Huang

Reward models (RMs) are the surrogate objectives in reinforcement learning from human feedback (RLHF), and their scores directly steer policy optimization. We show that standard RM training is vulnerable in data subsets where response quality depends only weakly on the context: such instances encourage the RM to ignore the context, leading to context neglect and degraded accuracy. To address this failure mode, we propose Distribution-Aware Reward Modeling (DARM), which augments the RM objective with a conditional mutual information regularizer that maximizes context and the predicted reward conditioned on the response. By explicitly preserving the sensitivity of reward signals to the prompting context, DARM reduces over-reliance on response-only features and improves robustness to contextual variation. Extensive experiments across in-distribution and out-of-distribution settings show that DARM trained RMs deliver more accurate and consistent scoring than strong baselines. We further evaluate its downstream impact in RLHF, where DARM produce better aligned policies. We also demonstrate the necessity of each DARM design component and the impact of key parameters on performance through ablation experiments.

[Long] DART: Disambiguation-Aware Reasoning for Video-guided Machine Translation ► *Underline Video*

Boyu Guan, Chuang Han, Yang Zhao and Chengqing Zong

Video-guided Machine Translation (VMT) seeks to enhance translation quality by incorporating contextual information derived from paired short video clips. However, many VMT samples are text-sufficient; even when visual information is needed, only minimal cues are required. Aiming to tackle these issues, we propose a novel framework DART (Disambiguation-Aware Reasoning for Video-guided Machine Translation). Reinforcement learning is used to incorporate multimodal large language models multimodal reasoning into VMT. The model dynamically switches between text-only processing and multimodal integration, contingent on the necessity of visual disambiguation. Furthermore, we present TVRF (Translation-oriented Video Relevance Filtering), a systematic pipeline for constructing training data based on multimodal relevance to translation. This pipeline filters samples where video information is translation-relevant, mitigating training collapse caused by video-irrelevant data in conventional VMT. Experimental results show that our approach improves multimodal information utilization in VMT, yielding gains in both translation quality and computational efficiency.

[Long] De-Anonymization at Scale via Tournament-Style Attribution ► *Underline Video*

Lirui Zhang and Huishuai Zhang

As LLMs rapidly advance and enter real-world use, their privacy implications are increasingly important. We study an authorship de-anonymization threat: using LLMs to link anonymous documents to their authors, potentially compromising settings such as double-blind peer review. We propose De-Anonymization at Scale (DAS), a large-language-model-based method for attributing authorship among tens of thousands of candidate texts. DAS uses a sequential progression strategy: it randomly partitions the candidate corpus into fixed-size groups, prompts an LLM to select the text most likely written by the same author as a query text, and iteratively re-queries the surviving candidates to produce a ranked top-k list. To make this practical at scale, DAS adds a dense-retrieval prefilter to shrink the search space and a majority-voting-style aggregation over multiple independent runs to improve robustness and ranking precision. Experiments on anonymized review data show DAS can recover same-author texts from pools of tens of thousands with accuracy well above chance, demonstrating a realistic privacy risk for anonymous platforms. On standard authorship benchmarks (Enron emails and blog posts), DAS also improves both accuracy and scalability over prior approaches, highlighting a new LLM-enabled de-anonymization vulnerability.

[Long] DEBAR: Mitigating Contextual Bias in Cross-Document Relation Extraction via Dual-Stream Decoupling ► *Underline Video*

Zhixuan Yang, Fu Zhang, Huangming Xu and Jingwei Cheng

Cross-document Relation Extraction (CodRE) requires reasoning over scattered evidence to identify relations between target entities across multiple documents. Existing methods indiscriminately fuse target entities and the intermediate bridge entities that link them into a unified

representation. This leads to intermediate evidence that often aligns with only one side of the entity pair, resulting in one-sided relation transfer contextual bias and incomplete reasoning chains. Moreover, these methods typically employ a global threshold to determine relation existence for all entity pairs, limiting the model's reasoning performance. To address these issues, we propose DEBAR (Dual-stream Entity Bias Reduction), a framework designed to explicitly decouple and preserve bidirectional bridge evidence, combined with a novel dynamic loss optimization objective. Specifically, DEBAR employs a bridge-aware input construction strategy and a dual-stream graph reasoning network to separately encode head and tail contexts, preventing semantic interference while capturing global dependencies through iterative message passing. Furthermore, we introduce a curriculum-aware ranking optimization objective that progressively tightens classification constraints to stabilize training and enforce discriminative decision boundaries. Experiments on the CodRE benchmarks show that DEBAR achieves state-of-the-art performance while effectively mitigating cross-document contextual bias. Moreover, extensive experiments on our proposed loss across backbones confirm its generalization, suggesting it as a reliable replacement for existing CodRE losses. Code is available at https://github.com/newyuyou/DEBAR.

[Long] DecIF: Improving Instruction-Following through Decomposition ▶ *Underline Video*

Tingfeng Hui, Pengyu Zhu, Bowen Ping, Ling Tang, Guanting Dong, Yaqi Zhang and Sen Su

We propose a novel data synthesis framework, DecIF, which automatically generates accurate and diverse instruction-following data from scratch for supervised fine-tuning (SFT) and reinforcement learning (RL), leveraging large language models (LLMs) and minimal external resources. By decomposing the data synthesis pipeline into fine-grained steps, DecIF achieves meticulous quality and diversity control over generated instruction-following data. Extensive experiments across both SFT and RL demonstrate DecIFs strong capability to flexibly synthesize accurate instruction-following data for both paradigms compared to comprehensive baselines. Further analysis demonstrates the framework's robustness, scalability, and computational efficiency in instruction-following data generation, while its modular design ensures straightforward implementation and reproducibility.

[Long] DecoCal: Decoding with Calibration in Diffusion Large Language Models ▶ *Underline Video*

Fan Xu, Huixuan Zhang and Xiaojun Wan

Diffusion Large Language Models (DLLMs) generate text via iterative masked-token denoising, supporting parallel prediction and bidirectional context modeling. Despite these advantages, decoding remains challenging: many tokens appear predictable early, yet single-step predictions are often unstable, exhibiting temporal oscillations or overconfidence, making it difficult to determine which tokens can be safely committed. To address these challenges, we propose DecoCal, a Decoding framework that explicitly performs Calibration of token-level confidence across diffusion steps and leverages the calibrated results to guide decoding decisions. Specifically, DecoCal aggregates historical predictions to maintain calibrated confidence, triggering unmasking only when a token is sufficiently stable, while a remasking mechanism allows revision of premature commitments. This calibration-based design enables early decoding of reliably converged tokens while deferring or correcting unstable ones, balancing reliability and speed. Experiments on multiple DLLMs and benchmarks show that DecoCal improves generation accuracy compared to existing strategies. Our results highlight the importance of temporal calibration in unlocking the full potential of diffusion-based language generation.

[Long] Decoding-Unlearning: Fact Forgetting via Entropy-Guided Inference ▶ *Underline Video*

Jingwen Pu, Mingjun Shi, Xinrui Ren, Yizhe Wang, Xinyu Zhang, Zhaokun Wang and Kun She

Large Language Models (LLMs) exhibit powerful capabilities but inevitably memorize sensitive information, raising privacy, copyright, and safety concerns. Existing LLM unlearning methods typically rely on updating model parameters. While effective, they are often limited in real-world scenarios: fine-tuning large-scale models is costly, may introduce potential irreversible risks, and depends on both forget and retain datasets, which are often difficult to obtain in full. To address these challenges, an ideal solution is to achieve unlearning at inference time. To this end, we propose SEGUE, a training-free, plug-and-play inference-time unlearning strategy. SEGUE employs a probe to detect queries involving forgettable concepts and applies entropy-guided decoding to suppress target knowledge, enabling controllable non-factual generation while preserving overall model capabilities. Experiments on the MUSE, RWKU, and WMDP datasets, covering copyright, entity, and potential-risk knowledge, show that SEGUE effectively balances sensitive knowledge suppression and generation quality, outperforming existing most inference-time unlearning methods.

[Short] Decoupling Generalization and Adaptation in Meta-Learning for Large Language Models ▶ *Underline Video*

Nitin Vetcha, Binqian Xu and Dianbo Liu

Fine-tuning large language models (LLMs) for downstream tasks remains expensive, even with parameter-efficient methods like Low-Rank Adaptation (LoRA). In this regard, meta-learning approaches such as Model-Agnostic Meta-Learning for LLMs (MAML-en-LLM) and Amortized Bayesian Meta-Learning for LoRA (ABMLL) have emerged as promising solutions for rapid downstream LLM adaptation. However, these methods fundamentally couple two distinct objectives: learning generalizable initializations and enabling efficient task adaptation. We argue that this coupling limits both the quality of learned representations and adaptation efficiency. In this paper, we introduce DeGAML-LLM (Decoupled Generalization and Adaptation in Meta-Learning for LLMs), a novel framework that explicitly separates these two objectives through dedicated parameter spaces. Specifically, we maintain a generalization module that learns task-agnostic representations across the task distribution, and an adaptation module that specializes in rapid task-specific adjustment. Extensive experiments on common-sense reasoning, mathematics, logic, social, medical and coding benchmarks across model scales demonstrate that DeGAML-LLM outperforms existing meta-learning and standard multi-task baselines.

[Long] Deep-Reporter: Deep Research for Grounded Multimodal Long-Form Generation ▶ *Underline Video*

Fangda Ye, Kuicai Dong, Zhifei Xie, Yuxin Hu, Yihang Yin, Shurui Huang, Shikai Dong, Chen Zhang, Jianzhu Bao and Shuicheng Yan

Recent agentic search frameworks enable deep research via iterative planning and retrieval, reducing hallucinations and enhancing factual grounding. However, they remain text-centric, overlooking the multimodal evidence that characterizes real-world expert reports. We introduce a pressing task: multimodal long-form generation. Accordingly, we propose Deep-Reporter, a unified agentic framework for grounded multimodal long-form generation. It orchestrates: (i) Agentic Multimodal Search and Filtering to retrieve and filter textual passages and information-dense visuals; (ii) Checklist-Guided Incremental Synthesis to ensure coherent image-text integration and optimal citation placement; and (iii) Recurrent Context Management to balance long-range coherence with local fluency. We develop a rigorous curation pipeline producing 8K high-quality agentic traces for model optimization. We further introduce M^2LongBench, a comprehensive testbed comprising 247 research tasks across 9 domains and a stable multimodal sandbox. It enables unified multimodal assessment, fair comparison, and accessible evaluation without commercial APIs. Extensive experiments demonstrate that long-form multimodal generation is a challenging task, especially in multimodal selection and integration, and effective post-training can bridge the gap. Our code is available at https://github.com/fangda-ye/Deep-Report.

[Long] Detecting RAG Extraction Attack via Dual-Path Runtime Integrity Game ▶ *Underline Video*

Yuanbo Xie, Yingjie Zhang, Yulin Li, Shouyou Song, Xiaokun Chen, Zhihan Liu, Liya Su and Tingwen Liu

Retrieval-Augmented Generation (RAG) systems augment large language models with external knowledge, yet introduce a critical security vulnerability: RAG Knowledge Base Leakage, wherein adversarial prompts can induce the model to divulge retrieved proprietary content. Recent studies reveal that such leakage can be executed through adaptive and iterative attack strategies (named RAG extraction attack), while effective countermeasures remain notably lacking. To bridge this gap, we propose CanaryRAG, a runtime defense mechanism inspired by stack canaries in software security. CanaryRAG embeds carefully designed canary tokens into retrieved chunks and reformulates RAG extraction defense as a dual-path runtime integrity game. Leakage is detected in real time whenever either the target or oracle path violates its expected canary behavior, including under adaptive suppression and obfuscation. Extensive evaluations against existing attacks demonstrate that CanaryRAG provides robust defense, achieving substantially lower chunk recovery rates than state-of-the-art baselines while imposing negligible impact on task performance and inference latency. Moreover, as a plug-and-play solution, CanaryRAG can be seamlessly integrated into arbitrary RAG pipelines without requiring retraining or structural modifications, offering a practical and scalable safeguard for proprietary data.

[Long] Detecting What Queries Seek: Steering LLM Safety with FFN Output Activation Monitoring ▶ *Underline Video*

Xiaohao Luo, Ying Wei and Rui Zhao

Recently, activation steering has attracted considerable attention as a low-cost approach to improving the safety of large language models (LLMs). However, most existing methods apply interventions indiscriminately, often causing excessive refusal of benign queries. Although recent works have begun to explore selective intervention, their intervention decisions typically rely on residual stream activations where information is highly entangled, resulting in limited discriminative power and unreliable interventions. To address this issue, we propose FFN-Guided activation steering (FGAS). Motivated by the observation that feed-forward networks (FFNs) in LLMs serve as core modules for knowledge storage, we propose leveraging FFN output activations as more discriminative signals for intervention, since these activations more explicitly reflect the intent of a query. For a given query, FGAS projects the corresponding FFN output activation into a low-dimensional subspace that effectively separates harmful and benign queries, and then makes precise intervention decisions by assessing its similarity to pre-constructed prototype activations representing harmful and benign classes. Extensive experiments demonstrate that FGAS achieves state-of-the-art defense performance against various jailbreak attacks, while nearly preserving the model's original performance on benign tasks.

[Long] DiningBench: A Hierarchical Multi-view Benchmark for Perception and Reasoning in the Dietary Domain ▶ *Underline Video*

Song Jin, Juntian Zhang, Xun Zhang, Zeying Tian, Fei Jiang, Guojun Yin, Wei Lin, Yong Liu and Rui Yan

Recent advancements in Vision-Language Models (VLMs) have revolutionized general visual understanding. However, their application in the food domain remains constrained by benchmarks that rely on coarse-grained categories, single-view imagery, and inaccurate metadata. To bridge this gap, we introduce DiningBench, a hierarchical, multi-view benchmark designed to evaluate VLMs across three levels of cognitive complexity: Fine-Grained Classification, Nutrition Estimation, and Visual Question Answering. Unlike previous datasets, DiningBench comprises 3,021 distinct dishes with an average of 5.27 images per entry, incorporating fine-grained "hard" negatives from identical menus and rigorous, verification-based nutritional data. We conduct an extensive evaluation of 29 state-of-the-art open-source and proprietary models. Our experiments reveal that while current VLMs excel at general reasoning, they struggle significantly with fine-grained visual discrimination and precise nutritional reasoning. Furthermore, we systematically investigate the impact of multi-view inputs and Chain-of-Thought reasoning, identifying five primary failure modes. DiningBench serves as a challenging testbed to drive the next generation of food-centric VLM research. All codes are released in https://github.com/meituan/DiningBench.

[Long] DisCal: Distribution-Aware Calibration for Mathematical Reasoning Under Character-Level Noisy Inputs ▶ *Underline Video*

Bo Zhang, Jiawei Zhang, Cong Gao, Bingxu Han, Minghao Hu, Jun Zhang, Yunbo Cao, Zhunchen Luo, Wen Yao, Guotong Geng and Zhong Wang

Although large reasoning models (LRMs) exhibit exceptional mathematical reasoning capabilities on clean inputs, their reasoning accuracy drops substantially in the presence of character-level noise such as typographical errors. Critically, their confidence estimates fail to reflect the corresponding decline in reasoning accuracy. While confidence calibration offers a principled solution, existing methods predominantly target clean inputs, leaving noisy scenarios largely unexplored. To address this gap, we propose DisCal (Distribution-aware Calibration), a confidence calibration framework for character-level noisy inputs. DisCal extracts uncertainty signals from both the empirical answer distribution and the models predictive distribution, and integrates them via a learned calibrator to produce well-calibrated confidence. Experiments across multiple mathematical reasoning benchmarks demonstrate that DisCal consistently outperforms existing calibration methods under noisy inputs, reducing Expected Calibration Error (ECE) by up to 39.21% and improving Area Under the Receiver Operating Characteristic Curve (AUROC) by up to 31.44%.

[Long] DiZiNER: Disagreement-guided Instruction Refinement via Simulating Pilot Annotation for Zero-shot Named Entity Recognition ▶ *Underline Video*

Siun Kim and Hyung-Jin Yoon

Large language models (LLMs) have advanced information extraction (IE) by enabling zero-shot and few-shot named entity recognition (NER), yet their generative outputs still show persistent and systematic errors. Despite progress through instruction fine-tuning, zero-shot NER still lags far behind supervised systems. These recurring errors mirror inconsistencies observed in early-stage human annotation processes that resolve disagreements through pilot annotation. Motivated by this analogy, we introduce DiZiNER (Disagreement-guided Instruction Refinement via Pilot Annotation Simulation for Zero-shot Named Entity Recognition), a framework that simulates the pilot annotation process, employing LLMs to act as both annotators and supervisors. Multiple heterogeneous LLMs annotate shared texts, and a supervisor model analyzes inter-model disagreements to refine task instructions. Across 18 benchmarks, DiZiNER achieves zero-shot SOTA results on 14 datasets, improving prior bests by +8.0 F1 and reducing the zero-shot to supervised gap by over +11 points. It also consistently outperforms its supervisor, GPT-5 mini, indicating that improvements stem from disagreement-guided instruction refinement rather than model capacity. Pairwise agreement between models shows a strong correlation with NER performance, further supporting this finding.

[Long] Do LLMs Capture Embodied Cognition and Cultural Variation? Cross-Linguistic Evidence from Demonstratives ▶ *Underline Video*

Yu Wang, Emmanuele Chersoni and Chu-Ren Huang

Do large language models (LLMs) truly acquire embodied cognition and cultural conventions from text? We introduce demonstratives, fundamental spatial expressions like this/that in English and / in Chinese, as a novel probe for grounded knowledge. Using 6,400 responses from 320 native speakers, we establish a human baseline: English speakers reliably distinguish proximal-distal referents but struggle with perspective-taking, while Chinese speakers switch perspectives fluently but tolerate distal ambiguity. In contrast, five state-of-the-art LLMs fail to inherently understand the proximal-distal contrast and show no cultural differences, defaulting to English-centric reasoning. Our study contributes (i) demonstratives as a new lens for evaluating embodied cognition and cultural conventions, (ii) empirical evidence of cross-cultural asymmetries in human interpretation, (iii) a new perspective on the egocentric-sociocentric debate, showing both orientations coexist but vary across languages, and (iv) a call to address individual variation in future model design.

[Long] DREAM: Deep Research Evaluation with Agentic Metrics ▶ *Underline Video*

Elad Ben Avraham, ChangHao Li, Ron Dorfman, Roy Ganz, Oren Nuriel, Amir Dudai, Aviad Aberdam, Noah Flynn, Elman Mansimov, Aditya Kalyanpur and Ron Litman

Deep Research Agents generate analyst-grade reports, yet evaluating them remains challenging due to the absence of a single ground truth and the multidimensional nature of research quality. Recent benchmarks propose distinct methodologies, yet they suffer from the Mirage of Synthesis, where strong surface-level fluency and citation alignment can obscure underlying factual and reasoning defects. We characterize this gap by introducing a taxonomy across four verticals that exposes a critical capability mismatch: static evaluators inherently lack the tool-use capabilities required to assess temporal validity and factual correctness. To address this, we propose DREAM (Deep Research Evaluation with Agentic Metrics), a framework that instantiates the principle of capability parity by making evaluation itself agentic. DREAM structures assessment through an evaluation protocol combining query-agnostic metrics with adaptive metrics generated by a tool-calling agent, enabling temporally aware coverage, grounded verification, and systematic reasoning probes. Controlled evaluations demonstrate DREAM is significantly more sensitive to factual and temporal decay than existing benchmarks, offering a scalable, reference-free evaluation paradigm.

[Long] Dual Activation-Weight Sparsity: A Training-Free Framework for Efficient Large Language Model Compression ▶ *Underline Video*

Luoyang Sun, Guangyan Li, Cheng Deng, Haifeng Zhang, Jian Zhao, Yongqiang Tang, Wensheng Zhang and Jun Wang

Large language models (LLMs) excel at natural language tasks but face deployment challenges due to computational demands. We introduce Dual Activation-Weight Sparsity (DAWS), a training-free framework that jointly exploits activation and weight sparsity through magnitude-based routing. Systematic analysis of pretrained transformers reveals two key observations: (1) the activation energy is concentrated in a few neurons, and (2) activation and weight sparsity patterns are complementary between attention and FFN layers. DAWS employs a three-tier routing strategy: high-magnitude activations pass through full-precision weights to preserve critical pathways, medium-magnitude activations use magnitude-pruned sparse weights for efficiency, and low-magnitude activations are directly discarded. Unlike prior work that uses activation-aware pruning methods like WANDA, our approach uses direct magnitude-based pruning, which we

show is more robust to sample-level variations. Experiments on Llama and Mistral models demonstrate that DAWS maintains >98% of dense model performance at 50% sparsity, outperforming WANDA, TEAL, and R-Sparse.

[Long] DVI-DTM: Dual-View Representation Learning for Interpretable Short Text Dynamic Topic Modeling ▶ *Underline Video*
Di Liu, Zheng Fang and Bin Wu
Dynamic topic modeling aims to capture topic evolution from temporal text corpora. However, existing methods face two major challenges when applied to short texts: semantic ambiguity and interpretation ambiguity. Semantic ambiguity arises from the sparsity of short texts and the neglect of temporal semantic shifts. Interpretation ambiguity refers to the latent topics that lack human-understandable descriptions. In this work, we propose a novel Dual-View representation learning-based Interpretable short text Dynamic Topic Model (DVI-DTM). To address semantic ambiguity, the Dual-View Representation Learning module is presented to learn robust document-topic distributions by aligning temporal-aware term view and sentence view representations of short texts. To tackle interpretation ambiguity, we introduce a GEA Topic Refiner that leverages LLM agents to generate topic descriptions and refine document-topic distributions through collaborative semantic reasoning. Furthermore, a Dual-Factor Ranking module is designed to capture the topic evolution through semantic relevance and temporal uniqueness. Comprehensive experiments demonstrate that DVI-DTM outperforms the state-of-the-art baselines in topic alignment and dynamic topic quality metrics while producing highly interpretable topic descriptions.

[Long] DVMap: Fine-Grained Pluralistic Value Alignment via High-Consensus Demographic-Value Mapping ▶ *Underline Video*
Pengyun Zhu, Yuqi Ren, Zhen Wang, Lei Yang and Deyi Xiong
Current Large Language Models (LLMs) typically rely on coarse-grained national labels for pluralistic value alignment. However, such macro-level supervision often obscures intra-country value heterogeneity, yielding a loose alignment. We argue that resolving this limitation requires shifting from national labels to multi-dimensional demographic constraints, which can identify groups with predictable, high-consensus value preference. To this end, we propose DVMap (High-Consensus Demographic-Value Mapping), a framework for fine-grained pluralistic value alignment. In this framework, we first present a demographic archetype extraction strategy to construct a high-quality value alignment corpus of 56,152 samples from the World Values Survey (WVS) by strictly retaining respondents with consistent value preferences under identical demographics. Over this corpus, we introduce a Structured Chain-of-Thought (CoT) mechanism that explicitly guides LLMs to reason about demographic-value correlations. Subsequently, we employ Group Relative Policy Optimization (GRPO) to achieve adaptive anchoring of value distributions. To rigorously evaluate generalization, we further establish a triple-generalization benchmark (spanning cross-demographic, cross-country, and cross-value) comprising 21,553 samples. Experimental results demonstrate that DVMap effectively learns the manifold mapping from demographics to values, exhibiting strong generalization and robustness. On cross-demographic tests, Qwen3-8B-DVMap achieves 48.6% accuracy, surpassing the advanced open-source LLM DeepSeek-v3.2 (45.1%). The source code and dataset are available at https://github.com/EnlightenedAI/DVMap.

[Long] EdgeFormer: Latency-Aware Collaborative Multi-Head Attention of Transformer Inference in Edge Networks ▶ *Underline Video*
Yiming Yao, Jianwei Niu, Bin Dai and Tao Ren
Recent breakthroughs in Transformer-based large models, have driven widespread tasks, yet their reliance on centralized cloud deployment raises significant privacy risks due to sensitive data exposure. While edge-based collaborative inference offers a privacy-preserving alternative, existing methods face critical limitations: static model partitioning cannot adapt to dynamic edge resource fluctuations, and rigid multi-head attention handling overlooks semantic-critical prioritization and parallelism. We propose EdgeFormer, a latency-aware framework for distributed Transformer inference in resource-constrained edge networks. EdgeFormer dynamically allocates model blocks across devices via efficiency-storage trade-off optimization and introduces collaborative Multi-Head Attention (cMHA), which distributes semantic-critical attention heads across devices while pruning redundant ones under real-time constraints. We further develop LiScore, a composite metric integrating attention diversity and latency costs, alongside a similarity-based retrieval method to reduce recomputation overhead. Extensive experiments demonstrate that EdgeFormer achieves up to 2.01 times inference acceleration over state-of-the-art baselines with leq1.06% accuracy loss, maintaining robustness under varying edge conditions.

[Long] Edit-Aware Reward Modeling for Chinese Grammatical Error Correction ▶ *Underline Video*
Yilin Li and Xiaojun Wan
While large language models have achieved remarkable success in various natural language processing tasks, their potential in grammatical error correction remains underexplored. Recent work has applied reinforcement learning with rule-based rewards to CGEC, but these approaches rely on coarse-grained binary signals (exact match or not) that fail to capture fine-grained quality distinctions among correction candidates. In this paper, we propose Edit-Aware Reward Model (EARM), a novel reward modeling framework that explicitly incorporates edit-awareness into preference learning for CGEC. EARM introduces a dual-granularity training objective that jointly optimizes sentence-level and token-level weighted Bradley-Terry ranking losses, where edit tokens receive higher importance weights. When integrated with GRPO, our approach achieves 61.29/63.08 on FCGEC/NaCGEC (single output), and 65.04/64.59 with best-of-16 reranking, surpassing previous best by 5.41 and 1.80 points. Extensive experiments demonstrate that learned edit-aware rewards significantly outperform rule-based alternatives for CGEC preference optimization.

[Long] EmoHarbor: Evaluating Personalized Emotional Support by Simulating the User's Internal World ▶ *Underline Video*

Jing Ye, Lu Xiang, Yaping Zhang and Chengqing Zong

Current evaluation paradigms for emotional support conversations tend to reward generic empathetic responses, yet they fail to assess whether the support is genuinely personalized to users unique psychological profiles and contextual needs. We introduce EmoHarbor, an automated evaluation framework that adopts a User-as-a-Judge paradigm by simulating the user's inner world. EmoHarbor employs a Chain-of-Agent architecture that decomposes users' internal processes into three specialized roles, enabling agents to interact with supporters and complete assessments in a manner similar to human users. We instantiate this benchmark using 100 real-world user profiles that cover diverse personality traits and situations, and define 10 evaluation dimensions of personalized support quality. Comprehensive evaluation of 20 advanced LLMs on EmoHarbor reveals a critical insight: while these models excel at generating empathetic responses, they consistently fail to tailor support to individual user contexts. This finding reframes the central challenge, shifting research focus from merely enhancing generic empathy to developing truly user-aware emotional support. EmoHarbor provides a reproducible and scalable framework to guide the development and evaluation of more nuanced and user-aware emotional support systems.

[Long] Empirical Analysis of Decoding Biases in Masked Diffusion Models ▶ *Underline Video*

Pengcheng Huang, Tianming Liu, Zhenghao Liu, Yukun Yan, Shuo Wang, Tong Xiao, Zulong Chen and Maosong Sun

Masked Diffusion Models (MDMs) have recently emerged as a promising non-autoregressive paradigm for sequence generation. However, their performance is highly sensitive to the choice of decoding strategy. In this work, we reveal that prevalent uncertainty-based decoding strategies induce two decoding biases in MDMs: rigid boundary bias and trivial token bias. These biases limit the model's reasoning ability and ultimately degrade generation quality. To address these challenges, we propose UNmasking Calibration for DecOding DEbiasing (UNCODE), a decoding calibration framework that regularizes uncertainty-based decoding by incorporating two complementary priors to shape global decoding trajectories and promote content informativeness. Extensive experiments on three advanced MDMs across seven reasoning- and planning-intensive benchmarks demonstrate that UNCODE consistently outperforms existing decoding strategies by more than 7%, while achieving performance comparable to autoregressive models of similar parameter scales. Our code will be made publicly available on GitHub.

[Long] Enhancing LLM-based Search Agents via Contribution Weighted Group Relative Policy Optimization ▶ *Underline Video*

Junzhe Wang, Zhiheng Xi, Yajie Yang, Hao Luo, Shihan Dou, Tao Gui and Qi Zhang

Search agents extend Large Language Models (LLMs) beyond static parametric knowledge by enabling access to up-to-date and long-tail information unavailable during pretraining. While reinforcement learning has been widely adopted for training such agents, existing approaches face key limitations: process supervision often suffers from unstable value estimation, whereas outcome supervision struggles with credit assignment due to sparse, trajectory-level rewards. To bridge this gap, we propose Contribution-Weighted GRPO (CW-GRPO), a framework that integrates process supervision into group relative policy optimization. Instead of directly optimizing process rewards, CW-GRPO employs an LLM judge to assess the retrieval utility and reasoning correctness at each search round, producing per-round contribution weights. These weights are used to rescale outcome-based advantages along the trajectory, enabling fine-grained credit assignment without sacrificing optimization stability. Experiments on multiple knowledge-intensive benchmarks show that CW-GRPO outperforms standard GRPO by 5.0% on Qwen3-8B and 6.3% on Qwen3-1.7B, leading to more effective search behaviors. Additional analysis reveals that successful trajectories exhibit concentrated contributions in specific rounds, providing empirical insight into search agent tasks. Our code is available at https://github.com/zsxmwjz/CW-GRPO.

[Industry] Enhancing Online Recruitment with Category-Aware MoE and LLM-based Data Augmentation ▶ *Underline Video*

Minping Chen, Bing Xu, Zulong Chen, Chuanfei Xu, Ying Zhou, Zui Tao and Zeyi Wen

Person-Job Fit (PJF) is a critical component for online recruitment. Existing approaches face several challenges, particularly in handling low-quality job descriptions and similar candidate-job pairs, which impair model performance. To address these challenges, this paper proposes a large language model (LLM) based method with two novel techniques: (1) LLM-based data augmentation, which polishes and rewrites low-quality job descriptions by leveraging chain-of-thought (COT) prompts, and (2) category-aware Mixture of Experts (MoE) that assists in identifying similar candidate-job pairs. This MoE module incorporates category embeddings to dynamically assign weights to the experts and learns more distinguishable patterns for similar candidate-job pairs. We perform offline evaluations and online A/B tests on our recruitment platform. Our method relatively surpasses existing methods by 2.40% in AUC and 7.46% in GAUC, and boosts click-through conversion rate (CTCVR) by 19.4% in online tests, saving millions of CNY in external headhunting expenses.

[Long] Enhancing the Transferability of Jailbreak Attacks on Large Language Models via Exploiting Reparameterization Invariance ▶ *Underline Video*

Ao Wang, Xinghao Yang, Yongshun Gong, Wei Liu, Bao-di Liu and Weifeng Liu

Jailbreak attacks serve as a pivotal technique for evaluating the safety alignment of Large language models. Current token-level attacks have shown remarkable efficacy on open-source models by leveraging gradient-based optimization. However, these attacks suffer from poor cross-model transferability, severely limiting their utility on proprietary ones. To address this limitation, we propose Reparameterization Invariance Gradient-based Jailbreak (RIGJ), a natural gradient based framework designed to improve cross-model transferability. Unlike prior token-level methods whose optimization paths are constrained by model-specific Euclidean geometry, RIGJ defines update directions according to differences in output distributions rather than parameter-space distances. Since language models are trained to capture similar dependency structures of natural language, their output distributions share common geometry across architectures, yielding intrinsically model-agnostic optimization trajectories and substantially stronger jailbreak transferability. Extensive experiments demonstrate superior

performance, increasing the cross-model Attack Success Rate and Average Harmfulness Score by 14.9 and 1.23, respectively. Our code is provided https://github.com/nohuma/AISafety_transfer_jailbreak_RIGJ_2026.

[Long] Erasing Without Remembering: Implicit Knowledge Forgetting in Large Language Models ▶ *Underline Video*

Huazheng Wang, Yongcheng Jing, Haifeng Sun, Yingjie Wang, Jingyu Wang, Jianxin Liao and Dacheng Tao

In this paper, we investigate knowledge forgetting in large language models with a focus on its generalisation, ensuring that models forget not only specific training samples but also related implicit knowledge. To this end, we begin by identifying a broader unlearning scope that includes both target data and logically associated samples, including rephrased, subject-replaced, relation-reversed, and one-hop reasoned data. We then conduct a rigorous evaluation of 15 state-of-the-art methods across three datasets, revealing that unlearned models still recall paraphrased answers and retain target facts in their intermediate layers. This motivates us to take a preliminary step toward more generalised implicit knowledge forgetting by proposing PERMU, a novel probability perturbation-based unlearning paradigm. PERMU simulates adversarial unlearning samples to eliminate fact-related tokens from the logit distribution, collectively reducing the probabilities of all answer-associated tokens. Experiments are conducted on a diverse range of datasets, including TOFU, Harry Potter, ZsRE, WMDP, and MUSE, using models ranging from 1.3B to 13B in scale. The results demonstrate that PERMU delivers up to a 50.40% improvement in unlearning vanilla target data while maintaining a 40.73% boost in forgetting implicit knowledge. Our code can be found in the supplementary material.

[Long] EthicMind: A Risk-Aware Framework for Ethical-Emotional Alignment in Multi-Turn Dialogue ▶ *Underline Video*

Jiawen Deng, Wei Li, Wentao Zhang, Ziyun Jiao and Fuji Ren

Intelligent dialogue systems are increasingly deployed in emotionally and ethically sensitive settings, where failures in either emotional attunement or ethical judgment can cause significant harm. Existing dialogue models typically address empathy and ethical safety in isolation, and often fail to adapt their behavior as ethical risk and user emotion evolve across multi-turn interactions. We formulate ethical-emotional alignment in dialogue as an explicit turn-level decision problem, and propose EthicMind, a risk-aware framework that implements this formulation in multi-turn dialogue at inference time. At each turn, EthicMind jointly analyzes ethical risk signals and user emotion, plans a high-level response strategy, and generates context-sensitive replies that balance ethical guidance with emotional engagement, without requiring additional model training. To evaluate alignment behavior under ethically complex interactions, we introduce a risk-stratified, multi-turn evaluation protocol with a context-aware user simulation procedure. Experimental results show that EthicMind achieves more consistent ethical guidance and emotional engagement than competitive baselines, particularly in high-risk and morally ambiguous scenarios.

[Findings] Evaluating Perspectival Biases in Cross-Modal Retrieval ▶ *Underline Video*

Teerapol Saengsukhiran, Peerawat Chomphooyod, Narabodee Rodjananant, Chompakorn Chaksangchaichot, Patawee Prakrankamanant, Witthawin Sripheanpol, Pak Lovichit, Sarana Nutanong and Ekapol Chuangsuwanich

Multimodal retrieval systems are expected to operate in a semantic space, agnostic to the language or cultural origin of the query. In practice, however, retrieval outcomes systematically reflect perspectival biases: deviations shaped by linguistic prevalence and cultural associations. We introduce the Cross-Cultural, Cross-Modal, Cross-lingual Multimodal (3XCM) benchmark to isolate these effects. Results from our studies indicate that, for image-to-text retrieval, models tend to favor entries from prevalent languages over those that are semantically faithful. For text-to-image retrieval, we observe a consistent "tugging effect" in the joint embedding space between semantic alignment and language-conditioned cultural association. When semantic representations are insufficiently resolved, particularly in low-resource languages, similarity is increasingly governed by culturally familiar visual patterns, leading to systematic association bias in retrieval. Our findings suggest that achieving equitable multimodal retrieval necessitates targeted strategies that explicitly decouple language from culture, rather than relying solely on broader data exposure. This work highlights the need to treat linguistic and cultural biases as distinct, measurable challenges in multimodal representation learning.

[Long] Evidence-Augmented Policy Optimization with Reward Co-Evolution for Long-Context Reasoning ▶ *Underline Video*

Xin Guan, Zijian Li, Shen Huang, Pengjun Xie, Jingren Zhou and Jiuxin Cao

While Reinforcement Learning (RL) has advanced LLM reasoning, applying it to long-context scenarios is hindered by sparsity of outcome rewards. This limitation fails to penalize ungrounded "lucky guesses," leaving the critical process of needle-in-a-haystack evidence retrieval largely unsupervised. To address this, we propose EAPO (Evidence-Augmented Policy Optimization). We first establish the Evidence-Augmented Reasoning paradigm, validating via Tree-Structured Evidence Sampling that precise evidence extraction is the decisive bottleneck for long-context reasoning. Guided by this insight, EAPO introduces a specialized RL algorithm where a reward model computes a Group-Relative Evidence Reward, providing dense process supervision to explicitly improve evidence quality. To sustain accurate supervision throughout training, we further incorporate an Adaptive Reward-Policy Co-Evolution mechanism. This mechanism iteratively refines the reward model using outcome-consistent rollouts, sharpening its discriminative capability to ensure precise process guidance. Comprehensive evaluations across eight benchmarks demonstrate that EAPO significantly enhances long-context reasoning performance compared to SOTA baselines.

[Long] Evolutionary Negative Module Pruning for Better LoRA Merging ▶ *Underline Video*

Anda Cao, Zhuo Gou, Yi Wang, Kaixuan Chen, Yu Wang, Can Wang, Mingli Song and Jie Song

Merging multiple Low-Rank Adaptation (LoRA) experts into a single backbone is a promising approach for efficient multi-task deployment.

While existing methods strive to alleviate interference via weight interpolation or subspace alignment, they rest upon the implicit assumption that all LoRA matrices contribute constructively to the merged model. In this paper, we uncover a critical bottleneck in current merging paradigms: the existence of negative modules—specific LoRA layers that inherently degrade global performance upon merging. We propose Evolutionary Negative Module Pruning (ENMP), a plug-and-play LoRA pruning method to locate and exclude these detrimental modules prior to merging. By leveraging an evolutionary search strategy, ENMP effectively navigates the discrete, non-differentiable landscape of module selection to identify optimal pruning configurations. Extensive evaluations demonstrate that ENMP consistently boosts the performance of existing merging algorithms, achieving a new state-of-the-art across both language and vision domains. Code is available at https://github.com/CaoAnda/ENMP-LoRAMerging.

[Long] EvoNarrator: Modeling Scientific Evolution for Feasible Hypothesis Generation ▶ *Underline Video*

Xiaoying Le, Pengfei Qian, Yuanzhao Zhai, Xu Zhang, Qian Liu, Dawei Feng and Bo Ding

Scientific discovery evolution does not emerge in isolation but stems from the structural deepening and recombination of existing functionalities. However, current automated hypothesis generation methods, constrained by the statistical co-occurrence nature of Large Language Models (LLMs), lack perception of temporal causality and the "evolutionary patterns" inherent in scientific development. Consequently, they often yield superficial combinations that are logically infeasible. To address this, we propose EvoNarrator, a framework for hypothesis generation based on evolutionary narratives. We first extract structured P-M-L-F (Problem, Method, Limitation, Future Work) quadruples from citation networks. Subsequently, we introduce the SocketMatch mechanism, which eliminates logical disconnects between methods and problems by assessing their deep semantic compatibility. Finally, utilizing three macro patterns—Chain, Divergence, and Convergence—we constrain the generation process within historically logical derivation paths. Furthermore, double-blind expert reviews yielded an average score of 4.80/5.00 across novelty, feasibility, theoretical, and Logical. Additionally, hindcasting experiments validated its predictive foresight. Crucially, ablation studies indicate that integrating evolutionary patterns facilitates a paradigm shift from conservative incrementalism to theoretically grounded structural innovation. The code is available at https://github.com/xiyii-star/EvoNarrator.

[Long] ExecVerify: White-Box RL with Verifiable Stepwise Rewards for Code Execution Reasoning ▶ *Underline Video*

Lingxiao Tang, He Ye, Zhaoyang Chu, Muyang Ye, Zhongxin Liu, Xiaoxue Ren and Lingfeng Bao

Code LLMs still struggle with code execution reasoning, especially in smaller models. Existing methods rely on supervised fine-tuning (SFT) with teacher-generated explanations, primarily in two forms: (1) input-output (I/O) prediction chains and (2) natural-language descriptions of execution traces. However, intermediate execution steps cannot be explicitly verified during SFT, so the training objective can reduce to merely matching teacher explanations. Moreover, training data is typically collected without explicit control over task difficulty. We introduce ExecVerify, which goes beyond text imitation by incorporating verifiable white-box rewards derived from execution traces, including next-statement prediction and variable value/type prediction. Our work first builds a dataset with multiple difficulty levels via constraint-based program synthesis. Then, we apply reinforcement learning (RL) to reward correct answers about both intermediate execution steps and final outputs, aligning the training objective with semantic correctness at each execution step. Finally, we adopt a two-stage training pipeline that first enhances execution reasoning and then transfers to code generation. Experiments demonstrate that a 7B model trained with ExecVerify achieves performance comparable to 32B models on code reasoning benchmarks and improves pass@1 by up to 5.9% on code generation tasks over strong post-training baselines.

[Findings] FABLE: Fine-grained Fact Anchoring for Unstructured Model Editing ▶ *Underline Video*

Peng Wang, Biyu Zhou, Xuehai Tang, Jizhong Han and Songlin Hu

Unstructured model editing aims to update models with real-world text, yet existing methods often memorize text holistically without reliable fine-grained fact access. To address this, we propose FABLE, a hierarchical framework that decouples fine-grained fact injection from holistic text generation. FABLE follows a two-stage, fact-first strategy: discrete facts are anchored in shallow layers, followed by minimal updates to deeper layers to produce coherent text. This decoupling resolves the mismatch between holistic recall and fine-grained fact access, reflecting the unidirectional Transformer flow in which surface-form generation amplifies rather than corrects underlying fact representations. We also introduce UnFine, a diagnostic benchmark with fine-grained question-answer pairs and fact-level metrics for systematic evaluation. Experiments show that FABLE substantially improves fine-grained question answering while maintaining state-of-the-art holistic editing performance. Our code is publicly available at https://anonymous.4open.science/r/FABLE-B59E.

[Findings] FAER: Benchmarking VLMs for Failure-Aware Embodied Reasoning ▶ *Underline Video*

Hao Song, Kaifeng Liu, Yuanxing Liu, Xiang Tian, Xuesong Wang, Chen Yifan, Weinan Zhang and Ting Liu

Failures are inevitable when embodied agents execute complex tasks. Visual-language models (VLMs) serve as the core component of embodied agents in perceiving the environment and making decisions. Assessing the capabilities of VLMs in detecting and reasoning about failures has become increasingly important. Previous work primarily considered low-level manipulation failures (e.g., 3cm grasp offsets), neglecting high-level failures arising during long-horizon task execution (e.g., object-dropping failure in the “clean room” task) by embodied agents. In this paper, we propose FAER, a failure-aware benchmark aiming to evaluate the performance of VLMs in terms of failure detection, failure categorization, failure description, and failure correction in long-horizon tasks. FAER comprises 3,323 episodes, spanning 3 scenes, 65 tasks, and 83 objects. We assess the performance of 16 widely utilized VLMs and 4 LLMs for FAER tasks. Experimental results show that nearly all VLMs, even GPT-4o, exhibit limited performance in failure detection with a high false negative rate, meaning that they tend to ignore abnormal events, revealing notable gaps in current models’ capacity to effectively handle failures.

[Long] Failure Modes in Multi-Hop QA: The Weakest Link Effect and the Recognition Bottleneck ▶ *Underline Video*
Meiru Zhang, Zaiqiao Meng and Nigel Collier
Despite scaling to massive context windows, Large Language Models (LLMs) struggle with multi-hop reasoning due to inherent position bias, which causes them to overlook information at certain positions. Whether these failures stem from an inability to locate evidence (recognition failure) or integrate it (synthesis failure) is unclear. We introduce Multi-Focus Attention Instruction (MFAI), a semantic probe to disentangle these mechanisms by explicitly steering attention towards selected positions. Across 5 LLMs on two multi-hop QA tasks (MuSiQue and NeoQA), we identify the "Weakest Link Effect": in our 18-document, 3-bucket setting, multi-hop reasoning performance collapses to the level of the least visible evidence, governed by absolute position rather than the linear distance between facts. While matched MFAI resolves recognition bottlenecks, improving accuracy by up to 11.49% in low-visibility positions, misleading MFAI yields divergent effects modulated by task topology: entity-centric tasks with vertical reasoning chains are vulnerable, whereas event-centric tasks with horizontal evidence structures are more resilient. Finally, we demonstrate that "thinking" models utilizing System-2 reasoning effectively locate and integrate the required information, matching gold-only baselines even in noisy, long-context settings. Supplementary experiments on 2WikiMultiHopQA, extended 3-4 hop counts, and a 32B model confirm these findings generalize across datasets, reasoning depths, and model scales.

[Long] FedProxy: Federated Fine-Tuning of LLMs via Proxy SLMs and Heterogeneity-Aware Fusion ▶ *Underline Video*
Tao Fan, GuoqiangMa GuoqiangMa, Yuanfeng SONG, Lixin Fan and Kai ChenQiang Yang
Federated fine-tuning of Large Language Models (LLMs) is obstructed by a trilemma of challenges: protecting LLMs intellectual property (IP), ensuring client privacy, and mitigating performance loss on heterogeneous data. Existing methods like Offsite-Tuning (OT) secure the LLMs IP by having clients train only lightweight adapters, yet our analysis reveals they suffer from a fundamental performance bottleneck, leaving a significant gap compared to centralized training. To bridge this gap, we introduce FedProxy, a new federated adaptation framework. FedProxy replaces weak adapters with a unified, powerful Proxy Small Language Model (SLM), compressed from the proprietary LLM, to serve as a high-fidelity surrogate for collaborative fine-tuning. Our framework systematically resolves the trilemma through a three-stage architecture: (i) Efficient Representation via server-guided compression to create a resource-friendly proxy; (ii) Robust Optimization through an interference-mitigating aggregation strategy to handle data heterogeneity; and (iii) Effortless Fusion via a training-free "plug-in" mechanism to integrate learned knowledge back into the LLM. Experiments show FedProxy significantly outperforms OT methods and approaches centralized performance, establishing a new benchmark for secure and high-performance federated LLM adaptation.

[Findings] FeedEval: Pedagogically Aligned Evaluation of LLM-Generated Essay Feedback ▶ *Underline Video*
Seongyeub Chu, Jongwoo Kim and Munyong Yi
Going beyond the prediction of numerical scores, recent research in automated essay scoring has increasingly emphasized the generation of high-quality feedback that provides justification and actionable guidance. To mitigate the high cost of expert annotation, prior work has commonly relied on LLM-generated feedback to train essay assessment models. However, such feedback is often incorporated without explicit quality validation, resulting in the propagation of noise in downstream applications. To address this limitation, we propose FeedEval, an LLM-based framework for evaluating LLM-generated essay feedback along three pedagogically grounded dimensions: specificity, helpfulness, and validity. FeedEval employs dimension-specialized LLM evaluators trained on datasets curated in this study to assess multiple feedback candidates and select high-quality feedback for downstream use. Experiments on the ASAP++ benchmark show that FeedEval closely aligns with human expert judgments and that essay scoring models trained with FeedEval-filtered high-quality feedback achieve superior scoring performance. Furthermore, revision experiments using small LLMs show that the high-quality feedback identified by FeedEval leads to more effective essay revisions. We release our code and curated datasets at: https://github.com/BBeeChu/FeedEval.git.

[Long] FineLAP: Taming Heterogeneous Supervision for Fine-grained Language-Audio Pretraining ▶ *Underline Video*
Xiquan Li, Xuenan Xu, Ziyang Ma, Wenxi Chen, Haolin He, Qiuqiang Kong and Xie Chen
Contrastively pretrained audio-language models (e.g., CLAP) excel at clip-level understanding but struggle with frame-level tasks. Existing extensions fail to exploit the varying granularity of real-world audio-text data, where massive clip-level textual descriptions coexist with limited frame-level annotations. This paper proposes Fine-grained Language-Audio Pretraining (FineLAP), a novel training paradigm that advances both clip- and frame-level alignment in CLAP with heterogeneous data. FineLAP introduces a dual-stream sigmoid loss with a cluster-based sampling strategy to jointly learn from clip- and frame-level supervision. To capture both global semantics and local details, FineLAP uses a decoupled audio projector on top of a self-supervised encoder. To alleviate the scarcity of temporally annotated data, we present FineLAP-100k, a large-scale synthetic SED dataset constructed through a scalable curation pipeline. Extensive experiments demonstrate that FineLAP achieves SOTA performance across multiple audio understanding tasks, including retrieval, classification, sound event detection, and text-to-audio grounding. Ablation studies further show that coarse- and fine-grained alignment are mutually beneficial, providing insights for building better audio-language models (ALMs).

[Short] FL-MSCL: A Unified Figurative Language Detection Model Driven by Multi-Type Signals and Contrastive Learning ▶ *Underline Video*
Shijia Lu, Fumiyo Fukumoto, Xiaoxi Huang and Yoshimi Suzuki
Figurative language recognition poses significant challenges in NLP, particularly when distinguishing between fine-grained rhetorical

categories such as metaphor, metonymy, and simile. This paper formulates the problem as a four-way sentence-level classification task and proposes FL-MSCL, a unified framework integrating prompt-based knowledge injection with supervised contrastive learning. Experiments across both unified and single-class benchmarks demonstrate that FL-MSCL achieves competitive performance compared to State-of-the-Art (SOTA) methods, indicating consistent advantages in cross-category generalization and category-specific detection.

[Long] Flow-Based Page Unique Semantic Mapping Architecture for Document Visual Question Answering ▶ *Underline Video*

Haosen Wang, Jing Xiao, Chaochao Du, Xiaowang Zhang and Zhiyong Feng

Document Visual Question Answering (DocVQA) aims to generate answers by jointly understanding the textual, layout, and visual elements within document images. Although end-to-end vision-based generative methods have reduced dependency on OCR, they still struggle to achieve precise evidence localization when page semantics are complex and highly similar. However, existing research lacks an in-depth theoretical analysis of the question-driven semantic representation space, failing to fundamentally address the distinguishability problem among semantically similar pages. To fill this theoretical gap, we propose and prove that, given a specific question, each page possesses a unique semantic representation, and there exists a bijective mapping between the page and its unique semantics. Based on this theoretical foundation, we introduce the Flow-Based Page Unique Semantic Mapping Architecture (FUMA), which reconstructs evidence localization from similarity-based retrieval into precise selection on unique semantics. FUMA employs fine-grained cross-modal attention to extract discriminative cues and utilizes flow-based reversible transformations with likelihood regularization to learn bijective mappings, ensuring that each page obtains a unique semantic representation. Moreover, a multi-expert collaboration mechanism complementarily models fine-grained multimodal information within each page, achieving robust answer generation. Experimental results demonstrate that FUMA significantly outperforms existing methods in both evidence localization and answer generation.

[Long] FocalOrder: Focal Preference Optimization for Reading Order Detection ▶ *Underline Video*

Fuyuan Liu, Dianyu Yu, He Ren, Nayu Liu, Xiaomian Kang, Delai Qiu, Fa Zhang, Genpeng Zhen, Shengping Liu, Jiaen Liang, Wei Huang, Yining Wang and Junnan Zhu

Reading order detection is the foundation of document understanding. Most existing methods rely on uniform supervision, implicitly assuming a constant difficulty distribution across layout regions. In this work, we challenge this assumption by revealing a critical flaw: Positional Disparity, a phenomenon where models demonstrate mastery over the deterministic start and end regions but suffer a performance collapse in the complex intermediate sections. This degradation arises because standard training allows the massive volume of easy patterns to drown out the learning signals from difficult layouts. To address this, we propose FocalOrder, a framework driven by Focal Preference Optimization (FPO). Specifically, FocalOrder employs adaptive difficulty discovery with exponential moving average mechanism to dynamically pinpoint hard-to-learn transitions, while introducing a difficulty-calibrated pairwise ranking objective to enforce global logical consistency. Extensive experiments demonstrate that FocalOrder establishes new state-of-the-art results on OmniDocBench v1.0 and Comp-HRDoc. Our compact model not only outperforms competitive specialized baselines but also significantly surpasses large-scale general VLMs. These results demonstrate that aligning the optimization with intrinsic structural ambiguity of documents is critical for mastering complex document structures.

[Long] Foresight Optimization for Strategic Reasoning in Large Language Models ▶ *Underline Video*

Jiashuo Wang, Duan Jiawen, Jian Wang, Kaitao Song, Chunpu Xu, Johnny K. W. Ho, Fenggang Yu, Johan F. Hoorn and Wenjie Li

Reasoning capabilities in large language models (LLMs) have generally advanced significantly. However, it is still challenging for existing reasoning-based LLMs to perform effective decision-making abilities in multi-agent environments, due to the absence of explicit foresight modeling. To this end, strategic reasoning, the most fundamental capability to anticipate the counterpart's behaviors and foresee its possible future actions, has been introduced to alleviate the above issues. Strategic reasoning is fundamental to effective decision-making in multi-agent environments, yet existing reasoning enhancement methods for LLMs do not explicitly capture its foresight nature. In this work, we introduce Foresight Policy Optimization (FoPO) to enhance strategic reasoning in LLMs, which integrates opponent modeling principles into policy optimization, thereby enabling explicit consideration of both self-interest and counterpart influence. Specifically, we construct two curated datasets, namely Cooperative RSA and Competitive Taboo, equipped with well-designed rules and moderate difficulty to facilitate a systematic investigation of FoPO in a self-play framework. Our experiments demonstrate that FoPO significantly enhances strategic reasoning across LLMs of varying sizes and origins. Moreover, models trained with FoPO exhibit strong generalization to out-of-domain strategic scenarios, substantially outperforming standard LLM reasoning optimization baselines.

[Long] Free Energy-Driven Reinforcement Learning with Adaptive Advantage Shaping for Unsupervised Reasoning in LLMs ▶ *Underline Video*

Yiming Huang, Zhenbo Shi, Xincheng Wen, Jichuan Zeng, Cuiyun Gao, Peiyi Han and Chuanyi Liu

Unsupervised reinforcement learning (RL) has emerged as a promising paradigm for enabling self-improvement in large language models (LLMs). However, existing unsupervised RL-based methods often lack the capacity to adapt to the model's evolving reasoning capabilities during training. Therefore, these methods can misdirect policy optimization in the absence of ground-truth supervision. To address this issue, we introduce FREIA, a novel RL-based algorithm built on two key innovations: (1) Free Energy-Driven Reward (FER) adapts rewards to balance consensus and exploration based on the Free Energy Principle. (2) Adaptive Advantage Shaping (AAS) adaptively adjusts learning signals based on the statistical characteristics of sampled rewards. Empirical evaluations on nine datasets across three reasoning tasks showcase that FREIA outperforms other unsupervised RL-based baselines. Notably, in mathematical reasoning tasks, FREIA surpasses other

methods by an average of 0.5 to 3.5 points in Pass@1 using the DeepSeek-R1-Distill-Qwen-1.5B model.

[Long] From Form to Logic: Masked Reconstruction and Reasoning Distillation for Short Video Fake News Detection ▶ *Underline Video*

Qingyan Wang, Lianwei Wu, Botao Wang, Kang Wang and Yaxiong Wang

The rapid growth of short video platforms has made multimodal fake news more prevalent. Existing detectors suffer from two major limitations: (I) global-alignment bias that overemphasizes holistic cross-modal matching and thus misses subtle, localized inconsistencies; and (II) LLM-based methods that leverage powerful generative reasoning to identify cognitive forgeries but inherently suffer from hallucinations and high inference latency. To overcome these limitations, we propose PCDD, a novel Perception-Cognition Dual-driven Detector that jointly observes the form and probes the logic for short video fake news detection. The perception stream exposes fine-grained cross-modal conflicts by amplifying localized inconsistencies into explicit discrepancies. The cognition stream transfers reasoning capabilities from LLMs to a lightweight student to mine cognitive forgeries, while reducing the risk of hallucinations and eliminating reliance on LLMs at inference. Experiments on real-world datasets show that PCDD consistently outperforms baselines, while improving interpretability and robustness in data scarcity scenarios. Our code is available at: https://github.com/SeinCore/PCDD.

[Long] From Inheritance to Saturation: Disentangling the Evolution of Visual Redundancy for Architecture-Aware MLLM Inference Acceleration ▶ *Underline Video*

Jiaqi Shi, Xulong Zhang, Yuechan Li, Xiaoyang Qu and Jianzong Wang

High-resolution Multimodal Large Language Models (MLLMs) face prohibitive computational costs during inference due to the explosion of visual tokens. Existing acceleration strategies, such as token pruning or layer sparsity, suffer from severe "backbone dependency", performing well on Vicuna or Mistral architectures (e.g., LLaVA) but causing significant performance degradation when transferred to architectures like Qwen. To address this, we leverage truncated matrix entropy to uncover a universal three-stage inference lifecycle, decoupling visual redundancy into universal Intrinsic Visual Redundancy (IVR) and architecture-dependent Secondary Saturation Redundancy (SSR). Guided by this insight, we propose HalfV, a framework that first mitigates IVR via a unified pruning strategy and then adaptively handles SSR based on its specific manifestation. Experiments demonstrate that HalfV achieves superior efficiency-performance trade-offs across diverse backbones. Notably, on Qwen25-VL, it retains 96.8% performance at a 4.1x FLOPs speedup, significantly outperforming state-of-the-art baselines. Our code is available at https://github.com/civilizwa/HalfV.

[Findings] From Outcome to Process: Optimizing MoE Load Balancing with MCTS ▶ *Underline Video*

Wenjun Ke, Hengyuan Xu, Ziyu Shang, Yao He, Jiahao Wang, Zijie Xu, Peng Wang, Yuhang Lou and Jiajun Liu

Mixture of Experts (MoE) dynamically routes inputs to specialized expert networks, enabling large language models to scale capacity with low inference overhead. To further improve MoEs parameter efficiency in resource-constrained scenarios, LoRA-MoE integrates LoRA for lightweight adaptation while preserving MoEs specialization. Despite these benefits, the effectiveness of LoRA-MoE still hinges on balanced expert utilization, where certain experts dominate activations while most remain underutilized. Existing balancing strategies focus on constraining the final distribution of expert usage, but overlook the routing decisions made at each layer. As a result, imbalances gradually accumulate across the routing hierarchy. To address this challenge, we propose LayerMoE, a novel three-stage framework that leverages process-level rewards to guide balanced expert routing. Specifically, to overcome the limitation of focusing only on final losses and ignoring intermediate routing, we introduce Monte Carlo Tree Search (MCTS)-based sampling that decomposes outcome-level supervision into layer-wise reward signals, guiding expert choices throughout the routing process. For efficiency, we organize Transformer layers into groups, which constrain the search space of MCTS and keep exploration overhead tractable while retaining the hierarchical structure. Extensive experiments on representative datasets (e.g., ARC, RACE, OBQA) show that applying LayerMoE consistently improves the performance of state-of-the-art LoRA-MoE baselines, yielding an average accuracy gain of 1.39%. Notably, the maximum improvement reaches 2.50%.

[Long] From TDMA to CDMA: A Multi-bit Watermark for Diffusion Language Models ▶ *Underline Video*

Baizhou Huang and Xiaojun Wan

While DLMs have emerged as an alternative to ARMs, robust content provenance mechanisms for this architecture remain unexplored. Existing multi-bit watermarking schemes, heavily reliant on the sequential context of ARMs, cannot be directly applied to DLMs. In this paper, we reframe the multi-bit watermarking problem through a novel Digital Signal Processing (DSP) lens. We draw an analogy between prior works and TDMA (Time Division Multiple Access) in telecommunications, revealing their inherent limitations. To overcome these limitations, we introduce CDMArk, the first multi-bit watermarking framework tailored for DLMs, orchestrating a paradigm shift from TDMA to CDMA (Code Division Multiple Access). Our method encodes the entire watermark message across all tokens holographically. We further provide rigorous statistical guarantees for the watermark detection process. Extensive experiments demonstrate that CDMArk achieves a new state-of-the-art Pareto frontier between imperceptibility and effectiveness.

[Findings] From Tokens to Steps: Verification-Aware Speculative Decoding for Efficient Multi-Step Reasoning ▶ *Underline Video*

Kiran Purohit, Ramasuri Narayanam and Soumyabrata Pal

Speculative decoding (SD) accelerates large language model inference by allowing a lightweight draft model to propose outputs that a stronger target model verifies. However, its token-centric nature allows erroneous steps to propagate. Prior approaches mitigate this using external reward models, but incur additional latency, computational overhead, and limit generalizability. We propose SpecGuard, a

verification-aware speculative decoding framework that performs step-level verification using only model-internal signals. At each step, SpecGuard samples multiple draft candidates and selects the most consistent step, which is then validated using an ensemble of two lightweight model-internal signals: (i) an attention-based grounding score that measures attribution to the input and previously accepted steps, and (ii) a log-probability-based score that captures token-level confidence. These signals jointly determine whether a step is accepted or recomputed using the target, allocating compute selectively. Experiments across a range of reasoning benchmarks show that SpecGuard improves accuracy by 3.6% while reducing latency by ~11%, outperforming both SD and reward-guided SD.

[Long] From Verbatim to Gist: Distilling Pyramidal Multimodal Memory via Semantic Information Bottleneck for Long-Horizon Video Agents ▶ *Underline Video*

Niu Lian, Yuting Wang, Hanshu Yao, Jinpeng Wang, Bin Chen, Yaowei Wang, Min Zhang and Shu-Tao Xia

While multimodal large language models have demonstrated impressive short-term reasoning, they struggle with long-horizon video understanding due to limited context windows and static memory mechanisms that fail to mirror human cognitive efficiency. Existing paradigms typically fall into two extremes: vision-centric methods that incur high latency and redundancy through dense visual accumulation, or text-centric approaches that suffer from detail loss and hallucination via aggressive captioning. To bridge this gap, we propose MM-Mem, a pyramidal multimodal memory architecture grounded in Fuzzy-Trace Theory. MM-Mem structures memory hierarchically into a Sensory Buffer, Episodic Stream, and Symbolic Schema, enabling the progressive distillation of fine-grained perceptual traces (verbatim) into high-level semantic schemas (gist). Furthermore, to govern the dynamic construction of memory, we derive a Semantic Information Bottleneck objective and introduce SIB-GRPO to optimize the trade-off between memory compression and task-relevant information retention. In inference, we design an entropy-driven top-down memory retrieval strategy. Extensive experiments across 4 benchmarks confirm that MM-Mem achieves state-of-the-art performance on both offline and streaming tasks, demonstrating robust generalization and validating the effectiveness of cognition-inspired memory organization. Code and associated configurations are publicly available at 'https://github.com/EliSpectre/MM-Mem'.

[Long] From Where Words Come: Efficient Regularization of Code Tokenizers Through Source Attribution ▶ *Underline Video*

Pavel Chizhov, Egor Bogomolov and Ivan P. Yamshchikov

Efficiency and safety of Large Language Models (LLMs), among other factors, rely on the quality of tokenization. A good tokenizer not only improves inference speed and language understanding but also provides extra defense against jailbreak attacks and lowers the risk of hallucinations. In this work, we investigate the efficiency of code tokenization, in particular from the perspective of data source diversity. We demonstrate that code tokenizers are prone to producing unused, and thus under-trained, tokens due to the imbalance in repository and language diversity in the training data, as well as the dominance of source-specific, repetitive tokens that are often unusable in future inference. By modifying the BPE objective and introducing merge skipping, we implement different techniques under the name Source-Attributed BPE (SA-BPE) to regularize BPE training and minimize overfitting, thereby substantially reducing the number of under-trained tokens while maintaining the same inference procedure as with regular BPE. This provides an effective tool suitable for production use.

[Long] GAM: Hierarchical Graph-based Agentic Memory for LLM Agents ▶ *Underline Video*

Zhaofen Wu, Hanrong Zhang, Fulin Lin, Wujiang Xu, Xinran Xu, Yankai Chen, Henry Peng Zou, Shaowen Chen, Weizhi Zhang, Xue Liu, Philip S. Yu and Hongwei Wang

To sustain coherent long-term interactions, Large Language Model (LLM) agents must navigate the tension between acquiring new information and retaining prior knowledge. Current unified stream-based memory systems facilitate context updates but remain vulnerable to interference from transient noise. Conversely, discrete structured memory architectures provide robust knowledge retention but often struggle to adapt to fluid narrative evolution. To address this, we propose GAM, a hierarchical Graph-based Agentic Memory framework that explicitly decouples memory encoding from consolidation to effectively resolve the conflict between rapid context perception and stable knowledge retention. By isolating ongoing dialogue in a event progression graph and integrating it into a topic associative network only upon semantic shifts, our approach minimizes interference while preserving long-term consistency. Additionally, we introduce a Graph-guided, Multi-factor Retrieval strategy to enhance context precision. Experiments on LoCoMo and LongDialQA benchmarks indicate that our method consistently outperforms state-of-the-art baselines in both reasoning accuracy and computational efficiency.

[SRW] Garden Path Recovery in Causal and Masked Language Models ▶ *Underline Video*

Sanjan Baitalik; Rajashik Datta

Garden-path sentences offer a controlled probe of incremental sentence processing because they require a reader to revise an initially plausible parse when a later region disambiguates the structure. We present an architecture-aware comparison of garden-path recovery in causal and masked language models using 100 ambiguous–control pairs (200 sentences) spanning three classic constructions: NP/Z, NP/S, and MV/RR. Causal models are evaluated with left-to-right word surprisal, whereas masked models are evaluated with pseudo-surprisal derived from masked language model scoring. Beyond the disambiguating region, we analyze cumulative excess surprisal, area-under-curve recovery summaries, and layer-wise hidden-state divergence between ambiguous and control counterparts. Across the headline-valid model set, causal models show larger disruption than masked models overall, with the clearest family-level difference on NP/Z constructions. GPT-2-medium yields the strongest NP/Z spike, while DeBERTa-v3-base shows little evidence of disruption on the same construction. Hidden-state analyses further reveal a qualitative family difference: excluding the embedding layer, masked models peak in very shallow layers, whereas GPT-2 and GPT-2-medium peak substantially deeper. These results suggest that decoder-only models exhibit stronger online

disruption at the point of syntactic revision, while bidirectional encoders appear comparatively buffered at the disambiguator, consistent with their access to right context. More broadly, the findings argue that garden-path evaluation should emphasize recovery dynamics, not merely end-state plausibility or task accuracy.

[Long] GASE: Graph-Aware Semantic Embedding Learning with Frozen LLMs for Text-Attributed Graphs ▶ *Underline Video*
Mingqiang Ding, Jianjun Li, Wenqi Qi, Zhibo Zhang and Yushen Fang
Large Language Models (LLMs) have shown strong potential for text-attributed graph (TAG) learning, yet effectively integrating LLM semantics with graph structural information remains challenging. Embeddings obtained from frozen LLMs lack topology awareness, while fine-tuning LLMs is often computationally expensive. Moreover, LLM embeddings are high-dimensional, and naively reducing dimensionality tends to destroy semantics. To address these issues, we propose GASE, a framework for learning Graph-Aware Semantic Embeddings using frozen LLMs. GASE consists of two key stages: First, we introduce a Training-Free Structure-Aware Semantic Extraction (TSSE) module. Through inter-layer semantic feedback and progressive masked attention, it efficiently compresses and propagates semantic context from neighboring nodes without updating LLM parameters. Second, we propose a Subspace Decomposition and Structural Injection (SDSI) strategy. Embeddings obtained from TSSE are decomposed into a semantic-rich subspace and a structural injection subspace, and structural signals are injected into the latter, which preserves original semantics while integrating graph information. Experiments demonstrate that GASE outperforms state-of-the-art baselines on node classification and achieves a 5x speedup over fine-tuning-based methods.

[Long] GBV-SQL: Guided Generation and SQL2Text Back-Translation Validation for Multi-Agent Text2SQL ▶ *Underline Video*
Daojun Chen, Xi Wang, Shenyuan Ren, Qingzhi Ma, Pengpeng Zhao and An Liu
While Large Language Models have significantly advanced Text2SQL generation, a critical semantic gap persists: syntactically valid queries can still misinterpret user intent. To mitigate this challenge, we propose GBV-SQL, a multi-agent framework that introduces Guided Generation with SQL2Text Back-translation Validation. In particular, a dedicated validator translates generated SQL back into natural language and checks whether its logic is aligned with the original question. Beyond the method itself, we also conduct a systematic audit of benchmark quality and introduce a typology of "Gold Errors" in Text2SQL datasets. Our analysis shows that benchmark issues can coexist with strong execution accuracy and can substantially affect evaluation outcomes. On the challenging BIRD benchmark, GBV-SQL achieves 63.23% execution accuracy, a 5.8% absolute improvement over the Deepseek-v3-based MAC-SQL setting. Under manually audited benchmark corrections, GBV-SQL reaches 96.5% (dev) and 97.6% (test) on Spider, and 90.42% on repaired BIRD dev, providing a diagnostic view of model behavior under improved gold quality. Our work contributes both a practical framework for semantic validation and an empirical analysis of benchmark integrity in Text2SQL evaluation.

[Long] GCA Framework: A GCC Countries-Grounded Dataset and Agentic Pipeline for Climate Decision Support ▶ *Underline Video*
Muhammad Umer Sheikh, Khawar shehzad, Salman Khan, Fahad Shahbaz Khan and Muhammad Haris Khan
Climate decision-making in the GCC states increasingly demands systems that can translate heterogeneous scientific and policy evidence into actionable guidance, yet general-purpose large language models (LLMs) remain weak both in region-specific climate knowledge and grounded interaction with geospatial and forecasting tools. We present the GCA framework, which unifies (i) GCA-DS, a curated multimodal dataset grounded in the GCC states, and (ii) Gulf Climate Agent (GCA), a tool-augmented agent for climate analysis. GCA-DS comprises nearly 200k question-answer pairs spanning governmental policies and adaptation plans, NGO and international frameworks, academic literature, and event-driven reporting on heatwaves, dust storms, and floods, complemented with remote-sensing inputs that couple imagery with textual evidence. Building on this foundation, the GCA agent orchestrates a modular tool pipeline grounded in real-time and historical signals and geospatial processing that produces derived indices and interpretable visualizations. Finally, we benchmark open and proprietary LLMs on climate tasks in the GCC states and show that domain fine-tuning and tool integration substantially improve reliability over general-purpose baselines.

[Long] Generalizing Trust: Weak-to-Strong Trustworthiness in Language Models ▶ *Underline Video*
Lillian Sun, Martin Pawelczyk, Zhenting Qi, Aounon Kumar and Himabindu Lakkaraju
As large language models continue to advance, ensuring their trustworthiness is critical. However, inaccessible real-world ground truth labels pose a significant challenge in high-stakes domains. Recent studies have highlighted weak-to-strong generalization, where a strong model trained only on a weak model's labels surpasses the weak model in task performance. Yet, whether critical trustworthiness properties such as robustness, fairness, and privacy can generalize similarly remains an open question. This is the first work to study this question by examining if a stronger model can enhance trustworthiness when fine-tuned on a weaker models labels, a paradigm we term weak-to-strong trustworthiness. To address this, we introduce two fundamental fine-tuning strategies that leverage trustworthiness regularization during the fine-tuning of the weak model and the weak-to-strong transfer. Our experimental evaluation on real-world datasets reveals that while some trustworthiness properties, such as fairness, adversarial robustness, and OOD robustness, show significant improvement in trustworthiness generalization when both models were regularized, others, like privacy, do not exhibit signs of weak-to-strong trustworthiness. Our results highlight the potential of weak-to-strong trustworthiness as a practical pathway for enhancing the trustworthiness of increasingly capable AI systems, even under imperfect real-world conditions.

[Long] Generating then Refining for Reliable Knowledge Base Question Answering ▶ *Underline Video*

Jianqi Gao, Hang Yu, Jian Cao, Ranran Bu, Jinghua Tang, Nengjun Zhu and Yonggang Zhang

Knowledge Base Question Answering (KBQA) aims to retrieve accurate answers to natural language queries by retrieving and reasoning over large-scale structured knowledge bases (KBs). Advanced semantic parsing-based methods promoted by large language models (LLMs) demonstrate superior performance by transforming questions into structured queries, i.e., logical forms (LFs). However, LFs generated by LLMs could be non-executable due to the inherent semantic hallucination issue of LLMs and the complex graph retrieval characteristics of the KBQA task. To address this challenge, we propose a novel "generate-verify-refine" framework, termed Action-Reflection-Integrated KBQA (ARI-KBQA) for reliable LF generation. ARI-KBQA introduces a dual-module cooperative architecture: First, an action generator is trained to produce initial query paths based on a hop-by-hop reasoning strategy. Then a reflection verifier dynamically validates path feasibility by interacting with the KBs. Consequently, ARI-KBQA filters out invalid LFs and provides semantic correction feedback to the action generator for iteratively refining LFs. Evaluations on standard KBQA benchmarks show that the proposed ARI-KBQA significantly enhances model performance with a reduced search space, especially in complex multi-hop query scenarios.

[Long] GeoRA: Geometry-Aware Low-Rank Adaptation for RLVR ▶ *Underline Video*

Jiaying Zhang, Lei Shi, Jiguo Li, Jun Xu, Jiuchong Gao, Jinghua Hao and Renqing He

Reinforcement Learning with Verifiable Rewards (RLVR) is a key paradigm for improving large-scale reasoning models. Unlike supervised fine-tuning (SFT), RLVR exhibits distinct optimization dynamics and are sensitive to the preservation of pre-trained geometric structures. However, existing parameter-efficient methods face key limitations in this regime. Low-rank adaptation methods, such as PiSSA, are primarily designed for Supervised Fine-Tuning (SFT) and do not account for the distinct optimization dynamics and geometric structures of RLVR. Conversely, directly fine-tuning the unstructured sparse parameter subspace favored by RLVR encounters efficiency bottlenecks on modern hardware. To address these challenges, we propose GeoRA (Geometry-Aware Low-Rank Adaptation), a low-rank adaptation method tailored for RLVR. Specifically, GeoRA exploits the anisotropic and compressible structure of RL update subspace, and extracts its principal directions via Singular Value Decomposition (SVD) to initialize low-rank adapters, while freezing residual components as a structural anchor during training. This design preserves the pre-trained structure and enables efficient dense computation. Experiments on Qwen and Llama models from 1.5B to 32B parameters show that GeoRA consistently outperforms strong low-rank baselines across RLVR settings in mathematics, medicine, and coding, while showing stronger generalization and less forgetting on out-of-domain tasks.

[Short] GOLEMcoref: A Multilingual Coreference Dataset of Fiction ▶ *Underline Video*

Andreas van Cranenburgh, Xiaoyan Yang, Alvanita Alvanita, Cecilia Nicole Di Domenico, Maria Ferragud, Arianna Graciotti, Byungjun Kim, Seonyeong Park, Noa Visser Solissa, Xiaoyu Zhou and Federico Pianzola

We present a multilingual coreference dataset of 827k tokens of fiction in 7 languages: Bahasa Indonesia, Chinese, Dutch, English, Italian, Korean, and Spanish. The dataset includes full stories of diverse lengths, ranging from 500 to 17k words. We discuss our annotation scheme focusing on characters and language-specific challenges we encountered. Finally we present evaluation results of a neural coreference system trained on our dataset. We show that jointly training a system across all languages provides a strong improvement over monolingually trained models. The dataset is available under a creative commons license in CoNLL-2012 and CorefUD format at https://github.com/GOLEM-lab/GOLEMcoref/

[Long] GQLBench: A Large-Scale Cross-Domain, Cross-Dialect Benchmark for NL2GQL ▶ *Underline Video*

Yanning Su, Yuhang Zhou, Yang Fang, Sen Liu, Guangnan Ye and Hongfeng Chai

Despite growing interest in NL2GQL, benchmarking progress has been constrained by the lack of resources that are simultaneously large-scale, cross-domain, and cross-dialect. To address this gap, we present GQLBench, a new benchmark built through an automated and scalable framework that integrates NL2SQL-to-NL2GQL conversion with graph-native data generation. GQLBench supports execution-based evaluation on both Cypher and ISO-GQL, covering hundreds of graph databases and over 20k natural language questions for each dialect. By combining converted data from mature NL2SQL resources with synthetic graph-specific queries, it captures both schema diversity from real-world relational sources and graph-native reasoning challenges, including long paths and cycles. Beyond overall performance comparison, GQLBench also enables fine-grained evaluation across dialects, graph patterns, and query complexity. Experiments on advanced LLMs show that even strong proprietary models struggle on GQLBench, with gemini-3-flash achieving only 35.40% average execution accuracy across the two dialects. Our data and code are available at https://github.com/qxssadf/GQLBench.

[Long] Graph Reasoning Paradigm: Structured and Symbolic Reasoning with Topology-Aware Reinforcement Learning for Large Language Models ▶ *Underline Video*

Runxuan Liu, Xianhao Ou, Xinyan Ma, Jiyuan Wang, Jiafeng Liang, Jiaqi Li, Tao He, Zheng Chu, Rongchuan Mu, Zekun Wang, Baoxin Wang, Dayong Wu, Ming Liu, Shijin Wang, Guoping Hu and Bing Qin

Long Chain-of-Thought (LCoT), achieved by Reinforcement Learning with Verifiable Rewards (RLVR), has proven effective in enhancing the reasoning capabilities of Large Language Models (LLMs). However, reasoning in current LLMs is primarily generated as plain text, where performing semantic evaluation on such unstructured data creates a computational bottleneck during training. Despite RLVR-based optimization, existing methods still suffer from coarse-grained supervision, reward hacking, high training costs, and poor generalization. To address these issues, we propose the Graph Reasoning Paradigm (GRP), which realizes structured and symbolic reasoning, implemented via graph-structured representations with step-level cognitive labels. Building upon GRP, we further design Process-Aware Stratified Clipping Group Relative Policy Optimization (PASC-GRPO), which leverages structured evaluation to replace semantic evaluation, achieves

process-aware verification through graph-structured outcome rewards, and mitigates reward hacking via stratified clipping advantage estimation. Experiments demonstrate significant improvements across mathematical reasoning and code generation tasks. Data, models, and code will be released later.

[Long] GRASPrune: Global Gating for Budgeted Structured Pruning of Large Language Models ▶ *Underline Video*

Ziyang Wang, Jiangfeng Xiao, Chuan Xiao, Ruoxiang Li, Rui Mao and Jianbin Qin

Large language models (LLMs) are expensive to serve because dense FFN blocks, multi-head attention, and KV caches dominate memory, making structured pruning a natural way to reduce serving costs under tight parameter and memory budgets. We present GRASPrune, a global budgeted structured pruning framework applied post-hoc to a pretrained model that jointly prunes FFN channels and attention KV head groups under a single global parameter budget. GRASPrune attaches lightweight learnable gates to prunable units and optimizes only these gates on a small unlabeled language-modeling calibration set, keeping all backbone weights frozen while enforcing the target sparsity at every step. A final budget-preserving scaling calibration reweights the surviving channels and heads to correct scale shifts introduced by pruning. On LLaMA-2-7B, GRASPrune removes 50% of parameters and achieves 12.18 perplexity on WikiText-2 while maintaining competitive average zero-shot accuracy on five downstream benchmarks, using a short calibration run of four epochs on 512 unlabeled sequences on a single NVIDIA A100 80GB GPU, all without any full-model fine-tuning.

[Findings] GUITester: Enabling GUI Agents for Exploratory Defect Discovery ▶ *Underline Video*

Yifei Gao, Jiang Wu, Xiaoyi Chen, Yifan Yang, Zhe Cui, Tianyi Ma, Jiaming Zhang and Jitao Sang

Exploratory GUI testing is essential for software quality but suffers from high manual costs. While Multi-modal Large Language Model (MLLM) agents excel in navigation, they fail to autonomously discover defects due to two core challenges: Goal-Oriented Masking, where agents prioritize task completion over reporting anomalies, and Execution-Bias Attribution, where system defects are misidentified as agent errors. To address these, we first introduce GUITestBench, the first interactive benchmark for this task, featuring 143 tasks across 26 defects. We then propose GUITester, a multi-agent framework that decouples navigation from verification via two modules: (i) a Planning-Execution Module (PEM) that proactively probes for defects via embedded testing intents, and (ii) a Hierarchical Reflection Module (HRM) that resolves attribution ambiguity through interaction history analysis. GUITester achieves an F1-score of 48.90% (Pass@3) on GUITestBench, outperforming state-of-the-art baselines (33.35%). Our work demonstrates the feasibility of autonomous exploratory testing and provides a robust foundation for future GUI quality assurance.

[Findings] Hard to Be Heard: Phoneme-Level ASR Analysis of Phonologically Complex, Low-Resource Endangered Languages ▶ *Underline Video*

V.S.D.S.Mahesh Akavarapu, Michael Daniel and Gerhard Jäger

We present a phoneme-level analysis of automatic speech recognition (ASR) for two low-resourced and phonologically complex East Caucasian languages, Archi and Rutul, based on curated and standardized speech-transcript resources totaling approximately 50 minutes and 1 hour 20 minutes of audio, respectively. Existing recordings and transcriptions are consolidated and processed into a form suitable for ASR training and evaluation. We evaluate several state-of-the-art audio and audio–language models, including wav2vec2, Whisper, and Qwen2-Audio. For wav2vec2, we introduce a language-specific phoneme vocabulary with heuristic output-layer initialization, which yields consistent improvements and achieves performance comparable to or exceeding Whisper in these extremely low-resource settings. Beyond standard word and character error rates, we conduct a detailed phoneme-level error analysis. We find that phoneme recognition accuracy strongly correlates with training frequency, exhibiting a characteristic sigmoid-shaped learning curve. For Archi, this relationship partially breaks for Whisper, pointing to model-specific generalization effects beyond what is predicted by training frequency. Overall, our results indicate that many errors attributed to phonological complexity are better explained by data scarcity. These findings demonstrate the value of phoneme-level evaluation for understanding ASR behavior in low-resource, typologically complex languages.

[Long] Harmonizing the Past, Present, and Future: A Null-Space Constrained Region-Specific Method for Continual Learning in LLMs ▶ *Underline Video*

Jinhui Chen, Shizhu He, Xingchang Yang, Huanxuan Liao, Yequan Wang, Xiangwen Liao, Wenhao Teng, Kang Liu and Jun Zhao

Enabling Large Language Models (LLMs) to evolve sustainably requires simultaneously preserving previously acquired knowledge (Past), effectively acquiring new task-specific skills (Present), and reserving sufficient parameter capacity for subsequent adaptation (Future). However, existing continual learning (CL) paradigms often prioritize immediate performance through dense updates, leading to catastrophic forgetting and rapid exhaustion of model capacity. To harmonize these conflicting demands, we draw inspiration from the brain's functional partitioning and propose the Null-Space Constrained Parameter Region Specificity Method (PaRSP). PaRSP establishes a dynamic "Task-Region Mapping" that distinguishes between specialized neurons and generalist neurons. By precisely localizing a sparse "functional core" for each task, PaRSP restricts updates to specific regions via null-space orthogonality, preserving the vast majority of the network as an immutable "long-term memory bank." This induced sparsity not only enhances plasticity via targeted adaptation and minimizes interference to ensure stability, but also strategically reserves substantial capacity, securing sustainability for future evolution. Extensive experiments validate PaRSP's state-of-the-art performance, particularly on Standard CL and Long Sequence benchmarks, effectively harmonizing the stability-plasticity-sustainability trade-off. Code is available at https://github.com/JinhuiBot/PaRSP

[Long] HCSpec: Two-Tier Horizontal Cascade Speculative Decoding for High-Efficiency Large Language Model Inference

► *Underline Video*

Yizhou Zhang, Siming Chen, Hao Ye and Erhu Feng

Speculative decoding accelerates large language model (LLM) inference by using a draft model to propose token candidates for parallel verification by the target model. However, current state-of-the-art self-distilled draft models adopt a homogeneous architecture across all drafting positions, failing to account for a critical empirical observation: the expected utility of drafting decays rapidly after the initial positions. To exploit this imbalance, we propose Two-tier Horizontal Cascade Speculative Decoding (HCSpec), a novel framework that organizes heterogeneous, position-specialized draft modules into a horizontal cascade. The first tier employs a dual-layer, dual-path transformer that enhances early-step fidelity by decoupling token-logit prediction from recurrent feature propagation, while the second tier adopts a lightweight single-layer transformer that deliberately trades marginal accuracy for improved efficiency at later drafting steps. Extensive experiments on Qwen series models and Llama3.1-8B-Instruct, across multiple tasks and diverse inference configurations, demonstrate that HCSpec consistently outperforms the previous state-of-the-art (EAGLE-3). It delivers 15-30% higher end-to-end speedup over EAGLE-3 and achieves up to 3.72x acceleration over vanilla autoregressive decoding. Our code is provided in the supplementary materials.

[Long] HEALing Entropy Collapse: Enhancing Exploration in Few-Shot RLVR via Hybrid-Domain Entropy Dynamics Alignment ► *Underline Video*

Zhanyu Liu, Qingguo Hu, Ante Wang, Chenqing Liu, Zhishang Xiang, Hui Li, Delai Qiu and Jinsong Su

Reinforcement Learning with Verifiable Reward (RLVR) has proven effective for training reasoning-oriented large language models, but existing methods largely assume high-resource settings with abundant training data. In low-resource scenarios, RLVR is prone to more severe entropy collapse, which substantially limits exploration and degrades reasoning performance. To address this issue, we propose Hybrid-domain Entropy dynamics ALignment (HEAL), a framework tailored for few-shot RLVR. HEAL first selectively incorporates high-value general-domain data to promote more diverse exploration. Then, we introduce Entropy Dynamics Alignment (EDA), a reward mechanism that aligns trajectory-level entropy dynamics between the target and general domains, capturing both entropy magnitude and fine-grained variation. Through this alignment, EDA not only further mitigates entropy collapse but also encourages the policy to acquire more diverse exploration behaviors from the general domain. Experiments across multiple domains show that HEAL consistently improves few-shot RLVR performance. Notably, using only 32 target-domain samples, HEAL matches or even surpasses full-shot RLVR trained with 1K target-domain samples.

[Long] HiEdit: Lifelong Model Editing with Hierarchical Reinforcement Learning ► *Underline Video*

Yangfan Wang, Tianyang Sun, Chen Tang, Jie Liu, Wei Cai and Jingchi Jiang

Lifelong model editing (LME) aims to sequentially rectify outdated or inaccurate knowledge in deployed LLMs while minimizing side effects on unrelated inputs. However, existing approaches typically apply parameter perturbations to a static and dense set of LLM layers for all editing instances. This practice is counter-intuitive, as we hypothesize that different pieces of knowledge are stored in distinct layers of the model. Neglecting this layer-wise specificity can impede adaptability in integrating new knowledge and result in catastrophic forgetting for both general and previously edited knowledge. To address this, we propose HiEdit, a hierarchical reinforcement learning framework that adaptively identifies the most knowledge-relevant layers for each editing instance. By enabling dynamic, instance-aware layer selection and incorporating an intrinsic reward for sparsity, HiEdit achieves precise, localized updates. Experiments on various LLMs show that HiEdit boosts the performance of the competitive RLEdit by an average of 8.48% with perturbing only half of the layers per edit. Our code is available at: https://github.com/yangfanww/hiedit.

[Long] Hierarchical Acoustic-Semantic Modeling: Modality Separation and Semantic Coherence for Full-Duplex SLMs ► *Underline Video*

Zhenyu Liu, Xuanyu Zhang Yunxin Li, Qixun Teng, Shenyuan Jiang, Haolan Chen, Minjun Zhao, Fanbo Meng, Yu Xu, Yancheng He, Baotian Hu, Haizhou Li and Min Zhang

Developing seamless, high-performance, native intelligent full-duplex Spoken Language Models (SLMs) remains a critical challenge and long-standing goal for the speech and NLP community. Despite notable progress, recent endeavors are fundamentally constrained by severe modality interference, which causes substantial knowledge degradation and compromises semantic integrityultimately making full-duplex SLMs feel unnatural and unintelligent. In this paper, through an exhaustive fine-grained analysis of model optimization dynamics, we uncover the root cause of such performance degradation, revealing that modality interference arises from inherent gradient conflicts between acoustic and semantic modeling when the two modalities are forced to share a deep parameter space. Guided by this key insight, we introduce Lychee-FD, a native end-to-end full-duplex framework designed to mitigate modality interference. Importantly, we propose a hierarchical parameter separation strategy that decouples conflicting modalities in deep layers while preserving cross-modality coherence via a dedicated semantic alignment channel. Extensive experiments on multiple full-duplex benchmarks demonstrate that our method significantly advances the state of the art, yielding substantial improvements in both speech intelligence (+7.4% on Spoken QA) and full-duplex interaction fluidity (+28.5% on FullDuplexBench 1.5) without compromising inference efficiency. To the best of our knowledge, this work is the first to achieve two key advances: 1) uncovering and elucidating the root cause of modality interference in full-duplex SLMs, and 2) designing an elegant hierarchical model together with a practical solution for seamless, high-performance, native intelligent full-duplex SLMs.

[Long] How Controllable Are Large Language Models? A Unified Evaluation across Behavioral Granularities ► *Underline Video*

Ziwen Xu, Kewei Xu, Haoming Xu, Haiwen Hong, Longtao Huang, Hui Xue, Ningyu Zhang, Yongliang Shen, Guozhou Zheng, Huajun Chen and Shumin Deng

Large Language Models (LLMs) are increasingly deployed in socially sensitive domains, yet their unpredictable behaviors, ranging from misaligned intent to inconsistent personality, pose significant risks. We introduce SteerEval, a hierarchical benchmark for evaluating LLM controllability across three domains: language features, sentiment, and personality. Each domain is structured into three specification levels: L1 (what to express), L2 (how to express), and L3 (how to instantiate), connecting high-level behavioral intent to concrete textual output. Using SteerBench, we systematically evaluate contemporary steering methods, revealing that control often degrades at finer-grained levels. Our benchmark offers a principled and interpretable framework for safe and controllable LLM behavior, serving as a foundation for future research.

[Long] How Should We Enhance the Safety of Large Reasoning Models: An Empirical Study ▶ *Underline Video*

Zhexin Zhang, Xian Qi Loye, Victor Shea-Jay Huang, Junxiao Yang, Qi Zhu, Shiyao Cui, Fei Mi, Lifeng Shang, Yingkang Wang, Hongning Wang and Minlie Huang

Large Reasoning Models (LRMs) have achieved remarkable success on reasoning-intensive tasks such as mathematics and programming. However, their enhanced reasoning capabilities do not necessarily translate to improved safety performance, and in some cases, may even degrade it. This raises an important research question: how should we enhance the safety of LRMs? In this paper, we present a comprehensive empirical study on how to enhance the safety of LRMs through Supervised Fine-Tuning (SFT). Our investigation begins with an unexpected observation: directly distilling safe responses from DeepSeek-R1 fails to significantly enhance safety. We analyze this phenomenon and identify five key risky patterns that contribute to it. We then demonstrate that explicitly addressing these issues during the data distillation process can lead to substantial safety improvements. Next, we explore whether a long and complex reasoning process is necessary for achieving safety. Interestingly, we find that simply using short or template-based reasoning process can attain comparable safety performance. These findings prompt a deeper reflection on the role of reasoning in ensuring safety. Finally, we conduct a comprehensive ablation study to reveal the impact of different training configurations. Overall, we hope our empirical study could provide a more holistic picture on enhancing the safety of LRMs.

[Demo] HuoziIME: An On-Device LLM-Enhanced Input Method for Deep Personalization ▶ *Underline Video*

Baocai Shan, Yuzhuang Xu and Wanxiang Che

Mobile input method editors (IMEs) are the primary interface for text input, yet they remain constrained to manual typing and struggle to produce personalized text. While lightweight large language models (LLMs) make on-device auxiliary generation feasible, enabling deeply personalized, privacy-preserving, and real-time generative IMEs poses fundamental challenges. To this end, we present HUOZIIME, a personalized on-device IME powered by LLM. We endow HUOZIIME with initial human-like prediction ability by post-training a base LLM on synthesized personalization data. Notably, a hierarchical memory mechanism is designed to continually capture and leverage user-specific input history. Furthermore, we perform systemic optimizations tailored to on-device LLM-based IME deployment, ensuring efficient and responsive operation under mobile constraints. Experiments demonstrate efficient on-device execution and high-fidelity memory-driven personalization. Code and package are available at https://github.com/Shan-HIT/HuoziIME.

[Findings] ICLAD: In-Context Learning with Comparison-Guidance for Audio Deepfake Detection ▶ *Underline Video*

Benjamin Shiue-Hal Chou, Yi Zhu and Surya Koppisetti

Audio deepfakes pose a significant security threat, yet current state-of-the-art (SOTA) detection systems do not generalize well to realistic in-the-wild deepfakes. We introduce a novel In-Context Learning paradigm with comparison-guidance for Audio Deepfake detection (ICLAD). The framework enables the use of audio language models (ALMs) for training-free generalization to unseen deepfakes and provides rich textual explanations on the detection outcome. At the core of ICLAD is a pairwise comparative reasoning strategy that guides the ALM to discover and filter hallucinations and deepfake-irrelevant acoustic attributes. The ALM works alongside a specialized deepfake detector, whereby a routing mechanism feeds out-of-distribution samples to the ALM. On in-the-wild datasets, ICLAD improves macro F1 over the specialized detector, with up to 2x relative improvement. Further analysis demonstrates the flexibility of ICLAD and its potential for deployment on recent open-source ALMs.

[Long] Illusions of Confidence? Diagnosing LLM Truthfulness via Neighborhood Consistency ▶ *Underline Video*

Haoming Xu, Ningyuan Zhao, Yunzhi Yao, Weihong Xu, Hongru WANG, Xinle Deng, Shumin Deng, Jeff Z. Pan, Huajun Chen and Ningyu Zhang

As Large Language Models (LLMs) are increasingly deployed in real-world settings, correctness alone is insufficient. Reliable deployment requires maintaining truthful beliefs under contextual perturbations. Existing evaluations largely rely on point-wise confidence like Self-Consistency, which can mask brittle belief. We show that even facts answered with perfect self-consistency can rapidly collapse under mild contextual interference. To address this gap, we propose Neighbor-Consistency Belief (NCB), a structural measure of belief robustness that evaluates response coherence across a conceptual neighborhood. To validate the efficiency of NCB, we introduce a new cognitive stress-testing protocol that probes outputs stability under contextual interference. Experiments across multiple LLMs show that the performance of high-NCB data is relatively more resistant to interference. Finally, we present Structure-Aware Training (SAT), which optimizes context-invariant belief structure and reduces long-tail knowledge brittleness by approximately 30%.

[Industry] Index-Time Prefix Injection for Multi-Tenant Retrieval: Improving Search Relevance Without Model Fine-

Tuning ▶ *Underline Video*

Vaibhav Varshney and Manjunatha Naik MC

Multi-tenant enterprise search platforms serve hundreds of customers through a single shared retrieval model. Fine-tuning on individual customer data is typically prohibited by contractual and regulatory constraints, and maintaining per-customer models does not scale. We present index-time prefix injection, a training-free method that improves retrieval relevance by prepending domain-descriptive natural-language prefixes to documents during indexing. For example, prepending "IT service management knowledge article:" to an IT knowledge base shifts its embeddings into a tighter, more domain-coherent region of the vector space. Prefixes are discovered through a tiered strategy: LLM-based generation from document samples when data policies allow, domain-expert curation when they do not, and a standardized prefix library as fallback. Deployed across 18 languages and 400+ customer instances, the approach yields 3-8% Hit@5 improvements with zero model training. A/B tests confirm a 4.2% CTR lift. We describe the system design, evaluation at scale, and deployment lessons including failure modes.

[Findings] InteracSPARQL: An Interactive System for SPARQL Query Refinement Using Natural Language Explanations ▶ *Underline Video*

Xiangru Jian, Zhengyuan Dong and M. Tamer Özsu

Current approaches for Natural Language to SPARQL (NL2SPARQL) generation primarily rely on one-turn, training-intensive models. While effective in specific settings, these models often lack generalizability and fail to provide transparency or mechanisms for error recovery in realistic scenarios. Additionally, prior interactive works are largely outdated and incompatible with modern large language model (LLM) workflows. In this paper, we introduce InteracSPARQL, a training-free interactive refinement pipeline that acts as a plug-and-play enhancement for existing SPARQL generation systems. Our approach integrates a set of efficient entity and property lookup tools within a self-correction loop, guided by a novel hybrid Natural Language Explanation (NLE) module. This module combines rule-based Abstract Syntax Tree (AST) parsing with LLM semantic enrichment to produce explanations that are both structurally accurate and linguistically fluent. We evaluate InteracSPARQL on standard benchmarks (QALD-9 and QALD-10), showing that our tool-augmented self-refinement significantly boosts the accuracy of base models without fine-tuning. Furthermore, human evaluation confirms that our structured explanations substantially improve user understanding and ability to correct queries compared to unstructured baselines.

[Long] Investigating More Explainable and Partition-Free Compositionality Estimation for LLMs: A Rule-Generation Perspective ▶ *Underline Video*

Ziyao Xu, Cong Wang and Houfeng Wang

Compositional generalization tests are often used to estimate the compositionality of LLMs. However, such tests have the following limitations: (1) they only focus on the output results without considering LLMs' understanding of sample compositionality, resulting in explainability defects; (2) they rely on dataset partition to form the test set with combinations unseen in the training set, suffering from combination leakage issues. In this work, we propose a novel rule-generation perspective for compositionality estimation for LLMs. It requires LLMs to generate a program as rules for dataset mapping and provides estimates of the compositionality of LLMs using complexity-based theory. The perspective addresses the limitations of compositional generalization tests and provides a new way to analyze the compositionality characterization of LLMs. We conduct experiments and analysis of existing advanced LLMs based on this perspective on a string-to-grid task, and find various compositionality characterizations and compositionality deficiencies exhibited by LLMs.

[Long] Investigating the Representation of Backchannels and Fillers in Fine-tuned Language Models ▶ *Underline Video*

Yu Wang, Leyi Lao, Langchu Huang, Gabriel Skantze, Yang Xu and Hendrik Buschmeier

Backchannels and fillers are important linguistic expressions in dialogue, but often treated as "noise" to be bypassed in modern transformer-based language models. Our work studies the representation of them in language models using three fine-tuning strategies. The models are trained on three dialogue corpora in English and Japanese, where backchannels and fillers are preserved and annotated, to investigate how fine-tuning can help LMs learn their representations. We first apply clustering analysis to the learnt representation of backchannels and fillers, and have found increased silhouette scores in representations from fine-tuned models, which suggests that fine-tuning enables LMs to distinguish the nuanced semantic variation in different backchannel and filler use. We also use natural language generation (NLG) metrics and qualitative analysis to confirm that the utterances generated by fine-tuned language models resemble human-produced utterances more closely. Our findings suggest the potentials of transforming general LMs into conversational LMs that are more capable of producing human-like languages adequately.

[Long] Is EEG-to-Text Feasible in Real-World Scenarios? An In-Depth Analysis Using a Neuropsychology-Inspired Benchmark ▶ *Underline Video*

Zihan Zhang, Yu Bao, Xiao Ding, Tianyi Jiang and Kai Xiong

Translating brain signals into text could restore communication for people with severe paralysis, yet practically usable systems to date rely on invasive electrocorticography (ECoG). Electroencephalography (EEG) offers a non-invasive alternative, and EEG-to-text (EEG2Text) has been widely explored. Interestingly, however, EEG2Text models generally rely on teacher-forcing evaluation; without it, they fail to generate meaningful decoding. This reliance prevents EEG2Text from being applied in real-world, non-academic settings. This has fueled numerous debates about whether EEG2Text is a meaningful direction, by extension, and whether EEG truly contains decodable linguistic information. Here, using a neuropsychology-informed paradigm, we find that existing EEG2Text benchmarks have neglected EEG instability, a flaw that

has confounded inference and sparked debate. Our experiments furnish key evidence for the feasibility of teacher-forcing-free EEG2Text decoding. Accordingly, we assemble the Corpus OF Eeg-To-Text (COFETT) using a 128-channel high-density EEG cap, providing a benchmark dedicated to evaluating EEG2Text models. In comparisons with multiple existing benchmarks, COFETT achieves SOTA ability to distinguish among model performances and enables robust, teacher-forcing-free evaluation, thereby opening a path toward practical EEG2Text applications. COFETT is open sourced in https://github.com/baoyudu/COFETT.

[Findings] Iterative Knowledge Graph Refinement and Integration for Medical Question Answering ► *Underline Video*

Haochen Zou and Yongli Wang

Large Language Models (LLMs) are challenged by generating hallucinations and factually incorrect responses, particularly in complex and specialized medical question answering (QA). Integrating knowledge graphs (KGs) through retrieval-augmented generation (RAG) methods has emerged as a promising direction. However, existing graph-based RAG methods heuristically retrieve and refine question-relevant subgraphs, potentially introducing redundant and noisy factual information that is difficult for LLMs to process, ultimately limiting reasoning capability. To incorporate a concise yet informative evidence subgraph, we propose an iterative medical QA framework. It optimizes graph-based RAG methods by selectively retrieving focused knowledge from KGs to construct a precise evidence subgraph and progressively pruning it utilizing structured feature representations. The targeted KG integration maintains coherent and reliable inference. Experiments on three medical QA benchmark datasets demonstrate that the framework achieves state-of-the-art performance against representative baseline competitors, highlighting the importance of efficient KG integration.

[Long] Jakiro: Boosting Speculative Decoding via Decoupled MoE ► *Underline Video*

Haiduo Huang, Fuwei Yang, Zhenhua Liu and Pengju Ren

Speculative decoding has emerged as a promising technique to accelerate large language model inference by employing a smaller draft model to predict multiple tokens, which are then verified in parallel by the larger target model. However, existing approaches face a fundamental limitation: candidates at the same tree layer share identical feature representations, constraining diversity and diminishing overall effectiveness. We identify this as an intra-layer coupling problem that limits prediction accuracy. To address this challenge, we propose Jakiro, which introduces decoupled Mixture of Experts (MoE) into the draft model, enabling different experts to generate diverse candidate tokens from distinct feature spaces. We further propose Contrastive-Enhanced Parallel Decoding (CEPD) that combines autoregressive and parallel decoding with a contrastive mechanism to reduce inference steps while maintaining accuracy. Extensive experiments across diverse models and tasks demonstrate that Jakiro achieves significant speedups over strong baselines, with particularly notable improvements in non-greedy decoding scenarios where token diversity is crucial.

[Demo] JointCoder: Exploring Automated ICD Coding on Real-World Chinese EHRs with a Multi-Agent Framework ► *Underline Video*

Kangjun Liu, Zhenyu Li, Jianlei Wang, Hongjiao Guan, Ying Lian, Guoqiang Chen, Tao Xin and Wenpeng Lu

Automated ICD coding is a critical task for standardizing clinical information from electronic health records (EHRs) and supporting downstream healthcare administration. However, existing automated ICD coding systems face several fundamental challenges. First, the majority of existing research focuses on English ICD tasks, with limited attention to Chinese-language clinical contexts due to the scarcity of publicly available Chinese ICD datasets. Second, most approaches primarily target disease coding, overlooking procedure coding as well as the multi-stage workflows followed in real-world clinical practice. Moreover, many recent methods rely heavily on closed-source large language models or substantial computational resources, which limits their scalability and deployability in clinical environments. To address these gaps, this paper proposes JointCoder, which includes a real-world Chinese ICD coding dataset and a multi-agent framework that reformulates automated ICD coding as a joint disease-procedure coding task. JointCoder explicitly models real-world clinical coding workflows through stage-wise agent collaboration. All agents are instantiated using locally deployed 1.7B-parameter models, enabling scalable and privacy-preserving deployment. Extensive experiments on real-world Chinese ICD coding datasets demonstrate JointCoders superiority over state-of-the-art baselines across all evaluation metrics. Our Github repository is available at https://github.com/KJOrigin/JointCoder. Our demo website is available at http://159.89.222.214:9028/. The demo video can be accessed at https://yunpan.qlu.edu.cn:443/link/CE86D849DD24EA9EE3D68AF2FD90D539.

[Long] KG-ViP: Bridging Knowledge Grounding and Visual Perception in Multi-modal LLMs for Visual Question Answering ► *Underline Video*

Zhiyang Li, Ao Ke, Yukun Cao and Xike Xie

Multi-modal Large Language Models (MLLMs) for Visual Question Answering (VQA) often suffer from dual limitations: knowledge hallucination and insufficient fine-grained visual perception. Crucially, we identify that commonsense graphs and scene graphs provide precisely complementary solutions to these respective deficiencies by providing rich external knowledge and capturing fine-grained visual details. However, prior works typically treat them in isolation, overlooking their synergistic potential. To bridge this gap, we propose KG-ViP, a unified framework that empowers MLLMs by fusing scene graphs and commonsense graphs. The core of the KG-ViP framework is a novel retrieval-and-fusion pipeline that utilizes the query as a semantic bridge to progressively integrate both graphs, synthesizing a unified structured context that facilitates reliable multi-modal reasoning. Extensive experiments on FVQA 2.0+ and MVQA benchmarks demonstrate that KG-ViP significantly outperforms existing VQA methods.

[Long] Know Your Place: Diagnosing Implicit Social Adaptation Failures in Chinese Large Language Models ▶ *Underline Video*

Yu Tian, Jie Xing, Ziming Li, Jiang Li, Zehua Duo, Tian Lan, Xu Liu, Guanglai Gao and Xiangdong Su

As large language models (LLMs) are increasingly deployed in dialogue systems and interactive agents, their social adaptation during natural interaction has drawn growing attention. While prior work shows strong social regulation under explicit role or style instructions, it remains unclear whether LLMs can spontaneously perceive and respond to implicit social differences without explicit prompts. Focusing on high-context Chinese interactions, we identify a robust phenomenon termed Social Agnosia, where LLMs fail to adequately perceive and accommodate implicit social power, affective arousal, and epistemic status during natural interaction. To diagnose this behavior, we propose C-ISA, a framework grounded in Communication Accommodation Theory that decomposes social adaptation into three approximately orthogonal dimensions, and conduct controlled comparisons across multiple Chinese LLMs under implicit and explicit conditions. Results show that while models substantially adjust linguistic strategies under explicit conditioning, they exhibit socially insensitive and homogenized responses in natural interaction, revealing a structural gap between spontaneous behavior and conditioned capability. The C-ISA dataset is publicly available at https://github.com/ty373/C-ISA.

[Long] Knowledge Injection Exists in MoE? Exploring Expert-Aware Contrast Decoding in MoE for Mitigating LLMs Hallucinations ▶ *Underline Video*

Xinyue Fang, Zhiliang Tian, Zhen Huang, Ziyi Pan, Zhihua Wen, Xi Wang, Quntian Fang and Dongsheng Li

Existing LLM hallucination mitigation methods, including prompt engineering and model optimization, either hardly alter models' internal knowledge or have poor cross-domain generalization. Contrastive decoding mitigates hallucinations by using layer-wise differences in LLMs. However, prior studies only explore transformer-based models (e.g., GPT), ignoring other effective frameworks like mixture-of-experts (MoE) models. Since MoE alters the traditional transformer architecture, we conduct empirical studies to investigate whether similar layer-wise differences exist in MoEs. Our results show that they do not exist in MoE with shared experts; nevertheless, across different MoEs, higher layers exhibit distinct expert activation patterns between factual and non-factual outputs. Building on these, we propose EAACD, an expert-aware adaptive contrast decoding that uses expert differences in MoE's higher layers to mitigate hallucinations on QA tasks. EAACD splits highlayer experts into a higher-reliability group and several lower-reliability groups based on their confidence and consistency. It contrasts the higher-reliability group's prediction with each lower-reliability group's prediction to calibrate the model's original predictions. To strengthen this contrast, EAACD amplifies hallucinations from lower-reliability experts via attention and masking to provide stronger negative references. EAACD outperforms all baselines on four datasets

[Long] Knowledge Poisoning Attacks on Medical Multi-Modal Retrieval-Augmented Generation ▶ *Underline Video*

Peiru Yang, Haoran Zheng, Tong Ju, Shiting Wang, Wanchun Ni, Jiajun Liu, Shangguang Wang, Yongfeng Huang and Tao Qi

Retrieval-augmented generation (RAG) is a widely adopted paradigm for enhancing LLMs in medical applications by incorporating expert multi-modal knowledge during generation. However, the underlying retrieval databases may naturally contain, or be intentionally injected with, adversarial knowledge, which can perturb model outputs and undermine system reliability. To investigate this risk, prior studies have explored knowledge poisoning attacks in medical RAG systems. Nevertheless, most of them rely on the strong assumption that adversaries possess prior knowledge of user queries, which is unrealistic in deployments and substantially limits their practical applicability. In this paper, we propose M^3Att, a knowledge-poisoning framework designed for medical multimodal RAG systems, assuming only limited distribution knowledge of the underlying database. Our core idea is to inject covert misinformation into textual data while using paired visual data as a query-agnostic trigger to promote retrieval. We first propose a unified framework that introduces imperceptible perturbations to visual inputs to manipulate retrieval probabilities. Besides, due to the prior medical knowledge in LLMs, naively poisoned medical content with explicit factual errors can be corrected during generation. Thus, we leverage the inherent ambiguity of medical diagnosis and design a covert misinformation injection strategy that degrades diagnostic accuracy while evading model self-correction. Experiments on five LLMs and datasets demonstrate that M^3Att consistently produces clinically plausible yet incorrect generations. Codes: https://anonymous.4open.science/r/M3Att.

[Long] Language on Demand, Knowledge at Core: Composing LLMs with Encoder-Decoder Translation Models for Extensible Multilinguality ▶ *Underline Video*

Mengyu Bu and Yang Feng

Large language models (LLMs) exhibit strong general intelligence, yet their multilingual performance remains highly imbalanced. Although LLMs encode substantial cross-lingual knowledge in a unified semantic space, they often struggle to reliably interface this knowledge with low-resource or unseen languages. Fortunately, pretrained encoder-decoder translation models already possess balanced multilingual capacity, suggesting a natural complement to LLMs. In this work, we propose XBridge, a compositional encoder-LLM-decoder architecture that offloads multilingual understanding and generation to external pretrained translation models, while preserving the LLM as an English-centric core for general knowledge processing and reasoning. To address the resulting representation misalignment across models, we introduce lightweight cross-model mapping layers together with an optimal transport-based alignment objective, enabling fine-grained semantic consistency for multilingual generation. Experiments on four LLMs across multilingual understanding, reasoning, and generation indicate that XBridge outperforms strong baselines, especially on low-resource and previously unseen languages, without retraining the LLM.

[Long] LBLLM: Lightweight Binarization of Large Language Models via Three-Stage Distillation ▶ *Underline Video*

Siqing Song, Chuang Wang, Yong Lang, Yi Yang and Xu-Yao Zhang

Deploying large language models (LLMs) in resource-constrained environments is hindered by heavy computational and memory requirements. We present LBLLM, a lightweight binarization framework that achieves effective W(1+1)A4 quantization through a novel three-stage quantization strategy. The framework proceeds as follows: (1) initialize a high-quality quantized model via PTQ; (2) quantize binarized weights, group-wise bitmaps, and quantization parameters through layer-wise distillation while keeping activations in full precision; and (3) training learnable activation quantization factors to dynamically quantize activations to 4 bits. This decoupled design mitigates interference between weight and activation quantization, yielding greater training stability and better inference accuracy. LBLLM, trained only using 0.016B tokens with a single GPU, surpasses existing state-of-the-art binarization methods on W2A4 quantization settings across tasks of language modeling, commonsense QA, and language understanding. These results demonstrate that extreme low-bit quantization of LLMs can be both practical and highly effective without introducing any extra high-precision channels or rotational matrices commonly used in recent PTQ-based works, offering a promising path toward efficient LLM deployment in resource-limited situations.

[Findings] Learning Dynamic Representations and Policies from Multimodal Clinical Time-Series with Informative Missingness ▶ *Underline Video*

Zihan Liang, Ziwen Pan and Ruoxuan Xiong

Multimodal clinical records contain structured measurements and clinical notes recorded over time, offering rich temporal information about the evolution of patient health. Yet these observations are sparse, and whether they are recorded depends on the patients latent condition. Observation patterns also differ across modalities, as structured measurements and clinical notes arise under distinct recording processes. While prior work has developed methods that accommodate missingness in clinical time series, how to extract and use the information carried by the observation process itself remains underexplored. We therefore propose a patient representation learning framework for multimodal clinical time series that explicitly leverages informative missingness. The framework combines (1) a multimodal encoder that captures signals from structured and textual data together with their observation patterns, (2) a Bayesian filtering module that updates a latent patient state over time from observed multimodal signals, and (3) downstream modules for offline treatment policy learning and patient outcome prediction based on the learned patient state. We evaluate the framework on ICU sepsis cohorts from MIMIC-III, MIMIC-IV, and eICU. It improves both offline treatment policy learning and adverse outcome prediction, achieving FQE 0.679 versus 0.528 for clinician behavior and AUROC 0.886 for post-72-hour mortality prediction on MIMIC-III.

[Long] Learning from the Irrecoverable: Error-Localized Policy Optimization for Tool-Integrated LLM Reasoning ▶ *Underline Video*

Qiao Liang, Yuke Zhu, Chao Ge, Lei Yang, Ying Shen, Bo Zheng and Sheng Guo

Tool-integrated reasoning (TIR) enables LLM agents to solve tasks through planning, tool use, and iterative revision, but outcome-only reinforcement learning in this setting suffers from sparse, delayed rewards and weak step-level credit assignment. In long-horizon TIR trajectories, an early irrecoverable mistake can determine success or failure, making it crucial to localize the first irrecoverable step and leverage it for fine-grained credit assignment. We propose Error-Localized Policy Optimization (ELPO), which localizes the first irrecoverable step via binary-search rollout trees under a fixed rollout budget, converts the resulting tree into stable learning signals through hierarchical advantage attribution, and applies error-localized adaptive clipping to strengthen corrective updates on the critical step and its suffix. Across TIR benchmarks in math, science QA, and code execution, ELPO consistently outperforms strong Agentic RL baselines under comparable sampling budgets, with additional gains in Pass@K and Major@K scaling, rollout ranking quality, and tool-call efficiency. Our code is publicly released for reproducibility at https://anonymous.4open.science/r/ELPO-7C19.

[Long] Learning from the Irrecoverable: Error-Localized Policy Optimization for Tool-Integrated LLM Reasoning ▶ *Underline Video*

Qiao Liang, Yuke Zhu, Chao Ge, Lei Yang, Ying Shen, Bo Zheng and Sheng Guo

Tool-integrated reasoning (TIR) enables LLM agents to solve tasks through planning, tool use, and iterative revision, but outcome-only reinforcement learning in this setting suffers from sparse, delayed rewards and weak step-level credit assignment. In long-horizon TIR trajectories, an early irrecoverable mistake can determine success or failure, making it crucial to localize the first irrecoverable step and leverage it for fine-grained credit assignment. We propose Error-Localized Policy Optimization (ELPO), which localizes the first irrecoverable step via binary-search rollout trees under a fixed rollout budget, converts the resulting tree into stable learning signals through hierarchical advantage attribution, and applies error-localized adaptive clipping to strengthen corrective updates on the critical step and its suffix. Across TIR benchmarks in math, science QA, and code execution, ELPO consistently outperforms strong Agentic RL baselines under comparable sampling budgets, with additional gains in Pass@K and Major@K scaling, rollout ranking quality, and tool-call efficiency. Our code is publicly released for reproducibility at https://anonymous.4open.science/r/ELPO-7C19.

[Long] Learning Invariant Modality Representation for Robust Multimodal Learning from a Causal Inference Perspective ▶ *Underline Video*

Sijie Mai and Shiqin Han

Multimodal affective computing aims to predict humans' sentiment, emotion, intention, and opinion using language, acoustic, and visual modalities. However, current models often learn spurious correlations that harm generalization under distribution shifts or noisy modalities. To address this, we propose a causal modality-invariant representation (CmIR) learning framework for robust multimodal learning. At its core, we introduce a theoretically grounded disentanglement method that separates each modality into 'causal invariant representation' and 'environment-specific spurious representation' from a causal inference perspective. CmIR ensures that the learned invariant representations retain stable predictive relationships with labels across different environments while preserving sufficient information from the raw inputs

via invariance constraint, mutual information constraint, and reconstruction constraint. Experiments across multiple multimodal benchmarks demonstrate that CmIR achieves state-of-the-art performance. CmIR particularly excels on outofdistribution data and noisy data, confirming its robustness and generalizability.

[Long] Learning to Edit Knowledge via Instruction-based Chain-of-Thought Prompting ▶ *Underline Video*

Jinhu Fu, Yan Bai, Longzhu He, Yihang Lou, Yanxiao Zhao, Li Sun and Sen Su

Large language models (LLMs) can effectively handle outdated information through knowledge editing. However, current approaches face two key limitations: (I) Poor generalization: Most approaches rigidly inject new knowledge without ensuring that the model can use it effectively to solve practical problems. (II) Narrow scope: Current methods focus primarily on structured fact triples, overlooking the diverse unstructured forms of factual information (e.g., news, articles) prevalent in real-world contexts. To address these challenges, we propose a new paradigm: teaching LLMs to edit knowledge via Chain of Thoughts (CoTs) reasoning (CoT2Edit). We first leverage language model agents for both structured and unstructured edited data to generate CoTs, building high-quality instruction data. The model is then trained to reason over edited knowledge through supervised fine-tuning (SFT) and Group Relative Policy Optimization (GRPO). At inference time, we integrate Retrieval-Augmented Generation (RAG) to dynamically retrieve relevant edited facts for real-time knowledge editing. Experimental results demonstrate that our method achieves strong generalization across six diverse knowledge editing scenarios with just a single round of training on three open-source language models.

[Findings] Learning to Extract Rational Evidence via Reinforcement Learning for Retrieval-Augmented Generation ▶ *Underline Video*

Xinping Zhao, Shouzheng Huang, Yan Zhong, Xinshuo Hu, Meishan Zhang, Baotian Hu and Min Zhang

Retrieval-Augmented Generation (RAG) effectively improves the accuracy of Large Language Models (LLMs). However, retrieval noises significantly undermine the quality of LLMs' generation, necessitating the development of denoising mechanisms. Previous works extract evidence straightforwardly without deep thinking, which may risk filtering out key clues and struggle with generalization. To this end, we propose EviOmni, which learns to extract rational evidence via reasoning first and then extracting. Specifically, EviOmni integrates evidence reasoning and evidence extraction into one unified trajectory, followed by knowledge token masking to avoid information leakage, optimized via on-policy reinforcement learning with verifiable rewards in terms of answer, length, and format. Extensive experiments on five benchmark datasets show the superiority of EviOmni, which provides compact and high-quality evidence, enhances the accuracy of downstream tasks, and supports both traditional and agentic RAG systems.

[Long] Learning to Think on Hypergraph: HyperCoT for Structure-Guided N-ary Knowledge Graph Completion ▶ *Underline Video*

Mengxue Yang, Jingming Li, Chun Yang, Jiaqi Zhu, Jiafan Li, Guanhua Zhang and Ying Li

N-ary knowledge graph completion (KGC) aims to infer missing components in facts with multiple entities under distinct semantic roles, commonly formulated as a knowledge hypergraph link prediction task. Most embedding-based approaches score individual hyperedges relying on enriched structural representations, but overlook intermediate propagation states containing complementary local and global structural evidence. Despite their capability to generate chain-of-thought (CoT) representations for the classical KGC task, large language models (LLMs) struggle with hypergraph structure involving multiple facts, while current hypergraph QA methods only provide LLMs with a single query signal rather than path-level evidence. These limitations hinder the transferability of existing methods, especially those leveraging LLMs, to solve the knowledge hypergraph link prediction problem. To bridge this gap, we propose HyperCoT, a structure-aware approach that models multi-hop structural reasoning as a depth-sensitive progressive evidence accumulation process. It constructs a Graphical Chain-of-Thought (Graph-CoT) by aggregating role-aware hyperedge states along strongly correlated reasoning paths, and injects the resulting path-level structural evidence into each token in query and candidate entities to prompt LLMs. Experiments on three real-world datasets demonstrate that HyperCoT consistently outperforms strong n-ary KGC baselines, particularly in high arity and structural sparsity scenarios, meanwhile yielding interpretable multi-hop reasoning traces.

[Long] Leibniz: Theory-of-Mind Driven Neuro-Symbolic Logical Reasoning via Multi-Agent Collaboration ▶ *Underline Video*

Yue Fan, Hu zhang, Yunxiao Zhao, Guangjun Zhang, Hao Zhan, Ru Li, Hongye Tan and Yuanlong Wang

Logical reasoning with large language models (LLMs) has made significant progress in recent years. However, existing methods still suffer from insufficient rule semantic grounding and weak rule application mechanisms, making it difficult to achieve precise understanding and effective utilization of rules in complex multi-step reasoning. To address this, we propose Leibniz, a theory-of-mind driven neuro-symbolic reasoning framework. Specifically, we construct a bidirectional reasoning model based on multi-agent collaboration, which characterizes the reasoning process from two complementary perspectives, namely the Evolution Agent and the Reduction Agent. The former transforms belief-unstable propositions into stable ones that are beneficial for proving the target conclusion. The latter performs reverse reduction from the target to resolve belief conflicts and distill new inferential insights. Both share a belief state space and achieve efficient collaborative reasoning through continual belief updating. We integrate natural language and symbolic representations throughout the reasoning process. The experimental results demonstrate that Leibniz surpasses existing state-of-the-art models in reasoning accuracy, and further analyses reveal its substantial advantages in reliability and flexibility.

[Long] Less Languages, Less Tokens: An Efficient Unified Logic Cross-lingual Chain-of-Thought Reasoning Framework ▶ *Underline Video*

Chenyuan Zhang, Qiguang Chen, Xie Chen, Zhuotao Tian, Bowen Xing, Meishan Zhang, Libo Qin, Baotian Hu and Min Zhang
Cross-lingual chain-of-thought (XCoT) with self-consistency markedly enhances multilingual reasoning, yet existing methods remain costly due to extensive sampling of full trajectories across languages. Moreover, multilingual LLM representations vary strongly by language, hindering direct feature comparisons and effective pruning. To address this, we introduce UL-XCoT, the first efficient unified logic cross-lingual reasoning framework that minimizes redundancy in token usage and latency, yielding the greatest efficiency under limited sampling budgets during inference. Specifically, UL-XCoT (1) achieves less languages by selecting, per query, a small candidate language set in a language-invariant unified logic space, (2) enables less tokens by monitoring logic-space trajectory dynamics during decoding to prune low-quality reasoning paths, and (3) aggregates the remaining high-quality trajectories via voting. Experiments on PolyMath across 18 languages with DeepSeek-R1-Distill-Qwen-7B demonstrate that UL-XCoT achieves competitive accuracy while sharply cutting over 50% decoding token cost and latency versus prior sampling baselines. It also delivers more stable gains on low-resource languages, underscoring consistently superior robustness where standard XCoT self-consistency method fails.

[Long] Leveraging Outline-Optimized Generative Interactions and Critique for Self-Refining Outlines with Reinforcement Learning ▶ *Underline Video*
Hengwei Liu, Haoyuan Ma, Qingqing Lyu, Daoxin Zhang, Yao Hu, Yongliang Shen, Yin Zhang and Weiming Lu
Long-form outline generation requires satisfying multiple competing objectives simultaneously: outlines must be engaging, well-organized, topically relevant, and comprehensive while maintaining logical consistency across hierarchical structures. Current approaches either rely on expensive multi-turn interactions with large language models or employ procedural refinement pipelines that cannot systematically learn from critique. We present Logic-RL, a framework that transforms critique-guided outline refinement into a learnable policy through reinforcement learning. Our approach constructs refinement trajectories from teacher demonstrations, synthesizes explicit reasoning chains that decompose the critique-revision process, and optimizes a refinement policy using group relative policy optimization with structure-aware rewards. Experiments on FreshWiki and WikiOutline demonstrate that Logic-RL achieves substantial improvements over strong baselines, with the 0.6B model obtaining 79.17% relative gain and the 1.7B model achieving 8.67% improvement in average rubric scores compared to the best existing methods. Further analysis reveals that learned refinement policies generalize across domains and can be iteratively applied, with quality continuing to improve through three refinement rounds before diminishing returns.

[Long] Libra-VLA: Achieving Learning Equilibrium via Asynchronous Coarse-to-Fine Dual-System ▶ *Underline Video*
Yifei Wei, Linqing Zhong, Yi Liu, Yuxiang Lu, Xindong He, Maoqing Yao and Guanghui Ren
Vision-Language-Action (VLA) models are a promising paradigm for generalist robotic manipulation by grounding high-level semantic instructions into executable physical actions. However, prevailing approaches typically adopt a monolithic generation paradigm, directly mapping visual-linguistic features to high-frequency motor commands in a flat, non-hierarchical fashion. This strategy overlooks the inherent hierarchy of robotic manipulation, where complex actions can be naturally modeled in a Hybrid Action Space, decomposing into discrete macro-directional reaching and continuous micro-pose alignment, severely widening the semantic-actuation gap and imposing a heavy representational burden on grounding high-level semantics to continuous actions. To address this, we introduce Libra-VLA, a novel Coarse-to-Fine Dual-System VLA architecture. We explicitly decouple the learning complexity into a coarse-to-fine hierarchy to strike a training equilibrium, while simultaneously leveraging this structural modularity to implement an asynchronous execution strategy. The Semantic Planner predicts discrete action tokens capturing macro-directional intent, while the Action Refiner conditions on coarse intent to generate high-frequency continuous actions for precise alignment. Crucially, our empirical analysis reveals that performance follows an inverted-U curve relative to action decomposition granularity, peaking exactly when the learning difficulty is balanced between the two sub-systems. With the asynchronous design, our approach offers a scalable, robust, and responsive solution for open-world manipulation.

[Long] Lightweight LLM Agent Memory with Small Language Models ▶ *Underline Video*
Jiaquan Zhang, Chaoning Zhang, Shuxu Chen, Zhenzhen Huang, Pengcheng Zheng, Zhicheng Wang, Ping Guo, Fan Mo, Sung-Ho Bae, Jie Zou, Jiwei Wei and Yang Yang
Although LLM agents can leverage tools for complex tasks, they still need memory to maintain cross-turn consistency and accumulate reusable information in long-horizon interactions. However, retrieval-based external memory systems incur low online overhead but suffer from unstable accuracy due to limited query construction and candidate filtering. In contrast, many systems use repeated large-model calls for online memory operations, improving accuracy but accumulating latency over long interactions. We propose LightMem, a lightweight memory system for better agent memory driven by Small Language Models (SLMs). LightMem modularizes memory retrieval, writing, and long-term consolidation, and separates online processing from offline consolidation to enable efficient memory invocation under bounded compute. We organize memory into short-term memory (STM) for immediate conversational context, mid-term memory (MTM) for reusable interaction summaries, and long-term memory (LTM) for consolidated knowledge, and uses user identifiers to support independent retrieval and incremental maintenance in multi-user settings. Online, LightMem operates under a fixed retrieval budget and selects memories via a two-stage procedure: vector-based coarse retrieval followed by semantic consistency re-ranking. Offline, it abstracts reusable interaction evidence and incrementally integrates it into LTM. Experiments show consistent gains across model scales, with an average F1 improvement of about 2.5 over A-MEM on LoCoMo, while achieving higher efficiency and low median latency (83 ms for retrieval and 581 ms end-to-end).

[Findings] LinguaGame: A Linguistically Grounded Game-Theoretic Paradigm for Multi-Agent Dialogue Generation ▶ *Underline Video*

Yuxiao Ye, Yiming Zhang, Yiran Ma, Huiyuan Xie, Huining Zhu and Zhiyuan Liu

Large Language Models (LLMs) have enabled Multi-Agent Systems (MASs) where agents interact through natural language to solve complex tasks or simulate multi-party dialogues. Recent work on LLM-based MASs has mainly focused on architecture design, such as role assignment and workflow orchestration. In contrast, this paper targets the interaction process itself, aiming to improve agents' communication efficiency by helping them convey their intended meaning more effectively through language. To this end, we propose LinguaGame, a linguistically-grounded game-theoretic paradigm for multi-agent dialogue generation. Our approach models dialogue as a signalling game over communicative intents and strategies, solved with a training-free equilibrium approximation algorithm for inference-time decision adjustment. Unlike prior game-theoretic MASs, whose game designs are often tightly coupled with task-specific objectives, our framework relies on linguistically informed reasoning with minimal task-specific coupling. Specifically, it treats dialogue as intentional and strategic communication, requiring agents to infer what others aim to achieve (intents) and how they pursue those goals (strategies). We evaluate our framework in simulated courtroom proceedings and debates, with human expert assessments showing significant gains in communication efficiency. We release our code and data on GitHub.

[Findings] LiveCANNBench: Benchmark SWE AI Coding for Ascend CANN ► *Underline Video*

Sijie Wang, Kai Zhao, Wee Peng Tay, Shuo Zhang, Chengwen Liu, Quanjiang Guo, Ren Junhao, Xin Li, Heng Lian, Jingdi Lei, Rui She, Huacan Wang and Ronghao Chen

AI coding has emerged as a core application of large language models (LLMs), evolving from single-file coding tasks towards complex software engineering (SWE) scenarios. Recent advances in agents have enabled multi-file, multi-language, and dependency-aware AI coding, significantly expanding the scope of AI-assisted software development. While a variety of benchmarks have been proposed to evaluate coding capabilities in general-purpose or GPU coding ecosystems such as CUDA and ROCm, systematic evaluation for Huawei Ascend CANN remains largely underexplored. In this work, we propose LiveCANNBench, an SWE-level benchmark designed for AI coding in the CANN software stack. LiveCANNBench is constructed from real-world CANN repositories and consists of over 400 task instances spanning multi-file, multi-language, and execution-aware coding challenges. Unlike existing static benchmarks that primarily focus on kernel-level code generation, LiveCANNBench adopts a live benchmarking paradigm, effectively mitigating data leakage and enabling more reliable evaluation of modern coding agents.

[Long] LLM-Powered Benchmark Factory: Reliable, Generic, and Efficient ► *Underline Video*

Peiwen Yuan, Shaoxiong Feng, Yiwei Li, Xinglin Wang, Yueqi Zhang, Jiayi Shi, Chuyi Tan, Boyuan Pan, Yao Hu and Kan Li

The rapid advancement of large language models (LLMs) has led to a surge in both model supply and application demands. To facilitate effective matching between them, generic and efficient benchmark generators that can construct high-quality benchmarks are widely needed. However, human annotators are constrained by inefficiency, and current LLM-based benchmark generators lack not only generalizability but also a comprehensive evaluation framework for validation and optimization. To fill this gap, we first establish an automated evaluation framework, structured around four dimensions and ten criteria. Under this framework, we carefully analyze the advantages and weaknesses of directly prompting LLMs as generic benchmark generators. On this basis, we introduce a series of methods to address the identified weaknesses and integrate them as BenchMaker. Experiments across multiple LLMs and tasks confirm that BenchMaker achieves comparable performance to human-annotated benchmarks on most metrics, highlighting its generalizability and validity. More importantly, it delivers highly consistent evaluation results across 21 LLMs (e.g., 0.969 Pearson correlation against MMLU-Pro on language understanding task), while incurring minimal overhead (e.g., $0.005 and 0.38 minutes per sample when using GPT-4o mini as generator).

[Long] LLM-VA: Resolving the Jailbreak-Overrefusal Trade-off via Vector Alignment ► *Underline Video*

Haonan Zhang, Dongxia Wang, Yi Liu, Kexin Chen and Wenhai Wang

Safety-aligned LLMs suffer from two failure modes: jailbreak (responding to harmful inputs) and over-refusal (declining benign queries). Existing vector steering methods adjust the magnitude of answer vectors, but this creates a fundamental trade-off—reducing jailbreak increases over-refusal and vice versa. We identify the root cause: LLMs encode the decision to respond (answer vector v_a) and the judgment of input safety (benign vector v_b) as nearly orthogonal directions, treating them as independent processes. We propose LLM-VA, which aligns v_a with v_b through closed-form weight updates, making the model's willingness to respond causally dependent on its safety assessment—without fine-tuning or architectural changes. Our method identifies vectors at each layer using SVMs, selects safety-relevant layers, and iteratively aligns vectors via minimum-norm weight modifications. Experiments on 12 LLMs demonstrate that LLM-VA achieves 11.45% higher F1 than the best baseline while preserving 95.92% utility, and automatically adapts to each model's safety bias without manual tuning. Code and models are available at https://hotbento.github.io/LLM-VA-Web/.

[Long] LLM-XTM: Enhancing Cross-Lingual Topic Models with Large Language Models ► *Underline Video*

Minh Chu Xuan, Tien-Phat Nguyen, Linh Ngo Van, Dinh Viet Sang, Nguyen Thi Ngoc Diep and Trung Le

Cross-lingual topic modeling aims to discover shared semantic structures across languages, yet existing models depend on sparse bilingual resources and often yield incoherent or weakly aligned topics. Recent LLM-based refinements improve interpretability but are costly, document-level, and prone to hallucination, with prior white-box approaches requiring inaccessible token probabilities. We propose LLM-XTM, a framework that integrates LLM-guided topic refinement with self-consistency uncertainty quantification, enabling black-box, stable, and scalable enhancement of cross-lingual topic models. Experiments on multilingual corpora show that LLM-XTM achieves superior topic coherence and alignment while reducing reliance on bilingual dictionaries and expensive LLM calls.

[Long] Long-Chain Reasoning Distillation via Adaptive Prefix Alignment ▶ *Underline Video*

Zhenghao Liu, Zhuoyang Wu, Xinze Li, Yukun Yan, Shuo Wang, Zulong Chen, Yu Gu, Ge Yu and Maosong Sun

Large Language Models (LLMs) have demonstrated remarkable reasoning capabilities, particularly in solving complex mathematical problems. Recent studies show that distilling long reasoning trajectories can effectively enhance the reasoning performance of small-scale student models. However, teacher-generated reasoning trajectories are often excessively long and structurally complex, making them difficult for student models to learn. This mismatch leads to a gap between the provided supervision signal and the learning capacity of the student model. To address this challenge, we propose Prefix-ALIGNment distillation (P-ALIGN), a framework that fully exploits teacher CoTs for distillation through adaptive prefix alignment. Specifically, P-ALIGN adaptively truncates teacher-generated reasoning trajectories by determining whether the remaining suffix is concise and sufficient to guide the student model. Then, P-ALIGN leverages the teacher-generated prefix to supervise the student model, encouraging effective prefix alignment. Experiments on multiple mathematical reasoning benchmarks demonstrate that P-ALIGN outperforms all baselines by over 3%. Further analysis indicates that the prefixes constructed by P-ALIGN provide more effective supervision signals, while avoiding the negative impact of redundant and uncertain reasoning components. All codes are available at https://github.com/NEUIR/P-ALIGN.

[Long] Looking at Radiology Report Generation through a Causal Lens: A Survey ▶ *Underline Video*

Satyam Kumar, Kaustubh Shivshankar Shejole and Pushpak Bhattacharyya

Automatic radiology report generation (RRG) has emerged as a promising approach to reduce clinicians' workload, yet existing systems are vulnerable to biases induced by spurious correlations across data, models, and evaluation pipelines. Such biases raise serious fairness concerns and may adversely affect patient care, making their mitigation critical in clinical settings. Leveraging causal inference to identify true cause-effect relationships can mitigate many biases and yield fair, reliable systems with clinically meaningful outputs. Existing surveys on RRG primarily emphasize deep learning approaches while overlooking the critical role of causality. This survey addresses this gap by analyzing bias across the RRG pipeline, formalizing RRG as a causal modeling problem, and reviewing representative causal techniques from the literature. Based on the level of intervention, we organize existing mitigation strategies into a three-tier taxonomy. We further examine commonly used public medical imaging datasets and evaluation metrics through a causal lens, revealing their biases and limitations in capturing causal alignment and clinical fidelity. To address these limitations, we advocate broader demographic coverage and causal-aware evaluation metrics to improve fairness and reliability, and identify important directions for future work.

[Long] LoPT: Lossless Parallel Tokenization Acceleration for Long Context Inference of Large Language Model ▶ *Underline Video*

Wei Shao, Lingchao Zheng, Pengyu Wang, Peizhen Zheng, Jun Li and Yuwei Fan

Long context inference scenarios have become increasingly important for large language models, yet they introduce significant computational latency. While prior research has optimized long-sequence inference through operators, model architectures, and system frameworks, tokenization remains an overlooked bottleneck. Existing parallel tokenization methods accelerate processing through text segmentation and multi-process tokenization, but they suffer from inconsistent results due to boundary artifacts that occur after merging. To address this, we propose LoPT, a novel Lossless Parallel Tokenization framework that ensures output identical to standard sequential tokenization. Our approach employs character-position-based matching and dynamic chunk length adjustment to align and merge tokenized segments accurately. Extensive experiments across diverse long-text datasets demonstrate that LoPT achieves significant speedup while guaranteeing lossless tokenization. We also provide theoretical proof of consistency and comprehensive analytical studies to validate the robustness of our method.

[Long] Losing our Tail, Again: (Un)Natural Selection & Multilingual LLMs ▶ *Underline Video*

Eva Vanmassenhove

Multilingual Large Language Models considerably changed how technologies influence language. While previous technologies could mediate or assist humans, there is now a tendency to offload the task of writing itself to these technologies, enabling models to change our languages more directly. While they provide us quick access to information and impressively fluent output, beneath their (apparent) sophistication lies a subtle, insidious threat: the gradual decline and loss of linguistic diversity. In this position paper, I explore how model collapse, with a particular focus on translation technology, can lead to the loss of linguistic forms, grammatical features, and cultural nuance. Model collapse refers to the consequences of self-consuming training loops, where automatically generated data (re-)enters the training data, leading to a gradual distortion of the data distribution and the underrepresentation of low-probability linguistic phenomena. Drawing on recent work in Computer Vision, Natural Language Processing and Machine Translation, I argue that the many tails of our linguistic distributions might be vanishing, and with them, the narratives and identities they carry. This paper is a call to resist linguistic flattening and to reimagine Natural Language Processing as a field that encourages, values and protects expressive multilingual diversity and creativity.

[Long] M^2PO: Multi-Perspective Multi-Pair Preference Optimization for Machine Translation ▶ *Underline Video*

Hao Wang, Linlong Xu, Heng Liu, Yangyang Liu, Xiaohu Zhao, Bo Zeng, Liangying Shao, Yichen Dong, Xinwei Wu, Jiang Zhou, Tianyu Dong, Xiangxiang Zeng, Longyue Wang and Weihua Luo

Aligning Large Language Models (LLMs) to human preferences is pivotal for Machine Translation (MT), yet current approaches are often hindered by misleading reward signals. Our analysis reveals that prevailing Quality Estimation (QE) models exhibit a systematic blind spot towards partial errors, specifically partial hallucinations and omissions, often favoring superficially fluent but unfaithful translations.

To address this, we propose M^2PO (Multi-Perspective Multi-Pair Preference Optimization), a data-centric framework for preference optimization in machine translation. First, to correct the bias towards fluency, M^2PO uses a multi-perspective alignment mechanism that decouples semantic fidelity from fluency, prioritizing faithfulness via a curriculum strategy. Second, with the bias corrected, partial errors fall between perfect and severely incorrect translations, making them inefficient to learn via standard best-versus-worst comparisons. We thus introduce a multi-pair objective that leverages the full candidate list to capture these fine-grained error signals. Experiments on WMT23, WMT24, and FLORES-200 show that M^2PO enables a 9B model to outperform leading open-source baselines and achieve parity with proprietary models like GPT-4o and Gemini-2.0-Flash, demonstrating significant potential for efficient, high-fidelity LLM-based translation.

[Long] MARD: Module-Aware Reasoning Distillation for Language Models with Adaptive Supervision ► *Underline Video*

Wenqi Yang, Jianjun Li, Zhibo Zhang, Mingqian Ding and Yushen Fang

Multi-step reasoning remains challenging for language models with limited capacity. While recent reasoning distillation approaches transfer chain-of-thought supervision from large teacher models, they typically apply uniform supervision across all Transformer components, overlooking the fact that different modules contribute unequally to reasoning. We propose Module-Aware Reasoning Distillation, a parameter-efficient framework that explicitly targets key Transformer components for effective reasoning transfer. Through systematic analysis, we identify the feed-forward network projections and the output projection of self-attention as primary bottlenecks for reasoning. Based on these findings, we introduce lightweight adapter modules at these components while freezing the backbone parameters, enabling focused and efficient distillation. Our approach adopts an offline distillation setting, where a strong teacher model provides reasoning trajectories in advance, and incorporates an adaptive supervision strategy that adjusts the strength of reasoning-related losses according to problem difficulty. Experiments on mathematical reasoning benchmarks demonstrate consistent improvements over strong baselines, and ablation studies confirm the importance of both module-aware placement and adaptive supervision.

[Long] MARS-RA: Rank Aggregation for Credit Assignment via Multimodal Comparisons in Embodied Multi-Agent Cooperation ► *Underline Video*

Dawei Wang, Di Zhao, Xinyuan Liu, Marci Chi Ma, Xiaoyang Liu, Chengming Zhou, Gary Ushaw and Richard Davison

Credit assignment is a fundamental challenge in cooperative multi-agent reinforcement learning, particularly in embodied AI settings characterized by limited and delayed feedback as well as dynamically changing numbers of active agents. We propose MARS-RA, a framework that reformulates credit assignment as a rank aggregation problem using contribution-based pairwise comparisons among agents generated by large multimodal models. This shift from absolute to relative estimation ensures robustness against noise and dynamic agent participation, converting comparison results into contribution scores for potential-based reward shaping. We provide theoretical justification for the convergence and robustness of the proposed framework, and show that Shapley values can be used as an interpretive reference. Experimental results on challenging tasks of different types indicate that MARS-RA can guide agents toward effective cooperation.

[Long] MCP-Flow: Facilitating LLM Agents to Master Real-World, Diverse and Scaling MCP Tools ► *Underline Video*

Wenhao Wang, Peizhi Niu, Zhao Xu, Zhaoyu Chen, Jian Du, Yaxin Du, Xianghe Pang, Keduan Huang, Yanfeng Wang, Qiang Yan and Siheng Chen

Large Language Models (LLMs) increasingly rely on external tools to perform complex, realistic tasks, yet their ability to utilize the rapidly expanding Model Contextual Protocol (MCP) ecosystem remains limited. Existing MCP research covers few servers, depends on costly manual curation, and lacks training support, hindering progress toward real-world deployment. To overcome these limitations, we introduce MCP-Flow, an automated web-agent-driven pipeline for large-scale server discovery, data synthesis, and model training. MCP-Flow collects and filters data from 1166 servers and 11536 tools, producing 68733 high-quality instruction-function call pairs and 6439 trajectories, far exceeding prior work in scale and diversity. Extensive experiments demonstrate MCP-Flow's effectiveness in driving superior MCP tool selection, function-call generation, and enhanced agentic task performance. MCP-Flow thus provides a scalable foundation for advancing LLM agents' proficiency in real-world MCP environments.

[Long] Measuring Social Bias in Vision-Language Models with Face-Only Counterfactuals from Real Photos ► *Underline Video*

Haodong Chen, Qiang Huang, Jiaqi Zhao, Qiuping Jiang, Xiaojun Chang and Jun Yu

Vision-Language Models (VLMs) are increasingly deployed in socially consequential settings, raising concerns about social bias driven by demographic cues. A central challenge in measuring such social bias is attribution under visual confounding: real-world images entangle race and gender with correlated factors such as background and clothing, obscuring attribution. We propose a face-only counterfactual evaluation paradigm that isolates demographic effects while preserving real-image realism. Starting from real photographs, we generate counterfactual variants by editing only facial attributes related to race and gender, keeping all other visual factors fixed. Based on this paradigm, we construct FOCUS, a dataset of 480 scene-matched counterfactual images across six occupations and ten demographic groups, and propose REFLECT, a benchmark comprising three decision-oriented tasks: two-alternative forced choice, multiple-choice socioeconomic inference, and numeric salary recommendation. Experiments on five state-of-the-art VLMs reveal that demographic disparities persist under strict visual control and vary substantially across task formulations. These findings underscore the necessity of controlled, counterfactual audits and highlight task design as a critical factor in evaluating social bias in multimodal models.

[Long] MECH: A Cost-Effective Multi-Task Cascade Framework for Classroom Opinion Evolution Recognition ► *Underline Video*

Yancui Li, Xiaoyu Zhou, Guoyi Miao and Fang Kong

Classroom discourse analysis is critical for tracing cognitive restructuring, yet existing research predominantly focuses on Dialogue Acts (DA), overlooking the deeper dimension of Opinion Evolution (OE). In this paper, we formally define the task of Classroom Opinion Evolution Recognition and introduce the Classroom Opinion Evolution Dataset (COED). Addressing the "Accuracy-Cost-Data" trilemma in real-world educational scenarios and the "overconfidence" failure mode of traditional confidence-based cascading systems on long-tail samples, we propose the Multi-task Enhanced Cascade Hybrid (MECH) framework. Grounded in the CODA (Continuous Opinions and Discrete Actions) theory, MECH conceptually translates the "Action-Opinion" dualism into a risk-aware routing mechanism. Instead of relying solely on prediction confidence, this mechanism utilizes high-risk argumentative DA signals derived from multi-task learning to construct a "semantic safety net" effectively routing implicit or ambiguous samples to a Large Language Model for reasoning. Experimental results demonstrate that MECH achieves a state-of-the-art accuracy of 78.55% while reducing API costs by 44.4%. Furthermore, the framework exhibits robustness in few-shot scenarios (using only 20% of data), offering a cost-effective and interpretable solution for large-scale educational dialogue analysis. Our code and data are available at https://github.com/ywh24284-code/MECH.

[Findings] Meeseeks: A Feedback-Driven, Iterative Self-Correction Benchmark evaluating LLMs' Instruction Following Capability ▶ *Underline Video*

Jiaming wang, Yunke Zhao, Peng Ding, Jun Kuang, Yibin Shen, Zhe Tang, Yilin Jin, ZongYu Wang, Xiaoyu Li and Xuezhi Cao

The capability to precisely adhere to instructions is a cornerstone for Large Language Models (LLMs) to function as dependable agents in real-world scenarios. However, confronted with complex prompts, LLMs frequently encounter difficulties in fulfilling all specified requirements within a single response. Drawing inspiration from recent advancements in Chain-of-Thought (CoT) prompting and self-correction methodologies, we introduce Meeseeks, a fully automated iterative instruction-following benchmark equipped with an integrated feedback mechanism. Meeseeks identifies erroneous components in model responses and provides corresponding feedback accurately, thereby iteratively guiding the model toward self-correction. The dataset contains over 700 curated instances annotated by 32 distinct capability tags in Chinese and English. Extensive experimental results reveal that different state-of-the-art commercial and open-source LLMs exhibit vastly disparate performance, and even after 20 turns of iterative feedback-driven self-correction, nearly all models demonstrate suboptimal performance. We conducted comprehensive analysis and uncovered numerous common issues prevalent in current state-of-the-art models, as well as several counterintuitive phenomena. Meeseeks has been open-sourced on https://github.com/ADoublLEN/Meeseeks.

[Long] MentalSeek-Dx: Towards Progressive Hypothetico-Deductive Reasoning for Real-world Psychiatric Diagnosis ▶ *Underline Video*

Xiao Sun

Mental health disorders represent a burgeoning global public health challenge. While Large Language Models (LLMs) have demonstrated potential in psychiatric assessment, their clinical utility is severely constrained by benchmarks that lack ecological validity and fine-grained diagnostic supervision. To bridge this gap, we introduce MentalDx Bench, the first benchmark dedicated to disorder-level psychiatric diagnosis within real-world clinical settings. Comprising 712 de-identified electronic health records annotated by board-certified psychiatrists under ICD-11 guidelines, the benchmark covers 76 disorders across 16 diagnostic categories. Evaluation of 18 LLMs reveals a critical paradigm misalignment: strong performance at coarse diagnostic categorization contrasts with systematic failure at disorder-level diagnosis, underscoring a gap between pattern-based modeling and clinical hypothetico-deductive reasoning.In response, we propose MentalSeek-Dx, a medical-specialized LLM trained to internalize this clinical reasoning process through supervised trajectory construction and curriculum-based reinforcement learning. Experiments on MentalDx Bench demonstrate that MentalSeek-Dx achieves state-of-the-art (SOTA) performance with only 14B parameters, establishing a clinically grounded framework for reliable psychiatric diagnosis. The dataset and code are available.

[Findings] microCLIP: Unsupervised CLIP Adaptation via Coarse-Fine Token Fusion for Fine-Grained Image Classification ▶ *Underline Video*

Sathira Silva, Eman Ali, Chetan Arora and Muhammad Haris Khan

Unsupervised adaptation of CLIP-based vision-language models (VLMs) for fine-grained image classification is challenging because pseudo-labels must be inferred from tiny local cues. CLIPs contrastive pretraining yields strong zero-shot transfer, but its coarse-grained [CLS] visual token often misses fine spatial details. Existing UA methods mainly enrich text prompts with large language model (LLM) descriptions while still relying on [CLS], which limits spatial precision. We propose microCLIP, a label-free self-training framework that jointly adapts visual features and LLM-derived text prototypes using fine-grained cues. First, a Saliency-Oriented Attention Pooling (SOAP) mechanism inside a lightweight TokenFusion module constructs a saliency-guided [FG] token from patch features and fuses it with the global [CLS] representation for coarse-fine alignment. Second, a two-headed LLM-derived classifier combines a frozen text head, used with multi-view CLIP features as a stable prior for pseudo-labels, with a learnable head initialized from the same LLM-derived descriptions. Finally, Dynamic Knowledge Aggregation convexly combines fixed CLIP/LLM priors with TokenFusion logits during self-training. Across 13 diverse classification benchmarks including fine-grained ones, microCLIP yields a mean improvement of +2.90% points over the strongest prior UA baseline while fine-tuning only layer norms and a tiny head.

[Findings] MiMIC: Mitigating Visual Modality Collapse in Universal Multimodal Retrieval While Avoiding Semantic Misalignment ▶ *Underline Video*

Juan Li, Chuanghao Ding, Xujie Zhang and Cam-Tu Nguyen

Universal Multimodal Retrieval (UMR) aims to map different modalities (e.g., visual and textual) into a shared embedding space for multi-modal retrieval. Existing UMR methods can be broadly divided into two categories: early-fusion approaches, such as Marvel, which projects visual features into the language model (LM) space for integrating with text modality, and late-fusion approaches, such as UniVL-DR, encode visual and textual inputs using separate encoders and obtain fused embeddings through addition. Our pilot study reveals that Marvel exhibits visual modality collapse, which is characterized by the models tendency to disregard visual features while depending excessively on textual cues. In contrast, although UniVL-DR is less affected by this issue, it is more susceptible to semantic misalignment, where semantically related content is positioned far apart in the embedding space. To address these challenges, we propose MiMIC, which introduces two key innovations: (1) a fusion-in-decoder architecture for effective multimodal integration, and (2) robust training through single-modality mix-in and random caption dropout. Experiments on the WebQA+ and EVQA+ datasets, where image in documents or queries might lack captions, indicate that MiMIC consistently outperforms both early- and late-fusion baselines.

[Long] MIND: From Passive Mimicry to Active Reasoning through Capability-Aware Multi-Perspective CoT Distillation ► *Underline Video*

Jin Cui, Jiaqi Guo, Jiepeng Zhou, Ruixuan Yang, Jiayi Lu, Jiajun Xu, Jiangcheng Song, Boran Zhao and Pengju Ren

While Large Language Models (LLMs) have emerged with remarkable capabilities in complex tasks through Chain-of-Thought (CoT) reasoning, practical resource constraints have sparked interest in transferring these abilities to smaller models. However, achieving both domain performance and cross-domain generalization remains challenging. Existing approaches typically restrict students to following a single golden rationale and treat different reasoning paths independently. Due to distinct inductive biases and intrinsic preferences, alongside the student's evolving capacity and reasoning preferences during training, a teacher's "optimal" rationale could act as out-of-distribution noise. This misalignment leads to a degeneration of the student's latent reasoning distribution, causing suboptimal performance. To bridge this gap, we propose MIND, a capability-adaptive framework that transitions distillation from passive mimicry to active cognitive construction. We synthesize diverse teacher perspectives through a novel "Teaching Assistant" network. By employing a novel Feedback-Driven Inertia Calibration mechanism, this network utilizes inertia-filtered training loss to align supervision with the students current adaptability, effectively enhancing performance while mitigating catastrophic forgetting. Extensive experiments demonstrate that MIND achieves state-of-the-art performance on both in-distribution and out-of-distribution benchmarks, and our sophisticated latent space analysis further confirms the mechanism of reasoning ability internalization.

[Long] Miner: Mining Intrinsic Mastery for Data-Efficient RL in Large Reasoning Models ► *Underline Video*

Shuyang Jiang, Yuhao Wang, Ya Zhang, Yanfeng Wang and Yu Wang

Current critic-free RL methods for large reasoning models suffer from severe inefficiency when training on positive homogeneous prompts (where all rollouts are correct), resulting in waste of rollouts due to zero advantage estimates. We introduce a radically simple yet powerful solution to Mine intrinsic mastery (Miner), that repurposes the policy's intrinsic uncertainty as a self-supervised reward signal, with no external supervision, auxiliary models, or additional inference cost. Our method pioneers two key innovations: (1) a token-level focal credit assignment mechanism that dynamically amplifies gradients on critical uncertain tokens while suppressing overconfident ones, and (2) adaptive advantage calibration to seamlessly integrate intrinsic and verifiable rewards. Evaluated across six reasoning benchmarks on Qwen3-4B and Qwen3-8B base models, Miner achieves state-of-the-art performance among the other four algorithms, yielding up to 4.58 absolute gains in Pass@1 and 6.66 gains in Pass@K compared to GRPO. Comparison with other methods targeted at exploration enhancement further discloses the superiority of the two newly proposed innovations. This demonstrates that latent uncertainty exploitation is both necessary and sufficient for efficient and scalable RL training of reasoning models. Code is available at https://github.com/pixas/Miner.

[Long] MIRTH: Mutual-Information Reasoning with Temporal Hubs for Vision-Language-Action Agents ► *Underline Video*

Hao Sun, Yu Song, TENG SHIYU, Ziwei Niu and Yen-wei Chen

VLA models have emerged as a powerful paradigm for transferring semantic knowledge from web-scale data to physical robotic control. However, current single-frame architectures suffer from intrinsic limitations: temporal myopia that discards historical dynamics, reasoning gaps between high-level instructions and low-level motor commands, and inference inefficiency due to autoregressive scalar decoding. In this work, we propose MIRTH, a unified framework designed to address these challenges. MIRTH augments a pretrained VLA backbone with three key innovations: (1) dual-scale temporal memory hubs that compress long-term scene evolution and short-term motion trends into compact embeddings; (2) latent reasoning tokens optimized via a mutual-information objective carving out a semantic plan space to align multimodal context with action trajectories; and (3) a parallel action decoding scheme that replaces autoregressive generation with vector-wise prediction to maximize control throughput. Extensive evaluations on the LIBERO simulation benchmark and a real-world LeRobot platform demonstrate that MIRTH achieves state-of-the-art performance and exhibiting emergent error recovery capabilities. We will release our code and collected datasets to facilitate reproducible research in embodied AI upon publication.

[Long] Mitigating Legal Hallucinations via Symbolic Constraints and Analogical Precedents ► *Underline Video*

Zixuan Huang, Yanxiang Ma, Luhan Wang, Yunke Wang, Duo Shi and Chang Xu

With the growing potential of large language models (LLMs) in the legal domain, domain-specific finetuning and retrieval-augmented generation (RAG) methods have received widespread attention. However, current methods still suffer from hallucination risk and failing to resolve semantic drift and adapt to varying citation numbers. To address this, we propose Authoritative and Accurate Lawyer (AALawyer), a complementary dual-retriever framework based on the Legal Syllogism and the nature of different legal data.

First, we introduce Symbolic Constrained Retrieval (SCR) for closed-set article retrieval, by constraining retrieval to the generative prediction. Second, we build Analogical Precedent Retrieval (APR) to retrieve open-set judicial precedents for reasoning with a newly collected large criminal dataset. Extensive experiments, including LawBench, our Hallucination Risk-Benchmark, and comprehensive ablation studies, demonstrate the effectiveness of AALawyer, which mitigates hallucinations while improving the explainability of legal reasoning.

[Long] Mitigating Selection Bias in Large Language Models via Permutation-Aware GRPO ▶ *Underline Video*

Jinquan Zheng, Jia Yuan, Jiacheng Yao, Chenyang Gu, Pujun Zheng and Guoxiu He

Large language models (LLMs) used for multiple-choice and pairwise evaluation tasks often exhibit selection bias due to non-semantic factors like option positions and label symbols. Existing inference-time debiasing is costly and may harm reasoning, while pointwise training ignores that the same question should yield consistent answers across permutations. To address this issue, we propose Permutation-Aware Group Relative Policy Optimization (PA-GRPO), which mitigates selection bias by enforcing permutation-consistent semantic reasoning. PA-GRPO constructs a permutation group for each instance by generating multiple candidate permutations, and optimizes the model using two complementary mechanisms: (1) cross-permutation advantage, which computes advantages relative to the mean reward over all permutations of the same instance, and (2) consistency-aware reward, which encourages the model to produce consistent decisions across different permutations. Experimental results demonstrate that PA-GRPO outperforms strong baselines across seven benchmarks, substantially reducing selection bias while maintaining high overall performance. The code is available on github (https://github.com/ECNU-Text-Computing/PA-GRPO).

[Long] MMSciCode: Real-world Evaluation of Multilingual Multi-Discipline Scientific Research Coding ▶ *Underline Video*

Xue Xia, Zheyuan Yang, Arman Cohan and Yilun Zhao

We introduce MMSciCode, a comprehensive expert-level, multilingual multi-discipline benchmark for evaluating foundation models in scientific code generation. It includes 624 expert-annotated research coding problems spanning six core scientific disciplines. Compared to prior benchmarks, MMSciCode features three key advancements. First, it challenges models to integrate domain-specific knowledge with algorithmic reasoning to implement core functions from research papers, moving beyond the isolated, general-purpose coding tasks typically assessed in current benchmarks. Second, each problem is meticulously annotated by domain experts through a rigorous paper-grounded process, with strict quality controls implemented to ensure dataset integrity and authenticity. Finally, each problem is equipped with comprehensive unit test suites and containerized environments, enabling reproducible and diagnostic evaluation of both functional correctness and domain validity. We conduct an extensive evaluation of 28 state-of-the-art foundation models and 2 agentic coding tools on MMSciCode. Our results reveal that even the best non-agentic model achieves only around 15% accuracy, while the top agentic coding tool reaches 32.2%, both still far below human expert performance of 68.8%. Through comprehensive error analyses and case studies, we identify substantial performance gaps between models and human experts, providing actionable insights for advancing expert-level scientific code generation.

[Long] MobileWorld: Benchmarking Autonomous Mobile Agents in Agent-User Interactive and MCP-Augmented Environments ▶ *Underline Video*

Quyu Kong, Xu Zhang, Zhenyu Yang, Nolan Gao, Chen Liu, Panrong Tong, Chenglin Cai, Hanzhang Zhou, Jianan Zhang, Liangyu Chen, Zhidan Liu, Steven Hoi and Yue Wang

While AndroidWorld has become the dominant mobile-use benchmark due to its reproducible environment and deterministic evaluation, recent agents achieving over 90% success rates indicate saturation and motivate the need for greater challenge. In addition, its environment lacks key application categories, such as e-commerce and enterprise communication, and does not reflect realistic mobile-use scenarios characterized by vague user instructions and hybrid tool usage. We introduce MobileWorld, a substantially more challenging benchmark with 201 tasks across 20 applications that reflects real-world usage through long-horizon, cross-application workflows requiring nearly twice as many steps (27.8 vs. 14.3) and featuring significantly more multi-app tasks (62.2% vs. 9.5%) than AndroidWorld. MobileWorld balances production-grade utility and reproducible evaluation using open-source alternatives to industry standards (e.g., Mattermost for Slack), enabling full observability through source code modification and direct database access. Beyond standard GUI manipulation, MobileWorld introduces novel task categories including agent-user interaction and Model Context Protocol (MCP)-augmented tasks for evaluating agents in user-aware, hybrid-tool scenarios. We develop a planner-executor framework with extended action spaces supporting user interactions and MCP calls. Results show a sharp performance drop from AndroidWorld, with the best agentic framework and end-to-end model achieving 51.7% and 20.9% success rates, respectively, highlighting substantial room for future research.

[Long] Modeling Human-Like Cognition for Stance Detection: Integrating Intuitive Judgment and Analytical Reasoning ▶ *Underline Video*

Zhaodan Zhang, Jin Zhang, Jiafeng Guo and Xueqi Cheng

Stance detection aims to identify the attitude expressed in text towards a given target, with applications in public opinion analysis and misinformation mitigation. Despite recent advances in large language models (LLMs), two key challenges remain: (1) spurious correlations between superficial features and stance labels, and (2) lack of cognitive modeling that simulates the transition from intuitive perception to deliberate reasoning. To address these issues, we propose Cognitive-Driven Stance Detection (CDSD), inspired by Kahnemans Dual-Process Theory. CDSD integrates fast intuitive judgment (System 1) and analytical reasoning (System 2), enhanced by three key modules: attention-based cognitive alignment to compare system focus, uncertainty-aware belief update using Bayesian inference, and self-doubt-triggered counterfactual reasoning for re-evaluation under low consistency or high uncertainty.

Experimental results on SEM16, P-Stance, and VAST show that CDSD outperforms state-of-the-art methods across multiple LLMs. Notably, CDSD exhibits strong robustness against textual perturbations such as emotional word removal and rhetorical restructuring. By integrating cognitive theory with NLP, our work provides a promising path toward more reliable and interpretable stance detection systems.

[Industry] Modular Monolingual Adaptation using Pretrained Language Models ▶ *Underline Video*

Nalin Kumar and Ondrej Dusek

Building monolingual language models (LMs) for low-resource languages typically relies on adapting pretrained language models (PLMs) by finetuning the whole model on the target language. This approach is widely favored over training from scratch, as it enables effective knowledge transfer. Additionally, prior work has shown that using a language-specific tokenizer can enhance the adaptability. In this work, we hypothesize that full model tuning is often unnecessary and propose a more modular approach. Specifically, we replace the tokens, freeze the corresponding embeddings, and tune the rest of the model. We use Scottish Gaelic, Irish, and Quechua for our experiments, with Quechua being a very low-resource language (8.5k training instances). Evaluation on natural language understanding (NLU) tasks – mask-filling, NER, and POS – shows that our proposed approach improves performance when adapting the models to low-resource languages. Additionally, we provide a comprehensive analysis of the effectiveness of training strategies, the choice of pretrained embeddings, and models.

[Long] More Than Meets the Eye: Measuring the Semiotic Gap in Vision-Language Models via Semantic Anchorage ▶ *Underline Video*

Wei He

Vision-Language Models (VLMs) excel at photorealistic generation, yet often struggle to represent abstract meaning such as idiomatic interpretations of noun compounds. To study whether high visual fidelity interferes with idiomatic compositionality under visual abstraction, we introduce DIVA, a controlled benchmark that replaces high-fidelity visual detail with schematic iconicity by generating paired, sense-anchored visualizations for literal and idiomatic readings. We further propose Semantic Alignment Gap (), an architecture-agnostic metric that quantifies divergence between literal and idiomatic visual grounding. We additionally introduce a directional signed bias b(t) to separately measure the direction and strength of literal preference. Evaluating 8 recent VLMs, we reveal a consistent Literal Superiority Bias: model scale alone does not resolve literal preference, and increased visual fidelity is associated with weaker symbolic alignment, suggesting cognitive interference from hyper-realistic imagery. Our findings suggest that improving compositional understanding requires iconographic abstraction of visual input and anchoring interpretation and generation in intended meaning.

[Long] MoRI: Learning Motivation-Grounded Reasoning for Scientific Ideation in Large Language Models ▶ *Underline Video*

Chenyang Gu, Jiahao Cheng, Meicong Zhang, Pujun Zheng, Jinquan Zheng and Guoxiu He

Scientific ideation aims to propose novel solutions within a given scientific context. Existing LLM-based agentic approaches emulate human research workflows, yet inadequately model scientific reasoning, resulting in surface-level conceptual recombinations that lack technical depth and scientific grounding. To address this issue, we propose MoRI (Motivation-grounded Reasoning for Scientific Ideation), a framework that enables LLMs to explicitly learn the reasoning process from research motivations to methodologies. The base LLM is initialized via supervised fine-tuning to generate a research motivation from a given context, and is subsequently trained under a composite reinforcement learning reward that approximates scientific rigor: (1) entropy-aware information gain encourages the model to uncover and elaborate high-complexity technical details grounded in ground-truth methodologies, and (2) contrastive semantic gain constrains the reasoning trajectory to remain conceptually aligned with scientifically valid solutions. Empirical results show that MoRI consistently outperforms strong commercial LLMs and complex agentic baselines across multiple dimensions, including novelty, technical rigor, and feasibility. The code is available on GitHub.

[SRW] Morphology-Aware Multi-Granularity Representation Learning for Agglutinative Languages ▶ *Underline Video*

zhonghao zhang; NA LIU; Jiajia Ma; Nier Wu; Guiping Liu

Low-resource agglutinative languages, characterized by rich morphological inflection and severe vocabulary sparsity in corpora, have long posed numerous challenges in the field of representation learning. Word-level representations preserve semantic integrity but struggle to handle sparse surface forms, whereas morpheme-level representations, though easier to learn, often lack holistic semantic information. Existing multi-granularity methods are typically modeled at the word and phrase levels, with very limited application to low-resource agglutinative languages. Focusing on the morphemes of agglutinative languages, this paper proposes MAGNet, a morphology-aware gated multi-granularity pre-training framework. At the morpheme granularity, this framework leverages morphological knowledge and integrates morpheme segmentation with morphological tagging to construct fine-grained representations. It further introduces a morphology-aware masked language modeling objective to facilitate the model in learning functional morphological regularities. Meanwhile, at the word granularity, a word-level encoder is employed to capture contextual semantics and maintain its semantic coherence.Finally, a gated fusion mechanism dynamically fuses representations of different granularities according to the context. Experiments conducted on two low-resource agglutinative languages, Mongolian and Turkish, for the tasks of dependency parsing and named entity recognition (NER) demonstrate that our method achieves consistent performance improvements over strong baseline models. Ablation studies further validate the complementary roles of morphological tagging and whole-word modeling in efficient representation learning.

[Long] MTA: Multi-Granular Trajectory Alignment for Large Language Model Distillation ▶ *Underline Video*

Pham Khanh Chi, Quoc Phong Dao, Thuat Nguyen, Linh Ngo Van, Trung Le and Thanh Hong Nguyen

Knowledge distillation is a key technique for compressing large language models (LLMs), but most existing methods align representations

at fixed layers or token-level outputs, ignoring how representations evolve across depth. As a result, the student is only weakly guided to capture the teachers internal relational structure during distillation, which limits knowledge transfer. To address this limitation, we propose Multi-Granular Trajectory Alignment (MTA), a framework that aligns teacher and student representations along their layer-wise transformation trajectory. MTA adopts a layer-adaptive strategy: lower layers are aligned at the word level to preserve lexical information, while higher layers operate on phrase-level spans (e.g., noun and verb phrases) to capture compositional semantics. We instantiate this idea through a Dynamic Structural Alignment loss that matches the relative geometry among semantic units within each layer. This design is motivated by empirical findings that Transformer representations become increasingly abstract with depth, and is also consistent with linguistic views in which higher-level meaning emerges through the composition of lower-level lexical units. We further incorporate a Hidden Representation Alignment loss to directly align selected teacher-student layers. Experiments show that MTA consistently outperforms state-of-the-art baselines on standard benchmarks, with ablations confirming the contribution of each component.

[Long] Multi-Granularity Semantic Revision for Large Language Model Distillation ► *Underline Video*

Xiaoyu Liu, Yun Zhang, Wei Li, Simiao Li, Xudong Huang, Hanting Chen, Yehui Tang, Jie Hu, Zhiwei Xiong and Yunhe Wang

Knowledge distillation is crucial for compressing Large Language Models (LLMs), enabling smaller student models to learn from larger teacher models. However, existing LLM distillation methods overly rely on student-generated outputs, which may introduce generation errors and misguide the distillation process. Moreover, existing distillation loss functions struggle to align the most informative part due to the complex output distributions of LLMs. To address these problems, we propose a multi-granularity semantic revision method for LLM distillation. At the sequence level, we propose a sequence correction and re-generation (SCRG) strategy. SCRG identifies error tokens by calculating the semantic cognitive difference between teacher and student outputs, corrects them using teacher-generated tokens, and re-generates the sequence to minimize errors. At the token level, we design a distribution adaptive clipping Kullback-Leibler (DAC-KL) loss, which uses a learnable sub-network to focus on semantically dense areas of the teacher's output, reducing the impact of redundant information. At the span level, we utilize span priors to compute probability correlations within sequences, ensuring consistency between teacher and student outputs to enhance semantic information transfer. Extensive experiments on models ranging from 0.1B to 13B parameters demonstrate the effectiveness of our approach compared to existing methods.

[Industry] Natural-Language Policies to Executable Decisions: An Interpretable Large Language Model Framework ► *Underline Video*

Ziqiang Zhang, Jing Ma, Zilong Wang, Jiayuan Chen, Yi Qiao, Dai Cheng, Yu He, Wei Zhang and Xiaoyu Shen

Pricing automation in large-scale tourism is challenging due to informal orders and rapidly evolving, open-ended policies. Rule engines are brittle, while unconstrained LLM agents lack the reliability and auditability required for financial decisions. We present a production LLM-powered pricing system with a strict decision boundary: LLMs perform structured extraction and bounded policy/path selection, while all numeric pricing (including total-price computation) is executed deterministically. Policies are compiled into interpretable condition trees, exposing auditable artifacts and supporting human-in-the-loop control; tree induction and path matching are improved via periodic fine-tuning on logged traces. Deployed at a municipal state-owned tourism enterprise across 7 scenic sites and 12 business categories with 1,500+ operators, the system processed 3,960 orders in six months, reduced the order management team from 15-20 to 3, and cut per-order handling time from ~10 minutes to < 2 minutes.

[Long] NaturalGAIA: A Verifiable Benchmark and Hierarchical Framework for Long-Horizon GUI Tasks ► *Underline Video*

Zihan Zheng, Tianle Cui, Taoran Wang, Fengtao Wang, Jiahui Pan, Lewei He and Qianglong Chen

Despite significant advances in LLM-driven GUI agents, the field remains constrained by the challenge of reconciling high-fidelity realism with verifiable evaluation accuracy. To address this, we introduce NaturalGAIA, a verifiable evaluation dataset grounded in real-world human GUI interaction intents. By decoupling logical causal pathways from linguistic narratives, it rigorously simulates natural human intent, characterized by cognitive non-linearity and contextual dependencies. Furthermore, we propose LightManus-Jarvis, a hierarchical collaborative framework where LightManus manages dynamic topological planning and context evolution, while Jarvis ensures execution precision via hybrid visual-structural perception. Experiments demonstrate that our approach achieves a Weighted Pathway Success Rate of 45.6%, significantly outperforming the state-of-the-art baseline (21.1%), while reducing token consumption by 75% and execution time by 76%. These results validate the efficacy of the macro-planning and micro-execution paradigm in handling complex naturalized tasks. Our code is publicly available at: https://anonymous.4open.science/r/NatureGAIA-721F/.

[Long] NaviMaster: Learning a Unified Policy for GUI and Embodied Navigation Tasks ► *Underline Video*

Zhihao Luo, Wentao Yan, Jingyu Gong, Min Wang, Zhizhong Zhang, Xuhong Wang, Yuan Xie and Xin Tan

Recent advances in Graphical User Interface (GUI) and embodied navigation have driven progress, yet these domains have largely evolved in isolation, with disparate datasets and training paradigms. In this paper, we observe that both tasks can be formulated as Markov Decision Processes (MDP), suggesting a foundational principle for their unification. Hence, we present NaviMaster, the first unified agent capable of unifying GUI navigation and embodied navigation within a single framework. Specifically, NaviMaster (i) proposes a visual-target trajectory collection pipeline that generates trajectories for both GUI and embodied tasks using a single formulation. (ii) employs a unified reinforcement learning framework on the mix data to improve generalization. (iii) designs a novel distance-aware reward to ensure efficient learning from the trajectories. Through extensive experiments on out-of-domain benchmarks, NaviMaster is shown to outperform state-of-the-art agents in GUI navigation, spatial affordance prediction, and embodied navigation. Ablation studies further

demonstrate the efficacy of our unified training strategy, data mixing strategy, and reward design. Resources will be released to the community.

[Long] NeuralFSM: Adaptive Multi-Agent Coordination via Learning Finite-State Execution Policy ▶ *Underline Video*
Jiye Wang, Yu Wang, Jianbin Li, Shiduo Yang, Kenan Guo and Yuanhe Zhao
LLM-powered multi-agent systems (MAS) have demonstrated strong performance on complex tasks. However, most existing approaches still rely on hand-crafted communication protocols or automatically designed communication topologies, which generalize poorly across tasks. We introduce NeuralFSM, a state-driven framework that formulates multi-agent problem solving as a finite-state execution process. NeuralFSM learns both the state transition distribution and inter-agent communication weights from interaction traces using a Temporal Coordination Controller. Rather than prioritizing explicit structure generation, the proposed framework uses task context to modulate transition and routing decisions, enabling flexible coordination without manual protocol design. To improve robustness against noisy or adversarial agents, we incorporate graph regularization during training and apply trust-aware message attenuation at runtime. Experiments on diverse benchmarks show that NeuralFSM consistently outperforms prior baselines by an average margin of 6.74%–19.39%, while substantially reducing token consumption. Moreover, NeuralFSM exhibits strong inherent robustness, which is further enhanced by the protection layer, resulting in only a 1.82% performance drop under attack.

[Long] New Terms, New Toxicity: Consensus-based Chinese Neologism Toxicity Detection via Search-Augmented LLMs ▶ *Underline Video*
Shiyao Cui, QingLin Zhang, Di Wang, Yida Lu, Zhexin Zhang, Jinhua Gao, Jinglin Yang, Min He, Han Qiu and Minlie Huang
Neologisms, emerging terms in meaning or form, can serve as new vehicles for toxic expression, like "" ("country girl") as a stigmatizing label targeting feminism. Such toxic neologisms appear benign but have evolved into toxic usage in public consensus, posing challenges to moderation systems and remaining underexplored. In this paper, we investigate how to detect implicit toxicity expressed via neologisms. We first propose a taxonomy that captures the origins and consensus-verification criteria of toxic neologisms, followed by the construction of a lexicon spanning widely observed risk categories. To capture toxicity grounded in public consensus, we introduce SeTox, a search-augmented framework that enables static large language models (LLMs) to incorporate real-time web context for neologism toxicity detection. Experiments show that SeTox, even with 3B-scale models, outperforms recent large-scale models, demonstrating its scalability to incorporate real-world knowledge for toxic neologism detection. Disclaimer: this paper has offensive contents that may be disturbing to some readers.

[Long] Nirvana: A Specialized Generalist Model With Task-Aware Memory Mechanism ▶ *Underline Video*
Yuhua Jiang, Shuang Cheng, Yihao Liu, Ermo Hua, Che Jiang, Weigao Sun, Yu Cheng, Feifei Gao, Biqing Qi and Bowen Zhou
Large Language Models (LLMs) excel at general language tasks but struggle in specialized domains. Specialized Generalist Models (SGMs) address this by preserving broad capabilities while adapting to target domains. However, existing architectures provide limited support for task-guided specialized memory mechanisms. In this work, we introduce Nirvana, an SGM featuring specialized memory, linear-time complexity, and test-time task information extraction. Central to Nirvana are: (1) Task-Aware Memory Trigger (Trigger), which treats each input as a self-supervised fine-tuning task and adjusts task-related parameters on the fly; and (2) Specialized Memory Updater (Updater), which dynamically consolidates task-relevant context. Nirvana matches or surpasses LLM baselines on general benchmarks and achieves the lowest perplexity across specialized domains including biomedicine, finance, and law. On the challenging task of Magnetic Resonance Imaging (MRI), we attach lightweight codecs to the frozen Nirvana backbone and fine-tune them on paired k-space signals and images. Nirvana achieves higher-fidelity reconstructions than conventional LLM-based models, with Trigger providing effective domain-specific adaptation. Ablation studies confirm that removing Trigger leads to substantial degradation across all tasks, underscoring its essential role in task-aware specialization. Models are available at https://huggingface.co/collections/YuhuaJiang/nirvana. Code is available at https://github.com/YuhuaJiang2002/Nirvana.

[Long] NiuTrans.LMT: Toward Inclusive and Scalable Multilingual Machine Translation with LLMs ▶ *Underline Video*
Yingfeng Luo, Ziqiang Xu, Yuxuan Ouyang, MuRun Yang, DingYang Lin, Kaiyan Chang, Tong Zheng, Bei Li, Peinan Feng, Quan Du, Tong Xiao and JingBo Zhu
Large language models have significantly advanced Multilingual Machine Translation (MMT), yet scaling to many languages while keeping quality robust across directions remains challenging. In this paper, we identify a failure mode of multilingual supervised fine-tuning (SFT) on multi-way parallel data: when such data are reused symmetrically around a pivot language (e.g., English), performance on reverse directions (X -> pivot) can drop substantially. We term this phenomenon Directional Degeneration and attribute it to excessive many-to-one mappings, which encourage shortcut learning. We propose Strategic Downsampling (SD), a simple yet effective method to mitigate this degeneration. In addition, we introduce Parallel Multilingual Prompting (PMP), which augments translation instructions with an auxiliary parallel sentence to promote cross-lingual transfer during training and enables optional test-time enhancement when auxiliary translations are available. We further develop NiuTrans.LMT (Large-scale Multilingual Translation, abbreviated as LMT), a Chinese-English-centric suite of multilingual translation models spanning four sizes (0.6B/1.7B/4B/8B) and covering 60 languages and 234 directions. Comprehensive evaluations show that LMT is competitive among open-source MMT systems, and that our 4B LMT model performs on par with or better than substantially larger baselines. We release our models and project resources to support inclusive and scalable MMT.

[Short] On the Rejection Criterion for Proxy-based Test-time Alignment ▶ *Underline Video*
Ayoub Hammal, Pierre Zweigenbaum and Caio Corro

Recent works proposed test-time alignment methods that rely on a small aligned model as a proxy that guides the generation of a larger base (unaligned) model. The implicit reward approach skews the large model distribution, whereas the nudging approach defers the generation of the next token to the small aligned model when the large base one is unconfident about its outcome. In this work, we first show that both approaches can be reduced to sampling from similar graphical models, where they differ only in the definition of a rejection criterion (or distribution). Moreover, we argue that the confidence criterion is ill-motivated due to linguistic phenomena like ambiguous phrasing. We propose a novel rejection criterion based on a conservative confidence bet. Experimentally, our novel approach outperforms previous work on several datasets.

[Long] OneRec-Think: In-Text Reasoning for Generative Recommendation ▶ *Underline Video*

Zhanyu Liu, Shiyao Wang, Xingmei Wang, Rongzhou Zhang, Jiaxin Deng, Honghui Bao, Jinghao Zhang, Wuchao Li, Pengfei Zheng, Xiangyu Wu, Yifei Hu, Qigen Hu, Xinchen Luo, Lejian Ren, Zixing Zhang, Qianqian Wang, Kuo Cai, Yunfan Wu, Hongtao Cheng, Zexuan Cheng, Lu Ren, Huanjie Wang, Yi Su, Ruiming Tang, Kun Gai and Guorui Zhou

The powerful generative capacity of Large Language Models (LLMs) has instigated a paradigm shift in recommendation. However, existing generative models (e.g., OneRec) operate as implicit predictors, critically lacking the capacity for explicit and controllable reasoning, a key advantage of LLMs. To bridge this gap, we propose OneRec-Think, a unified framework that seamlessly integrates dialogue, reasoning, and personalized recommendation. OneRec-Think incorporates: (1) Itemic Alignment: cross-modal Item-Textual Alignment for semantic grounding; (2) Reasoning Activation: Reasoning Scaffolding to activate LLM reasoning within the recommendation context; and (3) Reasoning Enhancement, where we design a recommendation-specific reward function that accounts for the multi-validity nature of user preferences. Experiments across public benchmarks show state-of-the-art performance. Moreover, our proposed "Think-Ahead" architecture enables effective industrial deployment, achieving a 0.159% gain in APP Stay Time and validating the practical efficacy of the model's explicit reasoning capability.

[Long] OptiCo: Adaptive Distributed Training Optimization via Collaborative Agent Reasoning ▶ *Underline Video*

Sheng Chen, Zhe Tang, Weixing Zhang, Fei Yang, Yuanyuan Wang, Tianlin Li and Yang Liu

Optimizing distributed training strategies for large-scale deep learning models remains a critical challenge in both industry and academia, demanding extensive domain expertise and manual tuning. Existing automated distributed training frameworks are plagued by over-reliance on prior profiling, poor generalization across models/hardware, and scalability constraints stemming from vast search spaces, impeding real-world applicability. To address these challenges, we propose OptiCo, a model-driven multi-agent framework that leverages Large Language Models (LLMs) to enable automatic and explainable distributed training strategy configuration. OptiCo orchestrates a team of reasoning-driven agents, through a shared Global Message Pool facilitating persistent memory and coordination. By employing inception prompting and Chain-Of-Thought (COT) reasoning, agents iteratively refine configurations, detect bottlenecks, analyze failures, and optimize resource utilization. Evaluated across 25+ configurations spanning diverse model architectures, GPU types and scales, OptiCo outperforms expert-designed strategies within 20 iterations, achieving an average performance improvement of 1.84%, with gains ranging from 0.08% to 8.65%. The source codes are avaiable at https://github.com/TangZhe96/OptiCo-public.

[Long] Optimizing RAG Rerankers with LLM Feedback via Reinforcement Learning ▶ *Underline Video*

Yuhang Wu, Xiangqing Shen, Fanfan Wang, Cangqi Zhou, Zhen Wu, Xinyu Dai and Rui Xia

Rerankers play a pivotal role in refining retrieval results for Retrieval-Augmented Generation. However, current reranking models are typically optimized on static human annotated relevance labels in isolation, decoupled from the downstream generation process. This isolation leads to a fundamental misalignment: documents identified as topically relevant by information retrieval metrics often fail to provide the actual utility required by the LLM for precise answer generation. To bridge this gap, we introduce ReRanking Preference Optimization (RRPO), a reinforcement learning framework that directly aligns reranking with the LLM's generation quality. By formulating reranking as a sequential decision-making process, RRPO optimizes for context utility using LLM feedback, thereby eliminating the need for expensive human annotations. To ensure training stability, we further introduce a reference-anchored deterministic baseline. Extensive experiments on knowledge-intensive benchmarks demonstrate that RRPO significantly outperforms strong baselines, including the powerful list-wise reranker RankZephyr. Further analysis highlights the versatility of our framework: it generalizes seamlessly to diverse readers (e.g., GPT-4o), integrates orthogonally with query expansion modules like Query2Doc, and remains robust even when trained with noisy supervisors.

[Long] OSCBench: Benchmarking Object State Change in Text-to-Video Generation ▶ *Underline Video*

Xianjing Han, Bin Zhu, Shiqi Hu, Franklin Mingzhe Li, Patrick Carrington, Roger Zimmermann and Jingjing Chen

Text-to-video (T2V) generation models have made rapid progress in producing visually high-quality and temporally coherent videos. However, existing benchmarks primarily focus on perceptual quality, text-video alignment, or physical plausibility, leaving a critical aspect of action understanding largely unexplored: object state change (OSC) explicitly specified in the text prompt. OSC refers to the transformation of an objects state induced by an action, such as peeling a potato or slicing a lemon. In this paper, we introduce OSCBench, a benchmark specifically designed to assess OSC performance in T2V models. OSCBench is constructed from instructional cooking data and systematically organizes action-object interactions into regular, novel, and compositional scenarios to probe both in-distribution performance and generalization. We evaluate six representative open-source and proprietary T2V models using both human user study and multimodal large language model (MLLM)-based automatic evaluation. Our results show that, despite strong performance on semantic and scene alignment, current T2V models consistently struggle with accurate and temporally consistent object state changes, especially in novel and

compositional settings. These findings position OSC as a key bottleneck in text-to-video generation and establish OSCBench as a diagnostic benchmark for advancing state-aware video generation models. Project page: https://hanxjing.github.io/OSCBench.

[Long] Par-ITA: Benchmarking Seq2Seq and LLMs on a Human-Supervised Parallel Corpus for Italian Hyperpartisan Neutralization ▶ *Underline Video*

Michele Joshua Maggini, Søren Fomsgaard, Michele Maestroni, Gaël Dias and Pablo Gamallo

Neutralizing hyperpartisan content is essential for mitigating online polarization, yet research has largely focused on English. We present Par-ITA, a curated subset from Semeval 2023 task 3, consisting in the first human-supervised parallel corpus for Italian hyperpartisan neutralization of 2,475 paragraph pairs. The dataset is constructed using a rigorous three-stage pipeline: (1) expert-led preliminary selection of LLMs for high-quality generation, (2) human-supervised data production with high editing rates (32-68%), and (3) post-hoc human validation. We establish extensive benchmarks for this task across seq2seq and decoder-only architectures, evaluating standard fine-tuning, Direct Preference Optimization (DPO), and in-context learning. Our analysis highlights that while DPO effectively maximizes neutrality scores in seq2seq models, automated evaluators like GPT-4o-mini exhibit systematic biases, specifically over-penalizing sensitive political topics compared to human experts. Par-ITA provides a foundational resource for non-English neutralization and a reproducible framework for developing high-quality datasets in subjective domains.

[Long] Parallel Test-Time Scaling for Latent Reasoning Models ▶ *Underline Video*

Runyang You, Yongqi Li, Meng Liu, Wenjie Wang, Liqiang Nie and Wenjie Li

Parallel test-time scaling (TTS) is a pivotal approach for enhancing large language models (LLMs), typically by sampling multiple token-based chains-of-thought in parallel and aggregating outcomes through voting or search. Recent advances in latent reasoning, where intermediate reasoning unfolds in continuous vector spaces, offer a more efficient alternative to explicit Chain-of-Thought, yet whether such latent models can similarly benefit from parallel TTS remains open, mainly due to the absence of sampling mechanisms in continuous space, and the lack of probabilistic signals for advanced trajectory aggregation. This work enables parallel TTS for latent reasoning models by addressing the above issues. For sampling, we introduce two uncertainty-inspired stochastic strategies: Monte Carlo Dropout and Additive Gaussian Noise. For aggregation, we design a Latent Reward Model (LatentRM) trained with step-wise contrastive objective to score and guide latent reasoning. Extensive experiments and visualization analyses show that both sampling strategies scale effectively with compute and exhibit distinct exploration dynamics, while LatentRM enables effective trajectory selection. Together, our explorations open a new direction for scalable inference in continuous spaces. Code and checkpoint are included as supplementary materials. GitHub Project: https://github.com/ModalityDance/LatentTTS

[Long] PASs-MoE: Mitigating Misaligned Co-drift among Router and Experts via Pathway Activation Subspaces for Continual Learning ▶ *Underline Video*

Zhiyan Hou, Haiyun Guo, Haokai Ma, Yandu Sun, Yonghui Yang and Jinqiao Wang

Continual instruction tuning (CIT) requires multimodal large language models (MLLMs) to adapt to a stream of tasks without forgetting prior capabilities. A common strategy is to isolate updates by routing inputs to different LoRA experts. However, existing LoRA-based Mixture-of-Experts (MoE) methods often jointly update the router and experts in an indiscriminate way, causing the routers preferences to co-drift with experts adaptation pathways and gradually deviate from early-stage input-expert specialization. We term this as Misaligned Co-drift, which blurs expert responsibilities and exacerbates forgetting. To address this, we introduce the pathway activation subspace (PASs), a LoRA-induced subspace that reflects which low-rank pathway directions an input activates in each expert, providing a capability-aligned coordinate system for routing and preservation. Based on PASs, we propose a fixed-capacity PASs-based MoE-LoRA method with two components: PAS-guided Reweighting, which calibrates routing using each experts pathway activation signals, and PAS-aware Rank Stabilization, which selectively stabilizes rank directions important to previous tasks. Experiments on a CIT benchmark show that our approach consistently outperforms a range of conventional continual learning baselines and MoE-LoRA variants in both accuracy and resistance to forgetting, without increasing model parameters. Our code is publicly available at https://github.com/yueluoshuangtian/PASs-MoE.

[Long] Please refuse to answer me! Mitigating Over-Refusal in Large Language Models via Adaptive Contrastive Decoding ▶ *Underline Video*

Yupeng Qi, Ziyu Lyu, Lixin Cui, Lu Bai and Feng Xia

Safety-aligned large language models (LLMs) often generate refusal responses to harmless queries due to the over-refusal problem. However, existing methods for mitigating over-refusal cannot maintain a low refusal ratio for harmless queries while keeping a high refusal ratio for malicious ones. In this paper, we analyze how system prompts with varying safety levels affect LLM refusal behaviors when facing over-refusal queries. A key observation is that, when LLMs suffer from the over-refusal issue, non-refusal tokens remain present in the next-token candidate list, but the model systematically fails to select them, despite the generation of refusal tokens. Based on this observation, we propose a training-free and model-agnostic approach, Adaptive Contrastive Decoding (AdaCD), to mitigate over-refusal while maintaining LLM safety. First, AdaCD compares the output distributions of the LLM with or without an extreme safety system prompt to refine the refusal token distribution. Second, we introduce an adaptive contrastive decoding strategy that dynamically incorporates or removes the refusal token distribution, adaptively boosting the probability of selecting refusal or non-refusal tokens. Experimental results on five benchmark datasets show that, on average, AdaCD reduces the refusal ratio for over-refusal queries by 10.35%, yet still increases the refusal ratio for malicious queries by 0.13%. Code is available at https://shorturl.at/Z31Oe.

[Findings] Position: From Noise to Signal to Selbstzweck - Reframing Human Label Variation in the Era of Post-training in NLP ▶ *Underline Video*

Shanshan Xu, Santosh T.Y.S.S and Barbara Plank

Human Label Variation (HLV) refers to legitimate disagreement in annotation that reflects the diversity of human perspectives rather than mere error. Long treated in NLP as noise to be eliminated, HLV has only recently been reframed as a signal for improving model robustness. With the rise of large language models (LLMs) and post-training methods such as human feedback-based alignment, the role of HLV has become increasingly consequential. Yet current preference-learning datasets routinely collapse multiple annotations into a single label, flattening diverse perspectives into artificial consensus. Preserving HLV is necessary not only for pluralistic alignment but also for sociotechnical safety evaluation, where model behavior must be assessed in relation to human interaction and societal context. This position paper argues that preserving HLV as an embodiment of human pluralism must be treated as a Selbstzweck, an intrinsic value in itself. We analyze the limitations of existing preference datasets and propose actionable strategies for incorporating HLV into dataset construction to better preserve pluralistic human values.

[Findings] Preference Estimation via Opponent Modeling in Multi-Agent Negotiation ▶ *Underline Video*

Yuta Konishi, Kento Yamamoto, Eisuke Sonomoto, Rikuho Takeda, Ryo Furukawa, Yusuke Muraki, Takafumi Shimizu, Kazuma Fukumura, Yuya Kanemoto, Takayuki Ito and Shiyao Ding

Automated negotiation in complex, multi-party and multi-issue settings critically depends on accurate opponent modeling. However, conventional numerical-only approaches fail to capture the qualitative information embedded in natural language interactions, resulting in unstable and incomplete preference estimation. Although Large Language Models (LLMs) enable rich semantic understanding of utterances, it remains challenging to quantitatively incorporate such information into a consistent opponent modeling. To tackle this issue, we propose a novel preference estimation method integrating natural language information into a structured Bayesian opponent modeling framework. Our approach leverages LLMs to extract qualitative cues from utterances and converts them into probabilistic formats for dynamic belief tracking. Experimental results on a multi-party benchmark demonstrate that our framework improves the full agreement rate and preference estimation accuracy by integrating probabilistic reasoning with natural language understanding.

[Findings] PrinciplismQA: A Philosophy-Grounded Approach to Assessing LLM-Human Clinical Medical Ethics Alignment ▶ *Underline Video*

Chang Hong, Minghao Wu, Qingying Xiao, Yuchi Wang, Xiang Wan, Guangjun Yu, Benyou Wang and Yan Hu

As medical LLMs transition to clinical deployment, assessing their ethical reasoning capability becomes critical. While achieving high accuracy on knowledge benchmarks, LLMs lack validated assessment for navigating ethical trade-offs in clinical decision-making where multiple valid solutions exist. Existing benchmarks lack systematic approaches to incorporate recognized philosophical frameworks and expert validation for ethical reasoning assessment. We introduce PrinciplismQA, a philosophy-grounded approach to assessing LLM clinical medical ethics alignment. Grounded in Principlism, our approach provides a systematic methodology for incorporating clinical ethics philosophy into LLM assessment design. PrinciplismQA comprises 3,648 expert-validated questions spanning knowledge assessment and clinical reasoning. Our expert-calibrated pipeline enables reproducible evaluation and models ethical biases. Evaluating recent models reveals significant ethical reasoning gaps despite high knowledge accuracy, demonstrating that knowledge-oriented training does not ensure clinical ethical alignment. PrinciplismQA provides a validated tool for assessing clinical AI deployment readiness.

[Findings] Privacy Risks of Intermediate Representations: Attribute Inference in Distributed LLM Inference ▶ *Underline Video*

YANG LYU, Jin CAO, Yang Xiao, Zhe Sun, Ben Niu, Fenghua Li and HUI LI

Distributed LLM inference avoids sending raw inputs by transmitting intermediate hidden states, a practice widely assumed to preserve privacy. We challenge this assumption and demonstrate that intermediate representations alone are sufficient to leak sensitive user attributes. This setting poses a fundamental obstacle for existing attribute inference attacks, which typically rely on auxiliary embedding-attribute pairs. To characterize this previously underexplored privacy risk, we reformulate attribute inference as zero-shot matching over candidate attributes directly in the intermediate representation space, and introduce a purely intermediate-representation-based attribute inference attack, termed IR-AIA. To address two structural challenges that hinder attribute inference from intermediate representations, we propose SG-APCR to address layer-dependent anisotropy in intermediate embeddings and a sliding-window similarity matching strategy to handle subword-level semantic fragmentation. Experiments across three LLMs and three real-world datasets show that sensitive attributes can be reliably inferred using only intermediate representations, achieving Top-1 accuracy of up to 0.997 on CMS, 0.980 on Skytrax, and 0.986 on ECHR. These results reveal that intermediate states commonly considered safe to share can expose sensitive personal attributes on their own.

[SRW] Processing Inconsistency Predicts Language Competence: LLM Evaluation Without Answer Labels on Turkic Languages ▶ *Underline Video*

Ilya Galyukshev; Ilseyar Alimova

Most languages lack labeled evaluation benchmarks for large language models (LLMs). Creating such benchmarks requires native speakers, domain expertise, and answer annotation - resources unavailable for the vast majority of languages. We investigate whether a model's internal processing signals - such as generation entropy and tokenizer statistics - correlate with its actual accuracy on a language, with the long-term goal of estimating language competence without labeled data. Our key observation is that for languages a model does not know,

both tokenizer segmentation and generation entropy become highly variable across questions, whereas for known languages they remain consistent. We call this the inconsistency hypothesis and test it on 11 instruction-tuned LLMs (1B-70B parameters) across 14 language-script varieties (12 Turkic plus English and Russian controls). Standard deviation outperforms the mean as an accuracy predictor in 21 of 25 greedy-pass family x measure combinations (binomial $p < 0.001$, though the measures are not fully independent); sampling-based measures show no such pattern (4 of 20). A three-feature proxy combining a tokenizer coverage ratio chars_std / fertility), entropy standard deviation, and the model's EN/RU benchmark accuracy achieves leave-one-language-out $R^2 = 0.75$ and Kendall = 0.72 on 154 model-language pairs, with the tokenizer ratio contributing most of the language-level signal ($R^2 = 0.34$). Removing the EN/RU anchor isolates an anchor-free component at $R^2 = 0.49$, which exceeds the model-mean baseline and captures language coverage but not model capability.

[Long] Projecting Out the Malice: A Global Subspace Approach to LLM Detoxification ► *Underline Video*

Zenghao Duan, Zhiyi yin, Zhichao Shi, Liang Pang, Shaoling Jing, Zihe Huang, Jiayi Wu, Yu Yan, Jingcheng Deng, Huawei Shen and Xueqi Cheng

Large language models (LLMs) exhibit exceptional performance but pose inherent risks of generating toxic content, restricting their safe deployment. While traditional methods (e.g., alignment) adjust output preferences, they fail to eliminate underlying toxic regions in parameters, leaving models vulnerable to adversarial attacks. Prior mechanistic studies characterize toxic regions as "toxic vectors" or "layer-wise subspaces", yet our analysis identifies critical limitations: i) Removed toxic vectors can be reconstructed via linear combinations of non-toxic vectors, demanding targeting of entire toxic subspace; ii) Contrastive objective over limited samples inject noise into layer-wise subspaces, hindering stable extraction. These highlight the challenge of identifying robust toxic subspace and removing them. Therefore, we propose GLOSS (GLobal tOxic Subspace Suppression), a lightweight method that mitigates toxicity by identifying and eliminating this global subspace from FFN parameters. Experiments on LLMs (e.g., Qwen3) show GLOSS achieves SOTA detoxification while preserving general capabilities without requiring large-scale retraining.

[Long] ProMedical: Hierarchical Fine-Grained Criteria Modeling for Medical LLM Alignment via Explicit Injection ► *Underline Video*

He GENG, Yangmin Huang, Lixian Lai, Qianyun Du, Hui Chu, Zhiyang He, Jiaxue Hu and Xiaodong Tao

Aligning Large Language Models (LLMs) with high-stakes medical standards remains a significant challenge, primarily due to the dissonance between coarse-grained preference signals and the complex, multi-dimensional nature of clinical protocols. To bridge this gap, we introduce ProMedical, a unified alignment framework grounded in fine-grained clinical criteria. We first construct ProMedical-Preference-50k, a dataset generated via a human-in-the-loop pipeline that augments medical instructions with rigorous, physician-derived rubrics. Leveraging this corpus, we propose the Explicit Criteria Injection paradigm to train a multi-dimensional reward model. Unlike traditional scalar reward models, our approach explicitly disentangles safety constraints from general proficiency, enabling precise guidance during reinforcement learning. To rigorously validate this framework, we establish ProMedical-Bench, a held-out evaluation suite anchored by double-blind expert adjudication. Empirical evaluations demonstrate that optimizing the Qwen3-8B base model via ProMedical-RM-guided GRPO yields substantial gains, improving overall accuracy by 22.3% and safety compliance by 21.7%, effectively rivaling proprietary frontier models. Furthermore, the aligned policy generalizes robustly to external benchmarks, demonstrating performance comparable to state-of-the-art models on UltraMedical. We publicly release our datasets, reward models, and benchmarks to facilitate reproducible research in safety-aware medical alignment.

[Industry] Proximity-Based Multi-Turn Optimization: Practical Credit Assignment for LLM Agent Training ► *Underline Video*

Yangyi Fang, Jiaye Lin, Xiaoliang Fu, Cong Qin, Chang Liu and Haolin Shi

Multi-turn LLM agents are becoming pivotal to production systems, spanning customer service automation, e-commerce assistance, and interactive task management, where accurately distinguishing high-value informative signals from stochastic noise is critical for sample-efficient training. In real-world scenarios, a failure in a trivial task may reflect random instability, whereas success in a high-difficulty task signifies a genuine capability breakthrough. Yet, existing group-based policy optimization methods rigidly rely on statistical deviation within discrete batches, frequently misallocating credit when task difficulty fluctuates. To address this issue, we propose Proximity-based Multi-turn Optimization (ProxMO), a practical and robust framework engineered specifically for the constraints of real-world deployment. ProxMO integrates global context via two lightweight mechanisms: success-rate-aware modulation dynamically adapts gradient intensity based on episode-level difficulty, while proximity-based soft aggregation derives baselines through continuous semantic weighting at the step level. Extensive evaluations on ALFWorld and WebShop benchmarks demonstrate that ProxMO yields substantial performance gains over existing baselines with negligible computational cost. Ablation studies further validate the independent and synergistic efficacy of both mechanisms. Crucially, ProxMO offers plug-and-play compatibility with standard GRPO frameworks, facilitating immediate, low-friction adoption in existing industrial training pipelines. Our implementation is available at: https://github.com/GithubX-F/ProxMO-RL.

[Long] Pruning Unsafe Tickets: A Resource-Efficient Framework for Safer and More Robust LLMs ► *Underline Video*

Wai Man Si, Mingjie Li, Michael Backes and Yang Zhang

Machine learning models are increasingly deployed in real-world applications, but even aligned models such as Mistral and LLaVA still exhibit unsafe behaviors inherited from pre-training. Current alignment methods like SFT and RLHF primarily encourage models to generate preferred responses, but do not explicitly remove the unsafe subnetworks that trigger harmful outputs. In this work, we introduce a resource-efficient pruning framework that directly identifies and removes parameters associated with unsafe behaviors while preserving

model utility. Our method employs a gradient-free attribution mechanism, requiring only modest GPU resources, and generalizes across architectures and quantized variants. Empirical evaluations on ML models show substantial reductions in unsafe generations and improved robustness against jailbreak attacks, with minimal utility loss. From the perspective of the Lottery Ticket Hypothesis, our results suggest that ML models contain unsafe tickets responsible for harmful behaviors, and pruning reveals safety tickets that maintain performance while aligning outputs. This provides a lightweight, post-hoc alignment strategy suitable for deployment in resource-constrained settings.

[Findings] PsychePass: Calibrating LLM Therapeutic Competence via Trajectory-Anchored Tournaments ▶ *Underline Video*

Zhuang Chen, Dazhen Wan, Zhangkai Zheng, Guanqun Bi, Xiyao Xiao, Binghang Li and Minlie Huang;

While large language models show promise in mental healthcare, evaluating their therapeutic competence remains challenging due to the unstructured and longitudinal nature of counseling. We argue that current evaluation paradigms suffer from an unanchored defect, leading to two forms of instability: process drift, where unsteered client simulation wanders away from specific counseling goals, and standard drift, where static pointwise scoring lacks the stability for reliable judgment. To address this, we introduce Ps, a unified framework that calibrates the therapeutic competence of LLMs via trajectory-anchored tournaments. We first anchor the interaction trajectory in simulation, where clients precisely control the fluid consultation process to probe multifaceted capabilities. We then anchor the battle trajectory in judgments through an efficient Swiss-system tournament, utilizing dynamic pairwise battles to yield robust Elo ratings. Beyond ranking, we demonstrate that tournament trajectories can be transformed into credible reward signals, enabling on-policy reinforcement learning to enhance LLMs' performance. Extensive experiments validate the effectiveness of PsychePass and its strong consistency with human expert judgments.

[Long] QiMeng-PRepair: Precise Code Repair via Edit-Aware Reward Optimization ▶ *Underline Video*

Changxin Ke, Rui Zhang, Jiaming Guo, Yuanbo Wen, Li Ding, Shuo Wang, Xuyuan Zhu, Xiong Peng, Di Huang, Zidong Du, Xing Hu, Qi Guo and Yunji Chen

Large Language Models (LLMs) achieve strong program repair performance but often suffer from over-editing, where excessive modifications overwrite correct code and hinder bug localization. We systematically quantify its impact and introduce precise repair task, which maximizes reuse of correct code while fixing only buggy parts. Building on this insight, we propose PRepair, a framework that mitigates over-editing and improves repair accuracy. PRepair has two components: Self-Breaking, which generates diverse buggy programs via controlled bug injection and min-max sampling, and Self-Repairing, which trains models with Edit-Aware Group Relative Policy Optimization (EA-GRPO) using an edit-aware reward to encourage minimal yet correct edits. Experiments show that PRepair improves repair precision by up to 31.4% under fix_1@1, a metric that jointly considers repair correctness and extent, and significantly increases decoding throughput when combined with speculative editing, demonstrating its potential for precise and practical code repair.

[Long] QuantumQA: Enhancing Scientific Reasoning via Physics-Consistent Dataset and Verification-Aware Reinforcement Learning ▶ *Underline Video*

Songxin Qu, Tai-Ping Sun, Yun-Jie Wang, Huan-Yu Liu, Cheng Xue, Xiao-Fan Xu, Han Fang, Yang Yang, Yu-Chun Wu, Guo-Ping Guo and Zhao-Yun Chen

Large language models (LLMs) show strong capabilities in general reasoning but typically lack reliability in scientific domains like quantum mechanics, which demand strict adherence to physical constraints. This limitation arises from the scarcity of verifiable training resources and the inadequacy of coarse feedback signals in standard alignment paradigms. To address the data challenge, we introduce QuantumQA, a large-scale dataset constructed via a task-adaptive strategy and a hybrid verification protocol that combines deterministic solvers with semantic auditing to guarantee scientific rigor. Building on this foundation, we propose the verification-aware reward model (VRM) tailored for Reinforcement Learning with Verifiable Rewards (RLVR), which employs an adaptive reward fusion (ARF) mechanism to dynamically integrate deterministic signals from a scientific execution suite (SES) with multidimensional semantic evaluations for precise supervision. Experimental results demonstrate that our method consistently outperforms baselines and general-purpose preference models. Notably, our optimized 8B model achieves performance competitive with proprietary models, validating that incorporating verifiable, rule-based feedback into the reinforcement learning loop offers a parameter-efficient alternative to pure scaling.

[Findings] RCTEA: Richness-guided Co-training for Temporal Entity Alignment ▶ *Underline Video*

Jiayun Li, Wen Hua, Shiqi Fan, Fengmei Jin, Haiyang Jiang and Xue Li

Temporal Entity Alignment (TEA), which aims to identify equivalent entities across Temporal Knowledge Graphs (TKGs), is crucial for integrating knowledge facts from multiple sources. However, existing TEA models often fail to capture the orthogonal yet complementary effect between structural and temporal features, and typically overlook the importance of information richness, a key factor for effective message passing in the neural feature encoders. To address these limitations, we propose a RCTEA framework that jointly models both structural and temporal aspects of the TKGs for entity alignment. Specifically, we design a richness-guided attention mechanism along with an adaptive weighting strategy to facilitate effective feature fusion. To ensure robust alignment despite noisy entity contexts, we introduce a dual-view neighborhood consensus algorithm that jointly refines the feature encoders to enforce local structural consistency of the predicted alignments. Extensive experiments demonstrate the superiority of RCTEA, achieving state-of-the-art performance on public TEA benchmarks.

[Long] Re^3: Relevance & Recency Retrieval for Mitigating Temporal Hallucination ▶ *Underline Video*

Jiawei Cao, Jie Ouyang, Mingyue Cheng, Zhaomeng Zhou, Chunli Liu, Yupeng Li, Zirui Liu and Shijin Wang

Retrieval-Augmented Generation (RAG) is a mainstream approach to mitigating hallucinations in Large Language Models (LLMs), yet in dynamic real-world scenarios, such as weather forecasting or evolving news events, existing retrievers suffer from both temporal-semantic misalignment and outdated-document interference. To address this, we propose Relevance & Recency Retrieval (Re^3), a novel framework that mitigates temporal hallucinations via two core components: a Time-Aware Dual Relevance Encoder that embeds heterogeneous temporal signals into the semantic space to ensure retrieval fidelity, and a Conflict-Aware Recency Filter that performs listwise arbitration to identify and suppress obsolete factual versions. To rigorously evaluate this setting, we introduce Re^2 Bench, a large-scale benchmark comprising over 1.3 million instances designed to assess system robustness in realistic environments where temporal constraints and conflicting factual versions coexist. Experiments on three public benchmarks and Re^2 Bench demonstrate that Re^3 consistently outperforms the strongest baselines by an average of 9.7% in generation accuracy, with gains of up to 25.2% on challenging dynamic tasks, while demonstrating robustness across diverse RAG settings.

[Long] Re^3: Relevance & Recency Retrieval for Mitigating Temporal Hallucination ▶ *Underline Video*

Jiawei Cao, Jie Ouyang, Mingyue Cheng, Zhaomeng Zhou, Chunli Liu, Yupeng Li, Zirui Liu and Shijin Wang

Retrieval-Augmented Generation (RAG) is a mainstream approach to mitigating hallucinations in Large Language Models (LLMs), yet in dynamic real-world scenarios, such as weather forecasting or evolving news events, existing retrievers suffer from both temporal-semantic misalignment and outdated-document interference. To address this, we propose Relevance & Recency Retrieval (Re^3), a novel framework that mitigates temporal hallucinations via two core components: a Time-Aware Dual Relevance Encoder that embeds heterogeneous temporal signals into the semantic space to ensure retrieval fidelity, and a Conflict-Aware Recency Filter that performs listwise arbitration to identify and suppress obsolete factual versions. To rigorously evaluate this setting, we introduce Re^2 Bench, a large-scale benchmark comprising over 1.3 million instances designed to assess system robustness in realistic environments where temporal constraints and conflicting factual versions coexist. Experiments on three public benchmarks and Re^2 Bench demonstrate that Re^3 consistently outperforms the strongest baselines by an average of 9.7% in generation accuracy, with gains of up to 25.2% on challenging dynamic tasks, while demonstrating robustness across diverse RAG settings.

[Long] ReasonAny: Incorporating Reasoning Capability to Any Model via Simple and Effective Model Merging ▶ *Underline Video*

Junyao Yang, Chen Qian, Wen Shen, Yong Liu, Jing Shao and Dongrui Liu

Large Reasoning Models (LRMs) with long chain-of-thought reasoning have recently achieved remarkable success. Yet, equipping domain-specialized models with such reasoning capabilities, referred to as "Reasoning + X", remains a significant challenge. While model merging offers a promising training-free solution, existing methods often suffer from a destructive performance collapse: existing methods tend to both weaken reasoning depth and compromise domain-specific utility. Interestingly, we identify a counter-intuitive phenomenon underlying this failure: reasoning ability predominantly resides in parameter regions with low gradient sensitivity, contrary to the common assumption that domain capabilities correspond to high-magnitude parameters. Motivated by this insight, we propose ReasonAny, a novel merging framework that resolves the reasoning-domain performance collapse through Contrastive Gradient Identification. Experiments across safety, biomedicine, and finance domains show that ReasonAny effectively synthesizes "Reasoning + X" capabilities, significantly outperforming state-of-the-art baselines while retaining robust reasoning performance.

[Long] Reasoning Fails Where Step Flow Breaks ▶ *Underline Video*

Xiaoyu Xu, Yulan Pan, Xiaosong Yuan, Zhihong Shen, Minghao Su, Yuanhao Su and Xiaofeng Zhang

Large reasoning models (LRMs) that generate long chains of thought now perform well on multi-step math, science, and coding tasks. However, their behavior is still unstable and hard to interpret, and existing analysis tools struggle with such long, structured reasoning traces. We introduce Step-Saliency, which pools attention–gradient scores into step-to-step maps along the question–thinking–summary trajectory. Across several models, Step-Saliency reveals two recurring information-flow failures: Shallow Lock-in, where shallow layers over-focus on the current step and barely use earlier context, and Deep Decay, where deep layers gradually lose saliency on the thinking segment and the summary increasingly attends to itself and the last few steps. Motivated by these patterns, we propose StepFlow, a saliency-inspired test-time intervention that adjusts shallow saliency patterns measured by Step-Saliency via Odds-Equal Bridge and adds a small step-level residual in deep layers via Step Momentum Injection. StepFlow improves accuracy on math, science, and coding tasks across multiple LRMs without retraining, indicating that repairing information flow can recover part of their missing reasoning performance.

[Findings] Reducing Hallucinations in LLMs via Factuality-Aware Preference Learning ▶ *Underline Video*

Sindhuja Chaduvula, Ahmed Y. Radwan, Azib Farooq, Yani Ioannou and Shaina Raza

Preference alignment methods such as RLHF and Direct Preference Optimization (DPO) improve instruction following, but they can also reinforce hallucinations when preference judgments reward fluency and confidence over factual correctness. We introduce F-DPO (Factuality-aware Direct Preference Optimization), a simple extension of DPO that uses only binary factuality labels. F-DPO (i) applies a label-flipping transformation that corrects misordered preference pairs so the chosen response is never less factual than the rejected one, and (ii) adds a factuality-aware margin that emphasizes pairs with clear correctness differences, while reducing to standard DPO when both responses share the same factuality. We construct factuality-aware preference data by augmenting DPO pairs with binary factuality indicators and synthetic hallucinated variants. Across seven open-weight LLMs (1B-14B), F-DPO consistently improves factuality and reduces hallucination rates relative to both base models and standard DPO. On Qwen3-8B, F-DPO reduces hallucination rates by 5x(from 0.424 to 0.084) while improving factuality scores by 50% (from 5.26 to 7.90). F-DPO also generalizes to out-of-distribution benchmarks: on

TruthfulQA, Qwen2.5-14B achieves +17% MC1 accuracy (0.500 to 0.585) and +49% MC2 accuracy (0.357 to 0.531). F-DPO requires no auxiliary reward model, token-level annotations, or multi-stage training.

[SRW] Reflection in the Dark: Exposing and Escaping the Black Box in Reflective Prompt Optimization ▶ *Underline Video*
Shiyan Liu, Qifeng Xia, Qiyun Xia, Yisheng Liu, Xinyu Yu and Rui Qu;
Automatic prompt optimization (APO) has emerged as a powerful paradigm for improving LLM performance without manual prompt engineering. Reflective APO methods such as GEPA iteratively refine prompts by diagnosing failure cases, but the optimization process remains black-box and label-free, leading to uninterpretable trajectories and systematic failure. We identify and empirically demonstrate four limitations: on GSM8K with a defective seed, GEPA degrades accuracy from 23.81% to 13.50%. We propose VISTA, a multi-agent APO framework that decouples hypothesis generation from prompt rewriting, enabling semantically labeled hypotheses, parallel minibatch verification, and interpretable optimization trace. A two-layer explore-exploit mechanism combining random restart and epsilon-greedy sampling further escapes local optima. VISTA recovers accuracy to 87.57% on the same defective seed and consistently outperforms baselines across all conditions on GSM8K and AIME2025.

[Long] REG: Retrieval via Emotion Similarity for Guiding Empathetic Dialogue Generation ▶ *Underline Video*
Xu Wang and Bo WangYang XiangYihong TangDongming ZhaoZifei YuYuexian Hou
Empathy relies on the cognitive capacity to relate to similar past experiences. Consequently, retrieval-based approaches utilize analogous exemplars to guide empathetic dialogue generation. However, existing methods prioritize semantic similarity over emotion characteristics, often leading to unempathetic responses. To address this, we propose REG, a framework that integrates four Emotion Attributes into the retrieval process to ensure explicit emotional alignment. Furthermore, to mitigate the noise and limited diversity caused by coarse-grained sentence-level attributes, we incorporate Token-level Retrieval for finer granularity and a Retrieval Candidate Augmentation strategy to enhance diversity. Empirical results on the EmpatheticDialogues dataset demonstrate that REG significantly outperforms baselines, offering a robust solution for empathetic generation.

[Long] Render-of-Thought: Rendering Textual Chain-of-Thought as Images for Visual Latent Reasoning ▶ *Underline Video*
Yifan Wang, Shiyu Li, Peiming Li, Xiaochen Yang, Zheng Wei and Yang Tang
Chain-of-Thought (CoT) prompting has achieved remarkable success in unlocking the reasoning capabilities of Large Language Models (LLMs). Although CoT prompting enhances reasoning, its verbosity imposes substantial computational overhead. Recent works often focus exclusively on outcome alignment and lack supervision on the intermediate reasoning process. These deficiencies obscure the analyzability of the latent reasoning chain. To address these challenges, we introduce Render-of-Thought (RoT), the first framework to reify the reasoning chain by rendering textual steps into images, making the latent rationale explicit and traceable. Specifically, we leverage the vision encoders of existing Vision Language Models (VLMs) as semantic anchors to align the vision embeddings with the textual space. This design ensures plug-and-play implementation without incurring additional pre-training overhead. Extensive experiments on mathematical and logical reasoning benchmarks demonstrate that our method achieves 3-4x token compression and substantial inference acceleration compared to explicit CoT. Furthermore, it demonstrates a competitive efficiency-accuracy Pareto exploration compared to other methods, validating the feasibility of this paradigm. Our code is available at https://github.com/TencentBAC/RoT

[Industry] ResoDiff-44k: High-Fidelity Cross-Lingual Speech and Singing Synthesis via Discrete Diffusion ▶ *Underline Video*
Gyanendra Das and Sai Satyam Jena
While large-scale generative speech models have achieved remarkable semantic coherence, industrial deployment remains constrained by a fidelity ceiling typically capped at lower sampling rates. A fundamental limitation is the reliance on intermediate mel-spectrograms, a low-dimensional bottleneck that discards phase and high-frequency information, causing artifacts in expressive scenarios like singing. In this work, we introduce ResoDiff-44k, a production-grade generative foundation model designed for cinema-quality, 44.1kHz audio synthesis. Departing from standard masked audio modeling and mel-spectrogram inversion, ResoDiff-44k leverages Discrete Diffusion over a pure Descript Audio Codec latent space. We pre-train ResoDiff-44k on a massive 150K -hour multilingual dataset to establish a robust acoustic prior, followed by targeted fine-tuning on a curated regional mixed-language and singing corpus. Our experiments demonstrate that replacing the standard prediction head with a discrete diffusion trajectory significantly reduces misalignment in long sequences. We report a double-blind subjective evaluation showing that ResoDiff-44k achieves a 4.6 Mean Opinion Score in 44.1kHz singing synthesis and a 71% reduction in character error rate on regional mixed-language prompts compared to strong baselines. The proposed pipeline offers a viable path for deploying high-fidelity, culturally adaptive conversational agents.

[Long] RFS-Guard: Detecting Reasoning Hallucinations via Cross-Phase Routing Focus in Large Reasoning Models ▶ *Underline Video*
Zihang Liu, Zhouhua Fang, Hui Liu, Zhiwei Liu, Yong Li and Haishuai Wang
Large reasoning models (LRMs) achieve strong performance on complex tasks by generating intermediate reasoning before the final answer, yet they remain prone to reasoning hallucinations such as subtle arithmetic or constraint-violation errors. Prior hallucination detectors often rely on external verification or local token-level signals, which are limited for LRMs and largely overlook whether the cross-phase information flow from reasoning to answering is structurally robust. We propose Routing Focus Score (RFS), a step-level indicator that measures how strongly cross-step attention routing aligns with semantic proximity derived from hidden-state cosine similarity. We further

design RFS-Guard, a lightweight hallucination detection framework based on RFS. Empirically, we observe that higher reasoning-answer RFS is consistently associated with higher hallucination risk, suggesting a routing-collapse failure mode where models might prefer self-confirmation loops and suppress the ability to audit their own generations. Experimental results across multiple domains and models demonstrate the superiority of RFS-Guard for detecting and localizing hallucinations in LRMs without requiring external tools or repeated sampling.

[Long] RISK: A Framework for GUI Agents in E-commerce Risk Management ▶ *Underline Video*

Renqi Chen, Zeyin Tao, Jianming Guo, Jingzhe Zhu, Yiheng Peng, Qingqing Sun, Tianyi Zhang and Shuai Chen

E-commerce risk management requires aggregating diverse, deeply embedded web data through multi-step, stateful interactions, which traditional scraping methods and most existing Graphical User Interface (GUI) agents cannot handle. These agents are typically limited to single-step tasks and lack the ability to manage dynamic, interactive content critical for effective risk assessment. To address this challenge, we introduce RISK, a novel framework designed to build and deploy GUI agents for this domain. RISK integrates three components: (1) RISK-Data, a dataset of 8,492 single-step and 2,386 multi-step interaction trajectories, collected through a high-fidelity browser framework and a meticulous data curation process; (2) RISK-Bench, a benchmark with 802 single-step and 320 multi-step trajectories across three difficulty levels for standardized evaluation; and (3) RISK-R1, a R1-style reinforcement fine-tuning framework considering four aspects: (i) Output Format Constraint, (ii) Single-step and (iii) Multi-step Level Reward, and (iv) Task Level Reweight. Experiments show that RISK-R1 achieves a 6.8% improvement in offline single-step and an 8.8% improvement in offline multi-step, using only 7.2% of the parameters of the SOTA baseline. Moreover, it attains a top task success rate of 70.5% in online evaluation. RISK provides a scalable, domain-specific solution for automating complex web interactions in e-commerce risk management. The code is available at https://github.com/RenqiChen/RISK-GUI.

[Long] Robust Membership Inference for Large Language Models under Adversarial Generative Corruption ▶ *Underline Video*

Yuanhong Huang, Huili Wang, Xueying Bai, Jinrui Wang, Jiajun Liu, Ziqin Wang, Wanchun Ni, Shangguang Wang and Tao Qi

Membership inference attack (MIA) has emerged as a promising tool for auditing the training data of LLMs, supporting data privacy and copyright protection. Most existing MIA methods rely on the assumption that LLMs assign higher confidence scores to training samples than to non-training ones. However, since LLMs generate text by sampling high-confidence tokens, they naturally produce AI-generated texts (AIGTs) that also satisfy this assumption. In this work, we empirically confirm that such AIGTs, regardless of whether they are generated by the target LLM, can lead existing MIAs to assign even higher membership likelihoods than those of true training samples, thereby significantly undermining their reliability. To address this challenge, we propose a robust membership inference framework for reliably identifying training data. Our method adopts a mixture-of-experts formulation to jointly model interactions across complementary features derived from multiple MIA methods and AIGT detectors, which can remain robust against adversarially generated samples. Furthermore, by leveraging expert components, our method provides explainable insights into the characteristics of member data. Experiments on various datasets and LLMs show that adversarial samples substantially degrade the performance of baselines, whereas our method preserves performance close to that of the unattacked setting. Codes and datasets are released at https://github.com/kong-hyh/MoMIA.

[Long] rSIM: Incentivizing Reasoning Capabilities of LLMs via Reinforced Strategy Injection ▶ *Underline Video*

Sijia Chen, Baochun Li and Di Niu

Large language models (LLMs) are post-trained through reinforcement learning (RL) to evolve into Reasoning Language Models (RLMs), where the hallmark of this advanced reasoning is "aha" moments when they start to perform strategies, such as self-reflection and deep thinking, within chain of thoughts (CoTs). Motivated by this, this paper proposes a novel reinforced strategy injection mechanism (rSIM), that enables any LLM to become an RLM by employing a small planner to guide the LLM's CoT through the adaptive injection of reasoning strategies. To achieve this, the planner (leader agent) is jointly trained with an LLM (follower agent) using multi-agent RL (MARL), based on a leader-follower framework and straightforward rule-based rewards. Experimental results show that rSIM enables Qwen2.5-0.5B to become an RLM and significantly outperform Qwen2.5-14B across mathematical, coding, and financial reasoning tasks. Moreover, the planner is generalizable: it only needs to be trained once and can be applied as a plug-in to substantially improve the reasoning capabilities of existing LLMs. In addition, the planner supports continual learning across various tasks, allowing its planning abilities to gradually improve and generalize to a wider range of problems. Our source code is available under the examples/rSIM of https://github.com/AgenticFinLab/eparl.

[Long] RubricHub: A Comprehensive and Highly Discriminative Rubric Dataset via Automated Coarse-to-Fine Generation ▶ *Underline Video*

Sunzhu Li, Jiale Zhao, Huimin Ren, Zhenlin Wei, Yang Zhou, Jingwen Yang, Shunyu Liu, Kaike Zhang and Chen Wei

Reinforcement Learning with Verifiable Rewards (RLVR) has driven substantial progress in reasoning-intensive domains like mathematics. However, optimizing open-ended generation remains challenging due to the lack of ground truth. While rubric-based evaluation offers a structured proxy for verification, existing methods suffer from scalability bottlenecks and coarse criteria, resulting in a supervision ceiling effect. To address this, we propose an automated Coarse-to-Fine Rubric Generation framework. By synergizing principle-guided synthesis, multi-model aggregation, and difficulty evolution, our approach produces comprehensive and highly discriminative criteria capable of capturing the subtle nuances. Based on this framework, we introduce RubricHub, a large-scale (110k) and multi-domain dataset. We validate its utility through a two-stage post-training pipeline comprising Rubric-based Rejection Sampling Fine-Tuning (RuFT) and Reinforcement Learning (RuRL). Experimental results demonstrate that RubricHub unlocks significant performance gains: our post-trained Qwen3-14B achieves state-of-the-art (SOTA) results on HealthBench (69.3), surpassing proprietary frontier models such as GPT-5.

[Long] SAFE-QAQ: End-to-End Slow-Thinking Audio-Text Fraud Detection via Reinforcement Learning ▶ *Underline Video*

Peidong Wang, Zhiming Ma, Xin Dai, YongKang Liu, Shi Feng, Xiaocui Yang, Wenxing Hu, Zhihao Wang, Mingjun Pan, Li Yuan and Daling Wang

Existing fraud detection methods predominantly rely on transcribed text, suffering from ASR errors and missing crucial acoustic cues like vocal tone and environmental context. This limits their effectiveness against complex deceptive strategies. To address these challenges, we first propose SAFE-QAQ, an end-to-end comprehensive framework for audio-based slow-thinking fraud detection. First, the SAFE-QAQ framework eliminates the impact of transcription errors on detection performance. Secondly, we propose rule-based slow-thinking reward mechanisms that systematically guide the system to identify fraud-indicative patterns by accurately capturing fine-grained audio details, through hierarchical reasoning processes. Besides, our framework introduces a dynamic risk assessment framework during live calls, enabling early detection and prevention of fraud. Experiments on the TeleAntiFraud-Bench demonstrate that SAFE-QAQ achieves dramatic improvements over existing methods in multiple key dimensions, including accuracy, inference efficiency, and real-time processing capabilities. Currently deployed and analyzing over 70,000 calls daily, SAFE-QAQ effectively automates complex fraud detection, reducing human workload and financial losses. Code: https://anonymous.4open.science/r/SAFE-QAQ.

[Findings] Safety Guardrails of Large Language Models Are Vulnerable to Value-Driven Adversarial Prompting ▶ *Underline Video*

Xiaohao Luo, Ying Wei and Zhijun Li

In the real world, the execution of a task often depends on the executor's recognition of its value. Motivated by this observation, we propose the value-driven jailbreak attack (VDJA), a simple and effective black-box jailbreak method against large language models (LLMs). VDJA first exploits the phenomenon that LLMs tend to agree with humans to induce LLMs to affirm the moral value of harmful tasks. During autoregressive generation, these value-endorsement tokens function as an implicit value prior, making LLMs more likely to accept and generate harmful content. Extensive experiments on five state-of-the-art (SOTA) LLMs demonstrate the superiority of VDJA. Using only a single query and without concealing harmful instructions, VDJA achieves an average attack success rate (ASR) of 91.8% on JailbreakBench and 95.2% on the AdvBench subset, showcasing SOTA jailbreak success rates and attack efficiency. Most importantly, our work suggests a previously underexplored vulnerability in the safety guardrails of LLMs, which highlights the urgent need to enhance their robustness.

[Long] Safety-Utility Conflicts Are Not Global: Surgical Alignment via Head-Level Diagnosis ▶ *Underline Video*

Wang Cai, Yilin Wen, Jinchang Hou, Du Su, Guoqiu Wang, Zhonghou Lv, Chenfu Bao and Yunfang Wu

Safety alignment in Large Language Models (LLMs) inherently presents a multi-objective optimization conflict, often accompanied by an unintended degradation of general capabilities. Existing mitigation strategies typically rely on global gradient geometry to resolve these conflicts, yet they overlook Modular Heterogeneity within Transformers, specifically that the functional sensitivity and degree of conflict vary substantially across different attention heads. Such global approaches impose uniform update rules across all parameters, often resulting in suboptimal trade-offs by indiscriminately updating utility sensitive heads that exhibit intense gradient conflicts. To address this limitation, we propose Conflict-Aware Sparse Tuning (CAST), a framework that integrates head-level diagnosis with sparse fine-tuning. CAST first constructs a pre-alignment conflict map by synthesizing Optimization Conflict and Functional Sensitivity, which then guides the selective update of parameters. Experiments reveal that alignment conflicts in LLMs are not uniformly distributed. We find that the drop in general capabilities mainly comes from updating a small group of "high-conflict" heads. By simply skipping these heads during training, we significantly reduce this loss without compromising safety, offering an interpretable and parameter-efficient approach to improving the safety-utility trade-off.

[Industry] SCAIR: Schema-Conditioned Agentic Iterative Reasoning for Enterprise Knowledge Graphs ▶ *Underline Video*

Prateek Chaturvedi, Yuqicheng Zhu, Hongkuan Zhou, Dongzhuoran Zhou, Yunjie He, Steffen Staab, Fei Du, Jie Tang and Evgeny Kharlamov

Knowledge Graph-based Retrieval-Augmented Generation (KG-RAG) enables natural language interaction with structured enterprise knowledge, yet existing agentic approaches that perform well on public benchmarks often fail to generalize to real-world enterprise Knowledge Graphs (KGs), which are dense, schema-driven, and operationally constrained. To address these limitations, we propose SCAIR (Schema-Conditioned Agentic Iterative Reasoning), a training-free framework that integrates structured planning with controlled iterative reasoning by injecting schema-conditioned structural priors and enforcing schema-aware traversal during multi-hop reasoning. Experiments on an enterprise-oriented benchmark constructed from a real-world Configuration Management DataBase (CMDB) demonstrate that SCAIR substantially improves performance over existing KG-RAG methods. Crucially, our study highlights that reliable enterprise graph reasoning cannot rely on generic agentic designs; instead, it must explicitly incorporate the target domain's structural and operational constraints into the reasoning process. We demonstrate that by aligning agent design with business logic, substantial performance gains can be achieved without the need for costly model retraining.

[Long] SCVQ: Sparse-Compensated Vector Quantization for Large Language Models ▶ *Underline Video*

Zixuan Zhou, Yujun Diao, Zicheng Kong, Dehua Ma, Zhenbo Xu, Pei Pei Li and Zhaofeng He

Large Language Models (LLMs) are primarily constrained by memory and bandwidth bottlenecks during deployment. Although Vector Quantization (VQ) has emerged as a promising solution, existing methods incur inference overhead due to massive codebook storage and intensive index lookups. Moreover, these methods typically suffer from non-negligible performance degradation under ultra-low bitwidth regimes. To bridge this gap, we propose Sparse-Compensated Vector Quantization (SCVQ), a novel framework designed for high-efficiency

LLM vector quantization. SCVQ introduces a salience-aware weighted K-means clustering scheme with symmetry constraints to reduces codebook size and indexing costs. Central to our approach is a unified structured representation that consolidates outliers, salient weights, and quantization residuals into a single sparse compensation matrix. This design effectively preserves critical model information while leveraging VQ-specific properties to enable efficient custom kernels. Extensive experiments across multiple benchmarks demonstrate SCVQ's superior performance. Specifically, SCVQ achieves a perplexity of 5.78 on WikiText-2 for LLaMA-2-7B at 2-bit quantization, while delivering a 1.4x end-to-end inference speedup over existing baselines.

[Long] SDAR-VL: Stable and Efficient Block-wise Diffusion for Vision-Language Understanding ▶ *Underline Video*

Shuang Cheng, Yuhua Jiang, Zineng Zhou, Dawei Liu, Tao Wang, Linfeng Zhang, Biqing Qi and Bowen Zhou

Block-wise discrete diffusion offers an attractive balance between parallel generation and causal dependency modeling, making it a promising backbone for vision-language modeling. However, its practical adoption has been limited by high training cost, slow convergence, and instability, which have so far kept it behind strong autoregressive (AR) baselines. We present SDAR-VL, the first systematic application of block-wise discrete diffusion to large-scale vision-language understanding (VLU), together with an integrated framework for efficient and stable training. This framework unifies three components: 1) Asynchronous Block-wise Noise Scheduling to diversify supervision within each batch; 2) Effective Mask Ratio Scaling for unbiased loss normalization under stochastic masking; and 3) a Progressive Beta Noise Curriculum that increases effective mask coverage while preserving corruption diversity. Experiments on 21 single-image, multi-image, and video benchmarks show that SDAR-VL consistently improves training efficiency, convergence stability, and task performance over conventional block diffusion. On this evaluation suite, SDAR-VL sets a new state of the art among diffusion-based vision-language models and, under matched settings, matches or surpasses strong AR baselines such as LLaVA-OneVision as well as the global diffusion baseline LLaDA-V, establishing block-wise diffusion as a practical backbone for VLU.

[Long] SearchGym: Bootstrapping Real-World Search Agents via Cost-Effective and High-Fidelity Environment Simulation ▶ *Underline Video*

Xichen Zhang, Ziyi He, Yinghao Zhu, Sitong Wu, Shaozuo Yu, Meng Chu, Wenhu Zhang, Haoru Tan and Jiaya Jia

Search agents have emerged as a pivotal paradigm for solving open-ended, knowledge-intensive reasoning tasks. However, training these agents via Reinforcement Learning (RL) faces a critical dilemma: interacting with live commercial Web APIs is prohibitively expensive, while relying on static data snapshots often introduces noise due to data misalignment. This misalignment generates corrupted reward signals that destabilize training by penalizing correct reasoning or rewarding hallucination. To address this, we propose SearchGym, a simulation environment designed to bootstrap robust search agents. SearchGym employs a rigorous generative pipeline to construct a verifiable knowledge graph and an aligned document corpus, ensuring that every reasoning task is factually grounded and strictly solvable. Building on this controllable environment, we introduce SearchGym-RL, a curriculum learning methodology that progressively optimizes agent policies through purified feedback, evolving from basic interactions to complex, long-horizon planning. Extensive experiments across the Llama and Qwen families demonstrate strong Sim-to-Real generalization. Notably, our Qwen2.5-7B-Base model trained within SearchGym surpasses the web-enhanced ASearcher baseline across nine diverse benchmarks by an average relative margin of 10.6%. Our results validate that high-fidelity simulation serves as a scalable and highly cost-effective methodology for developing capable search agents.

[Long] Seeing No Evil: Blinding Large Vision-Language Models to Safety Instructions via Adversarial Attention Hijacking ▶ *Underline Video*

Jingru Li, Wei Ren and Tianqing Zhu

Large Vision-Language Models (LVLMs) rely on attention-based retrieval of safety instructions to maintain alignment during generation. Existing attacks typically optimize image perturbations to maximize harmful output likelihood, but suffer from slow convergence due to gradient conflict between adversarial objectives and the model's safety-retrieval mechanism. We propose Attention-Guided Visual Jailbreaking, which circumvents rather than overpowers safety alignment by directly manipulating attention patterns. Our method introduces two simple auxiliary objectives: (1) suppressing attention to system-prompt tokens and (2) anchoring generation on adversarial image features. This simple yet effective push-pull formulation reduces gradient conflict by 45% and achieves 94.4% attack success rate on Qwen-VL (vs. 68.8% baseline) with 40% fewer iterations. At tighter perturbation budgets (epsilon=8/255), we maintain 59.0% ASR compared to 45.7% for standard methods. Mechanistic analysis reveals a failure mode we term safety blindness: successful attacks suppress system-prompt attention by 80%, causing models to generate harmful content not by overriding safety rules, but by failing to retrieve them.

[TACL] Self-Consistency Falls Short!: The Adverse Effects of Positional Bias on Long-Context Problems ▶ *Underline Video*

Adam Byerly and Daniel Khashabi

Self-consistency (SC) improves the performance of large language models (LLMs) across various tasks and domains that involve short content. However, does this support its effectiveness for long-context problems? We challenge the assumption that SC's benefits generalize to long-context settings, where LLMs often struggle with position bias—the systematic over-reliance on specific context regions—which hinders their ability to utilize information effectively from all parts of their context. Through comprehensive experimentation with varying state-of-the-art models, tasks, and SC formulations, we find that SC not only fails to improve but actively degrades performance on long-context tasks. This degradation is driven by persistent position bias, which worsens with longer context lengths and smaller model sizes but remains invariant to prompt format or task type. Unlike short-context tasks, where SC diversifies reasoning paths, long-context SC amplifies positional errors. These comprehensive results provide valuable insight into the limitations of current LLMs in long-context

understanding and highlight the need for more sophisticated approaches.

[Findings] Self-Guided Plan Extraction for Instruction-Following Tasks with Goal-Conditional Reinforcement Learning ► *Underline Video*

Zoya Volovikova, Nikita Sorokin, Dmitriy Lukashevskiy, Aleksandr Panov and Alexey Skrynnik

We introduce SuperIgor, a framework for instruction-following tasks. Unlike prior methods that rely on predefined subtasks, SuperIgor enables a language model to generate and refine high-level plans through a self-learning mechanism, reducing the need for manual dataset annotation. Our approach involves iterative co-training: an RL agent is trained to follow the generated plans, while the language model adapts and modifies these plans based on RL feedback and preferences. This creates a feedback loop where both the agent and the planner improve jointly. We validate our framework in environments with rich dynamics and stochasticity. Results show that SuperIgor agents adhere to instructions more strictly than baseline methods, while also demonstrating strong generalization to previously unseen instructions.

[Long] Self-Reflective Generation at Test Time ► *Underline Video*

Jian Mu, Qixin Zhang, Zhiyong Wang, Menglin Yang, Shuang Qiu, Chengwei Qin, Zhongxiang Dai and Yao Shu

Large language models (LLMs) increasingly solve complex reasoning tasks via long chain-of-thought, but their forward-only autoregressive generation process is fragile; early token errors can cascade, which creates a clear need for self-reflection mechanisms. However, existing self-reflection either performs revisions over full drafts or learns self-correction via expensive training, both fundamentally reactive and inefficient. To address this, we propose Self-Reflective Generation at Test Time (SRGen), a lightweight test-time framework that reflects before generating at uncertain points. During token generation, SRGen utilizes dynamic entropy thresholding to identify high-uncertainty tokens. For each identified token, it trains a specific corrective vector, which fully exploits the already generated context for a self-reflective generation to correct the token probability distribution. By retrospectively analyzing the partial output, this self-reflection enables more trustworthy decisions, thereby significantly reducing the probability of errors at highly uncertain points. Evaluated on challenging mathematical reasoning benchmarks and a diverse set of LLMs, SRGen can significantly strengthen model reasoning. Moreover, our findings position SRGen as a plug-and-play method that integrates reflection into the generation process for reliable LLM reasoning, achieving consistent gains with bounded overhead and can be combined with other training-time (e.g., RLHF) and test-time (e.g., SLOT) techniques.

[SRW] Semantic Contrastive Adaptation for Multimodal Figurative Language Understanding ► *Underline Video*

Ayaan Siddiqui

Understanding idiomatic and figurative language in images remains a fundamental challenge for vision-language models, as it requires reasoning beyond literal image-text alignment. Although large pretrained models such as CLIP and BLIP-2 perform well on literal recognition, they consistently fail on multimodal figurative benchmarks, often favoring visually salient but semantically literal interpretations. We show that this failure arises from a systematic literal alignment bias rather than limited model capacity. Motivated by this observation, we reformulate multimodal figurative understanding as a contrastive semantic deviation problem, where figurative images must be distinguished from visually plausible literal alternatives. We introduce a parameter-efficient adaptation of CLIP using Low-Rank Adaptation (LoRA) with hard literal negative mining, achieving targeted reshaping of multimodal representations without full fine-tuning. Experiments on the IRFL benchmark across idioms, metaphors, and similes demonstrate substantial improvements over zero-shot CLIP, BLIP- 2, ensemble-based, and knowledge-augmented baselines. Finally, we introduce FIGMENT, a multilingual figurative grounding evaluation spanning five idiom-rich languages, and show that the adapted model generalizes across languages despite being trained exclusively on English supervision.

[Long] Shuttle Between Symbolic Instructions and Neural Parameters of Large Language Models ► *Underline Video*

Wangtao Sun, Haotian Xu, Huanxuan Liao, Xuanqing Yu, Zhongtao Jiang, Shizhu He, Jun Zhao and Kang Liu

This paper notices that while symbolic instruction and neural parameters play different roles on steering LLMs' behavior, both instructions and parameters are the compression of task data, they are supposed be strongly correlated and can be learned to predict one from the other. Therefore, This paper proposes a novel neural network framework, SHIP (Shuttle between the Instructions and the Parameters), to model and learn the bi-directional mappings between the instructions and the parameters of LLMs. We verify that SHIP can effectively map one of the instructions/parameters to the other by evaluating it on the tasks of instruction deduction and induction. The results show that SHIP performs better than existing baseline methods in terms of deductive capabilities while significantly surpassing them in inductive capabilities. Moreover, SHIP can effectively combine the two mapping processes to perform excellent inductive reasoning. We further discuss how the latent fusing methods and latent dimensions affect SHIP's performance, and show SHIP can effectively generalize with pre-training. The code and data for this paper are released at https://anonymous.4open.science/r/Shuttle-Between-Instructions-Parameters

[Findings] Simple Role Assignment is Extraordinarily Effective for Safety Alignment ► *Underline Video*

Zhou Ziheng, Jiakun Ding, Zhaowei Zhang, Ruosen Gao, Ying Nian Wu, Demetri Terzopoulos, Yipeng Kang, Fangwei Zhong and Junqi Wang

Principle-based alignment often lacks context sensitivity and completeness. Grounded in Theory of Mind, we propose role conditioning as a compact alternative: social roles (e.g., mother, judge) implicitly encode both values and the cognitive schemas required to apply them. We introduce a training-free pipeline featuring a role-conditioned generator and iterative role-based critics for refinement. Across five model families, our approach consistently outperforms principle-based, Chain-of-Thought (CoT) and other baselines across benchmarks. Notably, it reduces unsafe outputs on the WildJailbreak benchmark from 81.4% to 3.6% with DeepSeek-V3. Not only for common safety benchmarks, it consistently applies for agentic safety tasks. These results establish role assignment as a powerful, interpretable paradigm for AI alignment

and LLM-as-a-Judge construction.

[Short] Situated Embedding Models for Context-Aware Dense Retrieval ▶ *Underline Video*
Junjie Wu, Jiangnan Li, Yuqing Li, Lemao Liu, Liyan Xu, Jiwei Li, Dit-Yan Yeung, Jie Zhou and Mo Yu
Retrieval-augmented generation (RAG) over long documents typically involves splitting the text into smaller chunks, which serve as the basic units for retrieval. However, due to dependencies across the original document, contextual information is often essential for accurately interpreting each chunk. To address this, prior work has explored encoding longer context windows to produce embeddings for longer chunks, yet their gains in retrieval and downstream tasks remain limited. This is because (1) longer chunks strain the capacity of embedding models due to the increased amount of information they must encode, and (2) many real-world applications still require returning localized evidence due to constraints on model or human bandwidth. To this end, we propose an alternative approach to this challenge by representing short chunks in a way that is conditioned on a broader context window to enhance retrieval performance - i.e., situating a chunks meaning within its context. We further show that existing embedding models are not well-equipped to encode such situated context effectively, and thus introduce a new training paradigm and develop the first situated embedding model. To evaluate our method, we curate a book-plot retrieval dataset specifically designed to assess situated retrieval capabilities. On this benchmark, our 1B-parameter model substantially outperforms state-of-the-art embedding models, including several with up to 7B parameters.

[Industry] Smarter, not Bigger: Fine-Tuned RAG-Enhanced LLMs for Automotive HIL Testing ▶ *Underline Video*
Chao Feng, Zihan Liu, Siddhant Gupta and Jan von der Assen
Hardware-in-the-Loop (HIL) testing is essential for automotive validation but suffers from fragmented and underutilized test artifacts. This paper presents HIL-GPT, an industry-deployed retrieval-augmented generation (RAG) system that integrates semantic retrieval with domain-adapted large language models to support test engineers in real-world HIL workflows. The system combines domain-specific embeddings to enable traceable retrieval of test cases and requirements under industrial latency and cost constraints. Through empirical evaluation, we show that compact, domain-adapted models can achieve a favorable trade-off among accuracy, latency, and cost compared to larger general-purpose models, challenging the assumption that larger models are always preferable in industrial NLP systems. An A/B user study further confirms that HIL-GPT improves perceived helpfulness, truthfulness, and satisfaction over general-purpose LLMs.

[Long] SOAPTriage: SOAP-Guided Multi-View Clinical Text Modeling Framework for Automated ESI Prediction ▶ *Underline Video*
Enming Wang and Jianlei Wang
Emergency departments (ED) rely on the Emergency Severity Index (ESI) to assess patient acuity and prioritize care, a process that is largely driven by clinical triage text. Despite recent progress in automated ESI prediction, two fundamental challenges remain: the scarcity of high-quality triage text data due to privacy and regulatory constraints and the lack of a clinically grounded triage framework capable of explicitly capturing the multidimensional structure of triage reasoning. To address these challenges, we draw inspiration from the clinically grounded SOAP paradigm, in which SOAP refers to Subjective, Objective, Assessment, and Plan and captures four complementary aspects of clinical reasoning. Building on this paradigm, we propose SOAPTriage, a SOAP-guided multi-view clinical text modeling framework for automated ESI prediction. To mitigate data scarcity, SOAPTriage introduces a Clinical Note Augmentation (CNA) module that generates natural-language triage notes from structured ED records, resulting in 15,393 augmented clinical notes derived from a real-world dataset. To incorporate clinical structure, SOAPTriage employs a SOAP-Guided Encoding (SGE) module that models patient conditions from four complementary SOAP perspectives, together with an adaptive SOAP-Aware Aggregation and Inference (SAAI) module that performs multi-view reasoning to infer ESI levels. Extensive experiments show that SOAPTriage consistently outperforms strong prompting-based, multi-agent, and encoder-based baselines, demonstrating the effectiveness of SOAP-guided multi-view clinical text modeling for automated emergency triage.

[Findings] Social Welfare Function Leaderboard: On the Emergence of LLM Agents as the Welfare Dictator ▶ *Underline Video*
Zhengliang Shi, Ruotian Ma, Jen-tse Huang, Xinbei Ma, Xingyu Chen, Mengru Wang, Qu Yang, Yue Wang, Fanghua Ye, Ziyang Chen, Shanyi Wang, CIXING LI, Wenxuan Wang, Zhaopeng Tu, Xiaolong Li, Zhaochun Ren and Liefeng Bo
Large language models (LLMs) are increasingly entrusted with high-stakes decisions that affect human welfare. However, the principles and values that guide these models when distributing scarce societal resources remain largely unexamined. To address this, we introduce the Social Welfare Function (SWF) Benchmark, a dynamic simulation environment in which an LLM acts as a dictator, distributing tasks to heterogeneous recipients with different returns on investment (ROI). The benchmark is designed to create a dilemma between maximizing collective efficiency (i.e., overall ROI) and ensuring distributive fairness (measured by the Gini coefficient). We evaluate 20 state-of-the-art LLMs. Our findings reveal several key insights, including: (i) LLMs' general ability, as measured by popular Arena leaderboards, misaligns with their allocation skills; (ii) Most LLMs exhibit a strong default utilitarian orientation, prioritizing overall productivity at the expense of inequality. (iii) Allocation behaviors are highly manipulated, easily perturbed by common persuasion strategies. These results highlight the risks of deploying current LLMs as societal decision-makers and underscore the need for specialized benchmarks and alignment for AI governance.

[Findings] SOP-Maze: Evaluating Large Language Models on Complicated Business Standard Operating Procedures ▶ *Underline Video*
Jiaming wang, Zhe Tang, Zehao Jin, Hefei Chen, Yilin Jin, Peng Ding, Xiaoyu Li and Xuezhi Cao
As large language models (LLMs) are widely deployed as domain-specific agents, many benchmarks have been proposed to evaluate their

ability to follow instructions and make decisions in real-world scenarios. However, business scenarios often involve complex standard operating procedures (SOPs), and the evaluation of LLM capabilities in such contexts has not been fully explored. To bridge this gap, we propose SOP-Maze, a benchmark constructed from real-world business data and adapted into a collection of 397 instances and 3422 subtasks from 23 complex SOP scenarios. We further categorize SOP tasks into two broad classes: Lateral Root System (LRS), representing wide-option tasks that demand precise selection; and Heart Root System (HRS), which emphasizes deep logical reasoning with complex branches. Extensive experiments reveal that nearly all state-of-the-art models struggle with SOP-Maze. We conduct a comprehensive analysis and identify three key error categories: (i) route blindness: difficulty following procedures; (ii) conversational fragility: inability to handle real dialogue nuances; and (iii) calculation errors: mistakes in time or arithmetic reasoning under complex contexts. The systematic study explores LLM performance across SOP tasks that challenge both breadth and depth, offering new insights for improving model capabilities. We have open-sourced our work on the anonymous link: https://github.com/meituan-longcat/SOP-Maze.

[Findings] Sounding vs. Being an Expert: Disentangling Authority, Register and Cultural Impact in Sycophantic LLMs ▶ *Underline Video*

Gabriele Maraia, Fabio Massimo Zanzotto and Leonardo Ranaldi

Large Language Models (LLMs) have been shown to exhibit sycophancy, a tendency to align with user assertions even when they conflict with facts. We frame sycophancy as a sociolinguistic phenomenon, disentangling two distinct drivers of credibility: explicit authority (credentials) and implicit authority (linguistic register). We introduce the Sycophancy Matrix, an adversarial evaluation framework that isolates these variables. Using a controlled subset of TruthfulQA, we evaluate open-weight models across English, Spanish, and Portuguese variants. Our findings reveal that models often conflate high register with truthfulness: for some architectures, sophisticated tone triggers deference more effectively than explicit expertise. Furthermore, we observe statistically significant variability across cultural variants of Spanish and Portuguese, supporting the hypothesis that LLMs internalise language-specific sociolinguistic norms and that sycophancy is not a purely technical deficit but an emergent property of multilingual training and alignment. Finally, we identify stable sycophancy fingerprints–domain-specific vulnerability profiles that persist across languages–suggesting that alignment artefacts are intrinsic to model families rather than linguistic context.

[Long] SPARKLE: A Structured and Plug-and-play Agentic Retrieval Policy for Adaptive RAG Models ▶ *Underline Video*

Jinyuan Fang, Zaiqiao Meng and Craig Macdonald

Adaptive retrieval-augmented generation (RAG) models offer an effective approach for integrating external knowledge. However, existing methods either rely on frozen large language models (LLMs) without explicit supervision or require costly LLM finetuning. Therefore, we propose SPARKLE, a structured and plug-and-play agentic retrieval policy where an additional proxy model is introduced to control the retrieval process. The proxy model leverages knowledge graph-based reasoning to make retrieval decisions in a structured manner, while operating independently of the retriever and the LLM. This plug-and-play design allows SPARKLE to generalise across different retrievers and LLMs. SPARKLE is optimised via reinforcement learning (RL), treating the retriever and the LLM as part of the environment. To enable more effective exploration during RL training, we further introduce a binary tree-structured rollout strategy. Experiments on three in-domain and four out-of-domain QA benchmarks show that SPARKLE outperforms state-of-the-art adaptive RAG baselines, achieving average improvements of 9.17% and 2.85%, respectively.

[Findings] SpiralThinker: Latent Reasoning through an Iterative Process with Text-Latent Interleaving ▶ *Underline Video*

Shengmin Piao and Sanghyun Park

Recent advances in large reasoning models have been driven by reinforcement learning and test-time scaling, accompanied by growing interest in latent rather than purely textual reasoning. However, existing latent reasoning methods lack mechanisms to ensure stable reasoning dynamics in latent space and a systematic way to interleave implicit and explicit reasoning. We introduce SpiralThinker, a unified framework that performs iterative updates over latent representations while enabling interleaved reasoning across latent and textual steps. At its core, SpiralThinker employs a progressive alignment objective and structured annotations to stabilize latent reasoning and maintain coherence with textual reasoning. Across mathematical, logical, and commonsense reasoning tasks, SpiralThinker achieves state-of-the-art performance among latent reasoning baselines. Detailed analyses reveal that both iteration and alignment are indispensable, the numbers of latent tokens and iterations exhibit dataset-specific optima, and appropriate alignment proves critical for an effective iterative process. Overall, SpiralThinker bridges iterative computation and latent reasoning, demonstrating that aligned iterative updates can reliably steer reasoning in the latent space.

[Long] SRA: Span Representation Alignment for Large Language Model Distillation ▶ *Underline Video*

Quoc Phong Dao, Hoang Son Nguyen, Pham Khanh Chi, Tung Nguyen, Linh Ngo Van, Nguyen Thi Ngoc Diep and Trung Le

Cross-Tokenizer Knowledge Distillation (CTKD) enables knowledge transfer between a large language model and a smaller student, even when they employ different tokenizers. While existing approaches mainly focus on token-level alignment strategies, which are often brittle and sensitive to discrepancies between tokenizers, we argue that the method of aggregating tokens into more robust representations before distillation is of equal importance. In this paper, we introduce SRA (Span Representation Alignment for Large Language Model Distillation), a novel framework that reframes CTKD through the physical lens of Multi-Particle Dynamical Systems. SRA shifts the fundamental unit of alignment from tokens to robust, tokenizer-agnostic spans. We model each span as a cluster of particles and represent its state by its Center of Mass (CoM) - an attention-weighted average that captures rich semantic information. We leverage the concept of span centers of mass

with attention-derived weighting to prioritize the most salient spans. In addition, we employ a geometric regularizer to preserve the structural integrity of the representation space and introduce aligned span logit distillation to enhance knowledge transfer across models. In challenging cross-architecture distillation experiments, SRA consistently and significantly outperforms state-of-the-art CTKD baselines, validating our physically-grounded approach.

[Long] Stop Fixating on Prompts: Reasoning Hijacking and Constraint Tightening for Red-Teaming LLM Agents ► *Underline Video*

Yanxu Mao, Peipei Liu, Tiehan Cui, Congying Liu, Mingzhe Xing and Datao You

With the widespread application of LLM-based agents across various domains, their complexity has introduced new security threats. Existing red-team methods mostly rely on modifying user prompts, which lack adaptability to new data and may impact the agent's performance. To address the challenge, this paper proposes the JailAgent framework, which completely avoids modifying the user prompt. Specifically, it implicitly manipulates the agents reasoning trajectory and memory retrieval with three key stages: Trigger Extraction, Reasoning Hijacking, and Constraint Tightening. Through precise trigger identification, real-time adaptive mechanisms, and an optimized objective function, JailAgent demonstrates outstanding performance in cross-model and cross-scenario environments.

[Long] Suggest-Verify-Revise: A Three-Stage Document-Level Event Causality Identification with Narrative Consistency ► *Underline Video*

Ya Su, Hu zhang, Dan Qiao, YuJie Wang, Yunxiao Zhao, Yue Fan, Shike Li, Ru Li and Hongye Tan

Document-level Event Causality Identification (DECI) aims to identify causal relations among multiple events within unstructured text. Existing methods predominantly rely on local semantic similarity for independent event-pair discrimination, thereby overlooking the influence of the overall narrative backbone in the propagation of causal dependencies and the role differentiation of events within multi-cause/multi-effect structures. Therefore, we propose a suggest-verify-revise approach for document-level Event Causality Identification with narrative consistency (SVRECI). In the suggest stage, we integrate multi-dimensional heuristic causal suggestions generated by an LLM with structural suggestions derived from hypergraph modeling to provide multi-source initial support for candidate event pairs. In the verify stage, we introduce a Topological Hawkes process to perform constrained verification of narrative propagation consistency among events. In the revise stage, we construct a dynamically evolving document-level causal graph and incorporate a structure-aware dual-level contrastive learning mechanism at both the event and event-pair levels, iteratively reducing noisy edges over multiple iterations. Experimental results on EventStoryLine and Causal-TimeBank datasets demonstrate that our approach outperforms previous methods.

[Long] Surprisal Minimisation over Goal-directed Alternatives Predicts Production Choice in Dialogue ► *Underline Video*

Tom Utting; Mario Giulianelli; Arabella Sinclair

We model utterance production as probabilistic cost-sensitive choice over contextual alternatives, using information-theoretic notions of cost. We distinguish between goal-directed alternatives that realise a fixed communicative intent and goal-agnostic alternatives defined only by contextual plausibility, allowing us to derive speaker- and listener-oriented interpretations of different cost measures. We present a procedure to generate both types of alternative sets using language models. Analysing production choices in open-ended dialogue under both deterministic and noisy cost minimisation, we find that surprisal minimisation relative to goal-directed alternatives provides the strongest explanation of production choices. Uniformity-based costs show weaker overall predictive power, but their influence increases markedly when evaluated relative to goal-agnostic alternatives, consistent with listener-oriented accounts. More broadly, our study suggests that alternative-conditioned optimisation with LM-generated alternatives provides a principled framework for studying speaker and listener pressures in naturalistic language production.

[Long] Tailored Primitive Initialization is the Secret Key to Reinforcement Learning ► *Underline Video*

Yihang Yao, Guangtao Zeng, Raina Wu, Yang Zhang, Ding Zhao, Zhang-Wei Hong and Chuang Gan

Reinforcement learning (RL) has emerged as a powerful paradigm for improving the reasoning capabilities of large language models (LLMs). Despite its success, RL faces fundamental challenges, including low sample efficiency and a strong dependence on the quality of the base model: while some models improve rapidly with limited RL updates, others require substantial training data to achieve meaningful gains. Recent studies suggest that the patterns of thinking tokens play a critical role in RL performance, and that supervised fine-tuning (SFT) on datasets exhibiting desirable reasoning patterns can reduce reliance on base models and better prepare LLMs for RL. However, how to automatically discover such patterns across tasks remains unclear. In this work, we describe thinking token patterns with reasoning primitives and argue that initializing LLMs with diverse, high-quality primitives is crucial for stable and efficient RL training. We propose Tailor, a pipeline that automatically discovers such reasoning primitives and curates SFT datasets to prepare LLMs for RL. Extensive experiments on mathematical and logical reasoning benchmarks demonstrate that Tailor consistently improves downstream RL performance, outperforming strong baselines, including methods with expert domain knowledge.

[Long] TALAS: Teacher-Anchored Layer Alignment with Adaptive Sharpness-Aware Minimization for Embedding Distillation ► *Underline Video*

Quoc Phong Dao, Hoang Son Nguyen, Pham Khanh Chi, Linh Ngo Van, Nguyen Thi Ngoc Diep, Thien Huu Nguyen and Trung Le

Knowledge Distillation (KD) has established itself as a pivotal technique for compressing large pre-trained language models. However, existing methods that force a student to strictly mimic the teachers sentence embeddings or internal features often incur prohibitive computational costs and yield suboptimal performance due to the inherent capacity gap. To address these challenges, we propose TALAS

(Teacher-Anchored Layer Alignment with Sharpness-aware minimization), a unified framework that synergizes hierarchical (multi-layer) alignment with robust optimization. First, we introduce a Teacher-Anchored mechanism that selectively distills final sentence embeddings only into the student's upper layers, thereby reducing overhead while respecting capacity constraints. Second, we bridge the semantic gap in lower layers via Layer-Aligned Self-Distillation, which propagates knowledge top-down using internal geometric relational constraints in the embedding space. Finally, to prevent the student from memorizing point-wise teacher noise, we integrate Adaptive Sharpness-Aware Minimization (ASAM) into the training objective, guiding the model towards flat minima for enhanced generalization. Empirical results on standard sentence embedding benchmarks demonstrate that TALAS consistently outperforms strong distillation baselines while achieving superior training efficiency in terms of computational cost and memory footprint.

[Long] TAMAS: Benchmarking Adversarial Risks in Multi-Agent LLM Systems ► *Underline Video*

Ishan Kavathekar, Hemang Jain, Ameya Rathod, Ponnurangam Kumaraguru and Tanuja Ganu

Large Language Models (LLMs) have demonstrated strong capabilities as autonomous agents through tool use, planning, and decision-making abilities, leading to their widespread adoption across diverse tasks. As task complexity grows, multi-agent LLM systems are increasingly used to solve problems collaboratively. However, safety and security of these systems remains largely under-explored. Existing benchmarks and datasets predominantly focus on single-agent settings, failing to capture the unique vulnerabilities of multi-agent dynamics and co-ordination. To address this gap, we introduce Threats and Attacks in Multi-Agent Systems (TAMAS), a benchmark designed to evaluate the robustness and safety of multi-agent LLM systems. TAMAS includes five distinct scenarios comprising 300 adversarial instances across six attack types and 211 tools, along with 100 harmless tasks. We assess system performance across ten backbone LLMs and three agent interaction configurations from Autogen and CrewAI frameworks, highlighting critical challenges and failure modes in current multi-agent deployments. Furthermore, we introduce Effective Robustness Score (ERS) to assess the tradeoff between safety and task effectiveness of these frameworks. Our findings show that multi-agent systems are highly vulnerable to adversarial attacks, underscoring the urgent need for stronger defenses. TAMAS provides a foundation for systematically studying and improving the safety of multi-agent LLM systems. Code and dataset is available at https://github.com/microsoft/TAMAS.

[Long] Taming Actor-Observer Asymmetry in Agents via Dialectical Alignment ► *Underline Video*

Bobo Li, Rui Wu, Zibo Ji, Meishan Zhang, Hao Fei, Min Zhang, Mong-Li Lee and Wynne Hsu

Large Language Model agents have rapidly evolved from static text generators into dynamic systems capable of executing complex autonomous workflows. To enhance reliability, multi-agent frameworks assigning specialized roles are increasingly adopted to enable self-reflection and mutual auditing. While such role-playing effectively leverages domain expert knowledge, we find it simultaneously induces a human-like cognitive bias known as Actor-Observer Asymmetry (AOA). Specifically, an agent acting as an actor (during self-reflection) tends to attribute failures to external factors, whereas an observer (during mutual auditing) attributes the same errors to internal faults. We quantify this using our new Ambiguous Failure Benchmark, which reveals that simply swapping perspectives triggers the AOA effect in over 20% of cases for most models. To tame this bias, we introduce ReTAS (Reasoning via Thesis-Antithesis-Synthesis), a model trained through dialectical alignment to enforce perspective-invariant reasoning. By integrating dialectical chain-of-thought with Group Relative Policy Optimization, ReTAS guides agents to synthesize conflicting viewpoints into an objective consensus. Experiments demonstrate that ReTAS effectively mitigates attribution inconsistency and significantly improves fault resolution rates in ambiguous scenarios.

[Long] TaxPraBen: A Scalable Benchmark for Structured Evaluation of LLMs in Chinese Real-World Tax Practice ► *Underline Video*

Gang Hu, Yating Chen, Haiyan Ding, Wang Gao, Jiajia Huang, Min Peng, Qianqian Xie and Kun Yue

While Large Language Models (LLMs) excel in various general domains, they exhibit notable gaps in the highly specialized, knowledge-intensive, and legally regulated Chinese tax domain. Consequently, while tax-related benchmarks are gaining attention, many focus on isolated NLP tasks, neglecting real-world practical capabilities. To address this issue, we introduce TaxPraBen, the first dedicated benchmark for Chinese taxation practice. It combines 10 traditional application tasks, along with 3 pioneering real-world scenarios: tax risk prevention, tax inspection analysis, and tax strategy planning, sourced from 14 datasets totaling 7.3K instances. TaxPraBen features a scalable structured evaluation paradigm designed through process of "structured parsing, field alignment extraction, numerical and textual matching", enabling end-to-end tax practice assessment while being extensible to other domains. We evaluate 19 LLMs based on Bloom's taxonomy. The results indicate significant performance disparities: all closed-source large-parameter LLMs excel, and Chinese LLMs like Qwen2.5 generally exceed multilingual LLMs, while the YaYi2 LLM, fine-tuned with some tax data, shows only limited improvement. TaxPraBen serves as a vital resource for advancing evaluations of LLMs in practical applications.

[Long] Teaching Language Models to Check Grounded Claim Factuality with Human Test-Taking Strategies ► *Underline Video*

Yuxuan Ye, Raul Santos-Rodriguez and Edwin Simpson

Grounded claim factuality checking is important for large language model (LLM) applications such as retrieval-augmented generation, as it helps users assess the correctness of generated outputs. Existing metrics using entailment classifiers require dataset-specific threshold tuning, while LLM-based approaches often use direct prompting, which underutilises the reasoning capabilities of LLMs. We address this by formulating grounded claim factuality checking as a true/false reading comprehension task and prompting LLMs with explicit test-taking strategies for efficient reasoning. Our method reduces token usage by over 80% compared to unguided open-ended reasoning, and achieves

competitive performance to more expensive alternatives across two factuality benchmarks, setting a new state of the art on one. To further reduce inference cost, we train small language models (SLMs) to replace LLMs in the checking pipeline. Using supervised fine-tuning (SFT) and a self-revision mechanism, the SLMs learn to improve their factuality judgements. Experimental results show that the resulting SLMs perform on par with strong baselines, combining low inference costs with generating supporting rationales to support interpretability. Code and datasets will be released upon acceptance.

[Long] The Digital Dunning-Kruger Effect: Decoupling Hallucinations via Geometric Hidden-state Observation for Semantic Truthfulness ▶ *Underline Video*

Yueheng Mao, Min Yu, Gengwang Li, Jianguo Jiang, Gang Li, Meng Zhang, Zhen Xu, Weiqing Huang and Ming Liu

Large Language Models (LLMs) often generate overconfident yet factually incorrect hallucinations. Current detection paradigms suffer from a trade-off between the high accuracy of computationally expensive black-box methods and the inability of white-box methods to detect stubborn hallucinations. To bridge this gap, we propose GHOST (Geometric Hidden-state Observation for Semantic Truthfulness), an efficient white-box framework for hallucination detection in LLMs. We primarily target confused hallucinations marked by internal reasoning instability, while also capturing stubborn hallucinations characterized by premature layer-wise convergence as a complementary signal. By integrating internal geometric dynamics with output probability distributions, GHOST constructs a high-dimensional feature space for non-linear truthfulness classification. Extensive evaluations on FinanceBench, RAGTruth, HaluEval, and PopQA show that GHOST outperforms white-box baselines and achieves competitive black-box performance while reducing computational overhead by over 90%, offering a robust solution for real-time detection.

[Long] The Dominance of Text Space: Unveiling the Asymmetric Nature of Cross-Modal Alignment in Large Language Models ▶ *Underline Video*

Linqing Chen, Hanmeng Zhong, Wentao Wu and Peng Zhou

Recent advancements in Multimodal Large Language Models (MLLMs) have largely been driven by aligning visual encoders with pre-trained Large Language Models (LLMs). While effective, the geometric nature of this alignment remains under-explored. Existing methods often assume a symmetric interaction between visual and textual modalities, implying that both spaces adapt to each other. In this paper, we challenge this assumption and propose the "Text Space as Anchor" hypothesis. We argue that the semantic space of LLMs is rigid, anisotropic, and dominant; thus, effective cross-modal alignment may be an asymmetric projection of visual features onto this pre-existing text manifold without distorting it. We identify a potential issue in current parameter-efficient tuning paradigms where task-specific visual adjustments inadvertently disrupt the projector's geometry, leading to "catastrophic forgetting" of the alignment mechanism itself. To address this, we introduce Anchor-Preserving Projection (APP), a novel method that regularizes the projector to maintain the geometric structure of the text embedding space via spectral filtering. Extensive experiments on 8 diverse cross-modal tasks and 3 pure language benchmarks demonstrate that APP preserves the LLM's inherent linguistic capabilities (e.g., MMLU, GSM8K) and reduces object hallucination significantly better than standard fine-tuning methods. We release our code.

[Demo] The olmOCR Project: Building Fully Open OCR using VLMs ▶ *Underline Video*

Jake Poznanski, Kyle Lo and Luca Soldaini

We present olmOCR, a fully open OCR system developed through iterative public releases and community feedback. The system combines a 7B vision-language model trained in two stages: supervised finetuning on 260K diverse PDF pages, followed by reinforcement learning with visual unit tests over synthetic documents. Visual unit tests are binary checks of structural fidelity, including tables and equations, and serve both as an interpretable evaluation framework and as direct optimization targets. We also introduce olmOCR-Bench, a benchmark of 1.4K challenging PDFs evaluated via visual unit tests, on which olmOCR achieves state-of-the-art performance among open systems and proprietary APIs at a fraction of the cost. We have deployed olmOCR at scale to 100M+ PDFs to curate pretraining data for Olmo 3. We share lessons from our open development process and release all models, data, and code across two major releases.

[SRW] Thesis Proposal: Self-Adaptive and Epistemic Uncertainty-Guided ASR of Dense Intra-Sentential Code-Switched Speech for African Low-Resource Languages ▶ *Underline Video*

Umar Baba Umar

Automatic Speech Recognition (ASR) has achieved strong performance for high-resource languages, but dense intra-sentential code-switched speech in African low-resource settings remains underexplored. Existing multilingual and pretrained ASR systems improve general recognition accuracy, yet they remain weak at switch regions, are sensitive to language imbalance during adaptation, and are typically evaluated with metrics that obscure switching-specific errors. This thesis proposes a self-adaptive and epistemic uncertainty-guided framework for African low-resource code-switched ASR, using Hausa–English (Engausa) and Hausa–Yor\{\}‘ub\{\}’a as case studies. The work investigates three linked questions: (1) how to design a linguistically informed code-switched corpus with explicit switch-region annotation and labeled/unlabeled partitions for adaptive learning, (2) whether epistemic uncertainty is systematically elevated around switch regions and can guide pseudo-label selection in semi-supervised training, and (3) whether switch-aware adaptation with auxiliary language identification and boundary supervision can reduce recognition errors without increasing catastrophic forgetting. The long-term goal is to develop scalable and data-efficient ASR systems that model code-switching as a structured linguistic phenomenon rather than as noise in multilingual African speech.

[Long] Think Before Writing: Feature-Level Multi-Objective Optimization for Generative Citation Visibility ▶ *Underline Video*
Zikang Liu and Peilan Xu
Generative answer engines expose content through selective citation rather than ranked retrieval, fundamentally altering how visibility is determined. This shift calls for new optimization methods beyond traditional search engine optimization. Existing generative engine optimization (GEO) approaches primarily rely on token-level text rewriting, offering limited interpretability and weak control over the trade-off between citation visibility and content quality. We propose FeatGEO, a feature-level, multi-objective optimization framework that abstracts webpages into interpretable structural, content, and linguistic properties. Instead of directly editing text, FeatGEO optimizes over this feature space and uses a language model to realize feature configurations into natural language, decoupling high-level optimization from surface-level generation. Experiments on GEO-Bench across three generative engines demonstrate that FeatGEO consistently improves citation visibility while maintaining or improving content quality, substantially outperforming token-level baselines. Further analyses show that citation behavior is more strongly influenced by document-level content properties than by isolated lexical edits, and that the learned feature configurations generalize across language models of different scales.

[SRW] Think Less, Code Better: Probing When Chain-of-Thought Hurts and How to Route Around It ▶ *Underline Video*
Rajarshi Ghoshal, Debadri Basak, Salma E. Abdelhalim and Pratibha K. Arora
Chain-of-Thought (CoT) prompting is the dominant strategy for eliciting step-by-step reasoning in LLMs, yet its effect on model behavior remains poorly understood. We present a controlled study of when reasoning augmentation helps versus hurts in code generation , a task requiring deductive reasoning under formal constraints , spanning three architectures: a 2x2 design with Qwen2.5-Coder-1.5B and DeepSeek-Coder-1.3B (base and instruction-tuned variants) on HumanEval, MBPP, and LiveCodeBench, plus CodeLlama-7B. Our key finding is that instruction tuning reverses CoT's effect on the same base architecture: CoT significantly improves Qwen base (+13.4%, p<0.001) but significantly degrades Qwen instruct (-15.2%, p<0.001). DeepSeek remains insensitive regardless of training regime (all p>0.2). To explain why, we apply layer-wise linear probing and find that all models encode prompt type (Direct vs. CoT) by Layer 1-4 (>90% accuracy) , universally, whether CoT helps or hurts. This reveals a key mechanistic insight: representation does not determine interpretation. The same internal signal for "this is a reasoning prompt" drives opposite behavioral outcomes depending on training regime. Building on this, we develop a probe-guided style router that selects from 12 prompt styles per problem via a single forward pass (84ms). The router is statistically indistinguishable from the best fixed style in 7/8 settings (McNemar's, all p_Best>0.1) and significantly outperforms CoT where CoT is most harmful (p=0.012, h=+0.40). Our findings argue for model-aware prompt strategies informed by internal representations rather than one-size-fits-all reasoning augmentation.

[Long] Thinking Alignment of Scenario-Oriented User Simulation ▶ *Underline Video*
Xiaoting Wu, Yi Huang, Chunyang Gao, Mengfei Guo, Jingyu Yao and Junlan Feng
Existing user simulators based on prompting to role-play or SFT are generally confined to imitating users textual utterances, without adequately considering the multi-faceted cognitive processes that underlie human decision-making during interactions. To facilitate better alignment with real human thinking patterns, we construct the LMSYS-UserThinking dataset, in which we augment 51k human-LLM conversations by reconstructing the users inner reasoning both during and at the end of each dialogue. Furthermore, to enhance controllability and situational coherence, we introduce scenario settings that describe the global context and user goals throughout multi-turn conversations. Using this dataset, we train user simulators called ThinkingUS on different base models. We evaluate our approach from both offline and online user simulation perspectives, ultimately demonstrating its effectiveness.

[Industry] ThLexGraph: A Production Knowledge Graph Unifying Thailands Legal System ▶ *Underline Video*
Pawitsapak Akarajaradwong, Sarana Nutanong and Chompakorn Chaksangchaichot;
Jurisdictionally bound domains, such as law, often lack standardized, machine-readable data formats, requiring foundational infrastructure before downstream applications can succeed. We present ThLexGraph, the first unified temporal knowledge graph for Thai legal data, integrating 3,840 laws (6,273 versions) with 87,394 Supreme Court decisions, updated daily. The graph encodes hierarchy, temporal versioning, cross-references, and sequential order, all extracted from unstructured official sources where no structured representation previously existed. A five-setting comparison on NitiBench-Tax isolates data infrastructure as the sole variable: graph-structured retrieval achieves Citation F1 of 0.812 versus 0.666 for practitioner-standard web search and 0.685 for flat vector retrieval, while searching a corpus 53x larger. Trace analysis of 820 agent-issued queries reveals that hierarchy traversal and cross-reference following, capabilities absent from generic retrieval, are exercised in 50% and 16% of questions, respectively. Our system demonstrates that structured modeling of hierarchy, temporal versioning, cross-references, and sequential order can overcome structural limitations of legal data published without standardized formats.

[Long] TimeSAF: Towards LLM-Guided Semantic Asynchronous Fusion for Time Series Forecasting ▶ *Underline Video*
Fan Zhang, Shiming Fan and Hua Wang
Despite the recent success of large language models (LLMs) in time-series forecasting, most existing methods still adopt a Deep Synchronous Fusion strategy, where dense interactions between textual and temporal features are enforced at every layer of the network. This design overlooks the inherent granularity mismatch between modalities and leads to what we term semantic perceptual dissonance: high-level abstract semantics provided by the LLM become inappropriately entangled with the low-level, fine-grained numerical dynamics of time series, making it difficult for semantic priors to effectively guide forecasting. To address this issue, we propose TimeSAF, a new framework based on hierarchical asynchronous fusion. Unlike synchronous approaches, TimeSAF explicitly decouples unimodal feature

learning from cross-modal interaction. It introduces an independent cross-modal semantic fusion trunk, which uses learnable queries to aggregate global semantics from the temporal and prompt backbones in a bottom-up manner, and a stage-wise semantic refinement decoder that asynchronously injects these high-level signals back into the temporal backbone. This mechanism provides stable and efficient semantic guidance while avoiding interference with low-level temporal dynamics. Extensive experiments on standard long-term forecasting benchmarks show that TimeSAF significantly outperforms state-of-the-art baselines, and further exhibits strong generalization in both few-shot and zero-shot transfer settings.

[Long] To Judge or Not to Judge: Can Large Language Models Leverage the Dispute Focus in Legal Judgment? ▶ *Underline Video*

Luoming Hu, Liang Yang, Jingjie Zeng and Zijie Xing

Civil judicial cases are highly complicated, posing significant challenges for Large Language Models (LLMs) for Legal Judgment Prediction (LJP). While judges manage this complexity through the dispute focus, a mechanism distilling cases into core issues, existing research largely overlooks this tool in favor of generic reasoning frameworks that lack authentic judicial logic. To bridge this gap, we first introduce FocalLaw, the first dataset aligning full-process Chinese civil judicial data through the dispute focus, comprising 1,000 high-quality cases across six causes of action. Building on this dataset, we examine LLMs' capability to utilize the dispute focus and uncover a counter-intuitive phenomenon: LLMs fail to leverage the dispute focus even with CoT and SFT, which we identify as the "Clerk Trap". To solve the problem, we propose FocalJudge, a novel framework that leverages the dispute focus to guide LLMs through a structured, judge-like cognitive workflow. Experimental results demonstrate the effectiveness of FocalJudge and offer valuable insights into the interpretability and reliability of LLMs in the legal domain.

[Long] Too Long, Do Re-weighting for Efficient LLM Reasoning Compression ▶ *Underline Video*

Zhong-Zhi Li, Xiao Liang, Zihao Tang, Lei Ji, Peijie Wang, Haotian Xu, Xing W, Haizhen Huang, Weiwei Deng, Yeyun Gong, Zhijiang Guo, Fei Yin and Cheng-Lin Liu

Large Language Models (LLMs) have recently achieved remarkable progress on complex reasoning tasks by leveraging extended Chain-of-Thought (CoT) techniques. These reasoning processes can be roughly categorized into System-1 (fast and intuitive) and System-2 (slow and deliberate) paradigms. However, excessive reliance on lengthy System-2-style reasoning during inference can produce extremely long outputs, thereby reducing efficiency. In this work, we propose Thinking Length Data Re-weighting (TLDR), that does not rely on sophisticated data annotations or interpolation between multiple models. We continuously balance the weights between the model's System-1 and System-2 data to eliminate redundant reasoning processes while preserving the model's reasoning capability. We validate our method across multiple base models, including Deepseek-R1-Distilled Qwen models, as well as on a diverse benchmarks with varying difficulty levels. Our method significantly reduces the number of output tokens by nearly 40% while maintaining the accuracy of the reasoning.

[Findings] ToolCPT: Improving Tool Utilization in LLM Agents via Continuous Pre-training ▶ *Underline Video*

Yifan Yang, Jinghui Lu, Yaping Deng, Ao Yang, Peijie Yu, Tinghao Yu and Feng Zhang

Autonomous agents powered by large language models (LLM-based agents) are capable of using off-the-shelf tools to interact with the environment, solve real-world problems, and boost work efficiency. However, current approaches to enhancing tool use for LLM-based agents primarily focus on post-training fine-tuning or test-time context extension. These methods overlook the fundamental tool knowledge acquisition during the early training phase, where models actually learn and internalize core knowledge representations, restricting model performance on out-of-distribution tool usage. To solve such a problem, we introduce enhancing tool knowledge for LLM-based agents during continuous pre-training (ToolCPT). We identify and bridge a key gap in current LLM training by shifting focus from tool-calling patterns to deep internalization of core tool-knowledge representations. We begin by curating 5.1 million code artifacts from large-scale, high-quality code repositories. These artifacts are selected based on a set of criteria that defines a usable "proxy agent tool", thereby forming a comprehensive agent tool library. For each proxy tool, we then create a detailed playbook covering implementation specifications, core functionalities, interaction protocols with other tools, and illustrative positive and negative examples. This process yields a large-scale tool knowledge corpus comprising 18 billion tokens, which is used to continuously pre-train our model. Experiments show our playbook-enhanced corpus catalyzes deep knowledge internalization, driving the model to notable performance gains on multiple standard benchmarks.

[Long] Toward Robust In-Context Learning: Leveraging Out-of-distribution Proxies for Target Inaccessible Demonstration Retrieval ▶ *Underline Video*

Hao Xu, rite Bo, Fausto Giunchiglia, Yingji Li and Rui Song

Although studies have demonstrated that Large Language Models (LLMs) can perform well on Out-of-Distribution (OOD) tasks, their advantage tends to diminish as the distribution shift becomes more severe. Consequently, researchers aim to retrieve distributionally similar and informative demonstrations from the available source domain to boost the inference capabilities of LLMs. However, in practical scenarios where the target domain is inaccessible, evaluating the unknown distribution is challenging, which indirectly impacts the quality of the selected demonstrations. To address this problem, we propose DOPA, a demonstration search framework that incorporates an OOD proxy to approximate the inaccessible target domain and guide the retrieval process. Building on proxy-based evaluation, DOPA further introduces a Mahalanobis distance-based global diversity constraint to ensure sufficient diversity among the retrieved demonstrations. Experimental results on multiple LLMs and tasks demonstrate that DOPA effectively enhances robustness in OOD settings.

[Long] Towards Fast and Accurate Modeling for Cross-Lingual Label Projection ▶ *Underline Video*

Thang Le, Huy Huu Nguyen, Anh Tuan Luu, Thamar Solorio and Thien Huu Nguyen

Information extraction (IE) systems rely on structured data for training, but such annotated data is highly imbalanced across languages, with low-resource languages receiving little attention. Label projection techniques aim to bridge this gap by transferring structured annotations from high-resource to low-resource languages. However, existing methods are either inaccurate or too slow for large-scale use. This work aims to address this problem by developing a more effective method that remains sufficiently efficient for large-scale projection. In particular, we propose to synthesize alignment sequence pairs and fine-tune an encoder model with span alignment objective, while controlling data influence during training. Experimental results across 50+ languages show that our framework consistently outperforms previous state-of-the-art methods while maintaining fast inference speed. In addition, we introduce EXP - the first benchmark for explicit evaluation of label projection, thereby reducing confounders and non-determinism in method assessment.

[Long] Towards Fine-Grained and Multi-Granular Contrastive Language-Speech Pre-training ▶ *Underline Video*

Yifan Yang, Bing Han, Hui Wang, Wei Wang, Ziyang Ma, Long Zhou, Zengrui Jin, Guanrou Yang, Tianrui Wang, Xu Tan and Xie Chen

Modeling fine-grained speaking styles remains challenging for language-speech representation pre-training, as existing speech-text models are typically trained with coarse captions or task-specific supervision, and scalable fine-grained style annotations are unavailable. We present FCaps, a large-scale dataset with fine-grained free-text style descriptions, encompassing 47k hours of speech and 19M fine-grained captions annotated via a novel end-to-end pipeline that directly grounds detailed captions in audio, thereby avoiding the error propagation caused by LLM-based rewriting in existing cascaded pipelines. Evaluations using LLM-as-a-judge demonstrate that our annotations surpass existing cascaded annotations in terms of correctness, coverage, and naturalness. Building on FCaps, we propose CLSP, a contrastive language-speech pre-trained model that integrates global and fine-grained supervision, enabling unified representations across multiple granularities. Extensive experiments demonstrate that CLSP learns fine-grained and multi-granular speech-text representations that perform reliably across global and fine-grained speech-text retrieval, zero-shot paralinguistic classification, and speech style similarity scoring, with strong alignment to human judgments. Code and dataset are publicly available at https://github.com/yfyeung/CLSP.

[Findings] Towards Preference Following in Tool Calling Language Agents ▶ *Underline Video*

Zhi-Yuan Chen, Siyu Lu, Qianlong Xie, Xingxing Wang and Yankai Lin

Large language model (LLM)-based agents have demonstrated remarkable capabilities in tool use, but their ability to follow user preferences when calling tools remains underexplored. To address this gap, we introduce APOLLO, a benchmark designed to evaluate agents' ability to identify personalized user preferences from interaction histories and to adhere to these preferences when calling tools to solve user queries. In APOLLO, user preferences expressed in the interaction history take two forms: explicit preferences stated directly, and implicit preferences conveyed through behaviors such as option selection and comparison. In addition, the benchmark includes two types of queries, reactive and proactive, which pose challenges for LLMs to ground user queries in the corresponding preferences. Using APOLLO, we evaluate and analyze both language models and reasoning models, and investigate the impact of different agent frameworks, such as Reflexion, on model performance. Experimental results show that current models still struggle to follow user preferences when calling tools. For instance, GPT-4o achieves only 51.16% accuracy on the benchmark. Furthermore, we develop a reinforcement learning-based approach to improve LLMs, achieving substantial performance gains on APOLLO. Our dataset and code are publicly available at https://github.com/zhiyuanc2001/APOLLO.

[Long] Towards Robust Real-World Spreadsheet Understanding with Multi-Agent Multi-Format Reasoning ▶ *Underline Video*

Houxing Ren, Mingjie Zhan, Zimu Lu, Ke Wang, Yunqiao Yang, Haotian Hou and Hongsheng Li

Spreadsheets are central to real-world applications such as enterprise reporting, auditing, and scientific data management. Despite their ubiquity, existing large language model based approaches typically treat tables as plain text, overlooking critical layout cues and visual semantics. Moreover, real-world spreadsheets are often massive in scale, exceeding the input length that LLMs can efficiently process. To address these challenges, we propose SpreadsheetAgent, a two-stage multi-agent framework for spreadsheet understanding that adopts a step-by-step reading and reasoning paradigm. Instead of loading the entire spreadsheet at once, SpreadsheetAgent incrementally interprets localized regions through multiple modalities, including code execution results, images, and LaTeX tables. The method first constructs a structural sketch and row/column summaries, and then performs task-driven reasoning over this intermediate representation in the Solving Stage. To further enhance reliability, we design a verification module that validates extracted structures via targeted inspections, reducing error propagation and ensuring trustworthy inputs for downstream reasoning. Extensive experiments on two spreadsheet datasets demonstrate the effectiveness of our approach. With GPT-OSS-120B, SpreadsheetAgent achieves 38.16% on Spreadsheet Bench, outperforming the ChatGPT Agent baseline (35.27%) by 2.89 absolute points. These results highlight the potential of SpreadsheetAgent to advance robust and scalable spreadsheet understanding in real-world applications. Code is available at https://github.com/renhouxing/SpreadsheetAgent.

[Findings] TOXIFRENCH: Benchmarking and Enhancing Language Models via CoT Fine-Tuning for French Toxicity Detection ▶ *Underline Video*

Axel Delaval, Shujian Yang, Haicheng Wang, Han Qiu and Jialiang LU

Detecting toxic content using language models is crucial yet challenging. While substantial progress has been made in English, toxicity detection in French remains underdeveloped, primarily due to the lack of culturally relevant, human-annotated, large-scale datasets. In this work, we release TOXIFRENCH, a dataset of 53,622 French online comments, together with a 1,388-sample balanced benchmark split for

systematic evaluation. The dataset is constructed via a semi-automated annotation pipeline that reduces manual labeling to only 10% through high-confidence LLM-based pre-annotation and human verification, while ensuring statistically near-perfect alignment with human-only annotation. We then benchmark a broad range of models and uncover a counterintuitive insight: Small Language Models (SLMs) often surpass larger models in robustness and generalization on this task. Motivated by this finding, we propose a novel Chain-of-Thought (CoT) fine-tuning strategy using a dynamic weighted loss that progressively emphasizes the model's final decision, significantly improving faithfulness. Our fine-tuned 4B model (Qwen3-4B) achieves state-of-the-art performance on the benchmark, improving its balanced accuracy by 10% over its baseline and achieving better performance than GPT-4o and DeepSeek-R1 on our benchmark, while successfully retaining cross-lingual capabilities.

[Long] TPS-Bench: Evaluating AI Agents Tool Planning & Scheduling Abilities in Compounding Tasks ▶ *Underline Video*

Hanwen Xu, Xuyao Huang, Yuzhe Liu and Zhijie Deng

Large language model (LLM) agents have exhibited strong problem-solving competence across domains like research and coding. Yet, it remains underexplored whether LLM agents can tackle compounding real-world problems that require a diverse set of tools to complete. Given a broad, heterogeneous tool repository, LLM agents must not only select appropriate tools based on task planning analysis but also strategically schedule the execution order to ensure efficiency. This paper introduces TPS-Bench to benchmark the ability of LLM agents in solving such problems that demand Tool Planning and Scheduling. TPS-Bench collects 200 compounding tasks of two difficulty levels, based on a tool repository containing hundreds of model context protocol (MCP) tools. In particular, each task is composed of multiple subtasks, such as web search, map navigation, calendar checking, etc., and each subtask can be completed by a basic tool. Our evaluation emphasizes both task completion rate and efficiency. The empirical studies on popular closed-source and open-source LLMs indicate that most models can perform reasonable tool planning, but differ in scheduling. For example, GLM-4.5 achieves an outperforming task completion rate of 64.72% with extensive sequential tool calls, hence suffering from significantly long execution time. By contrast, GPT-4o prioritizes parallel tool calls but achieves only a 45.08% completion rate. Considering reinforcement learning (RL) can be a viable way to improve the scheduling efficiency without compromising performance, we perform an initial study on Qwen3-1.7B and witness a 14% reduction in execution time alongside a 6% gain in task completion rate based on only 597 RL training samples.

[Long] TRAC: Teacher-Guided Token Reward with Adaptive Calibration for Robust Policy Optimization ▶ *Underline Video*

Sitong Wu, Haoru Tan, Xichen Zhang, Bin Xia, Wenhu Zhang, Xiaojuan Qi, Bei Yu and Jiaya Jia

Reinforcement Learning (RL) with sparse outcome rewards suffers from inefficient credit assignment in complex LLM reasoning tasks. While utilizing stronger LLMs as teachers to derive dense token-level supervision offers a cost-effective alternative to proprietary reward models, it relies on the flawed assumption that teachers are perfect oracles. In reality, teacher models exhibit capability limitations and uncertainty, producing noisy signals that make student policies susceptible to reward hacking. To address this, we propose Teacher Reward Adaptive Calibration (TRAC), a robust framework that filters noisy supervision by dynamically modulating teacher influence via a multi-granularity calibration mechanism. TRAC evaluates teacher reliability across three principled dimensions: problem-level expertise, trajectory-level discrimination, and token-level confidence. Furthermore, we integrate TRAC with Group Relative Policy Optimization (GRPO), formulating as TRAC-GRPO, which treats calibrated teacher-derived reward as an additive advantage reshaping term to ensure fair advantage estimation. Extensive experiments demonstrate that TRAC effectively mitigates teacher noise, significantly enhancing the reasoning capabilities and training stability of LLMs compared to standard baselines. The code will be available at: https://github.com/JIA-Lab-research/TRAC.

[Long] Translation with Thought: Difficulty-Adaptive Reasoning via Reinforcement Learning for Multi-Domain Machine Translation ▶ *Underline Video*

Yongshi Ye, Biao Fu, Chongxuan Huang, Yidong Chen and Xiaodong Shi

Multi-domain machine translation (MDMT) poses a unique challenge due to varying levels of linguistic complexity across domains. Inspired by human translators' ability to adapt reasoning effort based on difficulty, we propose TwT (Translation with Thought), a resource-rational framework that learns to modulate inference between intuitive and deliberate reasoning. TwT is trained in two stages: (1) supervised fine-tuning on difficulty-aware long chain-of-though traces distilled from DeepSeek-R1 and rewritten by GPT-4o to reflect human-like reasoning economy, and (2) reinforcement learning with a hybrid reward to optimize translation quality and reasoning efficiency. Evaluated on 15 benchmarks spanning in-domain and out-of-domain settings, as well as 3 seen and 59 unseen languages, with ablations across three backbone models, TwT-7B and TwT-14B outperform much larger SOTA reasoning models in translation quality, while reducing token usage by 32-60%. These results confirm that aligning translation behavior with cognitive principles enables robust generalization, high translation quality, and efficient reasoning in MDMT.

[Long] TransLLM: A Unified Multi-Task Large Language Model for Urban Transportation via Learnable Prompting ▶ *Underline Video*

Jiaming Leng, Yunying Bi, Chuan Qin, Zhenya Huang, Bing Yin, Haojie Ren, Yanyong Zhang and Chao Wang

Urban transportation systems require precise modeling of dynamic spatiotemporal patterns across diverse tasks, such as traffic forecasting, electric vehicle (EV) charging demand prediction, and taxi dispatch. Existing approaches suffer from two key limitations: traditional deep learning models are task-specific and lack generalization capabilities, whereas Large Language Models (LLMs) struggle with structured spatiotemporal data and numerical reasoning. To bridge this gap, we propose TransLLM, a unified multi-task framework that synergizes spatiotemporal encoding with LLM reasoning through learnable prompt composition. To enable LLMs to perceive complex

graph dependencies, we design a noise-augmented spatiotemporal encoder that projects structured signals into the LLM's embedding space. Furthermore, to overcome the rigidity of fixed prompt templates in heterogeneous traffic scenarios, we introduce an instance-level prompt routing mechanism trained via reinforcement learning. The framework operates by encoding spatiotemporal patterns into contextual representations, dynamically composing personalized prompts to guide LLM reasoning, and projecting the resulting representations through specialized output layers to generate task-specific predictions. Experiments on seven datasets and three tasks demonstrate that TransLLM outperforms many baselines, showing superior adaptability in both supervised and zero-shot settings with excellent generalization and robustness. Our code and data are available at https://github.com/lengjiaming/TransLLM.

[Long] TwiUSD: A Benchmark Dataset and Structure-Aware LLM Framework for User Stance Detection ▶ *Underline Video*

Fuqiang Niu, Zini Chen, Zhiyu Xie, Hu Huang, Qing Liao, Qianlong Wang, Genan Dai and Bowen Zhang

Political user-level stance detection is vital for analyzing polarization, yet progress is hindered by the scarcity of high-quality benchmarks integrating linguistic and social signals. Existing datasets, largely relying on noisy heuristic or distant supervision, limit model robustness and generalizability. To address this, we introduce TwiUSD, a large-scale, expert-annotated benchmark for political user-level stance detection with explicit social network structure. TwiUSD comprises 16,211 users and 47,757 tweets, labeled by domain experts using a protocol that integrates both user content and followee signals, ensuring high-quality annotations (kappa > 0.9). Building upon TwiUSD, we propose MRFG, a Multi-scale Relevance Filtering and Graph-aware framework that leverages large language models to filter stance-relevant followee content and adaptively routes features based on structural informativeness. This design enables robust stance prediction by jointly modeling semantic and relational cues. Extensive experiments show that MRFG significantly outperforms strong baselines, highlighting the importance of relevance filtering and structure-aware modeling.

[Long] Understanding the Behaviors of Environment-aware Information Retrieval ▶ *Underline Video*

Ruifeng Yuan, Chaohao Yuan, David Dai, Yu Rong, Hong Cheng, Hou Pong Chan and Chenghao Xiao

Recent retrieval-augmented generation (RAG) approaches have demonstrated strong capability in handling complex queries, yet current research overlooks a critical challenge: different retrievers require fundamentally different query formulation strategies for optimal performance. In this work, we present the first systematic analysis of how LLMs can learn to adapt their query formulation strategies for different retrievers via reinforcement learning (RL). Our empirical study reveals that RL effectively teaches an LLM to tailor its queries to specific retriever characteristics. We discover that different retrievers exhibit surprisingly distinct optimal query styles (e.g., descriptive vs. question-like), suggesting strategies learned for one retriever ineffective for another. We further show that performance can be enhanced by incorporating retriever-specific human guidance and by scaling model size. To facilitate learning over multi-retrieval-step trajectories, we introduce a branching-based rollout technique that improves training stability. Our work provides the first empirical evidence and actionable insights for building truly retriever-aware RAG systems. Code and resources are available at https://github.com/LCO-Embedding/Envs-aware-Information-Retrieval.

[Long] UniDataBench: Evaluating Data Analytics Agents Across Structured and Unstructured Data ▶ *Underline Video*

Han Weng, Zhou Liu, Yuanfeng Song, Xiaoming Yin, Xing Chen and Wentao Zhang

In real-world business environments, data is stored in a variety of sources, including structured relational databases, semi-structured databases, and unstructured files. The ability to extract reasonable insights across these diverse sources is integral to data-driven decision-making. Existing benchmarks, however, are limited in assessing agents' capabilities across these diverse data types. To address this gap, we introduce UniDataBench, a multi-source benchmark designed to evaluate the performance of data analytics agents in handling diverse data sources. Specifically, UniDataBench is constructed based on real-life industry analysis reports, employing a pipeline to synthesize data that aligns with authentic analytical trends. It encompasses diverse datasets spanning relational databases, CSV files, and NoSQL stores to reflect real-world business settings, and provides a unified framework for evaluating how effectively agents can explore multiple data formats, extract insights, and generate meaningful summaries and recommendations. Based on UniDataBench, we propose a novel LLM-based agent named ReActInsight, an autonomous agent that performs end-to-end analysis over diverse data sources by automatically discovering cross-source linkages, decomposing goals, and generating robust, self-correcting code to extract actionable insights. Our benchmark and agent together provide a framework for facilitating the development of data analytics agents in real-world applications.

[Long] UNIKIE-BENCH: Benchmarking Large Multimodal Models for Key Information Extraction in Visual Documents ▶ *Underline Video*

Yifan Ji, Zhipeng Xu, Zhenghao Liu, Zulong Chen, Qian Zhang, Zhibo Yang, Junyang Lin, Yu Gu, Ge Yu and Maosong Sun

Key Information Extraction (KIE) from real-world documents remains challenging due to substantial variations in layout structures, visual quality, and task-specific information requirements. Recent Large Multimodal Models (LMMs) have shown promising potential for performing end-to-end KIE directly from document images. To enable a comprehensive and systematic evaluation across realistic and diverse application scenarios, we introduce UNIKIE-BENCH, a unified benchmark designed to rigorously evaluate the KIE capabilities of LMMs. UNIKIE-BENCH consists of two complementary tracks: a constrained-category KIE track with scenario-predefined schemas that reflect practical application needs, and an open-category KIE track that extracts any key information that is explicitly present in the document. Experiments on 15 state-of-the-art LMMs reveal substantial performance degradation under diverse schema definitions, long-tail key fields, and complex layouts, along with pronounced performance disparities across different document types and scenarios. These findings underscore persistent challenges in grounding accuracy and layout-aware reasoning for LMM-based KIE. All codes and datasets are

available at https://github.com/NEUIR/UNIKIE-BENCH.

[Long] UniSRM: A Unified Speech Reward Model for Reasoning-Based Fine-grained Assessment ▶ *Underline Video*

Yuanyuan Wang, Dongchao Yang, Yayue Deng, Zhiyong Wu, Yiwen Guo, Helen Meng and Xixin Wu

Evaluating speech generation still relies heavily on human judgments, such as Mean Opinion Score (MOS), which are expensive, subjective, and difficult to reproduce at scale. While a few recent studies have begun to explore AudioLLM-based judge models, existing efforts typically target only a narrow set of scenarios (e.g., utterance-level quality or single-turn dialogue) and provide limited coverage of diverse speech generation tasks and evaluation dimensions. In this work, we propose UniSRM, a unified speech reward model that can support multi-dimensional, interpretable reward signals with reliable reasoning. To support training and evaluation, we introduce UniSRM-Data and UniSRM-Bench, covering speech evaluation tasks from utterance-level quality to context-level coherence. Based on this dataset, we present the unified speech reward model, UniSRM, with a two-stage pipeline that enables reasoning-based fine-grained assessment. Furthermore, we introduce Reasoning-Consistent Rewards to improve the reliability of the reasoning process. Experiments show that UniSRM delivers more reliable and human-aligned judgments across a broad range of speech evaluation tasks, offering a practical foundation for scalable and unified evaluation of speech quality.

[Long] Unveiling the Unknown: Open-Set Entity Typing via Two-Stage Generation ▶ *Underline Video*

Hu Chen, Binhan Yang and Wei Shen

Conventional fine-grained entity typing (FET) operates under the closed-set assumption, wherein all classified types are limited within a predefined type taxonomy derived from a knowledge base. As the world evolves, new entities of unknown types inevitably emerge in open environments, falling beyond the scope of the existing type taxonomy. To deal with this problem, in this paper, we investigate a novel and critical task: open-set entity typing (OSET), which aims to not only classify entity mentions within the known type taxonomy but also detect those outside it, termed as unknown-type instances. However, owing to the lack of exposure to unknown-type instances during training, existing FET models are susceptible to misclassify them as known types, limiting their practical effectiveness for this new OSET task. Moreover, manually collecting and annotating large-scale unknown-type instances is both time-consuming and labor-intensive in open environments. To mitigate this issue, we propose a two-stage generation model that automatically produces large-scale, high-quality and diverse pseudo unknown-type instances, beneficial for the tailor-designed unified open-set classifier to effectively distinguish between known and unknown types. Furthermore, an innovative unknown-aware hierarchical contrastive learning strategy is designed to facilitate a clear delineation between closely related known types and unknown types. Extensive experiments on two newly established benchmark datasets demonstrate that our proposed framework significantly surpasses all baselines in addressing the OSET task.

[Findings] VANE: Guiding High-Value Exploration in RLVR via Outcome-Process Novelty Shaping ▶ *Underline Video*

Xu He, Jialiang Guo, Fucheng Xiong, Haodong Zhao, xingyang li, Ke Zeng and Xunliang Cai

Reinforcement Learning with Verifiable Rewards (RLVR) frequently suffers from mode collapse due to the inherent sparsity of feedback signals. While strategies such as entropy regularization introduce randomness, they lack directionality. Simply incorporating diversity rewards is overly one-sided and fails to identify potential logical errors or hallucinations. To address these limitations, we propose VANE (Value-Aligned Novelty Exploration), a method that simultaneously quantifies novelty across the outcome space (via reward or solution divergence) and the semantic process space (via semantic process divergence). Moreover, VANE employs a value-alignment mechanism that symmetrically amplifies scarce, high-quality solutions while explicitly penalizing diverse yet erroneous reasoning paths. Extensive experiments on models such as Qwen2.5-Math-7B across eight benchmarks, encompassing both large-scale mathematical reasoning and out-of-distribution (OOD) tasks, demonstrate the effectiveness and generalization of the proposed method.

[Long] Verifiable LLM-Generated Text Detection via Projected Semantic-Structural Distributions ▶ *Underline Video*

Ruochong Xiong, Qien Li, Wangwang Lian, Yulong Wan, Hanlin Xue, Zhouxing Tan, Han Yang, Fengyu Lu and Junfei Liu

The widespread deployment of large language models (LLMs) makes detecting LLM-Generated text a critical security task. Existing methods, primarily relying on output probabilities from proxy models or single semantic features, suffer from distribution misalignment and limited interpretability. We observe that machine-generated text exhibits a directionally consistent systematic translation relative to human-written text within the joint semantic-structural space. Accordingly, we propose ProSSD, a statistical framework utilizing supervised subspace learning to extract compact features and construct conditional semantic distributions based on syntactic structures. By employing a likelihood ratio test, we derive a modified Mahalanobis distance, weighted by the Wasserstein distance, as the discriminative metric. Experiments demonstrate ProSSDs superior robustness and computational efficiency across cross-domain, cross-model, and adversarial scenarios. Furthermore, we reveal the phenomena of systematic semantic translation and semantic collapse in machine-generated text, offering interpretable statistical insights into LLM generation behaviors.

[Long] VerilogLAVD: LLM-Aided Pattern Generation for Verilog CWE Detection ▶ *Underline Video*

Xiang Long, Yingjie Xia, Li Kuang and Yao Wan

LLMs often fail in hardware vulnerability detection due to the intrinsic semantic concurrency of HDLs (Hardware Description Language), where vulnerabilities arise from the interaction of multiple concurrent execution statements rather than a single sequential execution path. To address the problem, we propose VerilogLAVD, a LLM-Aided Vulnerability Detection framework by generating executable Traversal Detection Patterns (TDPs), i.e. the rules describing how to find the evidence of vulnerabilities in Verilog HDL. We first introduce a Unified

Verilog Property Graph (VeriPG) that explicitly models parallel semantics by combining AST, CFG, and DDG. Furthermore, a semantic validation mechanism is designed to constrain and filter the LLM-generated TDPs. By executing these validated TDPs on VeriPG, our method produces stable and deterministic detection results. Experiments demonstrate that VerilogLAVD improves the F1 score by 133% compared to LLM-based methods. Furthermore, the framework successfully identifies real-world hardware vulnerabilities in open-source hardware design repositories.

[Long] VideoStir: Understanding Long Videos via Spatio-Temporally Structured and Intent-Aware RAG ▶ *Underline Video*

Honghao Fu, Miao Xu, Yiwei Wang, Dailing Zhang, Jun Liu and Yujun Cai

Scaling multimodal large language models (MLLMs) to long videos is constrained by limited context windows. While retrieval-augmented generation (RAG) is a promising remedy by organizing query-relevant visual evidence into a compact context, most existing methods (i) flatten videos into independent segments, breaking their inherent spatio-temporal structure, and (ii) depend on explicit semantic matching, which can miss cues that are implicitly relevant to the querys intent. To overcome these limitations, we propose VideoStir, a structured and intent-aware long-video RAG framework. It firstly structures a video as a spatio-temporal graph at clip level, and then performs multi-hop retrieval to aggregate evidence across distant yet contextually related events. Furthermore, it introduces an MLLM-backed intent-relevance scorer that retrieves frames based on their alignment with the querys reasoning intent. To support this capability, we curate IR-600K, a large-scale dataset tailored for learning frame-query intent alignment. Experiments show that VideoStir is competitive with state-of-the-art baselines without relying on auxiliary information, highlighting the promise of shifting long-video RAG from flattened semantic matching to structured, intent-aware reasoning. Codes and checkpoints are available at https://github.com/RomGai/VideoStir.

[Long] VIGIL: Defending LLM Agents Against Tool-Stream Injection via Verify-Before-Commit ▶ *Underline Video*

Junda lin, Zhaomeng Zhou, Zhi Zheng, Shuochen Liu, Tong Xu, Yong Chen and Enhong Chen

LLM agents operating in open environments face escalating risks from indirect prompt injection, particularly within the tool stream where manipulated metadata and runtime feedback hijack execution flow. Existing defenses encounter a critical dilemma as advanced models prioritize injected rules due to strict alignment while static protection mechanisms sever the feedback loop required for adaptive reasoning. To reconcile this conflict, we propose VIGIL, a framework that shifts the paradigm from restrictive isolation to a verify-before-commit protocol. By facilitating speculative hypothesis generation and enforcing safety through intent-grounded verification, VIGIL preserves reasoning flexibility while ensuring robust control. We further introduce SIREN, a benchmark comprising 959 tool stream injection cases designed to simulate pervasive threats characterized by dynamic dependencies. Extensive experiments demonstrate that VIGIL outperforms state-of-the-art dynamic defenses by reducing the attack success rate by over 22% while more than doubling the utility under attack compared to static baselines, thereby achieving an optimal balance between security and utility. Our code is available at: https://github.com/Touring-686/vigil.

[Long] WebAggregator: Enhancing Compositional Reasoning Capabilities of Deep Research Agent Foundation Models ▶ *Underline Video*

Rui Wang, Ce Zhang, Jun-Yu Ma, Jianshu Zhang, Hongru WANG, Yi Chen, Boyang XUE, Tianqing Fang, Zhisong Zhang, Hongming Zhang, Haitao Mi, Dong Yu and Kam-Fai Wong

The hallmark of Deep Research agents lies in compositional reasoning, the capacity to aggregate distributed, heterogeneous information into coherent logical insights. However, current agentic systems are often retrieval-heavy but reasoning-light, where success is predominantly determined by simple entity-seeking rather than the multi-step aggregation of scattered evidence. To address this, we propose a data synthesis pipeline WebAggregator, designed to shift the agentic paradigm from retrieval-centric to compositional aggregation. Our approach first employs Proactive Explorer to collect interconnected knowledge, then Compositional Logic Proposer to weave knowledge into complex questions using over 12 composition guidelines derived from a rigorous deconstruction of the Deep Research problem setting. Fine-tuning on this corpus fundamentally transforms agent behavior, fostering deliberate composition reasoning and reduced tool redundancy. The resulting WebAggregator-32B surpasses GPT-4.1 and matches Claude-3.7-Sonnet on GAIA, WebWalkerQA, and XBench. To address the lack of benchmarks that emphasize both reasoning and retrieval, we introduce the WebAggregatorQA testbed, which reveals that even with perfect retrieval, top-tier models still underperformed. These results demonstrate that compositional reasoning, not retrieval, is the true performance ceiling for next-generation research agents.

[Short] What Makes AI Research Replicable? Executable Knowledge Graphs as Scientific Knowledge Representations ▶ *Underline Video*

Yujie Luo, Zhuoyun Yu, Xuehai Wang, Yuqi Zhu, Ningyu Zhang, Lanning Wei, Lun Du, Da Zheng and Huajun Chen

Replicating AI research is a crucial yet challenging task for large language model (LLM) agents. Existing approaches often struggle to generate executable code, primarily due to insufficient background knowledge and the limitations of retrieval-augmented generation (RAG) methods, which fail to capture latent technical details hidden in referenced papers. Furthermore, previous approaches tend to overlook valuable implementation-level code signals and lack structured knowledge representations that support multi-granular retrieval and reuse. To overcome these challenges, we propose Executable Knowledge Graphs (xKG), a pluggable, paper-centric knowledge base that automatically integrates code snippets and technical insights extracted from scientific literature. When integrated into three agent frameworks with two different LLMs, xKG shows substantial performance gains (10.9% with o3-mini) on PaperBench, demonstrating its effectiveness as a general and extensible solution for automated AI research replication.

[Industry] When Good OCR Is Not Enough: Benchmarking OCR Robustness for Retrieval-Augmented Generation ▶ *Underline Video*

Lin Sun, wangdexian wangdexian, Jingang Huang, Linglin Zhang, Change Jia, Zhengwei Cheng and Xiangzheng Zhang

Industrial Retrieval-Augmented Generation (RAG) systems depend on optical character recognition (OCR) to transform visual documents into text. Existing OCR benchmarks rely on character-level metrics, which inadequately measure downstream RAG effectiveness under real-world conditions. We introduce an OCR benchmark for industrial RAG systems covering 11 challenging document types, including extreme layouts, high-resolution pages, complex or watermarked backgrounds, historical documents with non-standard reading orders, visually decorated text, and documents containing tables and mathematical formulas. Evaluating recent SOTA OCR models under a controlled OCR-first RAG pipeline shows clear performance degradation on realistic industrial documents despite strong conventional benchmark scores. We find that high OCR accuracy does not necessarily translate into strong downstream RAG performance: structural and semantic errors can cause substantial retrieval failures even when WER/CER remains low. Further analysis shows that this mismatch is category-dependent, arises through both retrieval-side and downstream generation-side failures, and remains stable across representative OCR-first pipeline choices.

[Long] When KV Cache Reuse Fails in Multi-Agent Systems: Cross-Candidate Interaction is Crucial for LLM Judges ▶ *Underline Video*

Sichu Liang, Zhenglin Wang, Chujiajia Chujiajia, Pengfei Xia, Hui Zang and Deyu Zhou

Multi-agent LLM systems routinely generate multiple candidate responses that are aggregated by an LLM judge. To reduce the dominant prefill cost in such pipelines, recent work advocates KV cache reuse across partially shared contexts and reports substantial speedups for generation agents. In this work, we show that these efficiency gains do not transfer uniformly to judge-centric inference. Across GSM8K, MMLU, and HumanEval, we find that reuse strategies that are effective for execution agents can severely perturb judge behavior: end-task accuracy may appear stable, yet the judge's selection becomes highly inconsistent with dense prefill. We quantify this risk using Judge Consistency Rate (JCR) and provide diagnostics showing that reuse systematically weakens cross-candidate attention, especially for later candidate blocks. Our ablation further demonstrates that explicit cross-candidate interaction is crucial for preserving dense-prefill decisions. Overall, our results identify a previously overlooked failure mode of KV cache reuse and highlight judge-centric inference as a distinct regime that demands dedicated, risk-aware system design.

[Findings] Who is the richest club in the championship? Detecting and Rewriting Underspecified Questions Improve QA Performance ▶ *Underline Video*

Yunchong Huang, Gianni Barlacchi and Sandro Pezzelle

Large language models (LLMs) perform well on well-posed questions, yet standard question-answering (QA) benchmarks remain far from solved. We argue that this gap is partly due to underspecified questions, that are queries whose interpretation cannot be uniquely determined without additional context. We introduce an LLM-based classifier to identify underspecified questions and apply it to several widely used QA datasets, finding that 16% to over 60% of benchmark questions are underspecified and that LLMs perform significantly worse on them. To isolate the effect of underspecification, we conduct a controlled rewriting experiment that serves as an upper-bound analysis, rewriting underspecified questions into fully specified variants while holding gold answers fixed. QA performance consistently improves under this setting, indicating that many apparent QA failures stem from question underspecification rather than model limitations. Our findings highlight underspecification as an important confound in QA evaluation and motivate greater attention to question clarity in benchmark design.

[Long] WIST: Web-Grounded Iterative Self-Play Tree for Domain-Targeted Reasoning Improvement ▶ *Underline Video*

Fangyuan Li, Pengfei Li, Shijie Wang, Junqi Gao, Jianxing Liu, Biqing Qi and Yuqiang Li

Recent progress in reinforcement learning with verifiable rewards (RLVR) offers a practical path to self-improving language models, but existing methods face a key trade-off: endogenous self-play can drift over iterations, while corpus-grounded approaches rely on curated data environments. We present WIST, a Web-grounded Iterative Self-play Tree framework for domain-targeted reasoning improvement that learns directly from the open-web without requiring any pre-arranged domain corpus. WIST incrementally expands a domain tree to structure exploration and retrieves and cleans path-consistent web evidence to construct a controllable training environment. It then performs Challenger-Solver self-play with verifiable rewards, and feeds learnability signals back to update node posteriors and guide subsequent exploration through an adaptive curriculum. Across four backbones, WIST consistently improves over the base models and typically outperforms both purely endogenous self-evolution and corpus-grounded self-play baselines, with the Overall gains reaching +9.8 (Qwen3-4B-Base) and +9.7 (OctoThinker-8B-Hybrid-Base). WIST is also domain-steerable: improving Qwen3-8B-Base by +14.79 in medicine and Qwen3-4B-Base by +5.28 on PhyBench. Ablations further confirm the importance of WISTs key components for stable open-web learning. Our Code is available at https://github.com/lfy-123/WIST.

[Long] Would LLMs be Good Historical Linguists and Chinese Dialect Learners? ▶ *Underline Video*

Yicheng Liu, Shumin Shi, Youchao Zhou and Xingchen Zhang

Large language models (LLMs) perform well on Standard Chinese but struggle with low-resource Chinese dialects due to substantial phonological divergence. We investigate whether incorporating Middle Chinese, the common historical ancestor of most of the modern Chinese dialects, can improve dialectal pronunciation modeling in a linguistically interpretable manner. We focus on two specific task

variants: (1) conditional sound change rule induction (a variant of Sound Law Induction, SLI), where models infer executable phonological transformation rules from Middle Chinese to modern dialects, and (2) sentence-level dialectal pronunciation transcription (a variant of Grapheme-to-Phoneme, G2P), requiring dialect-specific International Phonetic Alphabet (IPA) generation. We construct a multi-source dataset covering Middle Chinese and 12 modern Chinese dialects, including character-level correspondences, rule exemplars, and sentence-level IPA transcription. We adopt a parameter-efficient training framework combining LoRA-based supervised fine-tuning and reinforcement learning via Group Relative Policy Optimization (GRPO) for the first task. Across both tasks and a wide range of dialects and evaluation metrics, our approach achieves overall improvements over strong baselines, including DeepSeek-V3.2 and ChatGPT-5.2, while revealing variation across dialects. These results demonstrate the value of leveraging historical linguistic knowledge for modeling low-resource Chinese dialects.

[Short] Z3D: Zero-Shot 3D Visual Grounding from Images ► *Underline Video*

Nikita Drozdov, Andrey Lemeshko, Nikita Gavrilov, Anton Konushin, Danila Rukhovich and Maksim Kolodiazhnyi

3D visual grounding (3DVG) aims to localize objects in a 3D scene based on natural language queries. In this work, we explore zero-shot 3DVG from multi-view images alone, without requiring any geometric supervision or object priors. We introduce Z3D, a universal grounding pipeline that flexibly operates on multi-view images while optionally incorporating camera poses and depth maps. We identify key bottlenecks in prior zero-shot methods causing significant performance degradation and address them with (i) a state-of-the-art zero-shot 3D instance segmentation method to generate high-quality 3D bounding box proposals and (ii) advanced reasoning via prompt-based segmentation, which utilizes full capabilities of modern VLMs. Extensive experiments on the ScanRefer and Nr3D benchmarks demonstrate that our approach achieves state-of-the-art performance among zero-shot methods.

[Long] Zero-shot Jianzi Recognition as Structured Visual Information Extraction in Open Compositional Symbolic Systems ► *Underline Video*

Zehan Li, Fu Zhang, Zhijun Liu and Jingwei Cheng

Guqin () Jianzi () is an open and freely compositional tablature system that encodes performance actions rather than acoustic outcomes. Its automatic recognition remains largely unexplored, as conventional OCR assumes a closed and enumerable glyph set and struggles with Jianzis unbounded composition and manuscript-level variability. We introduce Zero-shot Jianzi Recognition, which formulates Jianzi recognition as vision-to-sequence prediction of canonical component sequences under a zero-shot split. To enable scalable supervision, we construct Synthetic-JZ from aligned online composition metadata. We then synthesize manuscript-like training images via component-wise style recomposition and manuscript-domain noise modeling, and fine-tune a vision-language model for end-to-end component sequence recognition. At inference time, a lightweight legality-guided correction module re-ranks decoding candidates, suppressing structural hallucinations without modifying the backbone. Experiments on two benchmarks show that our method achieves 63.02% sequence accuracy on Real-JZ, our manually annotated real-world Jianzi benchmark, surpassing Gemini-3-Pro by 35.11%. This result highlights the feasibility of reliable automated Jianzi recognition and its potential for large-scale digitization of historical Guqin Jianzi Pu manuscripts.

Session 8: Oral/Posters/Demos D - 09:00-10:30

Demo Session D

09:00-10:30 (Grand Hall)

[Demo] AInterviewer: A Platform for Designing and Conducting AI-led Qualitative Interviews ► *Underline Video*

Tobias Gårdhus, Nikolas Vitsakis, Fie Lejre Frederiksen, Anna Rogers and Hjalmar Bang Carlsen

There are now multiple proposals for systems based on Large Language Models(LLMs) to conduct automated qualitative interviews. This approach scales up qualitative interview techniques that have traditionally been constrained by the high costs of data collection. However, most of the current solutions rely on proprietary LLMs, which compromise reproducibility and data security. They also rely on LLMs for all interview tasks, which limits standardisation of question wording as well as control over question order. To address these issues, we introduce the AInterviewer platform, based on a multi-agent framework that combines controlled question administration of survey software with the flexibility of LLMs. AInterviewer can run with locally hosted models to ensure security and transparency. Our platform provides a web-based GUI supporting each phase of data collection: from interview guide design and pilot testing to interview distribution and data collection monitoring.

[Demo] Dash-M5H: An Interactive Dashboard for Multi-Modal, Multi-Model Mental Health Assessment ► *Underline Video*

Raymond Alavo, Xinyuan Zhang, Gemza Ademaj, Junhui Cai, Hyeokhyen Kwon, Robert Cotes, Gari Clifford and Ahmed Abbasi

We present Dash-M5H, an interactive dashboard for multi-modal, multi-model mental health assessment that helps clinicians and researchers jointly inspect multimodal behavioral data with multi-model signal outputs of recorded clinical interviews. Guided by signal detection and integrated sensemaking theories, Dash-M5H synchronizes transcript text, audio, and facial behavior (action units and gaze) to support overview-to-detail evidence tracing; and it integrates extracted signals (e.g., sentiment and facial activity) with a clinically grounded

VLM prediction pipeline that produces DSM-5-aligned depression predictions. Dash-M5H (https://dash-m5h.io) is implemented in a lightweight, browser-based stack (Quarto + Observable JS + D3), supports local data import and time-synced clinical annotation with export. We demonstrate Dash-M5H through a depression screening scenario, evaluate its note-taking and screening capabilities through a user experiment, and release a live demo (https://youtu.be/w3qCJ02k6bw) and code (https://github.com/nd-hal/M5H-Dashboard-VLM) to facilitate reproducible evaluation.

[Demo] FactSearch: An Interactive Fact Search and Claim-Level Verification System for Large Language Models ▶ *Underline Video*
Meng Fang and Harry Mackenzie
Large language models (LLMs) frequently generate factually incorrect or unverifiable statements, motivating tool-augmented verification systems that combine model reasoning with external evidence retrieval. For factuality evaluation to be scientifically reliable, verification pipelines must be controllable and reproducible: retrieval configuration and reasoning behavior should be explicitly configurable and stable across runs. In practice, many existing systems depend on commercial search APIs whose ranking policies and retrieval behaviors are opaque and externally controlled, introducing uncontrolled variability into evaluation. This makes it difficult to disentangle reasoning errors from retrieval effects. We present FactSearch, a reproducible, claim-level factuality verification system built on locally aggregated open-source search infrastructure. FactSearch decomposes model outputs into atomic factual claims, retrieves supporting evidence via a self-hosted meta-search engine, and performs modular verification within a fully configurable pipeline. By treating retrieval infrastructure as a first-class component, the system enables deterministic configurations and systematic analysis of retrieval-reasoning interactions. An interactive web interface supports transparent inspection and practical deployment. A demo video is available at: https://youtu.be/qbfA2vIxsiI. Code is at https://github.com/aialt/factsearch.

[Demo] Gardener: An Agentic AI System for Single-Cell RNA Sequence Analysis ▶ *Underline Video*
Junhan Liu, Zhenke Liu, Yongcheng Shi, Peilin Yu, Minxing Zhang and Jiapeng Zhang
Gardener is an interactive agentic system for single-cell RNA-seq (scRNA-seq) analysis that enables expert-steered, iterative workflows under strict data-residency requirements. Existing large language model (LLM)-based analysis agents commonly encode workflow progress as implicit conversational state and rely on cloud-centric execution, which hinders traceability and auditability and complicates keeping sensitive expression data on-device. Gardener grounds cloud-side reasoning in a local, on-device scientific engine and an Experiment Management Kernel (EMK) that externalizes analysis progress as persistent, immutable snapshots linked by lineage. This explicit state representation supports rollback, branching, and comparison of alternative analysis paths while reusing prior computation. Gardener enforces data isolation by design: cloud-hosted LLMs operate only on snapshot identifiers and sanitized summaries, while raw expression matrices and local artifacts remain on the users device. A local graphical user interface (GUI) provides human-in-the-loop steering and inspection of workflow state and outputs. Gardener is released as an open-source desktop application for macOS and Windows under the Apache License 2.0.

[Demo] Interpreto: An Explainability Library for Transformers ▶ *Underline Video*
Antonin Poché, Thomas Mullor, Gabriele Sarti, Frédéric Boisnard, Corentin Friedrich, Charlotte Claye, François Hoofd, Raphael Bernas, Nicholas Asher, Céline Hudelot and Fanny Jourdan;
Interpreto is an open-source Python library for interpreting HuggingFace language models, from early BERT variants to LLMs. It provides two complementary families of methods: attribution methods and concept-based explanations. The library bridges recent research and practical tooling by exposing explanation workflows through a unified API for both classification and text generation. A key differentiator is its end-to-end concept-based pipeline (from activation extraction to concept learning, interpretation, and scoring), which goes beyond feature-level attributions and is uncommon in existing libraries.

[Demo] MED-COPILOT: A Medical Assistant Powered by GraphRAG and Similar Patient Case Retrieval ▶ *Underline Video*
Shuheng Chen, Namratha Patil, Haonan Pan, Angel Hsing-Chi Hwang, Yao Du, Ruishan Liu and Jieyu Zhao;
Clinical decision-making requires synthesizing heterogeneous evidence, including patient histories, clinical guidelines, and trajectories of comparable cases. While large language models (LLMs) offer strong reasoning capabilities, they remain prone to hallucinations and struggle to integrate long, structured medical documents. We present MED-COPILOT, an interactive clinical decision-support system designed for clinicians and medical trainees, which combines guideline-grounded GraphRAG retrieval with hybrid semantic-keyword similar-patient retrieval to support transparent and evidence-aware clinical reasoning. The system builds a structured knowledge graph from WHO and NICE guidelines, applies community-level summarization for efficient retrieval, and maintains a 36,000-case similar-patient database derived from SOAP-normalized MIMIC-IV notes and Synthea-generated records. We evaluate our framework on clinical note completion and medical question answering, and demonstrate that it consistently outperforms parametric LLM baselines and standard RAG, improving both generation fidelity and clinical reasoning accuracy. The full system is available at https://huggingface.co/spaces/Cryo3978/Med_GraphRAG, enabling users to inspect retrieved evidence, visualize token-level similarity contributions, and conduct guided follow-up analysis. Our results demonstrate a practical and interpretable approach to integrating structured guideline knowledge with patient-level analogical evidence for clinical LLMs.

[Demo] OxyGent: Making Multi-Agent Systems Modular, Observable, and Evolvable via Oxy Abstraction ▶ *Underline Video*
Junxing Hu, Tianlong Li, Lei Yu and Ai Han;

Deploying production-ready multi-agent systems (MAS) in complex industrial environments remains challenging due to limitations in scalability, observability, and autonomous evolution. We present OxyGent, an open-source framework that enables modular, observable, and evolvable MAS via a unified Oxy abstraction, in which agents, tools, LLMs, and reasoning flows are encapsulated as pluggable atomic components. This Lego-like assembly paradigm supports scalable system composition and non-intrusive monitoring. To enhance observability, OxyGent introduces permission-driven dynamic planning that replaces rigid workflows with execution graphs generated at runtime, which provide adaptive visualizations. To support continuous evolution, the framework integrates OxyBank, an AI asset management platform that supports automated data backflow, annotation, and joint evolution. Empirical evaluations and real-world case studies show that OxyGent provides a robust and scalable foundation for MAS. OxyGent is publicly available at https://oxygent.jd.com/.

[Demo] ParseJargon: Personalized Real-time Jargon Support in Online Meetings ► *Underline Video*

Yifan Song, Wing Yee Au, Hon Yung Wong, Brian P. Bailey and Tal August

Effective interdisciplinary communication is frequently hindered by domain-specific terms. These terms, or jargon, are dependent on a listener's background, and rarely do listeners seek explanations due to distraction and social concerns. To address these concerns, we built ParseJargon, an interactive LLM-powered system providing real-time personalized jargon support tailored to users individual backgrounds in online meetings. We first evaluated the effectiveness of personalization in a controlled setting with human participants. By comparing ParseJargon against baseline (no support) and general-purpose (non-personalized) conditions, we found that ParseJargon provided more precise jargon identification, and enhanced participants' comprehension, engagement, and appreciation of colleagues' work. We then evaluated the potential for using ParseJargon in real-time meetings through a latency test.

Oral Session D: AI/LLM Agents 2

09:00-10:30 (Harbor B-C)

[Long] CAR-bench: Evaluating the Consistency and Limit-Awareness of LLM Agents under Real-World Uncertainty ► *Underline Video*

Johannes Kirmayr, Lukas Stappen and Elisabeth André

Existing benchmarks for Large Language Model (LLM) agents focus on task completion under idealistic settings but overlook reliability in real-world, user-facing applications. In domains, such as in-car voice assistants, users often issue incomplete or ambiguous requests, creating intrinsic uncertainty that agents must manage through dialogue, tool use, and policy adherence. We introduce CAR-bench, a benchmark for evaluating consistency, uncertainty handling, and capability awareness in multi-turn, tool-using LLM agents in an in-car assistant domain. The environment features an LLM-simulated user, domain policies, and 58 interconnected tools spanning navigation, productivity, charging, and vehicle control. Beyond standard task completion, CAR-bench introduces Hallucination tasks that test agents limit-awareness under missing tools or information, and Disambiguation tasks that require resolving uncertainty through clarification or internal information gathering. Baseline results reveal large gaps between occasional and consistent success on all task types. Even frontier reasoning LLMs achieve less than 50% pass rate on Disambiguation tasks due to premature actions, and frequently violate policies or fabricate information to satisfy user requests in Hallucination tasks, underscoring the need for more reliable and self-aware LLM agents in real-world settings.

[Long] FlowSearch: Advancing Deep Research with Dynamic Structured Knowledge Flow ► *Underline Video*

Yusong Hu, Runmin Ma, Yue Fan, Jinxin Shi, Zongsheng Cao, Yuhao Zhou, Jiakang Yuan, Shuaiyu Zhang, Shiyang Feng, Xiangchao Yan, Shufei Zhang, Wenlong Zhang, LEI BAI and Bo Zhang

Deep research is an inherently challenging task that demands both breadth and depth of thinking. It involves navigating diverse knowledge spaces and reasoning over complex, multi-step dependencies, which presents substantial challenges for agentic systems. To address this, we propose FlowSearch, a multi-agent framework that actively constructs and evolves a dynamic structured knowledge flow to drive subtask execution and reasoning. FlowSearch is capable of strategically planning and expanding the knowledge flow to enable parallel exploration and hierarchical task decomposition, while also adjusting the knowledge flow in real time based on feedback from intermediate reasoning outcomes and insights. FlowSearch achieves competitive performance on both general and scientific benchmarks, including GAIA, HLE, GPQA and TRQA, demonstrating its effectiveness in multi-disciplinary research scenarios and its potential to advance scientific discovery. The code will be available.

[Long] From Word to World: Can Large Language Models be Implicit Text-based World Models? ► *Underline Video*

Yixia Li, Hongru WANG, Jiahao Qiu, Zhenfei Yin, Dongdong Zhang, Cheng Qian, Zeping Li, Xiaoteng Ma, Guanhua Chen and Heng Ji

Agentic learning increasingly hinges on interaction, yet real-world experience is expensive, limited, and often irreversible at inference time. World models promise to mitigate these limitations, but it remains unclear whether large language models can actually serve as reliable world models, and deliver concrete benefits to downstream agents. We investigate these questions in text-based environments, a controlled testbed that reframes language modeling as next-state prediction under interaction. We propose a three-level framework to evaluate LLM-based world models: (i) fidelity and consistency, (ii) scalability and robustness, and (iii) agent utility. Across five representative environments, we show that sufficiently trained world models capture coherent environment dynamics, scale predictably with data and model capacity, and

unlock tangible agent improvements, for example, action verification boosts GPT-4o by 5.5% on WebShop, and warm-started RL achieves a 15% gain on SciWorld. Crucially, these benefits hinge on behavioral coverage and environment complexity, sharply characterizing when world modeling meaningfully advances agent learning.

[TACL] Goal Alignment in LLM-Based User Simulators for Conversational AI ▶ *Underline Video*
Shuhaib Mehri, Xiaocheng Yang, Takyoung Kim, Gokhan Tur, Shikib Mehri and Dilek Hakkani-Tür
User simulators are essential to conversational AI, enabling scalable agent development and evaluation through simulated interactions. While current Large Language Models (LLMs) have advanced user simulation capabilities, we reveal that they struggle to consistently demonstrate goal-oriented behavior across multi-turn conversations–a critical limitation that compromises their reliability in downstream applications. We introduce User Goal State Tracking (UGST), a novel framework that tracks user goal progression throughout conversations. Leveraging UGST, we present a three-stage methodology for developing user simulators that can autonomously track goal progression and reason to generate goal-aligned responses. Moreover, we establish comprehensive evaluation metrics for measuring goal alignment in user simulators, and demonstrate that our approach yields substantial improvements across two benchmarks (MultiWOZ 2.4 and tau-Bench). Our contributions address a critical gap in conversational AI and establish UGST as an essential framework for developing goal-aligned user simulators.

[Long] Grammar Search for Multi-Agent Systems ▶ *Underline Video*
Mayank Singh, Vikas Yadav, Shiva Krishna Reddy Malay, Shravan Nayak, Sai Rajeswar, Sathwik Tejaswi Madhusudhan and Eduardo Blanco
Automatic search for Multi-Agent Systems has recently emerged as a key focus in agentic AI research. Several prior approaches have relied on LLM-based free-form search over the code space. In this work, we propose a more structured framework that explores the same space through a fixed set of composable components. We show that, despite lacking the generative flexibility of LLMs during the candidate generation stage, our method outperforms prior approaches on a majority of evaluated benchmarks across two backbone LLMs and two domains: mathematics and question answering. Furthermore, our method offers additional advantages, including a more cost-efficient search process and the generation of modular, interpretable multi-agent systems.

[Long] HSCodeComp: A Realistic and Expert-level Agent Benchmark for Hierarchical Rule Application ▶ *Underline Video*
Tian Lan, Yiqian Yang, Qianghuai Jia, Li Zhu, Hui Jiang, Hang Zhu, Weihua Luo and Longyue Wang
Despite recent progress, existing agent benchmarks neglect a fundamental real-world capability: hierarchical rule application, a critical requirement in fields such as law and medicine where agents must reason from broad categories down to specific exceptions to reach rule-compliant decisions. This introduces significant challenges in resolving logical dependencies and disambiguating vague boundaries. To bridge this gap, we introduce HSCodeComp, a novel benchmark derived from e-commerce, requiring agents to assign a unique 10-digit Harmonized System (HS) Code to products by aligning their fuzzy attributes with strict tariff classification rules. HSCodeComp comprises 632 realistic products across 32 categories, featuring detailed yet noisy product information (titles, attributes, and images). The HS Codes are annotated by a panel of 26 tariff experts, strictly adhering to official rules and an empirical knowledge base, both of which we jointly open-source. Through a comprehensive evaluation of 23 LLMs, VLMs, and agents on HSCodeComp, we demonstrate that: 1) a substantial performance gap remains between state-of-the-art agents and human experts (46.8% vs. 95.0%); and 2) test-time scaling fails to close this gap. Further analysis reveals that 1) excessive reasoning steps frequently induce "reasoning drift," which degrades accuracy; and 2) agents are prone to premature decisions on high-level categories and reasoning hallucinations that lack factual grounding.

[TACL] In-N-Out: A Parameter-Level API Graph Dataset for Tool Agents ▶ *Underline Video*
Seungkyu Lee, Nalim Kim and Yohan Jo
Tool agents—LLM-based systems that interact with external APIs—offer a way to execute real-world tasks. However, as tasks become increasingly complex, these agents struggle to identify and call the correct APIs in the proper order. To tackle this problem, we investigate converting API documentation into a structured API graph that captures API dependencies and leveraging it for multi-tool queries that require compositional API calls. To support this, we introduce In-N-Out, the first expert-annotated dataset of API graphs built from two real-world API benchmarks and their documentation. Using In-N-Out significantly improves performance on both tool retrieval and multi-tool query generation, nearly doubling that of LLMs using documentation alone. Moreover, graphs generated by models fine-tuned on In-N-Out close 90% of this gap, showing that our dataset helps models learn to comprehend API documentation and parameter relationships. Our findings highlight the promise of using explicit API graphs for tool agents and the utility of In-N-Out as a valuable resource. We will release the dataset and code publicly.

[Long] OctoTools: A Multi-Agent Framework with Extensible Tools for Complex Reasoning ▶ *Underline Video*
Pan Lu, Bowen Chen, Sheng Liu, Rahul Thapa, Joseph Boen and James Zou
Solving complex reasoning tasks may involve visual understanding, domain knowledge retrieval, numerical calculation, and multi-step reasoning. Existing methods augment large language models (LLMs) with external tools but are restricted to specialized domains, limited tool types, or require additional training data. In this paper, we introduce OctoTools, a training-free, user-friendly, and easily extensible multi-agent framework designed to tackle complex reasoning across diverse domains. OctoTools introduces standardized tool cards to encapsulate tool functionality, a planner for both high-level and low-level planning, and an executor to carry out tool usage. We validate OctoTools generality across 16 diverse tasks (including MathVista, MMLU-Pro, MedQA, and GAIA-Text), achieving substantial average

accuracy gains of 9.3% over GPT-4o. Furthermore, OctoTools also outperforms AutoGen, GPT-Functions, and LangChain by up to 10.6% when given the same set of tools. Through comprehensive analysi, ablations, and robustness tests with compact backbones and noisy tool environments, OctoTools demonstrates advantages in task planning, effective tool usage, and multi-step problem solving. Code, demos, and visualization are publicly available at https://octotools.github.io/.

[Long] Reinforcement Learning for Self-Improving Agent with Skill Library ▶ *Underline Video*

Jiongxiao Wang, Qiaojing Yan, Yawei Wang, Yijun Tian, Soumya Smruti Mishra, Zhichao Xu, Megha Gandhi, Panpan Xu and Lin Lee Cheong

Large Language Model (LLM)-based agents have demonstrated remarkable capabilities in complex reasoning and multi-turn interactions but struggle to continuously improve and adapt when deployed in new environments. One promising approach is implementing skill libraries that allow agents to learn, validate, and apply new skills. However, current skill library approaches rely primarily on LLM prompting, making consistent skill library implementation challenging. To overcome these challenges, we propose a Reinforcement Learning (RL)-based approach to enhance agents' self-improvement capabilities with a skill library. Specifically, we introduce Skill Augmented GRPO for self-Evolution (SAGE), a novel RL framework that systematically incorporates skills into learning. The framework's key component, Sequential Rollout, iteratively deploys agents across a chain of similar tasks for each rollout. As agents navigate through the task chain, skills generated from previous tasks accumulate in the library and become available for subsequent tasks. Additionally, the framework enhances skill generation and utilization through a Skill-integrated Reward that complements the original outcome-based rewards. Experimental results on AppWorld demonstrate that SAGE, when applied to supervised-finetuned model with expert experience, achieves 8.9% higher Scenario Goal Completion while requiring 26% fewer interaction steps and generating 59% fewer tokens, substantially outperforming existing approaches in both accuracy and efficiency. Our code is available at https://github.com/amazon-science/SAGE.

Oral Session D: Financial Applications and Time Series

09:00-10:30 (Promenade)

[Long] FinCall-Surprise: A Large Scale Multi-modal Benchmark for Earning Surprise Prediction ▶ *Underline Video*

Dong Shu, Yanguang Liu, Huopu Zhang and Mengnan Du

Predicting corporate earnings surprises is a profitable yet challenging task, as accurate forecasts can inform significant investment decisions. However, progress in this domain has been constrained by a reliance on expensive, proprietary, and text-only data, limiting the development of advanced models. To address this gap, we introduce FinCall-Surprise (Financial Conference Call for Earning Surprise Prediction), the first large-scale, open-source, and multi-modal dataset for earnings surprise prediction. Comprising 2,688 unique corporate conference calls from 2019 to 2021, our dataset features word-to-word conference call textual transcripts, full audio recordings, and corresponding presentation slides. We establish a comprehensive benchmark by evaluating 26 state-of-the-art unimodal and multi-modal LLMs. Our findings reveal that (1) while many models achieve high accuracy, this performance is often an illusion caused by significant class imbalance in the real-world data. (2) Some specialized financial models demonstrate unexpected weaknesses in instruction-following and language generation. (3) Although incorporating audio and visual modalities provides some performance gains, current models still struggle to leverage these signals effectively. These results highlight critical limitations in the financial reasoning capabilities of existing LLMs and establish a challenging new baseline for future research.

[Long] FinChain: A Symbolic Benchmark for Verifiable Chain-of-Thought Financial Reasoning ▶ *Underline Video*

Zhuohan Xie, Daniil Orel, Rushil Thareja, Dhruv Sahnan, Hachem Madmoun, Fan Zhang, Debopriyo Banerjee, Georgi Nenkov Georgiev, Xueqing Peng, Lingfei Qian, Jimin Huang, Jinyan Su, Aaryamonvikram Singh, Rui Xing, Rania Elbadry, Chen Xu, Haonan Li, Fajri Koto, Ivan Koychev, Tanmoy Chakraborty, Yuxia Wang, Salem Lahlou, Veselin Stoyanov, Sophia Ananiadou and Preslav Nakov

Multi-step symbolic reasoning is essential for robust financial analysis; yet, current benchmarks largely overlook this capability. Existing datasets such as FinQA and ConvFinQA emphasize final numerical answers while neglecting the intermediate reasoning steps required for transparency and verification. To address this gap, we introduce FinChain, the first benchmark specifically designed for verifiable Chain-of-Thought evaluation in finance. FinChain spans 58 topics across 12 financial domains, each represented by parameterized symbolic templates with executable Python code that enable fully machine-verifiable reasoning and scalable, contamination-free data generation. To assess reasoning capacity, we propose ChainEval, a dynamic alignment measure that jointly evaluates both the final-answer correctness and the step-level reasoning consistency. Our evaluation of 26 leading LLMs reveals that even frontier LLMs exhibit clear limitations in symbolic financial reasoning, while domain-adapted and math-enhanced fine-tuned models can substantially narrow this gap. Overall, FinChain exposes persistent weaknesses in multi-step financial reasoning and provides a foundation for developing trustworthy, interpretable, and verifiable financial AI. This project is available at https://github.com/mbzuai-nlp/finchain.git.

[Long] FinChart-Bench: Benchmarking Financial Chart Comprehension in Vision-Language Models ▶ *Underline Video*

Dong Shu, Haoyang Yuan, Yuchen Wang, Yanguang Liu, Huopu Zhang and Mengnan Du

Large vision-language models (LVLMs) have made significant progress in chart understanding. However, financial charts, characterized by complex temporal structures and domain-specific terminology, remain notably underexplored. We introduce FinChart-Bench, the first

benchmark specifically focused on real-world financial charts. FinChart-Bench comprises 1,200 financial chart images collected from 2015 to 2024, each annotated with True/False (TF), Multiple Choice (MC), and Question Answering (QA) questions, totaling 7,016 questions. We conducted a comprehensive evaluation of 26 state-of-the-art LVLMs on FinChart-Bench. Our evaluation reveals critical insights: (1) the performance gap between open-source and closed-source models is narrowing, (2) performance degradation occurs in upgraded models within families, (3) many models struggle with instruction following, (4) both advanced models show significant limitations in spatial reasoning abilities, and (5) current LVLMs are not reliable enough to serve as automated evaluators. These findings highlight important limitations in current LVLM capabilities for financial chart understanding.

[Long] Is the Attention Matrix Really the Key to SelfAttention in Multivariate LongTerm Time Series Forecasting? ▶ *Underline Video*
Xinyu Li, Kexi Chen, Jiajie Shen, Ying Zheng, Hong Lu, Jin Zhao and Xin Wang
In multivariate long-term time series forecasting, it is widely believed that the effectiveness of self-attention arises from its attention matrix. We challenge this assumption with a counterintuitive finding: our experiments, conducted on three classic and three latest Transformer models, show that dot-product attention can be replaced by element-wise operations without token interaction, such as the addition and Hadamard product, while maintaining or even improving accuracy. This leads to our central hypothesis: the effectiveness of self-attention in this task stems not from the dynamic attention matrix, but from the multi-branch feature extraction inherent in the parallel projections to Query, Key, and Value matrices and their fusion. To validate this, we construct a simple multi-branch MLP that isolates the multi-branch mapping with element-wise operation structure from the Transformer and show that it achieves competitive performance. Our results indicate that the source of performance in self-attention has been misattributed, suggesting that the true benefit lies in the architectural principle of multi-branch mapping and fusion, not in the attention matrix.

[Long] STReasoner: Empowering LLMs for Spatio-Temporal Reasoning in Time Series via Spatial-Aware Reinforcement Learning ▶ *Underline Video*
Juntong Ni, Shiyu Wang, Qi He, Ming Jin and Wei Jin
Spatio-temporal reasoning in time series involves the explicit synthesis of temporal dynamics, spatial dependencies, and textual context. This capability is vital for high-stakes decision-making in systems such as traffic networks, power grids, and disease propagation. However, the field remains underdeveloped because most existing works prioritize predictive accuracy over reasoning. To address the gap, we introduce ST-Bench, a benchmark consisting of four core tasks, including etiological reasoning, entity identification, correlation reasoning, and in-context forecasting, developed via a network SDE-based multi-agent data synthesis pipeline. We then propose STReasoner, a model that integrates time series, graph structure, and text for explicit reasoning. To promote spatially grounded logic, we introduce S-GRPO, a reinforcement learning algorithm that rewards performance gains specifically attributable to spatial information. Experiments show that STReasoner achieves average accuracy gains between 17% and 135% at only 0.004x the cost of proprietary models and generalizes robustly to real-world data.

[Long] ZARA: Training-Free Motion Time-Series Reasoning via Evidence-Grounded LLM Agents ▶ *Underline Video*
Zechen Li, Baiyu Chen, Hao Xue and Flora D. Salim
Motion sensor time-series are central to Human Activity Recognition (HAR), yet conventional approaches are constrained to fixed activity sets and typically require costly parameter retraining to adapt to new behaviors. While Large Language Models (LLMs) offer promising open-set reasoning capabilities, applying them directly to numerical time-series often leads to hallucinations and weak grounding. To address this challenge, we propose ZARA (Zero-training Activity Reasoning Agents), a knowledge- and retrieval-augmented agentic framework for motion time-series reasoning in a training-free inference setting. Rather than relying on black-box projections, ZARA distills reference data into a statistically grounded textual knowledge base that transforms implicit signal patterns into verifiable natural-language priors. Guided by retrieved evidence, ZARA iteratively selects discriminative cues and performs grounded reasoning over candidate activities. Extensive experiments on eight benchmarks show that ZARA generalizes robustly to unseen subjects and across datasets, demonstrating strong transferability across heterogeneous sensor domains. These results mark a step toward trustworthy, plug-and-play motion understanding beyond dataset-specific artifacts. Our code is available at https://github.com/zechenli03/ZARA.

Oral Session D: Industry 4

09:00-10:30 (Balboa)

[Industry] LEAP: Layer-wise Exit-Aware Pretraining for Efficient Transformer Inference ▶ *Underline Video*
Shashank Kapadia, Deep Naryan Mishra, Sujal Reddy Alugubelli, Haoan Wang, Saipraveen Vabbilisetty, Rishi Bhatia and Anupriya Sharma
Layer-aligned distillation and convergence-based early exit represent two predominant computational efficiency paradigms for transformer inference; yet we establish that they exhibit fundamental incompatibility under standard deployment conditions for convergence-based early exit. Distillation objectives that align intermediate student layers to teacher representations suppress the representational convergence that early-exit mechanisms exploit, rendering such mechanisms ineffective on distilled models. We introduce LEAP (Layer-wise Exit-Aware Pretraining), an auxiliary training objective that reconciles this incompatibility. LEAP requires no architectural modifications; it augments

standard distillation with a single constraint ensuring intermediate layers approximate final-layer representations. LEAP-MiniLM achieves 1.61x measured wall-clock speedup (batch = 1, NVIDIA L4) at theta = 0.95, with 91.9% of samples exiting by layer 7 and 1.80x theoretical layer reduction, where standard distilled models achieve zero effective speedup. We validate across sentence similarity (STS-B: 0.760 +/- 0.006) and retrieval benchmarks (BEIR), providing operational guidance including latency measurements, decision thresholds, and deployment criteria.

[Industry] Memory-Efficient Structured Backpropagation for On-Device LLM Fine-Tuning ► *Underline Video*
Juneyoung Park, Yuri Hong, Seongwan Kim and Jaeho Lee
On-device fine-tuning enables privacy-preserving personalization of large language models, but mobile devices impose severe memory constraints, typically 6–12GB shared across all workloads. Existing approaches force a trade-off between exact gradients with high memory (MeBP) and low memory with noisy estimates (MeZO). We propose Memory-efficient Structured Backpropagation (MeSP), which bridges this gap by manually deriving backward passes that exploit LoRA's low-rank structure. Our key insight is that the intermediate projection h = xA can be recomputed during backward at minimal cost since rank r « d_in, eliminating the need to store it. MeSP achieves 49% average memory reduction compared to MeBP on Qwen2.5 models (0.5B–3B) while computing mathematically identical gradients. Our analysis also reveals that MeZO's gradient estimates show near-zero correlation with true gradients (cosine similarity ~0.001), explaining its slow convergence. MeSP reduces peak memory from 361MB to 136MB for Qwen2.5-0.5B, enabling fine-tuning scenarios previously infeasible on memory-constrained devices.

[Industry] MobileLLM-Flash: Latency-Guided On-Device LLM Design for Industry Scale Deployment ► *Underline Video*
Hanxian Huang, Igor Fedorov, Andrey Gromov, Bernard Beckerman, Naveen Suda, David Eriksson, Maximilian Balandat, Rylan Conway, Patrick Huber, Chinnadhurai Sankar, Ayushi Dalmia, Zechun Liu, Lemeng Wu, Tarek Elgamal, Adithya Sagar, Vikas Chandra and Raghuraman Krishnamoorthi
Real-time AI experiences call for on-device large language models (OD-LLMs) optimized for efficient deployment on resource-constrained hardware. The most useful OD-LLMs produce near-real-time responses and exhibit broad hardware compatibility, maximizing user reach. We present a methodology for designing such models using hardware-in-the-loop architecture search under mobile latency constraints. This system is amenable to industry-scale deployment: it generates models deployable without custom kernels and compatible with standard mobile runtimes like Executorch. Our methodology avoids specialized attention mechanisms and instead uses attention skipping for long-context acceleration. Our approach jointly optimizes model architecture (layers, dimensions) and attention pattern. To efficiently evaluate candidates, we treat each as a pruned version of a pretrained backbone with inherited weights, thereby achieving high accuracy with minimal continued pretraining. We leverage the low cost of latency evaluation in a staged process: learning an accurate latency model first, then searching for the Pareto-frontier across latency and quality. This yields MobileLLM-Flash, a family of foundation models (350M, 650M, 1.4B) for efficient on-device use with strong capabilities, supporting up to 8k context length. MobileLLM-Flash delivers up to 1.8x and 1.6x faster prefill and decode on mobile CPUs with comparable or superior quality. Our analysis of Pareto-frontier design choices offers actionable principles for OD-LLM design.

Oral Session D: Language Models 3

09:00-10:30 (Harbor E-F)

[Long] AdvancedIF: Rubric-Based Benchmarking and Reinforcement Learning for Advancing LLM Instruction Following ► *Underline Video*
Yun He, Wenzhe Li, Hejia Zhang, Songlin Li, Karishma Mandyam, Sopan Khosla, Yuanhao Xiong, Nanshu Wang, Xiaoliang Peng, Beibin Li, Shengjie Bi, Shishir G Patil, Qi Qi, Shengyu Feng, Julian Katz-Samuels, Richard Yuanzhe Pang, Sujan Kumar Gonugondla, Hunter Lang, Yue Yu, Yundi Qian, Maryam Fazel-Zarandi, Licheng Yu, Amine Benhalloum, Hany Hassan Awadalla and Manaal Faruqui
Recent progress in large language models (LLMs) has led to impressive performance on a range of tasks, yet advanced instruction following (IF), especially for complex, multi-turn, and system-prompted instructions, remains a significant challenge. Rigorous evaluation and effective training for such capabilities are hindered by the lack of high-quality, human-annotated benchmarks and reliable, interpretable reward signals. In this work, we introduce AdvancedIF, a comprehensive benchmark featuring over 1,600 prompts and expert-curated rubrics that assess LLMs ability to follow complex, multi-turn, and system-level instructions. We also open-source the evaluation script of AdvancedIF. We further propose RIFL (Rubric-based Instruction-Following Learning), a novel post-training pipeline that leverages rubric generation, a finetuned rubric verifier, and reward shaping to enable effective reinforcement learning for instruction following. Extensive experiments demonstrate that RIFL substantially improves the instruction-following abilities of LLMs, achieving a 6.7% absolute gain on AdvancedIF and strong results on public benchmarks. Our ablation studies confirm the effectiveness of each component in RIFL. This work establishes rubrics as a powerful tool for both training and evaluating advanced IF in LLMs, paving the way for more capable and reliable AI systems.

[Long] Knowledge is Not Enough: Injecting RL Skills for Continual Adaptation ► *Underline Video*
Pingzhi Tang, Yiding Wang and Muhan Zhang

Large Language Models (LLMs) face the "knowledge cutoff" challenge, where their frozen parametric memory prevents direct internalization of new information. While Supervised Fine-Tuning (SFT) is commonly used to update model knowledge, it often updates factual content without reliably improving the models ability to use the newly incorporated information for question answering or decision-making. Reinforcement Learning (RL) is essential for acquiring reasoning skills; however, its high computational cost makes it impractical for efficient online adaptation. We empirically observe that the parameter updates induced by SFT and RL are nearly orthogonal. Based on this observation, we propose Parametric Skill Transfer (PaST), a framework that supports modular skill transfer for efficient and effective knowledge adaptation. By extracting a domain-agnostic Skill Vector from a source domain, we can linearly inject knowledge manipulation skills into a target model after it has undergone lightweight SFT on new data. Experiments on knowledge-incorporation QA (SQuAD, LooGLE) and agentic tool-use benchmarks (ToolBench) demonstrate the effectiveness of our method. On SQuAD, PaST outperforms the state-of-the-art self-editing SFT baseline by up to 9.9 points. PaST further scales to long-context QA on LooGLE with an 8.0-point absolute accuracy gain, and improves zero-shot ToolBench success rates by +10.3 points on average with consistent gains across tool categories, indicating strong scalability and cross-domain transferability of the Skill Vector.

[Long] LEDOM: Reverse Language Model ► *Underline Video*

Xunjian Yin, Sitao Cheng, Yuxi Xie, Xinyu Hu, Li Lin, Xinyi Wang, Liangming Pan, William Yang Wang and Xiaojun Wan

Autoregressive language models are trained exclusively left-to-right. We explore the complementary factorization, training right-to-left at scale, and ask what reasoning patterns emerge when a model conditions on future context to predict the past. We train LEDOM, an open-source purely reverse autoregressive language model (2B/7B parameters, 435B tokens), and find it develops capabilities distinct from forward models, including abductive inference, question synthesis, and structural handling of the reversal curse. We then explore one application of the reverse model: combining forward likelihood P(y x) with reverse posterior P(x y) through noisy channel duality. We propose Reverse Reward, which reranks forward outputs using reverse posterior estimates, and prove that bidirectional scoring penalizes hallucinated reasoning chains whose backward reconstruction degrades. Reverse Reward yields gains of up to 6.6% on AIME 2024 and 15% on AMC 2023 across multiple strong baselines. We release all codes at https://github.com/Arvid-pku/LEDOM.

[Long] NeuReasoner: Towards Explainable, Controllable, and Unified Reasoning via Mixture-of-Neurons ► *Underline Video*

Haonan Dong, Kehan Jiang, Haoran Ye, Wenhao Zhu, Zhaolu Kang and Guojie Song

Large Reasoning Models (LRMs) have recently achieved remarkable success in complex reasoning tasks. However, closer scrutiny reveals persistent failure modes compromising performance and cost: I) Intra-step level, marked by calculation or derivation errors; II) Inter-step level, involving oscillation and stagnation; and III) Instance level, causing maladaptive over-thinking. Existing endeavors target isolated levels without unification, while their black-box nature and reliance on RL hinder explainability and controllability. To bridge these gaps, we conduct an in-depth white-box analysis, identifying key neurons (Mixture of Neurons, MoN) and their fluctuation patterns associated with distinct failures. Building upon these insights, we propose NeuReasoner, an explainable, controllable, and unified reasoning framework driven by MoN. Technically, NeuReasoner integrates lightweight MLPs for failure detection with a special token-triggered self-correction mechanism learned via SFT. During inference, special tokens are inserted upon failure detection to actuate controllable remedial behaviors. Extensive evaluations across six benchmarks, six backbone models (8B~70B) against nine competitive baselines, demonstrate that NeuReasoner achieves performance gains of up to 27.0% while reducing token consumption by 19.6% ~ 63.3%.

[Long] Omni-Embed-Audio: Leveraging Multimodal LLMs for Robust Audio-Text Retrieval ► *Underline Video*

HaeJun Yoo, Yongseop Shin, Insung Lee, Myoung-Wan Koo and Du-Seong Chang

Audio-text retrieval systems based on Contrastive Language-Audio Pretraining (CLAP) achieve strong performance on traditional benchmarks; however, these benchmarks rely on caption-style queries that differ substantially from real-world search behavior, limiting their assessment of practical retrieval robustness. We present Omni-Embed-Audio (OEA), a retrieval-oriented encoder leveraging multimodal LLMs with native audio understanding. To systematically evaluate robustness beyond caption-style queries, we introduce User-Intent Queries (UIQs), five formulations reflecting natural search behaviors: questions, commands, keyword tags, paraphrases, and exclusion-based negative queries. For negative queries, we develop a hard negative mining pipeline and propose discrimination metrics (HNSR, TFR) assessing models' ability to suppress acoustically similar distractors. Experiments on AudioCaps, Clotho, and MECAT show that OEA achieves comparable text-to-audio retrieval performance to state-of-the-art M2D-CLAP, while demonstrating clear advantages in two critical areas: (1) dominant text-to-text retrieval (+22% relative improvement), and (2) substantially superior hard negative discrimination (+4.3%p HNSR@10, +34.7% relative TFR@10), revealing that LLM backbones provide superior semantic understanding of complex queries.

[Long] PHOTON: Hierarchical Autoregressive Modeling for Lightspeed and Memory-Efficient Language Generation ► *Underline Video*

Yuma Ichikawa, Naoya Takagi, Takumi Nakagawa, Yuzi Kanazawa and Akira Sakai

Transformers operate as horizontal token-by-token scanners; at each generation step, attending to an ever-growing sequence of token-level states. This access pattern increases prefill latency and makes long-context decoding more memory-bound, as KV-cache reads and writes dominate inference time over arithmetic operations. We propose Parallel Hierarchical Operation for TOp-down Networks (PHOTON), a hierarchical autoregressive model that replaces horizontal scanning with vertical, multi-resolution context scanning. PHOTON maintains a hierarchy of latent streams: a bottom-up encoder compresses tokens into low-rate contextual states, while lightweight top-down decoders reconstruct fine-grained token representations in parallel. We further introduce recursive generation that updates only the coarsest latent

stream and eliminates bottom-up re-encoding. Experimental results show that PHOTON is superior to competitive Transformer-based language models regarding the throughput-quality trade-off, providing advantages in long-context and multi-query tasks. In particular, this reduces decode-time KV-cache traffic, yielding up to 10^3 x higher throughput per unit memory.

[Long] Prune as You Generate: Online Rollout Pruning for Faster and Better RLVR ▶ *Underline Video*
Haobo Xu, Sirui Chen, Ruizhong Qiu, Yuchen Yan, Chen Luo, Monica Cheng, Jingrui He and Hanghang Tong
Reinforcement Learning with Verifiable Rewards (RLVR) has significantly advanced the reasoning capabilities of Large Language Models (LLMs). However, methods such as GRPO and DAPO suffer from substantial computational cost, since they rely on sampling many rollouts for each prompt. Moreover, in RLVR the relative advantage is often sparse: many samples become nearly all-correct or all-incorrect, yielding low within-group reward variance and thus weak learning signals. In this paper, we introduce ARRoL (Accelerating RLVR via online RoLlout Pruning), an online rollout pruning method that prunes rollouts during generation while explicitly steering the surviving ones more correctness-balanced to enhance learning signals. Specifically, ARRoL trains a lightweight quality head on-the-fly to predict the success probability of partial rollouts and uses it to make early pruning decisions. The learned quality head can further weigh candidates to improve inference accuracy during test-time voting. To improve efficiency, we present a system design that prunes rollouts inside the inference engine and re-batches the remaining ones for log-probability computation and policy updates. Across GRPO and DAPO on Qwen-3 and LLaMA-3.2 models (1B-8B), ARRoL improves average accuracy by +2.30 to +2.99 while achieving up to 1.7x training speedup, and yielding up to +8.33 additional gains in average accuracy in test-time voting.

[Long] Threshold Differential Attention for Sink-Free, Ultra-Sparse, and Non-Dispersive Language Modeling ▶ *Underline Video*
Xingyue Huang, Xueying Ding, Mingxuan Ju, Yozen Liu, Neil Shah and Tong Zhao
Softmax attention struggles with long contexts due to structural limitations: the strict sum-to-one constraint forces attention sinks on irrelevant tokens, while probability mass disperses as sequence lengths increase. We tackle these problems with Threshold Differential Attention (TDA), a sink-free attention mechanism that achieves ultra-sparsity and improved robustness at longer sequence lengths without the computational overhead of projection methods or the performance degradation caused by noise accumulation of standard rectified attention. TDA applies row-wise extreme-value thresholding with a length-dependent gate, retaining only exceedances. Inspired by the differential transformer, TDA also subtracts an inhibitory view to enhance expressivity. Theoretically, we prove that TDA controls the expected number of spurious survivors per row to O(1) and that consensus spurious matches across independent views vanish as context grows. Empirically, TDA produces >99% exact zeros and eliminates attention sinks while maintaining competitive performance on standard and long-context benchmarks.

[Long] When One LLM Drools, Multi-LLM Collaboration Rules ▶ *Underline Video*
Shangbin Feng, Wenxuan Ding, Alisa Liu, Zifeng Wang, Weijia Shi, Yike Wang, Shannon Zejiang Shen, Xiaochuang Han, Hunter Lang, Chen-Yu Lee, Tomas Pfister, Yejin Choi and Yulia Tsvetkov
This position paper argues that in many realistic (i.e., complex, contextualized, subjective) scenarios, one LLM is not enough to produce a reliable output. We challenge the status quo of relying solely on a single general-purpose LLM and argue for multi-LLM collaboration to better represent the extensive diversity of data, skills, and people. We first posit that a single LLM underrepresents real-world data distributions, heterogeneous skills, and pluralistic populations, and that such representation gaps cannot be trivially patched by further training a single LLM. We then organize existing multi-LLM collaboration methods into a hierarchy, based on the level of access and information exchange, ranging from API-level, text-level, logit-level, to weight-level collaboration. Based on these methods, we highlight how multi-LLM collaboration addresses challenges that a single LLM struggles with, such as reliability, democratization, and pluralism. Finally, we identify the limitations of existing multi-LLM methods and motivate future work. We envision multi-LLM collaboration as an essential path toward compositional intelligence and collaborative AI development.

Oral Session D: Linguistic theories, Cognitive Modeling and Psycholinguistics 2

09:00-10:30 (Harbor H-I)

[Long] CoSToM: Causal-oriented Steering for Intrinsic Theory-of-Mind Alignment in Large Language Models ▶ *Underline Video*
Mengfan Li, Xuanhua Shi and Yang Deng
Theory of Mind (ToM), the ability to attribute mental states to others, is a hallmark of social intelligence. While large language models (LLMs) demonstrate promising performance on standard ToM benchmarks, we observe that they often fail to generalize to complex task-specific scenarios, relying heavily on prompt scaffolding to mimic reasoning. The critical misalignment between the internal knowledge and external behavior raises a fundamental question: Do LLMs truly possess intrinsic cognition, and can they externalize this internal knowledge into stable, high-quality behaviors? To answer this, we introduce CoSToM (Causal-oriented Steering for ToM alignment), a framework that transitions from mechanistic interpretation to active intervention. First, we employ causal tracing to map the internal distribution of ToM features, empirically uncovering the internal layers' characteristics in encoding fundamental ToM semantics. Building on this insight, we implement a lightweight alignment framework via targeted activation steering within these ToM-critical layers. Experiments

demonstrate that CoSToM significantly enhances human-like social reasoning capabilities and downstream dialogue quality.

[Long] CxMP: A Linguistic Minimal-Pair Benchmark for Evaluating Constructional Understanding in Language Models ▶ *Underline Video*

Miyu Oba and Saku Sugawara

Understanding language acquisition in language models remains an open question, yet many benchmarks focus on grammatical acceptability, with far less attention to interpreting meanings conveyed by grammatical forms. We introduce the Linguistic Minimal-Pair Benchmark for Evaluating Constructional Understanding in Language Models (CxMP), grounded in Construction Grammar, which treats form-meaning pairings (constructions) as fundamental linguistic units. It evaluates whether models interpret the semantic information implied by constructions, using a controlled minimal-pairs across nine types. Our results show that constructional understanding develops more gradually and remains limited for some constructions even in large language models (LLMs), whereas performance on grammatical acceptability emerges earlier, with shallow heuristics in CxMP exhibiting a U-shaped pattern. These findings highlight the need to broaden existing linguistic evaluations to capture meanings encoded in linguistic form.

[Long] Different types of syntactic agreement recruit the same units within large language models ▶ *Underline Video*

Daria Kryvosheieva, Andrea Gregor de Varda, Evelina Fedorenko and Greta Tuckute

Large language models (LLMs) can reliably distinguish grammatical from ungrammatical sentences, but how grammatical knowledge is represented within the model remains an open question. We investigate whether different syntactic phenomena recruit shared or distinct components in LLMs. Using a functional localization approach inspired by cognitive neuroscience, we identify the LLM units most responsive to 67 English syntactic phenomena in seven open-weight models. These units are consistently recruited across sentence instances and causally support model performance. Critically, different types of syntactic agreement (e.g., subject-verb, anaphor, determiner-noun) recruit overlapping sets of units, suggesting that agreement constitutes a meaningful functional category in LLMs. This pattern holds in Russian and Chinese, and further, across languages: in a cross-lingual analysis of 57 languages, syntactically similar languages share more units for subject-verb agreement. Taken together, these findings reveal that syntactic agreement, a critical marker of syntactic dependencies, constitutes a meaningful category within LLMs' representational spaces.

[Long] MauBERT: Universal Phonetic Inductive Biases for Few-Shot Acoustic Units Discovery ▶ *Underline Video*

Angelo Ortiz Tandazo, Manel Khentout, Youssef Benchekroun, Thomas Hueber and Emmanuel Dupoux

This paper introduces MauBERT, a multilingual extension of HuBERT that leverages articulatory features for robust cross-lingual phonetic representation learning. We continue HuBERT pre-training with supervision based on a phonetic-to-articulatory feature mapping in 55 languages. Our models learn from multilingual data to predict articulatory features or phones, resulting in language-independent representations that capture multilingual phonetic properties. Through comprehensive ABX discriminability testing, we show MauBERT models produce more context-invariant representations than state-of-the-art multilingual self-supervised learning models. Additionally, the models effectively adapt to unseen languages and casual speech with minimal self-supervised fine-tuning (10 hours of speech). This establishes an effective approach for instilling linguistic inductive biases in self-supervised speech models.

[Long] Memory efficiency and resource-rational encoding in sentence processing ▶ *Underline Video*

Weijie Xu, Brian Dillon and Richard Futrell

There is a growing consensus that, in order to serve as models of human language processing, language models (LMs) need to be constrained in their use of memory for context, the analogue to human working memory (WM). Here we take a novel yet simple approach to constraining WM in language models, in a way that reflects models of human cognition where memory is treated as a limited resource and deployed strategically. In order to capture this constraint on memory encoding, we inject noise into the hidden representations of Transformer-based LMs at tunable rates. Then we train the models with a hybrid objective, such that they learn to maximize the performance of next-word prediction subject to explicit constraints on the total encoding precision. We find that explicit WM constraints improve the model's alignment with human reading times. More importantly, we find that the need to manage encoding precision reshapes the nature of the models' context representations, making them more compressed and categorical. Our results show how resource-rational models of WM allocation can be implemented in neural models simply and successfully, and point to a dissociation between WM retrieval mechanisms and the underlying memory representations in models of human sentence processing.

[TACL] Psychometric Item Validation Using Virtual Respondents with Trait-Response Mediators ▶ *Underline Video*

Sungjib Lim, Woojung Song, Yohan Jo and Eun-Ju Lee

As psychometric surveys are increasingly used to assess the traits of large language models (LLMs), the need for scalable survey item generation suited for LLMs has also grown. A critical challenge here is ensuring the construct validity of generated items, i.e., whether they truly measure the intended trait. Traditionally, this requires costly, large-scale human data collection. To make it efficient, we present a framework for virtual respondent simulation using LLMs. Our central idea is to account for mediators: factors through which the same trait can give rise to varying responses to a survey item. By simulating respondents with diverse mediators, we identify survey items that robustly measure intended traits. Experiments on three psychological trait theories (Big5, Schwartz, VIA) show that our mediator generation methods and simulation framework effectively identify high-validity items. LLMs demonstrate the ability to generate plausible mediators from trait definitions and to simulate respondent behavior for item validation. Our problem formulation, metrics, methodology, and dataset open a new

direction for cost-effective survey development and a deeper understanding of how LLMs simulate human survey responses. We publicly release our dataset and code to support future work.

[Long] Systematicity between Forms and Meanings across Languages Supports Efficient Communication ▶ *Underline Video*
Doreen Osmelak, Yang Xu, Michael Hahn and Kate McCurdy
Languages vary widely in how meanings map to word forms. These mappings have been found to support efficient communication; however, this theory does not account for systematic relations within word forms. We examine how a restricted set of grammatical meanings (e.g. person, number) are expressed on verbs and pronouns across typologically diverse languages. Consistent with prior work, we find that verb and pronoun forms are shaped by competing communicative pressures for simplicity (minimizing the inventory of grammatical distinctions) and accuracy (enabling recovery of intended meanings). Crucially, our proposed model uses a novel measure of complexity (inverse of simplicity) based on the learnability of meaning-to-form mappings. This innovation captures fine-grained regularities in linguistic form, allowing better discrimination between attested and unattested systems, and establishes a new connection from efficient communication theory to systematicity in natural language.

[Long] Towards A Scanpath-Conditioned Surprisal Theory: Modeling Reader Information States ▶ *Underline Video*
Michael Mooney and Edmond S. L. Ho
Standard surprisal is typically computed from the linear text prefix, but human reading is non-linear and memory constrained: readers skip words, regress, and do not retain prior context perfectly. We propose a formulation of surprisal conditioned on a reader-specific accessible information state given by the scanpath history and memory dynamics, rather than by the written prefix alone. Prior context is treated as only probabilistically accessible at each fixation, allowing predictability to depend on both non-linear exposure and forgetting. We evaluate the approach on eye-tracking corpora using held-out log-likelihood over standard duration based reading measures. Across model variants, conditioning on accessible information states improves predictive fit over standard surprisal baselines. These results suggest that predictability in human reading is better characterized relative to the readers evolving accessible information state than to the written prefix alone.

[Long] Using Perspectival Words Is Harder Than Vocabulary Words for Humans , and Even More So for Multimodal Language Models ▶ *Underline Video*
Dota Tianai Dong, Yifan Luo, Po-Ya Angela Wang, Asli Ozyurek and Paula Rubio-Fernandez
Multimodal language models (MLMs) increasingly demonstrate human-like communication, yet their use of everyday perspectival words remains poorly understood. To address this gap, we compare humans and MLMs in their use of three word types, which we predict impose increasing cognitive demands: vocabulary (e.g., 'boat' or 'cup'), possessives (e.g., 'mine' vs. 'yours'), and demonstratives (e.g., 'this one' vs. 'that one'). Testing seven MLMs against human participants, we find that perspectival words are harder than vocabulary words for both groups. The gap is even larger for MLMs: while models approach human-level performance on using vocabulary, they exhibit clear deficits with possessives and even greater difficulties with demonstratives. Ablation analyses point to limitations in perspective-taking and spatial reasoning as key sources of these gaps in MLMs. Instruction-based prompting helps close the gap for possessives but still leaves demonstratives far below human performance. These results show that, unlike vocabulary, perspectival words pose a greater challenge in human communication, and this difficulty is further amplified in MLMs, revealing a crucial shortfall in their pragmatic and social-cognitive abilities.

Oral Session D: Low-resource Methods for NLP 1

09:00-10:30 (Harbor D)

[Long] A Survey of Reinforcement Learning for Large Language Models under Data Scarcity: Challenges and Solutions ▶ *Underline Video*
Zhiyin Yu, Yuchen Mou, Juncheng Yan, Junyu Luo, Chunchun Chen, Xing Wei, Yunhui Liu, Hongru Sun, Yuxing Zhang, Jun Xu, Yatao Bian, Ming Zhang, Wei Ye, Tieke He, Jie Yang, Guanjie Zheng, Zhonghai Wu, Bo Zhang, Lei Bai and Xiao Luo
Reinforcement learning (RL) has emerged as a powerful post-training paradigm for enhancing the reasoning capabilities of large language models (LLMs). However, reinforcement learning for LLMs faces substantial data scarcity challenges, including the limited availability of high-quality external supervision and the constrained volume of model-generated experience. These limitations make data-efficient reinforcement learning a critical research direction. In this survey, we present the first systematic review of reinforcement learning for LLMs under data scarcity. We propose a bottom-up hierarchical framework built around three complementary perspectives: the data-centric perspective, the training-centric perspective, and the framework-centric perspective. We develop a taxonomy of existing methods, summarize representative approaches in each category, and analyze their strengths and limitations. Our taxonomy aims to provide a clear conceptual foundation for understanding the design space of data-efficient RL for LLMs and to guide researchers working in this emerging area. We hope this survey offers a comprehensive roadmap for future research and inspires new directions toward more efficient and scalable reinforcement learning post-training for LLMs.

[Long] Beyond End-to-End: Dynamic Chain Optimization for Private LLM Adaptation on the Edge ► *Underline Video*

Yebo Wu, Jingguang Li, Chunlin Tian, Kahou Tam, Zhijiang Guo and Li Li

Federated fine-tuning enables privacy-preserving LLM adaptation but faces a critical bottleneck: the disparity between LLMs' high memory demands and edge devices' limited capacity. To break the memory barrier, we propose Chain Federated Fine-tuning (ChainFed), an innovative paradigm that forgoes end-to-end updates in favor of a sequential, layer-by-layer manner. It first trains the initial adapter to convergence, freezes its weights, and then proceeds to the next. This iterative train-and-freeze process forms an optimization chain, gradually enhancing the model's task-specific proficiency. ChainFed further integrates three core techniques: 1) Dynamic Layer Co-Tuning to bridge semantic gaps between sequentially tuned layers and facilitate information flow; 2) Globally Perceptive Optimization to endow each adapter with foresight beyond its local objective; 3) Function-Oriented Adaptive Tuning to automatically identify the optimal fine-tuning starting point. Extensive experiments on multiple benchmarks demonstrate the superiority of ChainFed over existing methods, boosting average accuracy by up to 46.46%.

[Long] Collision to Cognition: Hash-Driven Graph Construction for Efficient RAG ► *Underline Video*

Chuang Zhou, Zheng Yuan, Linhao Luo, Zhaozhuo Xu, Yilin Xiao, Junnan Dong, Siyu An, di yin, Xing Sun and Xiao Huang

Retrieval-Augmented Generation (RAG) has long been a promising paradigm for enhancing large language models (LLMs) with external knowledge. Traditional embedding-based methods for graph construction can capture semantic similarity but struggle to establish fine-grained, interpretable logical relationships. Recently, Graph-enhanced RAG (GraphRAG) has gained increasing popularity for its capability in modeling logical relationships. However, graph construction requires extensive token consumption for triple extraction and summarization, making it costly and slow. Accordingly, we propose MeshRAG, a novel framework for mining efficient structures via hashing to enhance RAG. We adopt an inductive paradigm in which global graph structure emerges from local hash collisions rather than explicit symbolic extraction. By replacing neural embedding search with lightweight and bitwise operations, MeshRAG automates a simple and rapid graph construction process. Furthermore, the hash collision mechanism provides transparent evidence for logical connections and retrieval decisions. Experimental results show that MeshRAG outperforms existing baselines, while its graph construction requires no GPU resources or token budget and can structure over ten thousand chunks in a few minutes.

[Long] ImCoref-CeS: An Improved Lightweight Pipeline for Coreference Resolution with LLM-based Checker-Splitter Refinement ► *Underline Video*

Kangyang Luo, Yuzhuo Bai, Shuzheng Si, Cheng Gao, Zhitong Wang, Yingli Shen, Wenhao Li, Zhu Liu, Yufeng Han, Jiayi Wu, Cunliang Kong and Maosong Sun

Coreference Resolution (CR) is a critical task in Natural Language Processing (NLP). Current research faces a key dilemma: whether to further explore the potential of supervised neural methods based on small language models, whose detect-then-cluster pipeline still delivers top performance, or embrace the powerful capabilities of Large Language Models (LLMs). However, effectively combining their strengths remains underexplored. To this end, we propose ImCoref-CeS, a novel framework that integrates an enhanced supervised model with LLM-based reasoning. First, we present an improved CR method (ImCoref) to push the performance boundaries of the supervised neural method by introducing a lightweight bridging module to enhance long-text encoding capability, devising a biaffine scorer to comprehensively capture positional information, and invoking a hybrid mention regularization to improve training efficiency. Importantly, we employ an LLM acting as a multi-role Checker-Splitter agent to validate candidate mentions (filtering out invalid ones) and coreference results (splitting erroneous clusters) predicted by ImCoref. Extensive experiments demonstrate the effectiveness of ImCoref-CeS, which achieves superior performance compared to existing state-of-the-art (SOTA) methods.

[Long] MoA: Heterogeneous Mixture of Adapters for Parameter-Efficient Fine-Tuning of Large Language Models ► *Underline Video*

Jie Cao

Recent studies integrate Low-Rank Adaptation (LoRA) and Mixture-of-Experts (MoE) to further enhance the performance of parameter-efficient fine-tuning (PEFT) methods in Large Language Model (LLM) applications. Existing methods employ homogeneous MoE-LoRA architectures composed of LoRA experts with either similar or identical structures and capacities. However, these approaches often suffer from representation collapse and expert load imbalance, which negatively impact the potential of LLMs. To address these challenges, we propose a heterogeneous Mixture-of-Adapters (MoA) approach. This method dynamically integrates PEFT adapter experts with diverse structures, leveraging their complementary representational capabilities to foster expert specialization, thereby enhancing the effective transfer of pre-trained knowledge to downstream tasks. MoA supports two variants: (i) Soft MoA achieves fine-grained integration by performing a weighted fusion of all expert outputs; (ii) Sparse MoA activates adapter experts sparsely based on their contribution, achieving this with negligible performance degradation. Experimental results demonstrate that heterogeneous MoA outperforms homogeneous MoE-LoRA methods in both performance and parameter efficiency. Our project is available at https://github.com/DCDmllm/MoA.

[Long] Neuron-Aware Active Few-Shot Learning for LLMs ► *Underline Video*

Zhuowei Chen, Liwei Chen, Christian Schunn, Raquel Coelho and Xiang Lorraine Li

Active Few-Shot Learning (AFSL) adapts LLMs to specialized domains by identifying the most valuable unlabeled samples for annotation and use as few-shot demonstrations, effectively reducing human annotation costs while promoting high performance. However, existing methods typically rely on output-level signals for the sample identification, such as predictive entropy or semantic similarities with test-time data based on external embeddings, which often overlook models' internal dynamics which could pinpoint specific knowledge gaps.

To bridge this gap, we propose NeuFS, a Neuron-Aware Active Few-Shot Learning framework that shifts the selection paradigm from output-level proxies to models' internal dynamics. NeuFS utilizes neuron activation patterns to represent sample directly, and includes a dual-criteria selection strategy that: (1) ensures few-shot sample diversity with neuron patterns for broader example coverage, while (2) prioritizing on identifying informative and challenging few-shot samples LLMs tend to hallucinate by quantifying neuron consensus. Experiments on three datasets demonstrate that NeuFS excels in both reasoning and text classification tasks, outperforming existing AFSL baselines. Ablation studies further highlight that internal neuron activations provide a more principled and effective selection signal than external embeddings, validating the superiority of the proposed NeuFS.

[Long] PRISP: Privacy-Safe Few-Shot Personalization via Lightweight Adaptation ▶ *Underline Video*

Junho Park, Dohoon Kim and Taesup Moon

Large language model (LLM) personalization aims to adapt general-purpose models to individual users. Most existing methods, however, are developed under data-rich and resource-abundant settings, often incurring privacy risks. In contrast, realistic personalization typically occurs after deployment under (i) extremely limited user data, (ii) constrained computational resources, and (iii) strict privacy requirements. We propose PRISP, a lightweight and privacy-safe personalization framework tailored to these constraints. PRISP leverages a Text-to-LoRA hypernetwork to generate task-aware LoRA parameters from task descriptions, and enables efficient user personalization by optimizing a small subset of task-aware LoRA parameters together with minimal additional modules using few-shot user data. Experiments on a few-shot variant of the LaMP benchmark demonstrate that PRISP achieves strong overall performance compared to prior approaches, while reducing computational overhead and eliminating privacy risks.

[Short] Skill-Aware Data Selection and Fine-Tuning for Data-Efficient Reasoning Distillation ▶ *Underline Video*

Lechen Zhang, Yunxiang Zhang, Wei Hu and Lu Wang

Large reasoning models such as DeepSeek-R1 and their distilled variants achieve strong performance on complex reasoning tasks. Yet, distilling these models often demands large-scale data for supervised fine-tuning (SFT), motivating the pursuit of data-efficient training methods. To address this, we propose a skill-centric distillation framework that efficiently transfers reasoning ability to weaker models with two components: (1) Skill-based data selection, which prioritizes examples targeting the student models weaker skills, and (2) Skill-aware fine-tuning, which encourages explicit skill decomposition during problem solving. With only 1,000 training examples selected from a 100K teacher-generated corpus, our method surpasses random SFT baselines by +1.6% on Qwen3-4B and +1.4% on Qwen3-8B across five mathematical reasoning benchmarks. Further analysis confirms that these gains concentrate on skills emphasized during training, highlighting the effectiveness of skill-centric training for efficient reasoning distillation.

[Short] Translation or Recitation? Calibrating Evaluation Scores for Machine Translation of Extremely Low-Resource Languages ▶ *Underline Video*

Danlu Chen, Ka Sing He, Jiahe Tian, Chenghao Xiao, Zhaofeng Wu, Taylor Berg-Kirkpatrick and Freda Shi

The landscape of extremely low-resource machine translation (MT) is characterized by perplexing variability in reported performance, often making results across different language pairs difficult to contextualize. For researchers focused on specific language groups, such as ancient languages, it is nearly impossible to determine if breakthroughs reported in other contexts (e.g., African or American languages) result from superior methodologies or are merely artifacts of benchmark collection. To address this, we introduce the FRED Difficulty Metrics, Fertility Ratio (F), Retrieval Proxy (R) Pre-training Exposure (E) and Corpus Diversity (D) , that serve as dataset-intrinsic metrics to contextualize reported scores. Our findings reveal that a significant portion of result variability is explained by train-test overlap and pre-training exposure rather than model capability. Additionally, we identify that underperforming XLR languages, particularly extinct and non-Latin indigenous languages, suffer from poor tokenization coverage (high token fertility), highlighting structural limitations of transfer learning for languages outside pre-trained models' representation space. By providing these indices alongside performance scores, we enable more transparent evaluation of cross-lingual transfer and provide a more reliable foundation for the XLR MT community.

Oral Session D: Question Answering 1

09:00-10:30 (Regatta)

[Long] Act as you think: Reinforcing Consistent Reasoning in Medical Visual Question Answering ▶ *Underline Video*

Songtao Jiang, Yuan Wang, Ruizhe Chen, Yan Zhang, Ruilin Luo, Bohan Lei, Yeying Jin, Sibo Song, ZhiBo Yang, Jimeng Sun, Jian Wu and Zuozhu Liu

While reinforcement learning from verifiable rewards (RLVR) has been proven highly effective for enhancing reasoning, its application to medical visual question answering (Med-VQA) is hampered by models producing reasoning inconsistent with either the visual evidence or the final answer. Our analysis reveals a critical flaw in RLVR training: it paradoxically encourages models to disregard visual evidence and generate answers that contradict their own reasoning. This degradation is most pronounced in specialized medical modalities (e.g., Fundus, Ultrasound) where base VLMs lack robust understanding, a failure we attribute to a flawed reward mechanism exacerbated by the scarcity of diverse training data. To tackle this, we introduce Med-Zero-17K, a large-scale dataset spanning over 30 modalities and 24

clinically relevant tasks, and the Multi-Consistency Reward (MCR) framework, which explicitly rewards both perceptual grounding and logical coherence. Extensive experiments validate our approach: integrating MCR into the RLVR framework delivers robust performance gains. This success stems from our crucial finding that rewarding internal consistency is significantly more effective than attempting to judge reasoning correctness. Furthermore, MCR proves highly versatile, exhibiting strong generalization across diverse VLM backbones, compatibility with RL algorithms like GRPO and DPO, and extending its effectiveness to 3D VQA tasks and R1-style training paradigms. Code and dataset will be released.

[Long] Deliberative Searcher: Improving LLM Reliability via Reinforcement Learning with Constraints ▶ *Underline Video*

Zhenyun Yin, Shujie Wang, Xuhong Wang, Xingjun Ma and Yingchun Wang

Large language models with search capabilities frequently exhibit miscalibrated confidence, producing incorrect answers with high certainty. We present Deliberative Searcher, a reasoning-primary framework that integrates search operations into chain-of-thought generation while maintaining explicit confidence calibration. Our method employs constrained reinforcement learning with adaptive Lagrangian multipliers to jointly optimize correctness and reliability. Experiments across five benchmarks demonstrate substantial improvements: our 7B model reduces average false-certain rates from 54% in baselines to 2%, while our 72B variant achieves competitive accuracy with closed-source models and reduces false-certain rates to 9%. The well-calibrated confidence scores also enable more efficient test-time compute: instead of standard majority voting, we use confidence-weighted aggregation and match the performance of 16-sample majority voting with only 4 samples, a 4x reduction in inference compute. These results establish calibrated confidence as a foundation for both trustworthy outputs and adaptive test-time compute, demonstrating the value of the proposed constrained RL framework in search-augmented language models.

[Long] ET-Agent: Incentivizing Effective Tool-Integrated Reasoning Agent via Behavior Calibration ▶ *Underline Video*

Yifei Chen, Guanting Dong and Zhicheng Dou

Large Language Models (LLMs) can extend their parameter knowledge limits by adopting the Tool-Integrated Reasoning (TIR) paradigm. However, existing LLM-based agent training framework often focuses on answers' accuracy, overlooking specific alignment for behavior patterns. Consequently, agent often exhibits ineffective actions during TIR tasks, such as redundant and insufficient tool calls. How to calibrate erroneous behavioral patterns when executing TIR tasks, thereby exploring effective trajectories, remains an open-ended problem. In this paper, we propose ET-Agent, a training framework for calibrating agent's tool-use behavior through two synergistic perspectives: Self-evolving Data Flywheel and Behavior Calibration Training. Specifically, we introduce a self-evolutionary data flywheel to generate enhanced data, used to fine-tune LLM to improve its exploration ability. Based on this, we implement an two-phases behavior-calibration training framework. It is designed to progressively calibrate erroneous behavioral patterns to optimal behaviors. Further in-depth experiments confirm the superiority of ET-Agent across multiple dimensions, including correctness, efficiency, reasoning conciseness, and tool execution accuracy. Our ET-Agent framework provides practical insights for research in the TIR field. Codes can be found in ~https://github.com/RUC-NLPIR/ET-Agent

[Long] G^2RPO-A: Guided Group Relative Policy Optimization with Adaptive Guidance ▶ *Underline Video*

Yongxin Guo, Wenbo Deng, Zhenglin Cheng and Xiaoying Tang

Reinforcement Learning with Verifiable Rewards (RLVR) has markedly enhanced the reasoning abilities of large language models (LLMs). Its success, however, largely depends on strong base models with rich world knowledge, yielding only modest improvements for small-size language models (SLMs). To address this limitation, we investigate Guided GRPO, which injects ground-truth reasoning steps into roll-out trajectories to compensate for SLMs inherent weaknesses. Through a comprehensive study of various guidance configurations, we find that naively adding guidance delivers limited gains. These insights motivate G^2RPO-A, an adaptive algorithm that automatically adjusts guidance strength in response to the models evolving training dynamics. Experiments on mathematical reasoning and code-generation benchmarks confirm that G^2RPO-A substantially outperforms vanilla GRPO. Our code and models at available at https://github.com/T-Lab-CUHKSZ/G2RPO-A.

[Long] Mitigating Context Interference for Reliable and Efficient Search Agents ▶ *Underline Video*

Boyang XUE, Bin Wu, Shuofei Qiao, Sheng Wang, Rui Wang, Yiming Du, Hongru WANG, Jeff Z. Pan, Emine Yilmaz, Kam-Fai Wong and Aldo Lipani

Recent research empowers Large Language Models (LLMs) as multi-turn search agents to iteratively retrieve and generate outputs until complex tasks are solved. However, the contexts of multi-turn search agents are lengthy and complex. For example, the retrieved set of documents in each turn would inevitably introduce irrelevant information that distracts LLMs, referring to context interference, potentially hindering the reliability and efficiency of search agents. Therefore, we conduct a systematic study on context interference in multi-turn search agents, focusing on investigating i) which parts of the context of search agents will contribute to the context interference, ii) how to refine the contexts of search agents to mitigate the interference, and iii) can incorporating context refinement into search agent training yield further improvements. We reveal that interference primarily arises from the latest retrieved documents. Based on the explored findings, we then introduce a distill-based context refiner to dynamically mitigate context interference for multi-turn search agents. Finally, we validate that incorporating context refinement into RL training pipelines of search agents can significantly enhance both reliability and efficiency. This study highlights the importance of mitigating context interference of search agents, inspiring a novel paradigm of "refine context and then generate" for AI agents.

[TACL] MoNaCo: More Natural and Complex Questions for Reasoning Across Dozens of Documents ▶ *Underline Video*
Tomer Wolfson, Harsh Trivedi, Mor Geva, Dan Roth, Tushar Khot, Ashish Sabharwal and Reut Tsarfaty
Automated agents, powered by Large language models (LLMs), are emerging as the go-to tool for querying information. However, evaluation benchmarks for LLM agents rarely feature natural questions that are both information-seeking and genuinely time-consuming for humans. To address this gap we introduce MoNaCo, a benchmark of 1,315 natural and time-consuming questions that require dozens, and at times hundreds, of intermediate steps to solve — far more than any existing QA benchmark. To build MoNaCo, we developed a decomposed annotation pipeline to elicit and manually answer real-world time-consuming questions at scale. Frontier LLMs evaluated on MoNaCo achieve at most 61.2% F1, hampered by low recall and hallucinations. Our results underscore the limitations of LLM-powered agents in handling the complexity and sheer breadth of real-world information-seeking tasks — with MoNaCo providing an effective resource for tracking such progress. The MoNaCo benchmark, codebase, prompts and models predictions are all publicly available at: https://tomerwolgithub.github.io/monaco

[Long] SeLaR: Selective Latent Reasoning in Large Language Models ▶ *Underline Video*
Renyu Fu and Guibo Luo
Chain-of-Thought (CoT) has become a cornerstone of reasoning in large language models, yet its effectiveness is constrained by the limited expressiveness of discrete token sampling. Recent latent reasoning approaches attempt to alleviate this limitation by replacing discrete tokens with soft embeddings (probability-weighted mixtures of token embeddings) or hidden states, but they commonly suffer from two issues: (1) global activation injects perturbations into high-confidence steps, impairing reasoning stability; and (2) soft embeddings quickly collapse toward the highest-probability token, limiting exploration of alternative trajectories. To address these challenges, we propose SeLaR (Selective Latent Reasoning), a lightweight and training-free framework. SeLaR introduces an entropy-gated mechanism that activates soft embeddings only at low-confidence steps, while preserving discrete decoding at high-confidence steps. Additionally, we propose an entropy-aware contrastive regularization that pushes soft embeddings away from the highest-probability token's direction, encouraging sustained exploration of multiple latent reasoning paths. Experiments on five reasoning benchmarks demonstrate that SeLaR consistently outperforms standard CoT and state-of-the-art training-free methods.

[Long] TRACE: Traversal Retrieval-Augmented Chain of Evidence for Document Understanding ▶ *Underline Video*
Liqi He, Zuchao Li, Hao Huang and Ping Wang
Early Long-context Document Visual Question Answering (DocVQA) methods struggle with preserving visual semantics or handling finite context windows. Conversely, recent RAG-based approaches suffer from "semantic gaps" and "structural disconnections" due to passive retrieval mechanisms that ignore logical dependencies. To address these challenges, we introduce TRACE (Traversal Retrieval-Augmented Chain of Evidence). By navigating a Bi-Layered Graph that encodes both physical adjacency and semantic relevance, TRACE transforms retrieval from static matching into adaptive evidence chain construction. Furthermore, we propose M5BookVQA, a benchmark designed to assess deep, multi-hop reasoning in books, addressing the limitations of existing datasets. Extensive experiments show that TRACE achieves an average accuracy improvement of 14.07% on M5BookVQA and exhibits robust generalization with a 13.38% gain across four established benchmarks. Our source code is available at https://github.com/shimurenhlq/TRACE.

[Long] When LLMs Read Tables Carelessly: Measuring and Reducing Data Referencing Errors ▶ *Underline Video*
Yuqing Yang, Qi Zhu, Zhen Han, Boran Han, Zhengyuan Shen, Shuai Wang, Vassilis N. Ioannidis and Huzefa Rangwala
While large language models (LLMs) perform well on table tasks, they still make data referencing errors (DREs), i.e., incorrectly citing or omitting table values, despite understanding the table structure. Beyond final-answer accuracy, DREs directly compromise the correctness and reliability of intermediate reasoning steps. Yet prior studies have only offered limited, small-scale analyses. In this work, we present the first systematic evaluation of tabular data referencing errors across different models and tasks. Our results show that DREs occur across all tested models (1.7B to 20B parameters). Furthermore, we demonstrate that incorporating data referencing as a critic significantly improves answer accuracy up to 12.0%, through critic-based filtering and rejection sampling. Finally, we trained a lightweight 4B-parameter critic model that achieves an average F1 score of 78.2% in detecting both in-distribution and out-of-distribution DREs, and effectively assists inference for larger models.

Oral Session D: Special Theme: Explainability of NLP Models 1

09:00-10:30 (Harbor G)

[Long] Assessing the Belief Consistency of Large Language Models on the Logical Conversation Process ▶ *Underline Video*
Tomoki Tsujimura, Matss Rikters, Masaki Asada, Shusaku Egami, Tatsuya Ishigaki, Ken Yano and Hiroya Takamura
To reliably interpret the evolving context of an LLM as a reasoning trace, the underlying belief of the LLM needs to transition consistently with the progression of the context. We focus on evaluating whether the beliefs held by a model remain consistent before and after the extension of the context. Previous research on consistency evaluation typically uses datasets with ground-truth answers, which is problematic because task-solving ability acts as a confounding factor, obscuring the direct evaluation of consistency. Furthermore, evaluating cases

where inconsistency stems from multiple errors poses difficulties. We propose a new evaluation method to assess the consistency of LLMs in a multiple-choice question answering format, designed so that any option chosen is correct, allowing for the evaluation of the proposed belief consistency. It also supports isolation of errors such as reasoning failures and biases. We reveal that the belief consistency does not improve solely with model size scaling, whereas continual pre-training on code and mathematics text improves it. Furthermore, models trained on code and mathematics text show a seemingly contradictory result of increased logical failures, indicating that belief consistency and superficial consistency are not necessarily directly linked.

[CL] Evaluation Framework for Highlight Explanations of Context Utilisation in Language Models ▶ *Underline Video*
Jingyi Sun, Pepa Atanasova, Sagnik Ray Choudhury, Sekh Mainul Islam and Isabelle Augenstein
Context utilisation, the ability of Language Models (LMs) to incorporate relevant information from the provided context when generating responses, remains largely opaque to users, who cannot determine whether models draw from parametric memory or provided context, nor identify which specific context pieces inform the response. Highlight explanations (HEs) offer a natural solution as they can point to the exact context pieces and tokens that influenced model outputs. However, no existing work evaluates their effectiveness in accurately explaining context utilisation. We address this gap by introducing the first gold standard HE evaluation framework for context attribution, using controlled test cases with known ground-truth context usage, thereby avoiding the limitations of existing indirect proxy evaluations. To demonstrate the frameworks broad applicability, we evaluate four HE methods, three established techniques and MechLight, a mechanistic interpretability approach we adapt for this task, across four context scenarios, four datasets, and five LMs. Overall, we find that MechLight performs best across all context scenarios. However, our findings reveal systematic failures in all methods: explanation accuracy degrades significantly with context length, and all methods exhibit strong positional biases in multi-document settings. Surprisingly, widely used gradient-based methods provide little value for understanding context usage. These results challenge HEs utility in retrieval-augmented generation, factual verification, and other applications. Our framework provides the foundation for developing accurate context attribution methods.

[Long] Faithful Serum: Mitigating the Faithfulness Gap in Textual Explanations of LLM Decisions via Attribution Guidance ▶ *Underline Video*
Bar Alon, Itamar Zimerman and Lior Wolf
Large language models (LLMs) achieve strong performance and have revolutionized NLP, but their lack of explainability keeps them treated as black boxes, limiting their use in domains that demand transparency and trust. A promising direction to address this issue is post-hoc text-based explanations, which aim to explain model decisions in natural language. Prior work has focused on generating convincing rationales that appear to be subjectively faithful, but it remains unclear whether these explanations are epistemically faithful - that is, whether they reflect the internal evidence the model actually relied on for its decision. In this paper, we first assess the epistemic faithfulness of LLM-generated explanations via counterfactuals and show that they are often unfaithful. We then introduce a training-free method, that enhances faithfulness by guiding explanation generation through attention-level interventions, informed by token-level heatmaps extracted via a faithful attribution method. This method significantly improves epistemic faithfulness across multiple models, benchmarks, and prompts. Our code is attached as supplementary material.

[Long] How Do LLMs and VLMs Understand Viewpoint Rotation Without Vision? An Interpretability Study ▶ *Underline Video*
Zhen Yang, Ping Jian, Zhongbin Guo, Zuming Zhang, Chengzhi Li, Yonghong Deng, Xinyue Zhang and Wenpeng Lu
Over the past year, spatial intelligence has drawn increasing attention. Many prior works study it from the perspective of visual-spatial intelligence, where models have access to visuospatial information from visual inputs. However, in the absence of visual information, whether linguistic intelligence alone is sufficient to endow models with spatial intelligence, and how models perform relevant tasks with text-only inputs still remain unexplored. Therefore, in this paper, we focus on a fundamental and critical capability in spatial intelligence from a linguistic perspective: viewpoint rotation understanding (VRU). Specifically, LLMs and VLMs are asked to infer their final viewpoint and predict the corresponding observation in an environment given textual description of viewpoint rotation and observation over multiple steps. We find that both LLMs and VLMs perform poorly on our proposed dataset while human can easily achieve 100% accuracy, indicating a substantial gap between current model capabilities and the requirements of spatial intelligence. To uncover the underlying mechanisms, we conduct a layer-wise probing analysis and head-wise causal intervention. Our findings reveal that although models encode viewpoint information in the hidden states, they appear to struggle to bind the viewpoint position with corresponding observation, resulting in a hallucination in final layers. Finally, we selectively fine-tune the key attention heads identified by causal intervention to improve VRU performance. Experimental results demonstrate that such selective fine-tuning achieves improved VRU performance while avoiding catastrophic forgetting of generic abilities.

[Long] Is Chain-of-Thought Really Not Explainability? Chain-of-Thought Can Be Faithful without Hint Verbalization ▶ *Underline Video*
Kerem Zaman and Shashank Srivastava
Recent work, using the Biasing Features metric, labels a CoT as unfaithful if it omits a prompt-injected hint that affected the prediction. We argue this metric adopts a narrow notion of faithfulness and confuses unfaithfulness with incompleteness, the lossy compression needed to turn distributed transformer computation into a linear natural language narrative. On multi-hop reasoning tasks with instruct-tuned and reasoning models, many CoTs flagged as unfaithful by Biasing Features are judged faithful by other metrics, exceeding 50% in some models. With a new faithful@k metric, we show that larger inference-time budgets greatly increase hint verbalization (up to 90% in some

settings), suggesting much apparent unfaithfulness is due to tight token limits. Using Causal Mediation Analysis, we further show that even non-verbalized hints can causally mediate prediction changes through the CoT. We therefore caution against relying solely on hint-based evaluations and advocate a broader interpretability toolkit, including causal mediation and corruption-based metrics. We do not claim all CoTs are faithful, only that the absence of hint words alone does not prove unfaithfulness.

[Long] Mapping the Circumplex of Affect: Geometric Analysis of Emotion Representations via Hyperspherical Contrastive Learning ▶ *Underline Video*

Yusuke Yamauchi and Akiko Aizawa

Psychological research has long utilized circumplex models to structure emotions, placing similar emotions adjacently and opposing ones diagonally. Although frequently used to interpret deep learning representations, these models are rarely directly incorporated into the representation learning of language models, leaving their geometric validity unexplored. This paper proposes a method to induce circular emotion representations within language model embeddings via contrastive learning on a hypersphere. We show that while this circular alignment offers superior interpretability and robustness against dimensionality reduction, it underperforms compared to conventional designs in high-dimensional settings and fine-grained classification. Our findings elucidate the trade-offs involved in applying psychological circumplex models to deep learning architectures.

[Long] Mechanistic Interpretability Should Prioritize Feature Consistency in Sparse Autoencoders ▶ *Underline Video*

Xiangchen Song, Aashiq Muhamed, Yujia Zheng, Lingjing Kong, Zeyu Tang, Mona T. Diab, Virginia Smith and Kun Zhang

Sparse Autoencoders (SAEs) are a prominent tool in mechanistic interpretability (MI) for decomposing neural network activations into interpretable features. However, the aspiration to identify a canonical set of features is challenged by the observed inconsistency of learned SAE features across different training runs, undermining reproducibility and complicating model comparison. We study run-to-run feature consistency in SAEs and argue that it should be reported as a standard evaluation axis alongside reconstruction and sparsity. We propose the Pairwise Dictionary Mean Correlation Coefficient (PW-MCC) as an assignment-based metric to quantify consistency and demonstrate that high levels are achievable (PW-MCC approx 0.80 for TopK SAEs on LLM activations) with appropriate architectural choices. Our contributions include: (i) theoretical grounding for strong consistency in the idealized setting of TopK SAEs; (ii) synthetic validation using a model organism, which verifies PW-MCC as a reliable proxy for ground-truth recovery; and (iii) empirical analysis on LLM activations, where PW-MCC correlates with the similarity of automatically generated natural-language feature explanations.

[Long] Rhetorical Questions in LLM Representations: A Linear Probing Study ▶ *Underline Video*

Louie Hong Yao, Vishesh Anand, Yuan Zhuang and Tianyu Jiang

Rhetorical questions are asked not to seek information but to persuade or signal stance. How large language models internally represent them remains unclear. We analyze rhetorical questions in LLM representations using linear probes on two social-media datasets with different discourse contexts, and find that rhetorical signals emerge early and are most stably captured by last-token representations. Rhetorical questions are linearly separable from information-seeking questions within datasets, and remain detectable under cross-dataset transfer, reaching AUROC around 0.7-0.8. However, we demonstrate that transferability does not simply imply a shared representation. Probes trained on different datasets produce different rankings when applied to the same target corpus, with overlap among the top-ranked instances often below 0.2. Qualitative analysis shows that these divergences correspond to distinct rhetorical phenomena: some probes capture discourse-level rhetorical stance embedded in extended argumentation, while others emphasize localized, syntax-driven interrogative acts. Together, these findings suggest that rhetorical questions in LLM representations are encoded by multiple linear directions emphasizing different cues, rather than a single shared direction.

[Long] Truth as a Trajectory: What Internal Representations Reveal About Large Language Model Reasoning ▶ *Underline Video*

Hamed Damirchi, Ignacio Meza De la Jara, Ehsan Abbasnejad, Afshar Shamsi, Zhen Zhang and Javen Shi

Existing explainability methods for Large Language Models (LLMs) typically treat hidden states as static points in activation space, assuming that correct and incorrect inferences can be separated using representations from an individual layer. However, these activations are saturated with polysemantic features, leading to linear probes learning surface-level lexical patterns rather than underlying reasoning structures. We introduce Truth as a Trajectory (TaT), which models the transformer inference as an unfolded trajectory of iterative refinements, shifting analysis from static activations to layer-wise geometric displacement. By analyzing displacement of representations across layers, TaT captures structural patterns in the evolution of inference that distinguish valid reasoning from spurious behavior. We evaluate TaT across dense and Mixture-of-Experts (MoE) architectures on benchmarks spanning commonsense reasoning, question answering, and toxicity detection. Without access to the activations themselves and using only changes in activations across layers, we show that TaT effectively mitigates reliance on static lexical confounds, outperforming conventional probing, and establishes trajectory analysis as a complementary perspective on LLM explainability.

Oral Session D: Speech Processing and Spoken Language Understanding 1

09:00-10:30 (Harbor A)

[Long] Difference in Task Performance on Sparse Speech Representations ► *Underline Video*

Wenjie Peng, Chen Chen and Thomas Hain

Learning speech representations that are useful for a variety of downstream tasks has received considerable attention, due to the outstanding properties of Self-Supervised Learning (SSL) trained models. Despite advancements in modelling methods, understanding the difference in task performance on representations is limited. Mainly motivated by the no-free-lunch theorem and speech production, this work investigates changes in task performance in sparse speech representations, providing interpretability analysis under the Information Bottleneck (IB) framework. Autoencoders with varying sparsity levels were trained using three SSL features, and evaluated on six tasks of SUPERB: Speech Enhancement (SE), Speaker Identification (SID), Speech Emotion Recognition (SER), Phone Recognition (PR), Automatic Speech Recognition (ASR) and Slot Filling (SF). Experiments show that: 1) different tasks manifest different degrees of sensitivity to the sparsity levels; 2) the optimal sparsity level for task performance varies; 3) the choice of SSL features has a limited impact on most tasks but with an exception of PR; 4) overall PR and ASR require more preservation of relevant information about the labels, while SID and SER demand more compression of irrelevant information, where the input quality can shift this trade-off to some degree. These findings can contribute to the design of a universal sparse speech representation learner.

[Long] Evaluating the Expressive Appropriateness of Speech in Rich Contexts ► *Underline Video*

Tianrui Wang, Ziyang Ma, Yizhou Peng, Haoyu Wang, Zhikang Niu, Zikang Huang, Yihao Wu, Yi-Wen Chao, Yu Jiang, Yuheng Lu, Guanrou Yang, Xuanchen Li, Hexin Liu, Chunyu Qiang, Cheng Gong, Yifan Yang, Tianchi Liu, Junyu Wang, Nana Hou, Meng Ge, Fuming You, Yang Wei, Zhongqian Sun, Hu Haifeng, Xiaobao Wang, Eng Siong Chng, Xie Chen, Longbiao Wang and Jianwu Dang

Evaluating expressive speech remains challenging, as existing methods mainly assess emotional intensity and overlook whether a speech sample is expressively appropriate for its contextual setting. This limitation hinders reliable evaluation of speech systems used in narrative-driven and interactive applications, such as audiobooks and conversational agents. We introduce CEAEval, a Context-rich framework for Evaluating Expressive Appropriateness in speech, which assesses whether a speech sample expressively aligns with the underlying communicative intent implied by its discourse-level narrative context. To support this task, we construct CEAEval-D, the first context-rich speech dataset with real human performances in Mandarin conversational speech, providing narrative descriptions together with fifteen dimensions of human annotations covering expressive attributes and expressive appropriateness. We further develop CEAEval-M, a model that integrates knowledge distillation, planner-based multi-model collaboration, adaptive audio attention bias, and reinforcement learning to perform context-rich expressive appropriateness evaluation. Experiments on a human-annotated test set demonstrate that CEAEval-M substantially outperforms existing speech evaluation and analysis systems.

[Long] From Naturalness to Norms: Interactional Cultural Competence for SpeechLMs ► *Underline Video*

Santosh Tokala

Spoken language models (SpeechLMs) are increasingly real-time conversational actors. Yet many culturally consequential aspects of spoken interaction are not primarily lexical. Across sociolinguistics, linguistic anthropology, and conversation analysis, meaning emerges through how talk is produced and coordinated—prosody, timing, turn-taking, overlap, backchannels, and repair—within situated speech events. A transcript can be semantically correct yet interactionally inappropriate because many culture-bearing signals are audible and sequential rather than textual. This position paper argues for a speech-first view of cultural competence as interactional competence: the ability of a spoken agent to participate appropriately in event-situated interaction with locally normative conduct, while allowing plural acceptable realizations. Here, appropriate does not imply generic human-likeness; in many applications, the desired behavior may instead be constrained, neutral, predictable, or tool-like under an application-specific interaction contract. We synthesize social-science foundations into a theory-derived taxonomy of culture-bearing signals in speech, identify interactional phenomena where transcript correctness fails to predict appropriateness, and ground the agenda in today's SpeechLM stacks and evaluation practice. We propose an evaluation framing that complements WER/MOS and broad capability suites by making speech events and interaction contracts explicit, diagnosing where modern pipelines lose interactional cues, and treating cultural appropriateness as a norm-conditioned target rather than generic "naturalness."

[Short] Full-Duplex-Bench-v2: A Multi-Turn Evaluation Framework for Duplex Dialogue Systems with an Automated Examiner ► *Underline Video*

Guan-Ting Lin, Shih-Yun Shan Kuan, Jiatong Shi, Kai-Wei Chang, Siddhant Arora, Shinji Watanabe and Hung-yi Lee

While full-duplex speech agents enable natural, low-latency interaction by speaking and listening simultaneously, their consistency and task performance in multi-turn settings remain underexplored. We introduce Full-Duplex-Bench-v2 (FDB-v2), a streaming framework that integrates with an automated examiner that enforces staged goals under two pacing setups (Fast vs. Slow). FDB-v2 covers four task families, Daily, Correction, Entity Tracking, and Safety, and reports turn-taking fluency, multi-turn instruction following, and task-specific competence. The framework is extensible, supporting both commercial APIs and open-source models. When we test full-duplex systems with FDB-v2, they often get confused when people talk at the same time, struggle to handle corrections smoothly, and sometimes lose track of who or what is being talked about. Through an open-source, standardized streaming protocol and a task set, FDB-v2 makes it easy to extend to new task families, allowing the community to tailor and accelerate evaluation of multi-turn full-duplex systems.

[Long] On the Emotion Understanding of Synthesized Speech ► *Underline Video*

Yuan Ge, Haishu Zhao, Aokai Hao, Junxiang Zhang, Bei Li, Xiaoqian Liu, Chenglong Wang, Jianjin Wang, Bingsen Zhou, Bingyu Liu,

Jingbo Zhu, Zhengtao Yu and Tong Xiao
Emotion is a core paralinguistic feature in voice interaction. It is widely believed that emotion understanding models learn fundamental representations that transfer to synthesized speech, making emotion understanding results a plausible reward or evaluation metric for assessing emotional expressiveness in speech synthesis. In this work, we critically examine this assumption by systematically evaluating Speech Emotion Recognition (SER) on synthesized speech across datasets, discriminative and generative SER models, and diverse synthesis models. We find that current SER models can not generalize to synthesized speech, largely because speech token prediction during synthesis induces a representation mismatch between synthesized and human speech. Moreover, generative Speech Language Models (SLMs) tend to infer emotion from textual semantics while ignoring paralinguistic cues. Overall, our findings suggest that existing SER models often exploit non-robust shortcuts rather than capturing fundamental features, and paralinguistic understanding in SLMs remains challenging.

[Long] POWSM: A Phonetic Open Whisper-Style Speech Foundation Model ▶ *Underline Video*
Chin-Jou Li, Kalvin Chang, Shikhar Bharadwaj, Eunjung Yeo, Kwanghee Choi, Jian Zhu, David R. Mortensen and Shinji Watanabe
Recent advances in spoken language processing have led to substantial progress in phonetic tasks such as automatic speech recognition (ASR), phone recognition (PR), grapheme-to-phoneme conversion (G2P), and phoneme-to-grapheme conversion (P2G). Despite their conceptual similarity, these tasks have largely been studied in isolation, each relying on task-specific architectures and datasets. In this paper, we introduce POWSM (Phonetic Open Whisper-style Speech Model), the first unified framework capable of jointly performing multiple phone-related tasks. POWSM enables seamless conversion between audio, text (graphemes), and phones, opening up new possibilities for universal and low-resource speech processing. Our model outperforms or matches specialized PR models of similar size (Wav2Vec2Phoneme and ZIPA) while jointly supporting G2P, P2G, and ASR. Our training data, models, and code are released to foster open science.

[Long] SegTune: Structured and Fine-Grained Control for Song Generation ▶ *Underline Video*
Yuejiao Wang, Zihao Ji, Pengfei Cai, Xu Li, Haorui Zheng, Zewen Song, Zhongliang Liu, Chen Zhang and Pengfei Wan
Recent advances in neural song generation have enabled high-quality synthesis from lyrics and global textual prompts. However, most systems fail to model temporally varying attributes of songs, severely limiting fine-grained control over musical structure and dynamics. To address this, we propose Segtune, a Diffusion Transformer-based framework enabling structured and fine-grained controllability by allowing users or large language models (LLMs) to specify local musical descriptions aligned to song segments. These segment prompts are temporally broadcast to corresponding time windows, while global prompts ensure stylistic coherence. To support precise lyric-to-music alignment, we introduce an LLM-based duration predictor that autoregressively generates sentence-level timestamps in LyRiCs format. We further construct a large-scale data pipeline for high-quality song collection with aligned lyrics and prompts, and propose new metrics to evaluate segment alignment and vocal consistency. Experiments demonstrate that Segtune outperforms existing baselines in both musicality and controllability. Visit our demo page for codes and more generated songs.

[Long] UniSonate: A Unified Model for Speech, Music, and Sound Effect Generation with Text Instructions ▶ *Underline Video*
Chunyu Qiang, Xiaopeng Wang, Kang Yin, Yuzhe Liang, Yuxin Guo, Ziyu Zhang, Teng Ma, Tianrui Wang, Cheng Gong, Yushen CHEN, Ruibo Fu, Longbiao Wang and Jianwu Dang
Generative audio modeling has largely been fragmented into specialized tasks, text-to-speech (TTS), text-to-music (TTM), and text-to-audio (TTA), each operating under heterogeneous control paradigms. Unifying these modalities remains a fundamental challenge due to the intrinsic dissonance between structured semantic representations (speech/music) and unstructured acoustic textures (sound effects). In this paper, we introduce UniSonate, a unified flow-matching framework capable of synthesizing speech, music, and sound effects through a standardized, reference-free natural language instruction interface. To reconcile structural disparities, we propose a novel dynamic token injection mechanism that projects unstructured environmental sounds into a structured temporal latent space, enabling precise duration control within a phoneme-driven Multimodal Diffusion Transformer (MM-DiT). Coupled with a multi-stage curriculum learning strategy, this approach effectively mitigates cross-modal optimization conflicts. Extensive experiments demonstrate that UniSonate achieves state-of-the-art performance in instruction-based TTS (WER 1.47%) and TTM (SongEval Coherence 3.18), while maintaining competitive fidelity in TTA. Crucially, we observe positive transfer, where joint training on diverse audio data significantly enhances structural coherence and prosodic expressiveness compared to single-task baselines.

[Long] When Misinformation Speaks and Converses: Rethinking Fact-Checking in Audio Platforms ▶ *Underline Video*
Chaewan Chun, Delvin Ce Zhang and Dongwon Lee
Audio platforms have evolved beyond entertainment. They have become central to public discourse, from podcasts and radio to WhatsApp voice notes and live streams. With millions of shows and hundreds of millions of listeners, audio platforms are now a major channel for misinformation. Yet existing fact-checking pipelines are mostly designed for written claims, overlooking the unique properties of spoken media. We argue that audio misinformation is not merely textual content with transcripts: it is structurally different because it is both spoken, carrying persuasive force through prosody, pacing, and emotion, and conversational, unfolding across turns, speakers, and episodes. These dual properties introduce verification difficulties that traditional methods rarely face. This position paper synthesizes evidence across modalities and platforms, examines datasets and methods, and highlights why existing pipelines fail on audio. We argue that advancing fact-checking requires rethinking verification pipelines around the spoken and conversational realities of audio.

Poster Session D (Coronado)

09:00-10:30 (Coronado)

[Findings] PeerCheck: Enhancing LLM-Generated Academic Reviews Towards Human-Level Quality ► *Underline Video*

Zeyuan Chen, Ziqing Yang, Yihan Ma, Michael Backes and Yang Zhang

As academic submissions grow, the traditional peer review process struggles to keep up, raising concerns about quality and fairness. A trend of using large language models (LLMs) for assistance has emerged. In this work, we take a critical step toward improving the quality of LLM-generated reviews. We propose the PeerCheck framework, which investigates LLM-human review differences (RQ1) and explores methods to increase LLM-human similarity (RQ2). We first analyzed the human-written reviews with reviews generated by GPT-4o, Claude-3.7-Sonnet, and DeepSeek-V3 and found that LLMs and humans focus on different terms, e.g., LLMs prioritize theory while humans emphasize methodology and experiments. We further adopt prompt engineering, such as Chain-of-Thought (CoT), and utilize retrieval-augmented generation (RAG) to enhance the LLM-generated reviews towards human-level quality. We find CoT significantly improves the human similarity of LLM reviews, while we also discover an unexpected "RAG paradox," i.e., experiments with RAG produce different results for various LLMs and, in some cases, even reduce review quality. Our comprehensive analysis of LLM-generated academic reviews illustrates both possibilities and limitations, contributing to a more effective, human-aligned review system.

[Long] A Survey of Large Language Models for Text-Guided Molecular Discovery: From Molecule Generation to Optimization ► *Underline Video*

Ziqing Wang, Kexin Zhang, Zihan Zhao, Yibo Wen, Abhishek Pandey, Han Liu and Kaize Ding

Large language models (LLMs) are introducing a paradigm shift in molecular discovery by enabling text-guided interaction with chemical spaces through natural language and symbolic notations, with emerging extensions to incorporate multi-modal inputs. To advance this emerging field, this survey provides an up-to-date and forward-looking review of the emerging use of LLMs for two central tasks: molecule generation and molecule optimization. We organize our survey around four fundamental challenges that have emerged as critical evaluation dimensions in recent studies: ensuring validity, enhancing synthesizability, achieving precise property control, and maximizing diversity. Based on this, we systematically analyze how current LLM learning paradigms are applied to tackle each challenge, revealing the distinct capabilities and inherent limitations of each approach. In addition, we include the commonly used datasets and evaluation protocols aligned with these challenges. We conclude by discussing future directions, positioning this survey as a resource for researchers working at the intersection of LLMs and molecular science. A continuously updated reading list is available at https://github.com/REAL-Lab-NU/Awesome-LLM-Centric-Molecular-Discovery.

[Long] AdaFuse: Adaptive Ensemble Decoding for Large Language Models ► *Underline Video*

Chengming Cui, Tianxin Wei, Ziyi Chen, Ruizhong Qiu, Zhichen Zeng, Zhining Liu, Xuying Ning, Duo Zhou and Jingrui He

Large language models (LLMs) exhibit complementary strengths arising from differences in pretraining data, model architectures, and decoding behaviors. Inference-time ensembling provides a practical way to combine these capabilities without retraining. However, existing ensemble approaches suffer from fundamental limitations. Most rely on fixed fusion granularity, which lacks the flexibility required for mid-generation adaptation and fails to adapt to different generation characteristics across tasks. To address these challenges, we propose AdaFuse, an adaptive ensemble decoding framework that dynamically selects semantically appropriate fusion units during generation. Rather than committing to a fixed granularity, AdaFuse adjusts fusion behavior on the fly based on the decoding context, with words serving as basic building blocks for alignment. To be specific, we introduce an uncertainty-based criterion to decide whether to apply ensembling at each decoding step. Under confident decoding states, the model continues generation directly. In less certain states, AdaFuse invokes a diversity-aware scaling strategy to explore alternative candidate continuations and inform ensemble decisions. This design establishes a synergistic interaction between adaptive ensembling and test-time scaling, where ensemble decisions guide targeted exploration, and the resulting diversity in turn strengthens ensemble quality. Experiments on open-domain QA, arithmetic reasoning, and machine translation demonstrate that AdaFuse consistently outperforms strong ensemble baselines, achieving an average relative improvement of 6.88%.

[Findings] Agree, Disagree, Explain: Decomposing Human Label Variation in NLI through the Lens of Explanations ► *Underline Video*

Pingjun Hong, Beiduo Chen, Siyao Peng, Marie-Catherine de Marneffe, Benjamin Roth and Barbara Plank

Natural Language Inference (NLI) datasets often exhibit human label variation. To better understand these variations, explanation-based approaches analyze the underlying reasoning behind annotators decisions. One such approach is the LiTEx taxonomy, which categorizes free-text explanations in English into reasoning categories. However, previous work applying LiTEx has focused on within-label variation: cases where annotators agree on the NLI label but provide different explanations. This paper broadens the scope by examining how annotators may diverge not only in the reasoning category but also in the labeling. We use explanations as a lens to analyze variation in NLI annotations and to examine individual differences in reasoning. We apply LiTEx to two NLI datasets and align annotation variation from multiple aspects: NLI label agreement, explanation similarity, and taxonomy agreement, with an additional compounding factor of annotators selection bias. We observe instances where annotators disagree on the label but provide similar explanations, suggesting that surface-level disagreement may mask underlying agreement in interpretation. Moreover, our analysis reveals individual preferences in

explanation strategies and label choices. These findings highlight that agreement in reasoning categories better reflects the semantic similarity of explanations than label agreement alone. Our findings underscore the richness of reasoning-based explanations and the need for caution in treating labels as ground truth.

[Findings] AlgBench: To What Extent Do Large Reasoning Models Understand Algorithms? ▶ *Underline Video*

Henan Sun, Kaichi Yu, Yuyao Wang, Bowen Liu, Xunkai Li, Rong-Hua Li, Nuo Chen and Jia Li

Reasoning ability has become a central focus in the advancement of Large Reasoning Models (LRMs). Although notable progress has been achieved on several reasoning benchmarks such as MATH500 and LiveCodeBench, existing benchmarks for algorithmic reasoning remain limited, failing to answer a critical question: Do LRMs truly master algorithmic reasoning? To answer this question, we propose AlgBench, an expert-curated benchmark that evaluates LRMs under an algorithm-centric paradigm. AlgBench consists of over 3,000 original problems spanning 27 algorithms, constructed by ACM algorithmic experts and organized under a comprehensive taxonomy, including Euclidean-structured, non-Euclidean-structured, non-optimized, local-optimized, global-optimized, and heuristic-optimized categories. Empirical evaluations on leading LRMs (e.g., Gemini-3-Pro, DeepSeek-v3.2-Speciale and GPT-o3) reveal substantial performance heterogeneity: while models perform well on non-optimized tasks (up to 92%), accuracy drops sharply to around 49% on globally optimized algorithms such as dynamic programming. Further analysis uncovers strategic over-shifts, wherein models prematurely abandon correct algorithmic designs due to necessary low-entropy tokens. These findings expose fundamental limitations of problem-centric reinforcement learning and highlight the necessity of an algorithm-centric training paradigm for robust algorithmic reasoning.

[Long] AlignCultura: Towards Culturally Aligned Large Language Models? ▶ *Underline Video*

Gautam Siddharth Kashyap, Mark Dras and Usman Naseem

Cultural alignment in Large Language Models (LLMs) is essential for producing contextually aware, respectful, and trustworthy outputs. Without it, models risk generating stereotyped, insensitive, or misleading responses that fail to reflect cultural diversity w.r.t Helpful, Harmless, and Honest (HHH) paradigm. Existing benchmarks represent early steps toward cultural alignment; yet, no benchmarks currently enables systematic evaluation of cultural alignment in line with UNESCOs principles of cultural diversity w.r.t HHH paradigm. Therefore, to address this gap, we built Align-Cultura, two-stage pipeline for cultural alignment. Stage I constructs CULTURAX, the HHH-English dataset grounded in the UNESCO cultural taxonomy, through Query Construction, which reclassifies prompts, expands underrepresented domains (or labels), and prevents data leakage with SimHash. Then, Response Generation pairs prompts with culturally grounded responses via two-stage rejection sampling. The final dataset contains 1,500 samples spanning 30 subdomains of tangible and intangible cultural forms. Stage II benchmarks CULTURAX on general-purpose models, culturally fine-tuned models, and open-weight LLMs (Qwen3-8B and DeepSeek-R1-Distill-Qwen-7B). Empirically, culturally fine-tuned models improve joint HHH by 4%-6%, reduce cultural failures by 18%, achieve 10%-12% efficiency gains, and limit leakage to 0.3%.

[Findings] ANCHOR: LLM-driven Subject Conditioning for Text-to-Image Synthesis ▶ *Underline Video*

Aashish Anantha Ramakrishnan, Sharon Huang and Dongwon Lee

Text-to-image (T2I) models have achieved remarkable progress in high-quality image synthesis, yet most benchmarks rely on simple, self-contained prompts, failing to capture the complexity of real-world captions. Human-written captions often involve multiple interacting subjects, rich contextual references, and abstractive phrasing, conditions under which current image-text encoders like CLIP struggle. To systematically study these deficiencies, we introduce ANCHOR, a large-scale dataset of 70K+ abstractive captions sourced from five major news media organizations. Analysis with ANCHOR reveals persistent failures in multi-subject understanding, context reasoning, and nuanced grounding. Motivated by these challenges, we propose Subject-Aware Fine-tuning (SAFE), which uses Large Language Models (LLMs) to extract key subjects and enhance their representation at the embedding-level. Experiments with contemporary models show that SAFE significantly improves image-caption consistency and human preference alignment, serving as a practical and scalable solution. The dataset and code will be released upon publication.

[Findings] Article and Comment Frames Shape the Quality of Online Comments ▶ *Underline Video*

Matteo Guida, Yulia Otmakhova, Eduard Hovy and Lea Frermann

Framing theory posits that how information is presented shapes audience responses, but computational work has largely ignored audience reactions. While recent work has shown that article framing systematically shapes the content of reader responses, this paper asks: does framing also affect response quality? Analyzing 1M comments across 2.7K news articles, we operationalize quality as comment health. We find that article frames significantly predict comment health while controlling for topic, and that comments that adopt the article frame are healthier than those that depart from it. Further, unhealthy top-level comments tend to generate more unhealthy responses, independent of the frame being used in the comment. Our results establish a link between framing theory and discourse quality, laying the groundwork for downstream applications. We illustrate this potential with a pro-active frame-aware LLM- based system to mitigate unhealthy discourse.

[Findings] Attention to Non-Adopters ▶ *Underline Video*

Kaitlyn Zhou, Kristina Gligori, Myra Cheng, Michelle S. Lam, Vyoma Raman, Boluwatife Aminu, Caeley Woo, Michael Brockman, Hannah Cha and Dan Jurafsky

Although language model-based chat systems are increasingly used in daily life, most Americans remain non-adopters of chat-based LLMs — as of June 2025, 66% had never used ChatGPT. At the same time, LLM development and evaluation rely mainly on data from adopters

(e.g., logs, preference data), focusing on the needs and tasks for a limited demographic group of adopters in terms of geographic location, education, and gender. In this position paper, we argue that incorporating non-adopter perspectives is essential for developing broadly useful and capable LLMs. We contend that relying on methods that focus primarily on adopters will risk missing a range of tasks and needs prioritized by non-adopters, entrenching inequalities in who benefits from LLMs, and creating oversights in model development and evaluation. To illustrate this claim, we conduct case studies with non-adopters and show: how non-adopter needs diverge from those of current users, how non-adopter needs point us towards novel reasoning tasks, and how to systematically integrate non-adopter needs via human-centered methods.

[Long] Beyond Itinerary Planning, A Real-World Benchmark for Multi-Turn and Tool-Using Travel Tasks ▶ *Underline Video*

Xiang Cheng, Yulan Hu, Xiangwen Zhang, Lu Xu, Lide Tan, Zheng Pan, Xin Li and Yong Liu

Travel planning is a natural real-world task to test large language models' (LLMs) planning and tool-use abilities. Although prior work has studied LLM performance on travel planning, existing settings still differ from real-world needs, mainly due to limited domain coverage, insufficient modeling of users' implicit preferences in multi-turn conversations, and a lack of evaluation of agents' capability boundaries. To mitigate these gaps, we propose TravelBench, a benchmark for truly real-world travel planning. We collect user queries, user preferences, and tools from real scenarios, and construct three subtasks—Single-Turn, Multi-Turn, and Unsolvable—to evaluate agents' three core capabilities in real settings: (1) solving problems independently, (2) interacting with users to elicit implicit preferences, and (3) recognizing the capability boundaries. To enable stable tool invocation and reproducible evaluation, we cache real tool-call results and build a sandbox environment which integrates ten travel-related tools, enabling agents to combine these tools to solve most practical travel planning problems. We evaluate multiple LLMs on TravelBench and find that even advanced models exhibit imbalanced performance across different capabilities. Our further systematic verification demonstrates the stability of the proposed benchmark. TravelBench provides a practical and reproducible benchmark to advance research on LLM agents for real-world travel planning.

[Findings] Beyond Screenshots: Evaluating VLMs' Understanding of UI Animations ▶ *Underline Video*

Chen Liang, Xirui Jiang, Naihao Deng, Eytan Adar and Anhong Guo

AI agents operating on user interfaces must understand how interfaces communicate state and feedback to act reliably. As a core communicative modality, animations are increasingly used in modern interfaces, serving critical functional purposes beyond mere aesthetics. Thus, understanding UI animation is essential for comprehensive interface interpretation. However, recent studies of Vision Language Models (VLMs) for UI understanding have focused primarily on static screenshots, leaving it unclear how well these models handle dynamic UI animations. To address this gap, we created AniMINT, a novel dataset of 300 densely annotated UI animation videos. We systematically evaluate state-of-the-art VLMs on UI animation understanding, including their abilities to perceive the animation effects, identify animation purposes, and interpret animation meaning. Our results show that VLMs can reliably detect primitive motion. However, their high-level animation interpretation remains inconsistent, with substantial gaps relative to human performance. Finally, we use Motion, Context, and Perceptual Cues (MCPC) to probe factors affecting VLM performance, revealing key bottlenecks and directions for future improvement.

[Findings] Budget-Aware Anytime Reasoning with LLM-Synthesized Preference Data ▶ *Underline Video*

Xuanming Zhang, Shwan Ashrafi, Aziza Mirsaidova, Amir H. Rezaeian, Miguel Ballesteros, Lydia Chilton, Zhou Yu and Dan Roth

We study the reasoning behavior of large language models (LLMs) under limited computation budgets. In such settings, producing useful partial solutions quickly is often more practical than exhaustive reasoning, which incurs high inference costs. Many real-world tasks, such as trip planning, require models to deliver the best possible output within a fixed reasoning budget. We introduce an anytime reasoning framework and the Anytime Index, a metric that quantifies how effectively solution quality improves as reasoning tokens increase. To further enhance efficiency, we propose an inference-time self-improvement method using LLM-synthesized preference data, where models learn from their own reasoning comparisons to produce better intermediate solutions. Experiments on NaturalPlan (Trip), AIME, and GPQA datasets show consistent gains across Grok-3, GPT-oss, GPT-4.1/4o, and LLaMA models, improving both reasoning quality and efficiency under budget constraints.

[Short] CaBSALLM: Efficient Context-Aware Batch Annotation of Conversational Streams with Large Language Models ▶ *Underline Video*

Mohammadsadegh Abolhasani, Reza Mousavi and Paul Jen-Hwa Hu

Analyses of parasocial cues in live-stream chats require accurate, efficient, and scalable annotation. However, manual annotation is tedious, and large language models (LLMs) often make mistakes when applying subjective, discourse-dependent labels. This study proposes Context-aware Batching for Stream Annotation with LLMs (CaBSALLM), an efficient pipeline that incorporates lightweight conversational context and a novel dynamic batching method to improve throughput and scalability. Compared with state-of-the-art pipelines, this generalizable approach is significantly more time- and cost-efficient while achieving comparable or better predictive performance and agreement.

[Findings] Can LLMs Really Judge? A Progressive Argumentation-Mining Framework for Distinguishing Understanding from Aggregation ▶ *Underline Video*

Fuyu Wang, Jiangtong Li, Kun Zhu and Changjun Jiang

Current evaluations of large language models (LLMs) mainly rely on dataset-based generation accuracy. However, generative correctness

does not guarantee the discriminative capability required to verify solutions, frequently masking an inability to distinguish valid reasoning from plausible errors. While multi-agent debate inherently entails judgment, we show that uncontrolled context growth and convergence to majority voting introduce significant noise, obscuring intrinsic model judgment. To address these limitations, we propose a progressive argumentation-mining diagnostic framework designed to explicitly control context and isolate discriminative behaviors. Instead of indiscriminate aggregation, our approach distills and retains only the single most well-supported rationale per answer, preventing context dilution while enforcing strict quality-based selection. Applying this framework reveals a fundamental cognitive divergence: models exhibit structural susceptibility to plausible misinformation in knowledge tasks, whereas in reasoning tasks they demonstrate latent discriminative potential that remains fragile under pressure. These findings underscore the fragility of discriminative capabilities, advocating for diagnostic methodologies that prioritize judgment stability over simple generation performance.

[Long] Can Persona-Prompted LLMs Emulate Subgroup Values? An Empirical Analysis of Generalisability and Fairness in Cultural Alignment ▶ *Underline Video*

Bryan Chen Zhengyu Tan, Zhengyuan Liu, Xiaoyuan Yi, Jing Yao, Xing Xie, Nancy F. Chen and Roy Ka-Wei Lee

Despite their global prevalence, many Large Language Models (LLMs) are aligned to a monolithic, often Western-centric set of values. This paper investigates the more challenging task of fine-grained value alignment: examining whether LLMs can emulate the distinct cultural values of demographic subgroups. Using Singapore as a case study and the World Values Survey (WVS), we examine the value landscape and show that even state-of-the-art models like GPT-4.1 achieve only 57.4% accuracy in predicting subgroup modal preferences. We construct a dataset of over 20,000 samples to train and evaluate a range of models. We demonstrate that simple fine-tuning on structured numerical preferences yields substantial gains, improving accuracy on unseen, out-of-distribution subgroups by an average of 17.4%. These gains partially transfer to open-ended generation. However, we find significant pre-existing performance biases, where models better emulate young, male, Chinese, and Christian personas. Furthermore, while fine-tuning improves average performance, it widens the disparity between subgroups when measured by distance-aware metrics. Our work offers insights into the limits and fairness implications of subgroup-level cultural alignment.

[Findings] CentaurTA: A Self-Improving Human-Agents Collaboration Framework for Thematic Analysis ▶ *Underline Video*

Lei Wang, Min Huang and Eduard Dragut

Qualitative analysis is essential for studying complex social and behavioral phenomena, yet existing large language model (LLM) approaches face key limitations. Fully automated pipelines often compromise methodological rigor, while fully manual coding remains costly and labor-intensive. Although recent work emphasizes human-AI collaboration, existing multi-agent systems focus primarily on theme-level outputs, provide limited human oversight, and overlook fine-grained, data-level coding quality. We introduce CentaurTA, an iterative, self-improving human-agent framework for scalable thematic analysis. CentaurTA places humans in the loop to oversee and guide analysis, using expert feedback as a persistent learning signal to drive prompt-level refinement. By combining structured human feedback with rubric-based evaluation, CentaurTA provides fine-grained supervision for both open coding and theme construction while preserving methodological rigor. Experiments across multiple datasets, baselines, and LLM families show that CentaurTA improves coding alignment and transparency, highlighting the central role of human feedback in reliable qualitative analysis. Our code and data are available at https://github.com/Tom-Owl/CentaurTA.

[Long] ChipSeek: Optimizing Verilog Generation via EDA-Integrated Reinforcement Learning ▶ *Underline Video*

Zhirong Chen, Kaiyan Chang, Zhuolin Li, Cangyuan Li, Xinyang He, Chujie Chen, Mengdi Wang, Haobo Xu, Yinhe Han, Huawei Li and Ying Wang

Large Language Models have emerged as powerful tools for automating Register-Transfer Level (RTL) code generation, yet they face critical limitations: existing approaches typically fail to simultaneously optimize functional correctness and hardware efficiency metrics such as Power, Performance, and Area (PPA). Methods relying on supervised fine-tuning commonly produce functionally correct but suboptimal designs due to the lack of inherent mechanisms for learning hardware optimization principles. Conversely, external post-processing techniques aiming to refine PPA performance after generation often suffer from inefficiency and do not improve the LLMs' intrinsic capabilities. To overcome these challenges, we propose ChipSeek, a novel hierarchical reward based reinforcement learning framework designed to encourage LLMs to generate RTL code that is both functionally correct and optimized for PPA metrics. Our approach integrates direct feedback from EDA simulators and synthesis tools into a hierarchical reward mechanism, facilitating a nuanced understanding of hardware design trade-offs. Through Curriculum-Guided Dynamic Policy Optimization (CDPO), ChipSeek enhances the LLM's ability to generate high-quality, optimized RTL code. Evaluations on standard benchmarks demonstrate ChipSeek's superior performance, achieving state-of-the-art functional correctness and PPA performance. Furthermore, it excels in specific optimization tasks, consistently yielding highly efficient designs when individually targeting fine-grained optimization goals such as power, delay, and area. The artifact is open-source in https://github.com/rong-hash/chipseek.

[Long] CL^2GEC: A Multi-Discipline Benchmark for Continual Learning in Chinese Literature Grammatical Error Correction ▶ *Underline Video*

Shang Qin, Jingheng Ye, Yinghui Li, Hai-Tao Zheng, Qi Li, Jinxiao Shan, Zhixing Li and Hong-Gee Kim

The growing demand for automated writing assistance in diverse academic domains highlights the need for robust Chinese Grammatical Error Correction (CGEC) systems that can adapt across disciplines. However, existing CGEC research largely lacks dedicated benchmarks

for multi-disciplinary academic writing, overlooking continual learning (CL) as a promising solution to handle domain-specific linguistic variation and prevent catastrophic forgetting. To fill this crucial gap, we introduce CL^2GEC, the first Continual Learning benchmark for Chinese Literature Grammatical Error Correction, designed to evaluate adaptive CGEC across multiple academic fields. Our benchmark includes 10,000 human-annotated sentences spanning 10 disciplines, each exhibiting distinct linguistic styles and error patterns. CL^2GEC focuses on evaluating grammatical error correction in a continual learning setting, simulating sequential exposure to diverse academic disciplines to reflect real-world editorial dynamics. We evaluate large language models under sequential tuning, parameter-efficient adaptation, and four representative CL algorithms, using both standard GEC metrics and continual learning metrics adapted to task-level variation. Experimental results reveal that regularization-based methods mitigate forgetting more effectively than replay-based or naive sequential approaches. Our benchmark provides a rigorous foundation for future research in adaptive grammatical error correction across diverse academic domains.

[Findings] CodeContests-O: Powering LLMs via Feedback-Driven Iterative Test Case Generation ▶ *Underline Video*

Jianfeng Cai, Jinhua Zhu, Ruopei Sun, Kangwen Zhao, Dongyun Xue, Mingxiao Feng, Wengang Zhou and Houqiang Li

The rise of reasoning models necessitates large-scale verifiable data, for which programming tasks serve as an ideal source. However, while competitive programming platforms provide abundant problems and solutions, high-quality test cases for verification remain scarce. Existing approaches attempt to synthesize test cases using Large Language Models (LLMs), but rely solely on the model's intrinsic generation capabilities without external feedback, frequently resulting in insufficiently diverse cases. To address this limitation, we propose a Feedback-Driven Iterative Framework for comprehensive test case construction. Specifically, our method leverages the LLM to generate initial test cases, executes them against known correct and incorrect solutions, and utilizes the failed results as feedback to guide the LLM in refining the test cases toward high fidelity and discriminability. We then apply this method to the CodeContests dataset to construct an optimized high-quality derivative, CodeContests-O. Evaluating against the entire pool of solutions (1.1 x 10^7 in total), our dataset achieves an average True Positive Rate (TPR) of 89.35% and True Negative Rate (TNR) of 90.30%, significantly outperforming the CodeContests and CodeContests+ by margins of 4.30% and 8.78%, respectively. Furthermore, fine-tuning the Qwen2.5-7B model on CodeContests-O results in a 9.52% improvement on LiveCodeBench (Pass@1). Experiments demonstrate the effectiveness of our framework and the quality of CodeContests-O.

[Long] Difficulty-Controllable Cloze Question Distractor Generation ▶ *Underline Video*

Seokhoon Kang, Yejin Jeon, Seonjeong Hwang and Gary Geunbae Lee

Multiple-choice cloze questions are commonly used to assess linguistic proficiency and comprehension. However, generating high-quality distractors remains challenging, as existing methods often lack adaptability and control over difficulty levels, and the absence of difficulty-annotated datasets further hinders progress. To address these issues, we propose a novel framework for generating distractors with controllable difficulty by leveraging both data augmentation and a multitask learning strategy. First, to create a high-quality, difficulty-annotated dataset, we introduce a two-way distractor generation process to produce diverse and plausible distractors. These candidates are filtered and then categorized by difficulty using an ensemble QA system. Second, this newly created dataset is used to train a difficulty-controllable generation model via multitask learning. Experimental results demonstrate that our method generates high-quality distractors across difficulty levels and substantially outperforms GPT-4o in aligning distractor difficulty with human perception.

[Findings] Do Language Models Think Consistently? A Study of Value Preferences Across Varying Response Lengths ▶ *Underline Video*

Inderjeet Nair and Lu Wang

Evaluations of LLMs ethical risks and value inclinations often rely on short-form surveys and psychometric tests, yet real-world use involves long-form, open-ended responses, leaving value-related risks and preferences in practical settings largely underexplored. In this work, we ask: Do value preferences inferred from short-form tests align with those expressed in long-form outputs? To address this question, we compare value preferences elicited from short-form reactions and long-form responses, varying the number of arguments in the latter to capture users' differing verbosity preferences. Analyzing five LLMs (llama3-8b, gemma2-9b, mistral-7b, qwen2-7b, and olmo-7b), we find (1) a weak correlation between value preferences inferred from short-form and long-form responses across varying argument counts, and (2) similarly weak correlation between preferences derived from any two distinct long-form generation settings. (3 Alignment yields only modest gains in the consistency of value expression. Further, we examine how long-form generation attributes relate to value preferences, finding that argument specificity negatively correlates with preference strength, while representation across scenarios shows a positive correlation. Our findings underscore the need for more robust methods to ensure consistent value expression across diverse applications.

[Long] Emotion-Wheel-Guided Audio-Referred Text Representation for Multimodal Emotion Recognition in Conversation ▶ *Underline Video*

Eunseon Seong, Harim Lee, Dahye Kim, Changhyun Kim and Dong-Kyu Chae

Multimodal Emotion Recognition in Conversation aims to identify emotions within a dialogue with multimodal data, including audio, visual, and textual features. While existing methods have made significant improvements, there are two fundamental limitations to be addressed. From the modality fusion perspective, current approaches treat all modalities as functionally equivalent during fusion, overlooking their distinct communicative roles and information capacities, in which text conveys explicit semantic meaning while audio provides paralinguistic cues. From the emotion label perspective, many works ignore the continuous structure of emotion characterized by psychological theory

and apply uniform penalties regardless of affective proximity. To address these limitations, we propose EMART, EMotion-Wheel-Guided Audio-Referred Text Representation for ERC, specifically focusing on audio and text modalities. First, we propose a modality-aware fusion strategy capturing linguistic features from text as the primary source and audio as a complementary component. Secondly, we propose an emotion-wheel-guided supervised contrastive loss to encode emotional proximity based on Russells circumplex model. Experimental results on IEMOCAP and MELD demonstrate outstanding performance. The code is available at: https://github.com/DILAB-HYU/EMART.git.

[Findings] Empathy Applicability Modeling for General Health Queries ▶ *Underline Video*

Shan Randhawa, Agha Ali Raza, Kentaro Toyama, Julie Hui and Mustafa Naseem

LLMs are increasingly being integrated into clinical workflows, yet they often lack clinical empathy, an essential aspect of effective doctor-patient communication. Existing NLP frameworks focus on reactively labeling empathy in doctors' responses but offer limited support for anticipatory modeling of empathy needs, especially in general health queries. We introduce the Empathy Applicability Framework (EAF), a theory-driven approach that classifies patient queries in terms of the applicability of emotional reactions and interpretations, based on clinical, contextual, and linguistic cues. We release a benchmark of real patient queries, dual-annotated by human annotators and GPT-4o. In the subset with human consensus, we also observe substantial human-GPT alignment. To validate EAF, we train classifiers on human-labeled and GPT-only annotations to predict empathy applicability, achieving strong performance and outperforming the heuristic and zero-shot LLM baselines. Error analysis highlights persistent challenges: implicit distress, clinical-severity ambiguity, and contextual hardship, underscoring the need for multi-annotator modeling, clinician-in-the-loop calibration, and culturally diverse annotation. EAF provides a framework for identifying empathy needs before response generation, establishes a benchmark for anticipatory empathy modeling, and enables supporting empathetic communication in asynchronous healthcare.

[Findings] Empirical Analysis of Task Mixture Effects in Small-scale Instruction Tuning: A Statistical Approach ▶ *Underline Video*

Jeesu Jung and Sangkeun Jung

The performance of large language models heavily depends on instruction tuning, especially on task types and mixture ratios. However, previous research has primarily focused on mixing tasks at fixed ratios, lacking a systematic and quantitative analysis of task-wise interactions across diverse tasks. Moreover, it has relied heavily on human labeling. To address these limitations, this study conducts empirical experiments on unlabeled instruction corpora, varying both the number and proportion of task combinations to identify effective mixtures. To minimize manual labeling, we automatically extract five representative tasks, programming, math problem solving, history question answering, grammar correction, and creative writing, using only a few seed instructions. Across 51 mixtures, we find that 1-2 task mixtures work best with small datasets, while synergistic 3-task mixtures excel with larger data. Task interactions reveal both synergy (e.g., programming + math) and interference (e.g., programming + creative writing). These results provide practical guidelines for mixture design tailored to model scale and data size.

[Findings] Enabling Stroke-Level Structural Analysis of Hieroglyphic Scripts without Language-Specific Priors ▶ *Underline Video*

Fuwen Luo, Zihao Wan, Ziyue Wang, Yaluo Liu, Pau Tong Lin Xu, Xuanjia Qiao, Xiaolong Wang, Peng Li and Yang Liu

Hieroglyphs, as logographic writing systems, encode rich semantic and cultural information within their internal structural composition. Yet, current advanced Large Language Models (LLMs) and Multimodal LLMs (MLLMs) usually remain structurally blind to this information. LLMs process characters as textual tokens, while MLLMs additionally view them as raw pixel grids. Both fall short to model the underlying logic of character strokes. Furthermore, existing structural analysis methods are often script-specific and labor-intensive. In this paper, we propose Hieroglyphic Stroke Analyzer (HieroSA), a novel and generalizable framework that enables MLLMs to automatically derive stroke-level structures from character bitmaps without handcrafted data. It transforms modern logographic and ancient hieroglyphs character images into explicit, interpretable line-segment representations in a normalized coordinate space, allowing for cross-lingual generalization. Extensive experiments demonstrate that HieroSA effectively captures character-internal structures and semantics, bypassing the need for language-specific priors. Experimental results highlight the potential of our work as a graphematics analysis tool for a deeper understanding of hieroglyphic scripts.

[Long] FAIRGAMER: Evaluating Social Biases in LLM-Based Video Game NPCs ▶ *Underline Video*

Bingkang Shi, Jen-tse Huang, Luo Long, Tianyu Zong, Hongzhu Yi, Yuanxiang Wang, Songlin Hu, Xiaodan Zhang and Zhongjiang Yao

Large Language Models (LLMs) have increasingly enhanced or replaced traditional Non-Player Characters (NPCs) in video games. However, these LLM-based NPCs inherit underlying social biases (e.g., race or class), posing fairness risks during in-game interactions. To address the limited exploration of this issue, we introduce FairGamer, the first benchmark to evaluate social biases across three interaction patterns: transaction, cooperation, and competition. FairGamer assesses four bias types, including class, race, age, and nationality, across 12 distinct evaluation tasks using a novel metric, FairMCV. Our evaluation of seven frontier LLMs reveals that: (1) models exhibit biased decision-making, with Grok-4-Fast demonstrating the highest bias (average FairMCV = 76.9%); and (2) larger LLMs display more severe social biases, suggesting that increased model capacity inadvertently amplifies these biases. We release FairGamer at https://github.com/BingkangShi/FairGamer to facilitate future research on NPC fairness.

[Long] FedMental: Evaluating Federated Learning for Mental Health Detection from Social Media Data ▶ *Underline Video*

Nuredin Ali Abdelkadir, Anjali Ratnam, Zeerak Talat and Stevie Chancellor

Social media text data are often used to train Machine Learning (ML) models to identify users exhibiting high-risk mental health behaviors.

However, sharing this sensitive data poses privacy risks and limits the growth of benchmark datasets. We comprehensively evaluate whether privacy-preserving ML techniques can enable safer data sharing while preserving performance. Specifically, we apply federated learning (FL) and Differentially Private FL for two widely-studied mental health prediction tasks: depression detection on X (Twitter) and suicide crisis detection on Reddit. We simulate realistic data-sharing scenarios by treating each user as a client in a non-IID setting, evaluating across different client fractions, aggregation strategies, and privacy budgets. While FL achieves comparable performance to centralized training (centralized F 1 = 85.63; best FL model F 1 = 83.16) on depression identification, we find that Differentially Private FL has a large performance-privacy trade-off (up to F 1 = 27.01 drop) even with low levels of noise (= 50). This is due to the distortion of highly informative yet sparse mental health linguistic markers related to mental health, like health topics and emotion words. This research empirically demonstrates the potential and limitations of current privacy preservation techniques for mental health inference tasks.

[Long] Fiction Flows: A Replication and Reinterpretation of Narrative Sequentiality ▶ *Underline Video*

Andrew Piper, Sil Hamilton, Haiqi Zhou and Federico Pianzola

Narrative flow emerges from the interplay between memory and expectation, shaping how stories are both produced and understood. To operationalize this construct, Sap et al. (2022) propose sequentiality, a language-model-based measure of sentence-level predictability, and report that imagined stories flow better than recalled ones. We conduct a large-scale replication across multiple language models, examine how modeling choices shape the original findings, and test generalization beyond crowdworker data using passages from published fiction and narrative non-fiction. Although the original contrast replicates under their initial formulation, it diminishes substantially under alternative specifications, suggesting that it reflects properties of the measurement setup rather than a stable feature of narrative flow. By contrast, fiction does appear to exhibit a robust sequentiality advantage over reality-bound genres under a minimal context-only formulation. However, mixed-effects analyses indicate that this advantage is not reducible to standard coherence measures, underscoring the need for further theoretical and empirical grounding of narrative flow.

[Findings] FIFA: Unified Faithfulness Evaluation Framework for Text-to-Video and Video-to-Text Generation ▶ *Underline Video*

Liqiang Jing, Viet Lai, Seunghyun Yoon, Trung Bui and Xinya Du

Video Multimodal Large Language Models (VideoMLLMs) have achieved remarkable progress in both Video-to-Text and Text-to-Video tasks. However, they often suffer from hallucinations, generating content that contradicts the visual input. Existing evaluation methods are limited to one task (V2T) and also fail to assess hallucinations in open-ended, free-form responses. To address this gap, we propose FIFA, a unified FaIthFulness evAluation framework that extracts comprehensive descriptive facts, models their semantic dependencies via a Spatio-Temporal Semantic Dependency Graph, and verifies them using VideoQA models. We further introduce, a tool-based correction framework that revises hallucinated content. Extensive experiments demonstrate that FIFA aligns more closely with human judgment than existing evaluation methods, and that ~ effectively improves factual consistency in both text and video generation.

[Findings] From Heard to Lived Opinions: Simulating Opinion Dynamics with Grounded LLM Agents in Economic Environments ▶ *Underline Video*

Ryuji Hashimoto, Masahiro Kaneko, Ryosuke Takata, Takehiro Takayanagi and Kiyoshi Izumi

Opinion dynamics (OD) studies how individual opinions evolve and generate collective patterns such as consensus and polarization. While recent work explores OD using populations of LLM-based agents focusing on opinion exchange, it typically does not incorporate individuals lived experiences, such as economic outcomes of past decisions, which play a critical role in shaping opinions. We propose a novel OD simulation framework that grounds LLM-based agents in an economic environment, allowing them to act and receive environmental feedback. Our simulations exhibit coherent OD at both individual and population levels: individual opinions follow structured trajectories shaped by economic experiences, with adverse conditions inducing opinion rigidity, while at the population level, collective opinions co-move with economic conditions, with inequality amplifying polarization and price instability driving larger distributional shifts. These results highlight the importance of grounding LLM-based agents in environments to capture collective OD.

[Long] From Selection to Refinement: Iterative Optimization for Instruction Data ▶ *Underline Video*

Hang Hu, Ziyan Liu, Rujie Wen, Ruihui Hou, Xueyan Wu, Mu Zhang, Jianxing Yu, Tong Ruan and Jingping Liu

Instruction tuning plays a crucial role in enhancing large language models (LLMs) to better understand complex user instructions. While various data selection and revision methods have been explored to optimize instruction tuning datasets, they face two main challenges: unreasonable pruning of potentially valuable low-quality data and the persistence of noise or semantic drift during revision. To address these issues, we propose a novel automated iterative framework for instruction data optimization. Our framework introduces Instruction Quality Differentiation to identify valuable high-quality and low-quality data across multiple dimensions. For low-quality data, we propose a Feedback-driven Iterative Refinement mechanism with an "evaluate-refine-review" process and design an Output Alignment module to improve data quality. Experiments on seven public benchmark datasets show that our framework outperforms state-of-the-art methods, achieving 2.09% and 2.60% improvements on the Alpaca and Dolly datasets, respectively, with high data efficiency. Our code and data are available at the anonymous link https: //github.com/surihuhang/From-Selection-to-Refinement–Iterative-Optimization-for-Instruction-Data.

[Findings] Generalization or Memorization? Multi-Agent vs. Baseline LLMs and AutoML Models for Tabular Classification ▶ *Underline Video*

Aida Sanatizadeh, Sorouralsadat Fatemi, Reza Mousavi and Ahmed Abbasi

Large Language Models (LLMs) are increasingly used for structured tabular data, yet it remains unclear whether their performance reflects genuine reasoning or memorization of pre-training corpora. We investigate this question through a rigorous, contamination-aware evaluation of a representative modular Multi-Agent LLM (MALLM) framework against state-of-the-art AutoML systems and established baselines (TABLET, TABLLM). We evaluate eleven binary classification tasks: five pre-cutoff benchmarks likely seen during LLM pre-training and six post-cutoff datasets released after the LLM knowledge cutoff. Results show a sharp performance dichotomy: MALLM achieves competitive or superior performance on pre-cutoff datasets but substantially underperforms AutoML on post-cutoff data, exhibiting poor calibration and high variance, especially on hard-to-classify instances. By contrast, AutoML models generalize consistently and align confidence more closely with instance hardness. These findings suggest that, despite agentic scaffolding, current LLMs cannot yet replace production-grade discriminative models for tabular classification, underscoring the need for contamination-free benchmarks to accurately assess tabular reasoning capabilities.

[Findings] Generative Interfaces for Language Models ▶ *Underline Video*

Jiaqi Chen, Yanzhe Zhang, Yutong Zhang, Yijia Shao and Diyi Yang

Large language models (LLMs) are increasingly seen as assistants, copilots, and consultants, capable of supporting a wide range of tasks through natural conversation. However, most systems remain constrained by a linear request-response format that often makes interactions inefficient in multi-turn, information-dense, and exploratory tasks. To address these limitations, we propose Generative Interfaces for Language Models, a paradigm in which LLMs respond to user queries by proactively generating user interfaces (UIs) that enable more adaptive and interactive engagement. Our framework leverages structured interface-specific representations and iterative refinements to translate user queries into task-specific UIs. For systematic evaluation, we introduce a multidimensional assessment framework that compares generative interfaces with traditional chat-based ones across diverse tasks, interaction patterns, and query types, capturing functional, interactive, and emotional aspects of user experience. Results show that generative interfaces consistently outperform conversational ones, with up to a 72% improvement in human preference. These findings clarify when and why users favor generative interfaces, paving the way for future advancements in human-AI interaction. Data and code are available at https://github.com/SALT-NLP/GenUI.

[Long] GeoArena: Evaluating Open-World Geographic Reasoning in Large Vision-Language Models ▶ *Underline Video*

Pengyue Jia, Yingyi Zhang, Xiangyu Zhao and Sharon Li

Geographic reasoning is a fundamental cognitive capability that requires models to infer plausible locations by synthesizing visual evidence with spatial world knowledge. Despite recent advances in large vision-language models (LVLMs), existing evaluation paradigms remain largely outcome-centric, relying on static datasets and predefined labels that are conceptually misaligned with open-world geographic inference. Such outcome-centric evaluations often focus exclusively on label matching, leaving the underlying linguistic reasoning chains as unexamined black boxes. In this work, we introduce GeoArena, a dynamic, human-preference-based evaluation framework for benchmarking open-world geographic reasoning. GeoArena reframes evaluation as a pairwise reasoning alignment task on in-the-wild images, where human judges compare model-generated explanations based on reasoning quality, evidence synthesis, and plausibility. We deploy GeoArena as a public platform and benchmark 17 frontier LVLMs using thousands of human judgments, which complements existing benchmarks and supports the development of geographically grounded, human-aligned AI systems. We further provide detailed analyses of model behavior, including reliability of human preferences and factors influencing judgments of geographic reasoning quality. We open-source GeoArena to foster future research.

[Long] GeoLaux: A Benchmark for Evaluating MLLMs' Geometry Performance on Long-Step Problems Requiring Auxiliary Lines ▶ *Underline Video*

Yumeng Fu, Jiayin Zhu, Lingling Zhang, Wenjun Wu, Bo Zhao, Shaoxuan Ma, Yushun Zhang and Jun Liu

Geometry problem solving (GPS) poses significant challenges for Multimodal Large Language Models (MLLMs) in diagram comprehension, knowledge application, long-step reasoning, and auxiliary line construction. However, current benchmarks lack fine-grained evaluation for long-step problems necessitating auxiliary construction. To address these limitations, we present GeoLaux, a fine-grained annotated dataset comprising 2186 calculation and proof problems. It features long-step reasoning (with an average solution length of 6.51 steps, maximum of 24 steps) and auxiliary line construction (required in 41.8% of problems). Building on the dataset, we conduct a comprehensive five-dimensional evaluation of 23 leading MLLMs. The evaluation yields three pivotal findings: First, models perform significantly worse on long-step problems compared to short-step ones, with 18 models exhibiting a performance drop of over 50%. Second, it is crucial to enhance models' understanding, awareness, and proficiency in auxiliary line construction, which is vital for overall geometric reasoning. Third, limited answer hints effectively improve process correctness, whereas explicit answers lead models to neglect intermediate reasoning steps. These findings position GeoLaux both to benchmark MLLMs geometry reasoning abilities and to guide their improvement. Data and code are available at https://github.com/Candice-yu/GeoLaux

[Long] GeoRC: A Benchmark for Geolocation Reasoning Chains ▶ *Underline Video*

Mohit Talreja, Joshua Diao, Jim James, Radu Casapu, Tejas Santanam, Ethan Mendes, Alan Ritter, Wei Xu and James Hays

Vision Language Models (VLMs) are good at recognizing the global location of a photograph – their geolocation prediction accuracy rivals the best human experts. But many VLMs are startlingly bad at explaining which image evidence led to their prediction, even when their location prediction is correct. The reasoning chains produced by VLMs frequently hallucinate scene attributes to support their location prediction (e.g. phantom writing, imagined infrastructure, misidentified flora). In this paper, we introduce the first benchmark for geolocation

reasoning chains. We focus on the global location prediction task in the popular GeoGuessr game which draws from Google Street View spanning more than 100 countries. We collaborate with expert GeoGuessr players, including the reigning world champion, to produce 800 "ground truth" reasoning chains for 500 query scenes. These expert reasoning chains address hundreds of different discriminative visual attributes such as license plate shape, architecture, and soil properties to name just a few. We evaluate LLM-as-a-judge and VLM-as-a-judge strategies for scoring VLM-generated reasoning chains against our expert reasoning chains and find that Qwen 3 LLM-as-a-judge correlates best with human scoring. Our benchmark reveals that while large, closed-source VLMs such as Gemini and GPT 5 rival human experts at prediction locations, they still lag behind human experts when it comes to producing auditable reasoning chains. Open weights VLMs such as Llama and Qwen catastrophically fail on our benchmark – they perform only slightly better than a baseline in which an LLM hallucinates a reasoning chain with oracle knowledge of the photo location but no visual information at all. We believe the gap between human experts and VLMs on this task points to VLM limitations at extracting fine-grained visual attributes from high resolution images. We open source our benchmark for the community to use.

[Long] GrACE: A Generative Approach to Better Confidence Elicitation and Efficient Test-Time Scaling in Large Language Models ▶ *Underline Video*

Zhaohan Zhang, Ziquan Liu and Ioannis Patras

Assessing the reliability of Large Language Models (LLMs) by confidence elicitation is a prominent approach to AI safety in high-stakes applications, such as healthcare and finance. Existing methods either require expensive computational overhead or suffer from poor calibration, making them impractical and unreliable for real-world deployment. In this work, we propose GrACE, a Generative Approach to Confidence Elicitation that enables scalable and reliable confidence elicitation for LLMs. GrACE adopts a novel mechanism in which the model expresses confidence by the similarity between the last hidden state and the embedding of a special token appended to the vocabulary, in real-time. We fine-tune the model for calibrating the confidence with targets associated with accuracy. Extensive experiments show that the confidence produced by GrACE achieves the best discriminative capacity and calibration on open-ended generation tasks without resorting to additional sampling or an auxiliary model. Moreover, we propose two confidence-based strategies for test-time scaling with GrACE, which not only improve the accuracy of the final decision but also significantly reduce the number of required samples, highlighting its potential as a practical solution for deploying LLMs with reliable, on-the-fly confidence estimation. The code is available at: https://github.com/petezone/Grace.

[Long] HistLens: Mapping Idea Change across Concepts and Corpora ▶ *Underline Video*

Yi Jing, Weiyun Qiu, Yihang Peng and Zhifang Sui

Language change both reflects and shapes social processes, and the semantic evolution of foundational concepts provides a measurable trace of historical and social transformation. Despite recent advances in diachronic semantics and discourse analysis, existing computational approaches often (i) concentrate on a single concept or a single corpus, making findings difficult to compare across heterogeneous sources, and (ii) remain confined to surface lexical evidence, offering insufficient computational and interpretive granularity when concepts are expressed implicitly. We propose HistLens, a unified, SAE-based framework for multi-concept, multi-corpus conceptual-history analysis. The framework decomposes concept representations into interpretable features and tracks their activation dynamics over time and across sources, yielding comparable conceptual trajectories within a shared coordinate system. Experiments on long-span press corpora show that HistLens supports cross-concept, cross-corpus computation of patterns of idea evolution and enables implicit concept computation. By bridging conceptual modeling with interpretive needs, HistLens broadens the analytical perspectives and methodological repertoire available to social science and the humanities for diachronic text analysis.

[Long] HypoEval: Hypothesis-Guided Evaluation for Natural Language Generation ▶ *Underline Video*

Mingxuan Li, Hanchen Li and Chenhao Tan

Large language models (LLMs) have demonstrated great potential for automating the evaluation of natural language generation. Previous frameworks of LLM-as-a-judge fall short in two ways: they either use zero-shot setting without consulting any human input, which leads to low alignment, or fine-tune LLMs on labeled data, which requires a non-trivial number of samples. Moreover, previous methods often provide little reasoning behind automated evaluations. In this paper, we propose HypoEval, Hypothesis-guided Evaluation framework, which first uses a small corpus of human evaluations to generate more detailed rubrics for human judgments and then incorporates a checklist-like approach to combine LLM's assigned scores on each decomposed dimension to acquire overall scores. With only 30 human evaluations, HypoEval achieves state-of-the-art performance in alignment with both human rankings (Spearman correlation) and human scores (Pearson correlation), on average outperforming G-Eval by 11.86% and fine-tuned Llama-3.1-8B-Instruct with at least 3 times more human evaluations by 11.95%. Furthermore, we conduct systematic studies to assess the robustness of HypoEval, highlighting its effectiveness as a reliable and interpretable automated evaluation framework.

[Findings] I Don't Need Solution. I Need Emotional Support: Empathetic LLMs based on Emotional Validation ▶ *Underline Video*

Suhyune Son, Jungwoo Lim, Myunghoon Kang, Seongtae Hong, Yuna Hur, Evelyn H. Zi and Heuiseok Lim

Empathy plays a crucial role in prosocial behavior and supportive human interactions. According to emotional validation theory, effective empathetic conversations require observing and reflecting on the help-seekers situation before offering emotional support and guidance. While recent advancements in large language models (LLMs) have enabled fluent and coherent dialogue generation, our preliminary study reveals that existing LLMs struggle to generate emotional support response. Instead, they tend to offer repetitive solutions without sufficiently considering the emotional needs of help-seekers. To address this limitation, we propose EVA: empathetic LLMs with Emotional

VAlidation. EVA enhances empathetic response generation through a two-stage training process: empathy acquisition and emotional validation alignment. For the emotional validation alignment, we introduce the Emotional Validation Aware Dataset (EVAD), which is annotated with levels of emotional validation theory as conversations progress. Additionally, we propose EVAEval, a novel evaluation metric designed to assess whether a model-generated response aligns with emotional validation theory. Experimental results demonstrate that the EVA method significantly improves empathetic response generation, achieving superior performance in both automatic and human evaluations. Furthermore, comprehensive analyses confirm that the EVA method effectively mitigates patterned responses while ensuring adherence to emotional validation principles.

[Findings] Imperfectly Cooperative Human-AI Interactions: Comparing the Impacts of Human and AI Attributes in Simulated and User Studies ▶ *Underline Video*

Myke C. Cohen, Mingqian Zheng, Neel Bhandari, Hsien-Te Kao, Xuhui Zhou, Daniel Nguyen, Laura Cassani, Maarten Sap and Svitlana Volkova

AI design characteristics and human personality traits each impact the quality and outcomes of human-AI interactions. However, their relative and joint impacts are underexplored in imperfectly cooperative scenarios, where people and AI only have partially aligned goals and objectives. This study compares a purely simulated dataset comprising 2,000 simulations and a parallel human subjects experiment involving 290 human participants to investigate these effects across two scenario categories: (1) hiring negotiations between human job candidates and AI hiring agents; and (2) human-AI transactions wherein AI agents may conceal information to maximize internal goals. We examine user Extraversion and Agreeableness alongside AI design characteristics, including Adaptability, Expertise, and chain-of-thought Transparency. Our causal discovery analysis extends performance-focused evaluations by integrating scenario-based outcomes, communication analysis, and questionnaire measures. Results reveal divergences between purely simulated and human study datasets, and between scenario types. In simulation experiments, personality traits and AI attributes were comparatively influential. Yet, with actual human subjects, AI attributes – particularly transparency – were much more impactful. We discuss how these divergences vary across different interaction contexts, offering crucial insights for the future of human-centered AI agents.

[Findings] Is Your Language Model Ready for Monetization Decisions? ▶ *Underline Video*

Jialu Gao, Hua Li, Tracy Ortman, Yeyun Gong and Jian Jiao

Large language models (LLMs) are increasingly deployed in monetization-driven systems such as search engines, advertising platforms, and e-commerce services, where decision making is shaped by complex interactions among user intent, advertiser objectives, and platform constraints. Despite rapid progress, existing benchmarks primarily focus on shopping-centric scenarios and user-facing data, capturing only a limited subset of real-world monetization pipelines and overlooking intermediate decision stages and robustness considerations. In this work, we introduce MonBench, a high-quality multi-task benchmark designed to evaluate LLMs in realistic monetization contexts. The benchmark is constructed from large-scale production data collected from multiple search engines, including both intermediate candidate pools and user-visible outcomes, better reflecting the distributional characteristics of real monetization systems. MonBench covers key capability dimensions such as intent understanding, commercial matching, and user behavior modeling, and adopts a unified multiple-choice formulation to enable systematic comparison across models. We further propose a comprehensive evaluation protocol that measures both performance and robustness. We evaluate a diverse set of state-of-the-art LLMs and conduct detailed task-level analyses. Our results reveal monetization-specific behaviors, including gaps between relevance optimization and broader decision-making capabilities, as well as differences in robustness across model families. These findings provide new insights into the strengths and limitations of current LLMs and highlight the need for richer domain-specific supervision in monetization-oriented applications.

[Long] KinyaProp: Fine-Grained Propaganda Annotation in Kinyarwanda ▶ *Underline Video*

Manzi Fabrice Niyigaba, Ivory Yang and Soroush Vosoughi

Propaganda is a widely used approach for shaping public opinion and disseminating misinformation in news media. While it has recently gained significant attention within the NLP community, research on fine grained propaganda detection remains heavily concentrated in high resource languages. To bridge this gap, we introduce KinyaProp, the first fine-grained propaganda dataset of its kind for Kinyarwanda and, to our knowledge, the first such resource created for a Bantu language. Using this dataset, we evaluate whether state-of-the-art LLMs can function as reliable annotators in a genuinely low resource and culturally grounded setting. Our results show that current multilingual LLMs do not reliably approximate human annotation behavior. Instead, they behave as conservative annotators whose performance is largely limited to lexically explicit cues, substantially under-identifying propaganda and exhibiting extremely low and unstable performance on discourse-level techniques. Our findings highlight an important limitation of recent successes in LLM based annotation reported for high resource languages, demonstrating that such results do not readily transfer to low resource settings, where scalable annotation would be most valuable. We release KinyaProp to support future research on fine grained propaganda detection and to enable more robust evaluation of multilingual models in underrepresented languages.

[Long] Knowledge Beyond Language: Bridging the Gap in Multilingual Machine Unlearning Evaluation ▶ *Underline Video*

Kyomin Hwang, Hyeonjin Kim, Sangyeon Cho and Nojun Kwak

While LLMs are increasingly used in commercial services, they pose privacy risks such as leakage of sensitive personally identifiable information (PII). For LLMs trained on multilingual corpora, Multilingual Machine Unlearning (MMU) aims to remove information across multiple languages. However, prior MMU evaluations fail to capture such cross-linguistic distribution of information, being largely limited

to direct extensions of per-language evaluation protocols. To this end, we propose two metrics to evaluate the information spread across languages: the Knowledge Separability Score (KSS) and the Knowledge Persistence Score (KPS). KSS measures the overall unlearning quality across multiple languages, while KPS more specifically aims to assess consistent removal of information among different language pairs. We evaluated various unlearning methods in the multilingual setting with these metrics and conducted comprehensive analyses. Through our investigation, we provide insights into unique phenomena exclusive to MMU and offer a new perspective on MMU evaluation.

[Long] KnowMe-Bench: Benchmarking Person Understanding for Lifelong Digital Companions ▶ *Underline Video*

Tingyu Wu, Zhisheng Chen, Ziyan Weng, Shuhe Wang, Shuo Zhang, Sen Hu, Silin Wu, Qizhen Lan, Huacan Wang and Ronghao Chen

Existing long-horizon memory benchmarks mostly use multi-turn dialogues or synthetic user histories, which makes retrieval performance an imperfect proxy for person understanding. We present Knowme-Bench, a publicly releasable benchmark built from long-form autobiographical narratives, where actions, context, and inner thoughts provide dense evidence for inferring stable motivations and decision principles. Knowme-Bench reconstructs each narrative into a flashback-aware, time-anchored stream and evaluates models with evidence-linked questions spanning factual recall, subjective state attribution, and principle-level reasoning. Across diverse narrative sources, retrieval-augmented systems mainly improve factual accuracy, while errors persist on temporally grounded explanations and higher-level inferences, highlighting the need for memory mechanisms beyond retrieval.

[Long] Language Models Dont Know What You Want: Evaluating Personalization in Deep Research Needs Real Users ▶ *Underline Video*

Nishant Balepur, Malachi Hamada, Varsha Kishore, Sergey Feldman, Amanpreet Singh, Pao Siangliulue, Joseph Chee Chang, Eunsol Choi, Jordan Lee Boyd-Graber and Aakanksha Naik

Deep Research (DR) tools (e.g. OpenAI DR) help researchers cope with ballooning publishing counts. Such tools can synthesize scientific papers to answer researchers' queries, but lack understanding of their users. We change that in MyScholarQA (MySQA), a personalized DR tool that: 1) infers a profile of a user's research interests; 2) proposes personalized actions for a user's input query; and 3) writes a multi-section report for the query that follows user-approved actions. We first test MySQA with NLP's standard protocol: we design a benchmark of synthetic users and LLM judges, where MySQA beats baselines in citation metrics and personalized action-following. However, we suspect this process does not cover all aspects of personalized DR users value, so we interview users in an online version of MySQA to unmask them. We reveal nine nuanced errors of personalized DR undetectable by our LLM judges, and we study qualitative feedback to form lessons for future DR design. In all, we argue for a pillar of personalization that easy-to-use LLM judges can lead NLP to overlook: real progress in personalization is only possible with real users.

[Long] LeCoDe: A Benchmark Dataset for Interactive Legal Consultation Dialogue Evaluation ▶ *Underline Video*

Weikang Yuan, Kangsong Song, Zhuoren Jiang, Junjie Cao, Yujie Zhang, Jun Lin, Kun Kuang, Ji Zhang and Xiaozhong Liu

Legal consultation is essential for safeguarding individual rights and ensuring access to justice, yet remains costly and inaccessible to many individuals due to the shortage of professionals. While recent advances in Large Language Models (LLMs) offer a promising path toward scalable, low-cost legal assistance, current systems fall short in handling the interactive and knowledge-intensive nature of real-world consultations. To address these challenges, we introduce LeCoDe, a multi-turn benchmark dataset constructed from publicly available real-world legal consultation content and carefully processed into a de-identified, structured research resource for evaluating and advancing research on LLMs in legal consultation settings. LeCoDe contains 3,696 multi-turn consultation cases with 110,008 dialogue turns. The dataset is further enriched through expert annotation, including key facts, fact importance, and advice summaries. Furthermore, we propose a comprehensive evaluation framework that assesses LLMs' consultation capabilities in terms of (1) clarification capability and (2) professional advice quality. This unified framework incorporates 12 metrics across two dimensions. Through extensive experiments on various general and domain-specific LLMs, our results reveal significant challenges in this task, with even state-of-the-art models like GPT-4 achieving only 35.9% recall for clarification and 59.1% overall score for advice quality, highlighting the complexity of professional consultation scenarios. Based on these findings, we further explore several strategies to enhance LLMs' legal consultation abilities. Our benchmark contributes to advancing research in legal domain dialogue systems, particularly in simulating more real-world user-expert interactions. The resource is available at https://github.com/PiLab-ZJU/LeCoDe.

[Long] LiveCLKTBench: Towards Reliable Evaluation of Cross-Lingual Knowledge Transfer in Multilingual LLMs ▶ *Underline Video*

Pei-Fu Guo, Yun-Da Tsai, Chun-Chia Hsu, Kai-Xin Chen, Ya An Tsai, Kai-Wei Chang, Nanyun Peng, Mi-Yen Yeh and Shou-De Lin

Evaluating cross-lingual knowledge transfer in large language models (LLMs) is challenging, as correct answers in a target language may arise either from genuine transfer or from prior exposure during pre-training. We present LiveCLKTBench, an automated generation pipeline specifically designed to isolate and measure cross-lingual knowledge transfer. Our pipeline identifies self-contained, time-sensitive knowledge entities from real-world domains, filters them based on temporal occurrence, and verifies them against the models knowledge. The documents of these valid entities are then used to generate factual questions, which are translated into multiple languages to evaluate transferability across linguistic boundaries. Using LiveCLKTBench, we evaluate several LLMs across five languages and observe that cross-lingual transfer is strongly influenced by linguistic distance and often asymmetric across language directions. While larger models improve transfer, the gains diminish with scale and vary across domains. These findings provide new insights into multilingual transfer and demonstrate the value of LiveCLKTBench as a reliable benchmark for future research.

[Long] LiveFact: A Dynamic, Time-Aware Benchmark for LLM-Driven Fake News Detection ▶ *Underline Video*

Cheng Xu, Changhong Jin, Yingjie Niu, Nan Yan, Yuke Mei, Shuhao Guan, Liming Chen and Tahar Kechadi

The rapid development of Large Language Models (LLMs) has transformed fake news detection and fact-checking tasks from simple classification to complex reasoning. However, evaluation frameworks have not kept pace. Current benchmarks are static, making them vulnerable to benchmark data contamination (BDC) and ineffective at assessing reasoning under temporal uncertainty. To address this, we introduce LiveFact a continuously updated benchmark that simulates the real-world "fog of war" in misinformation detection. LiveFact uses dynamic, temporal evidence sets to evaluate models on their ability to reason with evolving, incomplete information rather than on memorized knowledge. We propose a dual-mode evaluation: Classification Mode for final verification and Inference Mode for evidence-based reasoning, along with a component to monitor BDC explicitly. Tests with 22 LLMs show that open-source Mixture-of-Experts models, such as Qwen3-235B-A22B, now match or outperform proprietary state-of-the-art systems. More importantly, our analysis finds a significant "reasoning gap." Capable models exhibit epistemic humility by recognizing unverifiable claims in early data slices-an aspect traditional static benchmarks overlook. LiveFact sets a sustainable standard for evaluating robust, temporally aware AI verification.

[Long] LLMs (Almost) Never Abstain Under Medical Uncertainty ▶ *Underline Video*

Alessio Cocchieri, Luca Ragazzi, Giuseppe Tagliavini and Gianluca Moro

Medical multiple-choice question answering (MCQA) benchmarks implicitly assume that large language models (LLMs) should always commit to an answer. However, in clinical practice, uncertainty is pervasive and abstaining is often the safest action. We introduce MedQAbstain, a benchmark explicitly designed to evaluate medical abstention under uncertainty. MedQAbstain repurposes standard medical MCQA datasets by removing the gold answer and introducing an explicit "I abstain" option, framed as a safety-critical decision with clinical consequences. The benchmark supports systematic analysis across abstention regimes, distractor complexity, and input modalities, and elicits self-reported model confidence to study calibration. Across all settings, we find that state-of-the-art LLMs systematically overcommit, rarely abstaining even when the question itself is hidden. These results reveal a fundamental mismatch between LLM behavior and clinical norms, highlighting abstention as a critical but overlooked dimension of medical decision-making evaluation.

[Short] LOGICAL-COMMONSENSEQA: A Benchmark for Logical Commonsense Reasoning ▶ *Underline Video*

Obed Junias and Maria Leonor Pacheco

Commonsense reasoning often involves evaluating multiple plausible interpretations rather than selecting a single atomic answer, yet most benchmarks rely on single-label evaluation, obscuring whether statements are jointly plausible, mutually exclusive, or jointly implausible. We introduce LOGICAL-COMMONSENSEQA, a benchmark that reframes commonsense reasoning as logical composition over pairs of atomic statements using plausibility-level operators (AND, OR and NEITHER/NOR). Evaluating instruction-tuned, reasoning-specialized, and fine-tuned models under zero-shot, few-shot, and chain-of-thought prompting, we find that while models perform reasonably on conjunctive and moderately on disjunctive reasoning, performance degrades sharply on negation-based questions. LOGICAL-COMMONSENSEQA exposes fundamental reasoning limitations and provides a controlled framework for advancing compositional commonsense reasoning.

[Long] Lost in Simulation: LLM-Simulated Users are Unreliable Proxies for Human Users in Agentic Evaluations ▶ *Underline Video*

Preethi Seshadri, Samuel Cahyawijaya, Ayomide Odumakinde, Sameer Singh and Seraphina Goldfarb-Tarrant

Agentic benchmarks increasingly rely on LLM-simulated users to scalably evaluate agent performance, yet the robustness, validity, and fairness of this approach remain unexamined. Through a user study with participants across the United States, India, Kenya, and Nigeria, we investigate whether LLM-simulated users serve as reliable proxies for real human users in evaluating agents on tau-Bench retail tasks. We find that user simulators lack robustness, with agent success rates varying up to 9 percentage points across different user LLMs. Furthermore, simulated users systematically miscalibrate performance, underestimating success on challenging tasks while overestimating moderately difficult ones. African American Vernacular English (AAVE) speakers experience consistently worse success rates and calibration errors than Standard American English (SAE) speakers, with disparities compounding significantly with age. We also find simulated users to be a differentially effective proxy for different populations, performing worst for AAVE and Indian English. Additionally, simulated users introduce conversational artifacts and surface different failure patterns than human users. These findings demonstrate that current evaluation practices risk misrepresenting agent capabilities across diverse user populations and may obscure real-world deployment challenges.

[Findings] LQM: Linguistically Motivated Multidimensional Quality Metrics for Machine Translation ▶ *Underline Video*

Samar M. Magdy, Fakhraddin Alwajih, Abdellah EL MEKKI, Wesam El-Sayed and Muhammad Abdul-Mageed

Existing MT evaluation frameworks, including automatic metrics and human evaluation schemes such as Multidimensional Quality Metrics (MQM), are largely language-agnostic. However, they often fail to capture dialect- and culture-specific errors in diglossic languages (e.g., Arabic), where translation failures stem from mismatches in language variety, content coverage, and pragmatic appropriateness rather than surface form alone.We introduce LQM: Linguistically Motivated Multidimensional Quality Metrics for MT. LQM is a hierarchical error taxonomy for diagnosing MT errors through six linguistically grounded levels: sociolinguistics, pragmatics, semantics, morphosyntax, orthography, and graphetics (Figure 1). We construct a bidirectional parallel corpus of 3,850 sentences (550 per variety) spanning seven Arabic dialects (Egyptian, Emirati, Jordanian, Mauritanian, Moroccan, Palestinian, and Yemeni), derived from conversational, culturally rich content. We evaluate six LLMs in a zero-shot setting and conduct expert span-level human annotation using LQM, producing 6,113 labeled error spans across 3,495 unique erroneous sentences, along with severity-weighted quality scores. We complement this analysis

with an automatic metric (spBLEU). Though validated here on Arabic, LQM is a language-agnostic framework designed to be easily applied to or adapted for other languages. LQM annotated errors data, prompts, and annotation guidelines are publicly available at https://github.com/UBC-NLP/LQM_MT

[Long] Mem-Gallery: Benchmarking Multimodal Long-Term Conversational Memory for MLLM Agents ▶ *Underline Video*

Yuanchen Bei, Tianxin Wei, Xuying Ning, Yanjun Zhao, Zhining Liu, Xiao Lin, Yada Zhu, Hendrik Hamann, Jingrui He and Hanghang Tong

Long-term memory is a critical capability for multimodal large language model (MLLM) agents, particularly in conversational settings where information accumulates and evolves over time. However, existing benchmarks either evaluate multi-session memory in text-only conversations or assess multimodal understanding within localized contexts, failing to evaluate how multimodal memory is preserved, organized, and evolved across long-term conversational trajectories. Thus, we introduce Mem-Gallery, a new benchmark for evaluating multimodal long-term conversational memory in MLLM agents. Mem-Gallery features high-quality multi-session conversations grounded in both visual and textual information, with long interaction horizons and rich multimodal dependencies. Building on this dataset, we propose a systematic evaluation framework that assesses key memory capabilities along three functional dimensions: memory extraction and test-time adaptation, memory reasoning, and memory knowledge management. Extensive benchmarking across twelve memory systems reveals several key findings, highlighting the necessity of explicit multimodal information retention and memory organization, the persistent limitations in memory reasoning and knowledge management, as well as the efficiency bottleneck of current models. Our benchmark and dataset are available at https://github.com/YuanchenBei/Mem-Gallery.

[Long] MINTQA: A Multi-Hop Question Answering Benchmark for Evaluating LLMs on New and Long-tail Knowledge ▶ *Underline Video*

Jie He, Nan Hu, Wanqiu Long, Jiaoyan Chen and Jeff Z. Pan

Retrieval-Augmented Generation (RAG) enhances large language models (LLMs) by integrating external knowledge, enabling them to tackle knowledge-intensive tasks. However, limited research has explored how LLMs effectively leverage RAG techniques for multi-hop question answering (QA), particularly when handling knowledge with with varying degrees of familiarity. In this paper, we introduce MINTQA (Multi-hop Question Answering on New and Tail Knowledge), a benchmark designed to evaluate multi-hop QA performance on questions involving 10,479 question-answer pairs for evaluating old/new knowledge and 17,887 pairs for assessing popular/unpopular knowledge, with each question equipped with corresponding sub-questions and answers. This benchmark primarily evaluates the multi-hop reasoning ability of LLMs and their capacity to handle knowledge with varying levels of familiarity during the reasoning process. We evaluate 22 state-of-the-art LLMs using three distinct QA strategies: LLM-based parameterized knowledge QA, direct RAG-enhanced QA, and multi-hop RAG-enhanced QA. Our experiments reveal key challenges in how LLMs handle knowledge with different familiarity and offer insights into improving their multi-hop reasoning capabilities when combined with RAG techniques.

[Long] MMAC: A Multilingual, Multimodal Alignment Framework for Cultural Grounding Evaluation ▶ *Underline Video*

Weihua Zheng, Zhengyuan Liu, Tanmoy Chakraborty, Weiwen Xu, Xiaoxue Gao, Bryan Chen Zhengyu Tan, Bowei Zou, Chang Liu, Yujia Hu, Xing Xie, Xiaoyuan Yi, Jing Yao, Chaojun Wang, Long Li, Rui Liu, Huiyao Liu, Koji Inoue, Ryuichi Sumida, Tatsuya Kawahara, Fan Xu, Lingyu Ye, Wei Tian, Dongjun Kim, Jimin Jung, Jaehyung Seo, Nadya YukiWangsajaya, Pham Minh Duc, Ojasva Saxena, Palash Nandi, Xiyan Tao, Wiwik Karlina, Tuan Luong, Keertana Arun Vasan, Roy Ka-Wei Lee and Nancy F. Chen

The global deployment of Large Language Models (LLMs) underscores the urgent need to evaluate their cultural alignment. However, assessing genuine "cultural awareness" across modalities (text, vision, speech) and languages remains a significant challenge. To comprehensively investigate this domain, we propose MMAC, a systematic framework that encompasses a tri-modally aligned cultural benchmark creation pipeline and a five-dimensional evaluation protocol to assess cross-country awareness disparities, evaluate cross-lingual and cross-modal consistency, and verify cultural knowledge generalization and grounding validity. Given the prevailing Western cultural bias in current models, we focus on 8 Asian countries as our dataset foundation to more acutely reveal potential cultural deficiencies in LLMs. Our dataset, MMAC-bench, features 27,000 human-curated questions across 10 languages. Crucially, it is the first dataset aligned at the input level across text, image, and speech, enabling direct cross-modal transfer tests. Each question consists of multiple-choice options accompanied by open-ended generated explanations, where 79% require multi-step reasoning grounded in cultural context, moving beyond simple memorization. We probe the causes of modal divergence, offering insights into fostering culturally robust MLLMs.

[Findings] MMTabReal: Real-World Benchmark for Multimodal Table Understanding ▶ *Underline Video*

Prasham Titiya, Jainil Trivedi, Chitta Baral and Vivek Gupta

Multimodal tables i.e. tabular layouts interleaved with charts, maps, icons, and color encodings are ubiquitous in real applications yet remain difficult for Multimodal Large Language Models (MLLMs). Despite advances in text and image understanding, systematic evaluation of table-centric multimodal reasoning is limited. We introduce MMTabReal, a MultiModal Table Benchmark, human-curated suite of 500 real-world tables paired with 4021 question-answer pairs. MMtabReal spans four question types, five reasoning categories, and eight structural archetypes. Evaluations of state-of-the-art models reveal substantial gaps, especially in visual grounding, spatial alignment, and multi-step inference, with 20-40% performance drops relative to existing benchmarks. These results highlight the need for architectures that more tightly fuse vision with tabular structure and support explicit numeric/logical operations. MMtabReal is released for evaluation only, providing a rigorous, reproducible testbed that reflects the linguistic, structural, and reasoning complexity of real-world multimodal tables. Code and data are available at: https://coral-lab-asu.github.io/mmtabreal/

[Findings] Multi-Agent Comedy Club: Investigating Community Discussion Effects on LLM Humor Generation ► *Underline Video*
Shiwei Hong, Lingyao Li, Ethan Z. Rong, Chenxinran Shen and Zhicong Lu
Prior work has explored multi-turn interaction and feedback for LLM writing, but evaluations still largely center on prompts and localized feedback, leaving persistent public reception in online communities underexamined. We test whether broadcast community discussion improves stand-up comedy writing in a controlled multi-agent sandbox: in the discussion condition, critic and audience threads are recorded, filtered, stored as social memory, and later retrieved to condition subsequent generations, whereas the baseline omits discussion. Across 50 rounds (250 paired monologues) judged by five expert annotators using A/B preference and a 15-item rubric, discussion wins 75.6% of instances and improves Craft/Clarity (Delta = 0.440) and Social Response (Delta = 0.422), with occasional increases in aggressive humor.

[Findings] Not all ANIMALs are equal: metaphorical framing through source domains and semantic frames ► *Underline Video*
Yulia Otmakhova, Matteo Guida and Lea Frermann
Metaphors are powerful framing devices, yet their source domains alone do not fully explain the specific associations they evoke. We argue that the interplay between source domains and semantic frames determines how metaphors shape understanding of complex issues, and present a computational framework that allows to derive salient discourse metaphors through their source domains and semantic frames. Applying this framework to climate change news, we uncover not only well-known source domains but also reveal nuanced frame-level associations that distinguish how the issue is portrayed. In analyzing immigration discourse across political ideologies, we demonstrate that liberals and conservatives systematically employ different semantic frames within the same source domains, with conservatives favoring frames emphasizing uncontrollability and liberals choosing neutral or more "victimizing" semantic frames. Our work bridges conceptual metaphor theory and linguistics, providing the first NLP approach for discovery of discourse metaphors and fine-grained analysis of differences in metaphorical framing.

[Long] Omni-RewardBench: Toward a Comprehensive Evaluation of Generative Reward Models Across Modalities ► *Underline Video*
Chi-Min Chan, Yujin Zhou, Pengcheng Wen, Boqin Yin, Jiaming Ji, Juntao Dai, Wei Xue, Sirui Han and Yike Guo
The rise of Omni-modality Large Language Models (OLLMs) capable of jointly processing text, audio, and visual inputs marks a major step toward general intelligence. Ensuring their alignment with human preferences requires effective Omni-modality Reward Models (ORMs), which serve as surrogates for human judgment to guide OLLMs behavior. However, ORMs evaluation remains underdeveloped in the previous literature. Existing benchmarks are largely text-centric or limited to bimodal tasks, restricting comprehensive assessment for ORMs. To bridge this gap, we introduce Omni-RewardBench, the first benchmark for comprehensive evaluation of ORMs across modalities. In short, our contributions are threefold: (1) a hybrid automatic-annotation and human-verification pipeline to construct high-quality evaluation data; (2) extensive experiments on 20+ models, including inherently omni-modal and modality-bridged systems. Our experimental results demonstrate that current OLLMs fall short as reward models, revealing several common failure modes such as perception failure, modality dominance failure, and cross-modal fusion failure; and (3) strong correlations between Omni-RewardBench scores and downstream performance (IID r = 0.94, OOD r = 0.72), validating its reliability as a predictor of real-world capability and alignment quality.

[Findings] ParaCook: On Time-Efficient Planning for Multi-Agent Systems ► *Underline Video*
Shiqi Zhang, Xinbei Ma, Yunqing Xu, Zouying Cao, Pengrui Lu, Haobo Yuan, Tiancheng Shen, Zhuosheng Zhang, Hai Zhao and Ming-Hsuan Yang
Large Language Models (LLMs) exhibit strong reasoning abilities for planning long-horizon, real-world tasks, yet existing agent benchmarks focus on task completion while neglecting time efficiency in parallel and asynchronous operations. To address this, we present ParaCook, a benchmark for time-efficient collaborative planning. Inspired by the Overcooked game, ParaCook provides an environment for various challenging interaction planning of multi-agent systems that are instantiated as cooking tasks, with a simplified action space to isolate the core challenge of strategic parallel planning. Through a comprehensive evaluation of state-of-the-art LLMs, we find that current approaches achieve suboptimal plans, which struggle with parallel actions or coordination. Our analysis also reveals LLMs' potential on abstract tasks where they can focus on high-level parallel optimization. ParaCook provides a scalable evaluation framework with adjustable complexity, establishing a foundation for developing and assessing time efficiency-aware multi-agent planning.

[Long] Perception, Understanding and Reasoning: A Multimodal Benchmark for Video Fake News Detection ► *Underline Video*
Cui Yakun, Peng Qi, Fushuo Huo, Hang Du, Weijie Shi, Juntao Dai, Zhenghao Zhu and Sirui Han
The advent of multi-modal large language models (MLLMs) has greatly advanced research on video fake news detection (VFND) tasks. Existing benchmarks typically focus on the detection accuracy, while failing to provide fine-grained assessments for the entire detection process. To address these limitations, we introduce POVFNDB (Process-oriented Video Fake News Detection Benchmark), a process-oriented benchmark comprising 10 tasks designed to systematically evaluate MLLMs' perception, understanding, and reasoning capabilities in VFND. This benchmark contains 36,240 human-annotated question-answer (QA) in structured or open-ended formats, spanning 15 distinct evaluation dimensions that characterize different aspects of the video fake news detection process.Using POVFNDB, we conduct comprehensive evaluations on both proprietary and open-source MLLMs. Moreover, We fine-tune Qwen2.5VL-7B-Instruct on a reasoning dataset generated by our proposed POVFND-CoT, a chain-of-thought method that utilizes rationales from evaluation results and rationale validation. The resulting model achieves sota performance on VFND.

[Findings] POLAR: A Benchmark for Multilingual, Multicultural, and Multi-Event Online Polarization ▶ *Underline Video*

Usman Naseem, Robert Geislinger, Juan Ren, Sarah Kohail, Rudy Alexandro Garrido Veliz, P Sam Sahil, Yiran Zhang, Idris Abdulmumin, Marco Antonio Stranisci, Özge Alacam, Cengiz Acarturk, Aisha Jabr, Saba Anwar, Abinew Ali Ayele, Simona Frenda, Alessandra Teresa Cignarella, Elena Tutubalina, Oleg Rogov, Aung Kyaw Htet, Xintong Wang, Surendrabikram Thapa, Kritesh Rauniyar, Tanmoy Chakraborty, MD ARFEEN ZEESHAN, DHEERAJ KODATI, Satya keerthi, Sahar Moradizeyveh, Firoj Alam, Md Arid Hasan, Syed Ishtiaque Ahmed, Ye Kyaw Thu, Shantipriya Parida, Ihsan Ayyub Qazi, Lilian Diana Awuor Wanzare, Nelson Odhiambo Onyango, Clemencia Siro, Jane Wanjiru Kimani, Ibrahim Said Ahmad, Adem Chanie Ali, Martin Semmann, Chris Biemann, Shamsuddeen Hassan Muhammad and Seid Muhie Yimam

Online polarization poses a growing challenge for democratic discourse, yet most computational social science research remains monolingual, culturally narrow, or event-specific. We introduce POLAR, a multilingual, multicultural, and multi-event dataset with over 110K instances in 22 languages drawn from diverse online platforms and real-world events. Polarization is annotated along three axes, namely detection, type, and manifestation, using a variety of annotation platforms adapted to each cultural context. We conduct two main experiments: (1) fine-tuning six pretrained small language models; and (2) evaluating a range of open and closed large language models in few-shot and zero-shot settings. Results show that while most models perform well on binary polarization detection, they achieve substantially lower performance when predicting polarization types and manifestations. These findings highlight the complex, highly contextual nature of polarization and underscore the need for robust, adaptable approaches in NLP and computational social science. All resources will be released to support further research and effective mitigation of digital polarization globally.

[Findings] PolicyLLM: Towards Excellent Comprehension of Public Policy for Large Language Models ▶ *Underline Video*

Han Bao, Penghao Zhang, Yue Huang, Zhengqing Yuan, Yanchi Ru, SU RUI, Yujun Zhou, Xiangqi Wang, Kehan Guo, Nitesh V Chawla, Yanfang Ye and Xiangliang Zhang

Large Language Models (LLMs) are increasingly integrated into real-world decision-making, including in the domain of public policy. Yet, their ability to comprehend and reason about policy-related content remains underexplored. To fill this gap, we present PolicyBench, the first large-scale bilingual benchmark evaluating policy comprehension, comprising 21K cases across a broad spectrum of policy areas, capturing the diversity and complexity of real-world governance. Following Bloom's taxonomy, the benchmark assesses three core capabilities: (1) Memorization: factual recall of policy knowledge, (2) Understanding: conceptual and contextual reasoning, and (3) Application: problem-solving in real-life policy scenarios. Building on this benchmark, we further propose PolicyMoE, a domain-specialized Mixture-of-Experts (MoE) model with expert modules aligned to each cognitive level. The proposed models demonstrate stronger performance on application-oriented policy tasks than on memorization or conceptual understanding, and yields the highest accuracy on structured reasoning tasks. Our results reveal key limitations of current LLMs in policy understanding and suggest paths toward more reliable, policy-focused models

[Findings] PrefIx: Understand and Adapt to User Preference in Human-Agent Interaction ▶ *Underline Video*

Jialin Li, Zhenhao Chen, Hanjun Luo and Hanan Salam

LLM-based agents can complete tasks correctly yet still frustrate users through poor interaction patterns, such as excessive confirmations, opaque reasoning, or misaligned pacing. Current benchmarks evaluate task accuracy but overlook how agents interact: whether they infer preferences from implicit cues, adapt dynamically, or maintain fine-grained interaction quality. We introduce, a configurable environment that evaluates both what agents accomplish and how they interact. Central to ~is the Interaction-as-a-Tool (IaaT) paradigm, which treats interaction behaviors as structured tool calls, unifying them with existing evaluation frameworks. We define 31 preference settings across 14 attributes and formalize user experience (UX) as a core metric alongside task accuracy. A composite LLM-as-a-Judge mechanism across seven UX dimensions achieves strong aggregate reliability (ICC > 0.79), high internal consistency (alpha = 0.943), and human correlation (= 0.52-0.78). Preference-aware agents show 7.6% average UX improvement and 18.5% gain in preference alignment.

[Findings] Prompt Injection in Automated Résumé Screening with Large Language Models: Single and Multi-Injection Settings ▶ *Underline Video*

Preet Baxi, Jiannan Xu, Jane Yi Jiang and Stefanus Jasin

Large language models (LLMs) are increasingly used to screen and rank job applicants, creating incentives for candidates to strategically manipulate algorithmic hiring systems. We study prompt injection in automated résumé screening, defined as subtle self-promotional text that introduces no new qualifications but is designed to influence LLM evaluations. Using controlled experiments, we show that prompt injection reliably improves applicant rankings when résumé quality is homogeneous and few candidates inject. However, its effectiveness rapidly diminishes as more candidates inject, collapsing when manipulation becomes widespread. When candidate quality is heterogeneous, prompt injection is less effective on average, but can occasionally allow lower-quality candidates to outrank higher-quality ones, raising fairness concerns. Overall, LLM-based screening is most vulnerable when manipulation is rare and candidate quality differences are small.

[Long] Putting Captions to the Test: Evaluating Video Caption Quality through Multiple-Choice Question Answering ▶ *Underline Video*

Zizhen Wang, Bo Feng, Zhengfeng Lai, Shiyu Li, Yang Lu, Meng Cao, Ping Huang and Simon Wang

Evaluating video captioning remains a critical challenge for Visual Large Language Models (VLLMs). Existing metrics primarily rely

on matching generated text against ground-truth references. This paradigm suffers from the "one-to-many" nature of video description, where high-quality captions are often penalized for lexical mismatches or valid shifts in visual focus. Furthermore, such assessments are typically one-dimensional, failing to provide a fine-grained analysis of caption quality. To address this, we redefine caption quality through the lens of information fidelity: A caption must maximize the coverage of salient visual information while ensuring strict factuality. We introduce CapQuiz, a novel reference-free benchmark that assesses captions based on their utility in answering human-verified, fine-grained, multiple-choice questions derived from the video. CapQuiz features a hierarchical taxonomy of 10 question types (spanning Descriptive and Inferential categories) across 24 diverse video domains. Extensive experiments demonstrate that CapQuiz correlates significantly better with human judgments than existing metrics and offers interpretable insights into model performance. We will release the benchmark to facilitate reproducible research.

[Long] Ranking Reasoning LLMs under Test-Time Scaling ▶ *Underline Video*

Mohsen Hariri, Michael Hinczewski, Jing Ma and Vipin Chaudhary

Test-time scaling evaluates reasoning LLMs by sampling multiple outputs per prompt, but ranking models in this regime remains underexplored. We formalize dense benchmark ranking under test-time scaling and introduce Scorio, a library that implements statistical ranking methods such as paired-comparison models, item response theory (IRT) models, voting rules, and graph- and spectral-based methods. Across 20 reasoning models on four Olympiad-style math benchmarks (AIME'24, AIME'25, HMMT'25, and BrUMO'25; up to N = 80 trials), most full-trial rankings agree closely with the Bayesian gold standard Bayes_U@80 (mean Kendall's tau_b = 0.93-0.95), and 19-34 methods recover exactly the same ordering. In the single-trial regime, the best methods reach tau_b approx 0.86. Using greedy decoding as an empirical prior (Bayes_R0@N) reduces variance at N = 1 by 16-52%, but can bias rankings when greedy and stochastic sampling disagree. These results identify reliable ranking methods for both high- and low-budget test-time scaling. We release Scorio as an open-source library at https://github.com/mohsenhariri/scorio.

[Long] RARE: Redundancy-Aware Retrieval Evaluation Framework for High-Similarity Corpora ▶ *Underline Video*

Hanjun Cho and Jay-Yoon Lee

Existing QA benchmarks typically assume distinct documents with minimal overlap, yet real-world retrieval-augmented generation (RAG) systems operate on corpora such as financial reports, legal codes, and patents, where information is highly redundant and documents exhibit strong inter-document similarity. This mismatch undermines evaluation validity: retrievers can be unfairly undervalued even when they retrieve documents that provide sufficient evidence, because redundancy across documents is not accounted for in evaluation. On the other hand, retrievers that perform well on standard benchmarks often generalize poorly to real-world corpora with highly similar and redundant documents. We present RARE (Redundancy-Aware Retrieval Evaluation), a framework for constructing realistic benchmarks by (i) decomposing documents into atomic facts to enable precise redundancy tracking and (ii) enhancing LLM-based data generation with CRRF. RAG benchmark data usually requires multiple quality criteria, but LLMs often yield trivial outputs. CRRF scores criteria separately and fuses decisions by rank, improving the reliability of generated data. Applying RARE to Finance, Legal, and Patent corpora, we introduce RedQA, where a strong retriever baseline drops from 66.4% PerfRecall@10 on 4-hop General-Wiki to 5.0–27.9% PerfRecall@10 at 4-hop depth, revealing robustness gaps that current benchmarks fail to capture. RARE enables practitioners to build domain-specific RAG evaluations that faithfully reflect real-world deployment conditions.

[Findings] Real or Robotic? Assessing Whether LLMs Accurately Simulate Qualities of Human Responses in Human-LLM Dialogue ▶ *Underline Video*

Jonathan Ivey, Shivani Kumar, Jiayu Liu, Hua Shen, Sushrita Rakshit, Rohan Raju, Haotian Zhang, Aparna Ananthasubramaniam, Junghwan Kim, Bowen Yi, Dustin Wright, Abraham Israeli, Anders Giovanni Møller, Lechen Zhang and David Jurgens

Building datasets for dialogue tasks is expensive and time-consuming, requiring recruitment, training, and data collection from study participants. In response, much recent work has sought to use large language models (LLMs) to simulate both human-human and human-LLM interactions, as they have been shown to generate convincingly human-like text in many settings. However, how well do LLM-based simulations reflect real human dialogue? In this work, we answer this question by generating a large-scale dataset of 100,000 paired LLM-LLM and human-LLM dialogues from the WildChat dataset and quantifying how well the LLM simulations align with their human counterparts. Overall, we find relatively low alignment between simulations and human interactions, with systematic differences in multiple textual properties, including style and conversational dynamics. Further, we find that models perform similarly in simulating English, Chinese, and Russian dialogues. Our results also suggest that LLMs only simulate a narrow range of the overall distribution of human dialogue, as they perform better on the subset of humans who write similarly to the LLM's own style.

[Long] Reframing Responsibility: Framing-Aware Event Causality Identification ▶ *Underline Video*

Jin Zhao and Jiayi Yao

Causal explanations in political narratives are often framed and contested. Different sources may explain the same event by assigning responsibility to different actors and expressing varying levels of certainty. Standard Event Causality Identification (ECI) focuses on detecting causal links and does not capture these distinctions. We introduce Framing-Aware Event Causality Identification (FrECI), a framing-aware extension of ECI that models causal explanations as structured claims including responsibility targets, evaluative framing, source type, and epistemic modality grounded in established framing theories. We construct a multilingual dataset aligned across English, Chinese, and Arabic narratives using shared event anchors. We evaluate FrECI using prompt-based large language model baselines and supervised neural

models. Results show that prompt-based baselines struggle to recover complete framed causal claims, while joint supervised models perform substantially better. Finally, we demonstrate that FrECI enables quantitative analysis of divergent causal attribution across narratives.

[Findings] Reheat Nachos for Dinner? Evaluating AI Support for Cross-Cultural Communication of Neologisms ▶ *Underline Video*

Dayeon Ki, Yu Hou, Rachel Rudinger, Hal Daumé III, Marine Carpuat and Fumeng Yang

Neologisms and emerging slang are central to daily conversation, yet challenging for non-native speakers (NNS) to interpret and use appropriately in cross-cultural communication with native speakers (NS). NNS increasingly make use of Artificial Intelligence (AI) tools to learn these words. We study the utility of such tools in mediating an informal communication scenario through a human-subjects study (N=234): NNS participants learn English neologisms with AI support, write messages using the learned word to an NS friend, and judge contextual appropriateness of the neologism in two provided writing samples. Using both NS evaluator-rated communicative competence of NNS-produced writing and NNS' contextual appropriateness judgments, we compare three AI-based support conditions: AI Definition, AI Rewrite into simpler English, AI Explanation of meaning and usage, and Non-AI Dictionary for comparison. We show that AI Explanation yields the largest gains over no support in NS-rated competence, while contextual appropriateness judgments show indifference across support. NNS participants' self-reported perceptions tend to overestimate NS ratings, revealing a mismatch between perceived and actual competence. We further observe a significant gap between NNS- and NS-produced writing, highlighting the limitations of current AI tools and informing design for future tools.

[Short] ReproEvalCard: A Reporting Standard for Reproducible Evaluation of LLM Pipelines ▶ *Underline Video*

Priyaranjan Pattnayak and Apoorv Bhatia

Evaluation of modern large language model (LLM) systems increasingly relies on multi-stage pipelines such as retrieval-augmented generation, tool-using agents, and prompt chains. Reproducing reported evaluation results for these systems often requires evaluation-specific artifacts beyond model weights and datasets, including prompts, judge configurations, retrieval snapshots, and intermediate traces, yet their availability has not been systematically examined. We introduce ReproEvalCard, a lightweight reporting standard that specifies the minimal artifacts required to reproduce and validate evaluations of LLM pipelines. To motivate this standard, we audit 55 pipeline-based LLM papers published between 2022 and 2025 and quantify the availability of reproducibility-critical evaluation artifacts. We find that randomness controls are missing in 75% of papers and intermediate execution traces in 61%, substantially limiting evaluation reproducibility. We further demonstrate ReproEvalCard through a worked example and provide a concise checklist for authors and reviewers, aiming to improve reproducibility and comparability in LLM evaluation.

[Long] ReRec: Reasoning-Augmented LLM-based Recommendation Assistant via Reinforcement Fine-tuning ▶ *Underline Video*

Jiani Huang, Shijie Wang, Liang-Bo Ning, Wenqi Fan and Li Qing

With the rise of LLMs, there is an increasing need for intelligent recommendation assistants that can handle complex queries and provide personalized, reasoning-driven recommendations. LLM-based recommenders show potential but face challenges in multi-step reasoning, underscoring the need for reasoning-augmented systems. To address this gap, we propose ReRec, a novel reinforcement fine-tuning (RFT) framework designed to improve LLM reasoning in complex recommendation tasks. Our framework introduces three key components: (1) Dual-Graph Enhanced Reward Shaping, integrating recommendation metrics like NDCG@K with Query Alignment and Preference Alignment Scores to provide fine-grained reward signals for LLM optimization; (2) Reasoning-aware Advantage Estimation, which decomposes LLM outputs into reasoning segments and penalizes incorrect steps to enhance reasoning of recommendation; and (3) Online Curriculum Scheduler, dynamically assess query difficulty and organize training curriculum to ensure stable learning during RFT. Experiments demonstrate that ReRec outperforms state-of-the-art baselines and preserves core abilities like instruction-following and general knowledge. Our codes are available at https://anonymous.4open.science/r/ReRec/.

[Long] Rethinking Composed Image Retrieval Evaluation: A Fine-Grained Benchmark from Image Editing ▶ *Underline Video*

Tingyu Song, Yanzhao Zhang, Mingxin Li, Zhuoning Guo, Dingkun Long, Pengjun Xie, Siyue Zhang, Yilun Zhao and Shu Wu

Composed Image Retrieval (CIR) is a pivotal and complex task in multimodal understanding. Current CIR benchmarks typically feature limited query categories and fail to capture the diverse requirements of real-world scenarios. To bridge this evaluation gap, we leverage image editing to achieve precise control over modification types and content, enabling a pipeline for synthesizing queries across a broad spectrum of categories. Using this pipeline, we construct EDIR, a novel fine-grained CIR benchmark. EDIR encompasses 5,000 high-quality queries structured across five main categories and fifteen subcategories. Our comprehensive evaluation of 13 multimodal embedding models reveals a significant capability gap; even state-of-the-art models (e.g., RzenEmbed and GME) struggle to perform consistently across all subcategories, highlighting the rigorous nature of our benchmark. Through comparative analysis, we further uncover inherent limitations in existing benchmarks, such as modality biases and insufficient categorical coverage. Furthermore, an in-domain training experiment demonstrates the feasibility of our benchmark. This experiment clarifies the task challenges by distinguishing between categories that are solvable with targeted data and those that expose intrinsic limitations of current model architectures.

[Findings] Rolling Out Data Quality Overnight, without losing the plot: A Multi-Agent System for Speech Data Quality Management ▶ *Underline Video*

Rishabh Kumar, Abhinav Painuli, Chriss Philip Saji, Devesh Soni, Amrith Krishna and Ganesh Ramakrishnan

Quality management when creating large-scale speech datasets is essential for building reliable downstream models, yet verification pipelines

are often brittle, domain-specific, and expertise-intensive. We introduce SpeechQM-Agent, a natural language-driven agentic framework that compiles user requirements into dependency-aware DAG workflows over modular tools for audio, transcript, and metadata verification. A central planner LLM enforces prerequisites and supports execution-time replanning (e.g., re-running failed steps or swapping tools), reducing manual pipeline engineering and improving robustness across heterogeneous vendor formats and multilingual settings. We also release SpeechQM-Dataset, a multilingual benchmark with controlled, vendor-inspired quality artifacts spanning 24 verification tasks. Across experiments, SpeechQM-Agent attains 80-90% agreement with expert verification while requiring <20% of the cost and time of manual QC, and we further validate transfer to real vendor-supplied corpora. Planner LLM comparisons highlight fidelity-efficiency trade-offs.

[Long] RubricBench: Aligning Model-Generated Rubrics with Human Standards ▶ *Underline Video*

Junyi Zhou, Qiyuan Zhang, Yufei Wang, Fuyuan Lyu, Yidong Ming, Can Xu, Qingfeng Sun, Kai Zheng, Peng Kang, Xue Liu and Chen Ma

As Large Language Model (LLM) alignment evolves from simple completions to complex, highly sophisticated generation, Reward Models are increasingly shifting toward rubric-guided evaluation to mitigate surface-level biases. However, the community lacks a unified benchmark to assess this evaluation paradigm, as existing benchmarks lack both the discriminative complexity and the ground-truth rubric annotations required for rigorous analysis. To bridge this gap, we introduce RubricBench, a curated benchmark with 1,147 pairwise comparisons specifically designed to assess the reliability of rubric-based evaluation. Our construction employs a multi-dimensional filtration pipeline to target hard samples featuring nuanced input complexity and misleading surface bias, augmenting each with expert-annotated, atomic rubrics derived strictly from instructions. Comprehensive experiments reveal a substantial capability gap between human-annotated and model-generated rubrics, indicating that even state-of-the-art models struggle to autonomously specify valid evaluation criteria, lagging considerably behind human-guided performance.

[Long] S2S-Arena: Evaluating Paralinguistic Instruction Following in Speech-to-Speech Models ▶ *Underline Video*

Feng Jiang, Zhiyu Lin, Yiyang Liu, Liumeng Xue, Fan Bu, Yuhao Du, Xiangying Chen, Benyou Wang and Haizhou Li

Recent advances in large language models (LLMs) have fundamentally reshaped speech-to-speech (S2S) systems, enabling increasingly natural spoken interaction. However, existing benchmarks still rely heavily on text-based evaluation and largely ignore paralinguistic cues such as prosody, emotion, and speaker traits, which are central to expressive and human-like communication. We introduce S2S-Arena, a speech-native benchmark for evaluating instruction-following S2S models with explicit assessment of both semantic understanding and paralinguistic expression. S2S-Arena features a four-level interaction protocol that systematically probes models under increasing paralinguistic complexity, a two-stage data construction pipeline that produces 1,243 speech samples spanning 100+ real-world tasks, and an arena-style evaluation framework that enables reference-free, pairwise comparison directly in the speech modality. Benchmarking 10 state-of-the-art S2S systems over 1,000+ comparisons reveals substantial performance gaps (especially under complex paralinguistic demands) between current academic and industrial systems. Our analysis further identifies key design factors governing expressive instruction following, providing actionable insights for building more natural, robust, and human-aligned speech agents.

[Long] SAD: A Large-Scale Strategic Argumentative Dialogue Dataset ▶ *Underline Video*

YongKang Liu, Jiayang Yu, Mingyang Wang, Yiqun Zhang, Ercong Nie, Shi Feng, Daling Wang, Kaisong Song and Hinrich Schuetze

Argumentation generation has attracted substantial research interest due to its central role in human reasoning and decision-making. However, most existing argumentative corpora focus on non-interactive, single-turn settings, either generating arguments from a given topic or refuting an existing argument. In practice, however, argumentation is often realized as multi-turn dialogue, where speakers defend their stances and employ diverse argumentative strategies to strengthen persuasiveness. To support deeper modeling of argumentation dialogue, we present the first large-scale Strategic Argumentative Dialogue dataset, SAD, consisting of 392,822 examples. Grounded in argumentation theories, we annotate each utterance with five strategy types, allowing multiple strategies per utterance. Unlike prior datasets, SAD requires models to generate contextually appropriate arguments conditioned on the dialogue history, a specified stance on the topic, and targeted argumentation strategies. We further benchmark a range of pretrained generative models on SAD and present in-depth analysis of strategy usage patterns in argumentation.

[Findings] Scaling Evaluation-Time Compute with Reasoning Models as Evaluators ▶ *Underline Video*

Seungone Kim, Ian Wu, Jinu Lee, Xiang Yue, Seongyun Lee, Minkyeong Moon, Carolin Lawrence, Kiril Gashteovski, Julia Hockenmaier, Graham Neubig and Sean Welleck

Language model (LM) evaluators that generate chain-of-thought (CoT) reasoning are widely used for the assessment of LM responses. Simultaneously, increasing LMs' "thinking" time through scaling test-time compute has proven to be an effective technique for solving challenging problems in domains such as math and code. This raises a natural question: can an LM's evaluation capability also be improved by scaling test-time compute? To answer this, we investigate employing reasoning models - LMs that natively generate long CoT reasoning - as evaluators. We explore scaling evaluation-time compute by using reasoning models to evaluate both the overall candidate response (i.e., outcome evaluation) and the individual reasoning steps within it (i.e., process evaluation). We observe that evaluator performance improves monotonically with the number of reasoning tokens generated, mirroring trends seen in LM reasoning. Furthermore, we use these more accurate evaluators to rerank multiple generations, and demonstrate that spending more compute at evaluation time can be as effective as increasing compute during generation for improving an LM's problem-solving performance.

[Findings] SciImpact: A Multi-Dimensional, Multi-Field Benchmark for Scientific Impact Prediction ▶ *Underline Video*

Hangxiao Zhu, Yuyu Zhang, Ping Nie and Yu Zhang
The rapid growth of scientific literature calls for automated methods to assess and predict research impact. Prior work has largely focused on citation-based metrics, leaving limited evaluation of models capability to reason about other impact dimensions. To this end, we introduce SciImpact, a large-scale, multi-dimensional benchmark for scientific impact prediction spanning 19 fields. SciImpact captures various forms of scientific influence, ranging from citation counts to award recognition, media attention, patent reference, and artifact adoption, by integrating heterogeneous data sources and targeted web crawling. It comprises 215,928 contrastive paper pairs reflecting meaningful impact differences in both short- (e.g., Best Paper Award) and long-term settings (e.g., Nobel Prize). We evaluate 11 widely used large language models (LLMs) on SciImpact. Results show that off-the-shelf models show substantial variability across dimensions and fields, while multi-task supervised fine-tuning consistently enables smaller LLMs (e.g., 4B) to markedly outperform much larger models (e.g., 30B) and surpass powerful closed-source LLMs (e.g., o4-mini). These results establish SciImpact as a challenging benchmark and demonstrate its value for multi-dimensional, multi-field scientific impact prediction. Our project homepage is https://flypig23.github.io/sciimpact-homepage/.

[Long] Seeing the Whole Elephant: A Benchmark for Failure Attribution in LLM-based Multi-Agent Systems ▶ *Underline Video*
Mengzhuo Chen, Junjie Wang, Fangwen Mu, Yawen Wang, Zhe Liu, Huanxiang Feng and Qing Wang
Failure attribution, i.e., identifying the responsible agent and decisive step of a failure, is particularly challenging in LLM-based multi-agent systems (MAS) due to their natural-language reasoning, nondeterministic outputs, and intricate interaction dynamics. A reliable benchmark is therefore essential to guide and evaluate attribution techniques. Yet existing benchmarks rely on partially observable traces that capture only agent outputs, omitting the inputs and context that developers actually use when debugging. We argue that attribution should be studied under full execution observability, aligning with real-world developer-facing scenarios where complete traces, rather than only outputs, are accessible for diagnosis. To this end, we introduce TraceElephant, a benchmark designed for failure attribution with full execution traces and reproducible environments. We then systematically evaluate failure attribution techniques across various configurations. Specifically, full traces improve attribution accuracy by up to 76.5% over a partial-observation counterpart, confirming that missing inputs obscure many failure causes. TraceElephant provides a foundation for follow-up failure attribution research, promoting evaluation practices that reflect real-world debugging and supporting the development of more transparent MASs.

[Findings] SenseJudge: Human-Centric Preference-Driven Judgment Framework ▶ *Underline Video*
Rui Li, Junfeng liu, Xiangwen Kong and Zhifang Sui
Large Language Models (LLMs) as judges across various scenarios such as assessing model responses is becoming an increasingly accepted paradigm. However, existing judgment approaches often rely on trained judgers using fixed preference data, which tend to overlook diverse user preferences and struggle to adapt to real-world human-AI dialogue scenarios. To address these limitations, we propose SenseJudge, a customizable judgment framework driven by human preferences and SenseBench, a diverse and challenging instruction following benchmark derived from real-world multi-turn interactions. We applied the automatic judgment framework and benchmark to two tasks: 1) LLMs as personalized judges, and 2) model ranking. We conducted extensive experiments, and the results demonstrate that the SenseJudge framework surpasses other judgment methods and models in the LLMs-as-personalized-judges task and achieves model ranking that aligns with real human sense. Additionally, we conducted analyses on position bias and consistency, alongside ablation studies, which affirmed the robustness of SenseJudge.

[Long] ServImage: An Image Generation and Editing Benchmark from Real-world Commercial Imaging Services ▶ *Underline Video*
Fengxian Ji, Jingpu Yang, Zirui Song, Lang Gao, Junhong Liang, Zhenhao Chen, Jinghui Zhang and Xiuying Chen
Recent image generation and editing models demonstrate robust adherence to instructions and high visual quality on academic benchmarks. However, their performance on paid, real-world design projects remains uncertain. We introduce ServImage, a benchmark that explicitly correlates model outputs with economic value in commercial design projects. ServImage consists of (i) ServImageBench: a dataset of 1.07k paid commercial design tasks and 2.05k designer deliverables totaling over $295k, covering portrait, product, and digital content, along with 33k candidate images and 33k human annotations. (ii) ServImageScore: an integrated scoring system that combines three quality dimensions: baseline requirements fulfilment, visual execution quality, and commercial necessity satisfaction. These three dimensions are designed to characterize the factors that drive human payment decisions and indicate whether an image is commercially acceptable. (iii) ServImageModel: under this scoring system, we propose a payment prediction model trained on the human-annotated candidate images, achieving 82.00% accuracy in predicting human payment decisions and producing calibrated payment probabilities. ServImage provides a comprehensive foundation for assessing the commercial viability of image generation models and offers a scalable resource for future research on economically grounded vision systems ~Github.

[Long] SILO-BENCH: A Scalable Environment for Evaluating Distributed Coordination in Multi-Agent LLM Systems ▶ *Underline Video*
Yuzhe Zhang, Feiran Liu, Yi Shan, Xinyi Huang, Xin Yang, Yueqi Zhu, Xuxin Cheng, Cao Liu, Ke Zeng, Terry Jingchen Zhang and Wenyuan Jiang
Large language models are increasingly deployed in multi-agent systems to overcome context limitations by distributing information across agents. However, whether LLM-based agents can reliably coordinate when each observes only a fragment of the global problem remains unclear. Existing benchmarks often prescribe agent roles or interaction patterns, conflating coordination ability with role-based priors. We introduce SILO-BENCH, a role-free benchmark for evaluating free-form collaboration under information silos. The benchmark

comprises 30 algorithmic tasks with exact ground-truth answers, organized into 3 complexity levels based on optimal communication complexity: aggregation, mesh, and global shuffle. To systematically probe coordination capabilities, we instantiate 54 configurations by varying 3 communication protocols, 6 agent scales and 3 frontier LLMs, conducting 1,620 experiments. We evaluate agent behavior along three dimensions: Success Rate, Token Consumption, and Communication Density. Our experiments reveal a fundamental Communication-Reasoning Gap: agents communicate actively, yet fail to translate interaction into effective distributed computation. Performance collapses as complexity increases, with Level-III tasks achieving zero success beyond 50 agents. These findings demonstrate that current LLMs cannot escape information silos through coordination alone. SILO-BENCH provides a foundation for tracking progress toward genuinely collaborative multi-agent systems. The code is available at https://github.com/jwyjohn/acl26-silo-bench.

[Long] SimPBL: A Multi-Agent Framework for Project-Based Learning ▶ *Underline Video*

Daniel Zhang-Li, Joy Jia Yin Lim, Binglin Liu, Shangqing Tu, Zijun Yao, Hao Peng, Jifan Yu, Haoxuan Li, Zhanxin Hao, Ye He, Zekun Li, Jiangyi Wang, Lei Hou, Bin Xu, Xin Cong, Zhiyuan Liu, Huiqin Liu, Yu Zhang and Juanzi Li

Project-Based Learning (PBL) is an important learning method that promotes understanding and acquiring practical skills through training learners through a project. However, effective PBL often requires sustained orchestration and collaboration, but existing LLM-based learning tools provide partial assistance without explicitly modeling these roles, and overly comprehensive help provided by LLM can reduce learner autonomy. We propose SimPBL, a multi-agent framework with an orchestrator agent that provides adaptive scaffolding from interaction logs and collaborator agents that support project work through boundary-aware collaboration. We conduct comprehensive evaluation to study the effectiveness of SimPBL, where we observe a 14% improvement in learner examination score. Results from extensive studies further highlights the ability of SimPBL to manage learning behavior and improve learning experience. Code and materials are available at https://anonymous.4open.science/r/SimPBL-D5B8.

[Long] Social Story Frames: Contextual Reasoning about Narrative Intent and Reception ▶ *Underline Video*

Joel Mire, Maria Antoniak, Steven Wilson, Zexin Ma, Achyutarama Ganti, Andrew Piper and Maarten Sap

Reading stories evokes rich interpretive, affective, and evaluative responses, such as inferences about narrative intent or judgments about characters. Yet, computational models of reader response are limited, preventing nuanced analyses. To address this gap, we introduce SocialStoryFrames, a formalism for distilling plausible inferences about reader response, such as perceived author intent, explanatory and predictive reasoning, affective responses, and value judgments, using conversational context and a taxonomy grounded in narrative theory, linguistic pragmatics, and psychology. We develop two models, SSF-Generator and SSF-Classifier, validated through human surveys (N=382 participants) and expert annotations, respectively. We conduct pilot analyses to showcase the utility of the formalism for studying storytelling at scale. Specifically, applying our models to SSF-Corpus, a curated dataset of 6,140 social media stories from diverse contexts, we characterize the frequency and interdependence of storytelling intents, and we compare and contrast narrative practices (and their diversity) across communities. By linking fine-grained, context-sensitive modeling with a generic taxonomy of reader responses, SocialStoryFrames enable new research into storytelling in online communities.

[Long] SubTokenTest: A Practical Benchmark for Real-World Sub-token Understanding ▶ *Underline Video*

Shuyang Hou, Yi Hu and Muhan Zhang

Recent advancements in large language models (LLMs) have significantly enhanced their reasoning capabilities. However, they continue to struggle with basic character-level tasks, such as counting letters in words—a problem rooted in their tokenization process. While existing benchmarks have highlighted this weakness through basic character operations, such failures are often dismissed due to lacking practical relevance. Yet, many real-world applications, such as navigating text-based maps or interpreting structured tables, rely heavily on precise sub-token understanding. In this regard, we introduce SubTokenTest, a comprehensive benchmark that assesses sub-token understanding through practical, utility-driven tasks. Our benchmark includes ten tasks across four domains and isolates tokenization-related failures by decoupling performance from complex reasoning. We provide a comprehensive evaluation of nine advanced LLMs. Additionally, we investigate the impact of test-time scaling on sub-token reasoning and explore how character-level information is encoded within the hidden states.

[Long] TabReX: Tabular Referenceless eXplainable Evaluation ▶ *Underline Video*

Tejas Anvekar, Junha Park, Aparna Garimella and Vivek Gupta

Evaluating the quality of tables generated by large language models (LLMs) remains an open challenge: existing metrics either flatten tables into text, ignoring structure, or rely on fixed references that limit generalization. We present TabReX, a reference-less, property-driven framework for evaluating tabular generation via graph-based reasoning. TabReX converts both source text and generated tables into canonical knowledge graphs, aligns them through an LLM-guided matching process, and computes interpretable, rubric-aware scores that quantify structural and factual fidelity. The resulting metric provides controllable trade-offs between sensitivity and specificity, yielding human-aligned judgments and cell-level error traces. To systematically asses metric robustness, we introduce TabReX-Bench, a large-scale benchmark spanning six domains and twelve planner-driven perturbation types across three difficulty tiers. Empirical results show that TabReX achieves the highest correlation with expert rankings, remains stable under harder perturbations, and enables fine-grained model-vs-prompt analysis establishing a new paradigm for trustworthy, explainable evaluation of structured generation systems.

[Findings] Telling Speculative Stories to Help Humans Imagine the Harms of Healthcare AI ▶ *Underline Video*

Xingmeng Zhao, Tongnian Wang, Dan Schumacher, Veronica Rammouz and Anthony Rios
Artificial intelligence (AI) is rapidly transforming healthcare, enabling the fast development of tools such as stress monitors, wellness trackers, and mental health chatbots. However, this rapid and low-barrier development can also introduce risks, including bias, privacy violations, and unequal access, especially when systems overlook real-world contexts, diverse user needs, and cultural settings. Many recent approaches use AI to identify such risks automatically, but this can reduce human engagement in understanding how harms arise, who they affect, and which stakeholder needs remain unspoken. We present a human-centered ethical foresight framework that generates speculative user stories and supports multi-agent discussions to help people reflect on potential benefits and harms of healthcare AI before deployment. In a user study, participants who engaged with stories identified a broader range of harms, distributing their responses more evenly across all 17 harm types, whereas those who did not engage with stories focused primarily on privacy and well-being (79.1%). Overall, our findings suggest that storytelling helps people anticipate potential risks and benefits and reflect more broadly on how AI systems may affect different users, contexts, and often unspoken needs.

[Short] Temporal Leakage in Search-Engine Date-Filtered Web Retrieval: A Retrospective Forecasting Case Study ▶ *Underline Video*
Ali El Lahib, Ying-Jieh Xia, Zehan Li, Yuxuan Wang and Xinyu Pi
Search-engine date filters are widely used to enforce pre-cutoff retrieval in retrospective evaluations of search-augmented forecasters. We show this approach is unreliable across two major search engines: auditing Google Search's before: filter and DuckDuckGo's date-range filter, we find that at least one retrieved page contains major post-cutoff leakage for 71% of questions on Google and 81% on DuckDuckGo, and the answer is directly revealed for 41% and 55%, respectively. Using gpt-oss-120b to forecast with these leaky documents, we demonstrate inflated prediction accuracy (Brier score 0.10 vs. 0.24 with leak-free documents). We characterize recurring leakage mechanisms, including updated articles, related-content modules, unreliable metadata, and absence-based signals, and argue that date-restricted search on these engines is insufficient for credible retrospective evaluation. We recommend stronger retrieval safeguards or evaluation on frozen, time-stamped web snapshots.

[Long] The GDN-CC Dataset: Automatic Corpus Clarification for AI-enhanced Democratic Citizen Consultations ▶ *Underline Video*
Pierre-Antoine Lequeu, Léo Labat, Laurène Cave, Gaël Lejeune, François Yvon and Benjamin Piwowarski
LLMs are ubiquitous in modern NLP, and while their applicability extends to texts produced for democratic activities such as online deliberations or large-scale citizen consultations, ethical questions have been raised for their usage as analysis tools. We continue this line of research with two main goals: (a) to develop resources that can help standardize citizen contributions in public forums at the pragmatic level, and make them easier to use in topic modeling and political analysis; (b) to study how well this standardization can reliably be performed by small, open-weights LLMs, i.e. models that can be run locally and transparently with limited resources. Accordingly, we introduce Corpus Clarification as a preprocessing framework for large-scale consultation data that transforms noisy, multi-topic contributions into structured, self-contained argumentative units ready for downstream analysis. We present GDN-CC, a manually-curated dataset of 1,231 contributions to the French Grand Débat National, comprising 2,285 argumentative units annotated for argumentative structure and manually clarified. We then show that finetuned Small Language Models match or outperform LLMs on reproducing these annotations, and measure their usability for an opinion clustering task. We finally release GDN-CC-large, an automatically annotated corpus of 240k contributions, the largest annotated democratic consultation dataset to date.

[Findings] Thunder-KoNUBench: A Corpus-Aligned Benchmark for Korean Negation Understanding ▶ *Underline Video*
Sungmok Jung, Yeonkyoung So, Joonhak Lee, Sangho Kim, Yelim Ahn and Jaejin Lee
Although negation is known to challenge large language models (LLMs), benchmarks for evaluating negation understanding, especially in Korean, are scarce. We conduct a corpus-based analysis of Korean negation and show that LLM performance degrades under negation. We then introduce Thunder-KoNUBench, a sentence-level negation understanding benchmark that reflects the empirical distribution of Korean negation phenomena. Evaluating 47 LLMs on Thunder-KoNUBench, we analyze the effects of model size and instruction tuning, and perform error analysis to better understand model behavior. We further show that fine-tuning on Thunder-KoNUBench improves negation understanding and broader contextual comprehension in Korean.

[Long] ToolScope: Enhancing LLM Agent Tool Use through Tool Merging and Context-Aware Filtering ▶ *Underline Video*
Marianne Menglin Liu, Daniel Garcia, Fjona Parllaku, Vikas Upadhyay, Fahad Shah and Dan Roth
Large language model (LLM) agents rely on external tools to solve complex tasks, but real-world toolsets often contain redundant tools with overlapping names and descriptions, introducing ambiguity and reducing selection accuracy. LLMs also face strict input context limits, preventing efficient consideration of large toolsets. To address these challenges, we propose ToolScope, which includes: (1) ToolScopeMerger with Auto-Correction to automatically audit and fix tool merges, reducing redundancy, and (2) ToolScopeRetriever to rank and select only the most relevant tools for each query, compressing toolsets to fit within context limits without sacrificing accuracy. Evaluations on three state-of-the-art LLMs and three open-source tool-use benchmarks show gains of 8.38% to 38.6% in tool selection accuracy, demonstrating ToolScopes effectiveness in enhancing LLM tool use.

[Long] Toward Robust Evaluation for Multilingual Grammatical Error Correction: Can Large Language Models Replace Human References? ▶ *Underline Video*
Alla Rozovskaya and Dan Roth

A standard method for evaluating grammatical error correction systems severely underestimates performance, as it compares outputs against a small, fixed set of human references, despite the large space of possible valid corrections. Prior research has shown that using a closest-gold reference - i.e., a human reference generated with respect to the system output rather than the original text - yields more accurate performance estimates. Yet, producing such references for each system individually is costly. We introduce an automated method for generating closest-gold references by prompting a large language model (LLM) with system outputs. We find that performance scores computed using automatic closest-gold references correlate well with human closest-golds, whereas standard reference-based evaluations show weak or no correlation. Building on this insight, we use both fixed human references and closest-gold references generated by Claude and Llama to compare the performance of supervised models and GPT-4 across 14 benchmarks spanning 12 languages. Consequently, while prior work has shown that GPT-4 appears to lag behind traditional models, we demonstrate that this is due to the failures of the standard evaluation method that systematically underestimates GPT-4 performance more severely than that of supervised models. We show that a more appropriate evaluation approach, based on the closest gold method, reveals that GPT-4 outperforms traditional state-of- the-art models on almost all languages.

[Findings] Towards Reliable Paper Contributions Annotation in the ACL Rolling Review ► *Underline Video*

Julien Aubert-Béduchaud, Florian Boudin, Akiko Aizawa, Béatrice Daille and Richard Dufour

With the rapid growth of scientific publications, researchers struggle to efficiently assess the relevance of numerous papers. Identifying the types of contributions an article makes can help readers quickly grasp its significance. The ACL Rolling Review (ARR) introduced a typology requiring authors to specify their contributions to improve review quality and fairness. However, the current typology lacks clear definitions and guidance, leading to inconsistent labeling and raising concerns about its reliability. Our re-annotation campaign reveals substantial disagreement between authors and domain experts. Moreover, the predictions of large language models (LLMs), when compared with expert annotations, tend to be close to those provided by the authors. These findings suggest a potential path toward better annotation reliability within the ARR process.

[Long] TRACE: Evaluating Execution Efficiency of LLM-Based Code Translation ► *Underline Video*

Zhihao Gong, Zeyu Sun, Dong Huang, Qingyuan Liang, Jie M. Zhang and Dan Hao

While Large Language Models (LLMs) have substantially improved the functional correctness of code translation, the critical dimension of execution efficiency remains overlooked. We present Trace, the first benchmark to explicitly assess efficiency in LLM-translated code. Trace includes 1,000 efficiency-critical tasks across C++, Java, and Python, each augmented with stress tests that reveal efficiency disparities often overlooked by small-scale tests. Using Trace, we conduct an extensive evaluation of 28 representative LLMs and highlight several key insights: 1) Correctness and efficiency are often misaligned: the correctness leader Claude-Sonnet-4-Think achieves only moderate time efficiency, outperformed by smaller open-source LLMs such as Qwen2.5-Coder-14B-Instruct. 2) Inefficiency is both prevalent and patterned: 23.5% of correct translations suffer from notable inefficiency, mainly arising from algorithm implementation discrepancy (11.9%), language construct mismatch (66.4%), and resource management inefficiency (21.7%). 3) Inference-time prompt strategies bring only modest improvements, indicating that simple prompting alone is insufficient to improve translation efficiency. Together, our results establish execution efficiency as an essential dimension of code translation and position Trace as a principled foundation for efficiency-oriented evaluation. Our code and data are available at: https://github.com/Albert-Gong/TRACE.

[Long] Uncovering Intervention Opportunities for Suicide Prevention with Language Model Assistants ► *Underline Video*

Jaspreet Ranjit, Hyundong Justin Cho, Claire J. Smerdon, Yoonsoo Nam, Myles Phung, Jonathan May, John R. Blosnich and Swabha Swayamdipta

The National Violent Death Reporting System (NVDRS) documents suicides in the United States. In a demanding public health data pipeline, annotators manually extract structured information from death investigation records following extensive codebooks (i.e. annotation guidelines) painstakingly developed by experts. In this work, we facilitate data-driven insights from the NVDRS data to support the development of novel suicide interventions by leveraging language models (LM) as assistants to these (a) data annotators and (b) experts. We find that LM predictions match existing data annotations about 85% of the time across 50 NVDRS variables. Where the LM disagrees with existing annotations, our expert review identifies that 38% of these instances reveal inconsistencies between narratives and structured data. Finally, we introduce a human-in-the-loop algorithm that helps experts efficiently build and refine codebooks for new variables by having them only focus on providing feedback for incorrect LM predictions. We apply our algorithm to a real-world case study, and find that about 96K narratives contain evidence of victim interactions with legal professionals, which surfaces a substantial opportunity for upstream intervention that is not captured in the original structured data. Our findings provide evidence that LMs can serve as effective assistants to public health researchers who handle sensitive data in high-stakes scenarios.

[Long] Wait! Theres a Way Out: A Decision Mechanism for Forecasting Conversational Derailment ► *Underline Video*

Laerdon Kim, Vivian Nguyen and Cristian Danescu-Niculescu-Mizil

Forecasting conversational derailment is the task of predicting, as the conversation unfolds, whether it will eventually derail into personal attacks. Since forecasting models operate in an online fashion, they must decide whether to "trigger" an alert after each utterance, for example, to notify participants or a moderator that the conversation is at risk of derailing. Existing approaches make this decision solely based on the estimated likelihood of derailment given the preceding utterances, implicitly assuming that the conversations future trajectory is fixed. As a result, they ignore the possibility of future recovery and incur an unnecessarily high rate of false positives. In this work we propose a method for decoupling the decision to trigger from derailment likelihood estimation. Our approach is inspired by the first human

baseline on this task, which shows that humans achieve dramatically lower false positive rates by selectively deferring their decision to trigger when they anticipate that tension is likely to subside. We operationalize this insight with a deferral mechanism that uses forward-looking simulations to assess whether a tense moment admits plausible paths to recovery. Incorporating this mechanism into a state-of-the-art forecasting model substantially reduces false positives without sacrificing forecasting accuracy. More broadly, this work highlights the value of treating decision-making as a first-class component of forecasting systems.

[Long] WildFeedback: Aligning LLMs With In-situ User Interactions And Feedback ▶ *Underline Video*

Taiwei Shi, Zhuoer Wang, Longqi Yang, Ying-Chun Lin, Zexue He, Mengting Wan, Pei Zhou, Sujay Kumar Jauhar, Sihao Chen, Shan Xia, Hongfei Zhang, Jieyu Zhao, Xiaofeng Xu, Xia Song and Jennifer Neville

As large language models (LLMs) continue to advance, aligning these models with human preferences has emerged as a critical challenge. Traditional alignment methods, relying on human or LLM annotated datasets, are limited by their resource-intensive nature, inherent subjectivity, misalignment with real-world user preferences, and the risk of feedback loops that amplify model biases. To overcome these limitations, we introduce WildFeedback, a novel framework that leverages in-situ user feedback during conversations with LLMs to create preference datasets automatically. Given a corpus of multi-turn user-LLM conversation, WildFeedback identifies and classifies user feedback to LLM responses between conversation turns. The user feedback is then used to create examples of preferred and dispreferred responses according to users' preference. Our experiments demonstrate that LLMs fine-tuned on WildFeedback dataset exhibit significantly improved alignment with user preferences, as evidenced by both traditional benchmarks and our proposed checklist-guided evaluation. By incorporating in-situ feedback from actual users, WildFeedback addresses the scalability, subjectivity, and bias challenges that plague existing approaches, marking a significant step toward developing LLMs that are more responsive to the diverse and evolving needs of their users.

[Findings] Something Just Like TRuST *: Toxicity Recognition of Span and Target ▶ *Underline Video*

Berk Atil, Namrata Sureddy and Rebecca J. Passonneau

Toxic language includes content that is offensive, abusive, or that promotes harm. Progress in preventing toxic output from large language models (LLMs) is hampered by inconsistent definitions of toxicity. We introduce TRuST, a large-scale dataset that unifies and expands prior resources through a carefully synthesized definition of toxicity, and corresponding annotation scheme. It consists of ~300k annotations, with high-quality human annotation on ~11k. To ensure high-quality, we designed a rigorous, multi-stage human annotation process, and evaluated the diversity of the annotators. Then we benchmarked state-of-the-art LLMs and pre-trained models on three tasks: toxicity detection, identification of the target group, and of toxic words. Our results indicate that fine-tuned PLMs outperform LLMs on the three tasks, and that current reasoning models do not reliably improve performance. TRuST constitutes one of the most comprehensive resources for evaluating and mitigating LLM toxicity, and other research in socially-aware and safer language technologies.

Poster Session D (Grand Hall)

09:00-10:30 (Grand Hall)

[Long] NavA^3: Understanding Any Instruction, Navigating Anywhere, Finding Anything ▶ *Underline Video*

Lingfeng Zhang, Xiaoshuai Hao, Yingbo Tang, Haoxiang Fu, Xinyu Zheng, Pengwei Wang, Zhongyuan Wang, Wenbo Ding and Shanghang Zhang

Embodied navigation is a fundamental capability of embodied intelligence, enabling robots to move and interact within physical environments. However, existing navigation tasks primarily focus on predefined object navigation or instruction following, which significantly differs from human needs in real-world scenarios involving complex, open-ended scenes. To bridge this gap, we introduce a challenging long-horizon navigation task that requires understanding high-level human instructions and performing spatial-aware object navigation in real-world environments. Existing embodied navigation methods struggle with such tasks due to their limitations in comprehending high-level human instructions and localizing objects with an open vocabulary. In this paper, we propose NavA^3, a hierarchical framework divided into two stages: global and local policies. In the global policy, we leverage the reasoning capabilities of Reasoning-VLM to parse high-level human instructions and integrate them with global 3D scene views. This allows us to reason and navigate to regions most likely to contain the goal object. In the local policy, we have collected a dataset of 1.0 million samples of spatial-aware object affordances to train the NaviAfford model (PointingVLM), which provides robust open-vocabulary object localization and spatial awareness for precise goal identification and navigation in complex environments. Extensive experiments demonstrate that NavA^3 achieves SOTA results in navigation performance and can successfully complete long-horizon navigation tasks across different robot embodiments in real-world settings, paving the way for universal embodied navigation. The dataset and code will be made available.

[TACL] A Survey on Memory-Efficient Fine-Tuning for Large Language Models ▶ *Underline Video*

Yeachan Kim, Mingyu Lee and SangKeun Lee

Fine-tuning large language models (LLMs) is a crucial process to align them with human intentions, yet this process remains memory-intensive, varying across tasks and model architectures. These huge and variable memory costs complicate scaling and deployment of LLMs, especially on limited hardware. However, existing surveys on memory efficiency are often either superficial or too narrow in scope, typically

focusing on specific subfields. To address this gap, this survey presents the first systematic review of memory-efficient fine-tuning (MEFT) tailored for LLMs. To structure the research landscape, we first categorize existing approaches by their optimization environments (i.e., model itself and systems) and further classify model-based approaches by their specific optimization targets. We also discuss evaluation strategies for assessing MEFT methods and provide empirical analyses. By highlighting challenges and future directions based on current methods, this survey aims to serve as a practical guide for developing MEFT methods.

[SRW] A11y-Compressor: A framework for Enhancing the Efficiency of GUI Agent Observations through Visual Context Reconstruction and Redundancy Reduction ▶ *Underline Video*

Michito Takeshita, Takuro Kawada, Takumi Ohashi, Shunsuke Kitada and Hitoshi Iyatomi;

AI agents that interact with graphical user interfaces (GUIs) require effective observation representations for reliable grounding. The accessibility tree is a commonly used text-based format that encodes UI element attributes, but it suffers from redundancy and lacks structural information such as spatial relationships among elements. We propose A11y-Compressor, a framework that transforms linearized accessibility trees into compact and structured representations. Our implementation, Compressed-a11y, applies a lightweight and structured transformation pipeline with modal detection, redundancy reduction, and semantic structuring. Experiments on the OSWorld benchmark show that Compressed-a11y reduces input tokens to 22% of the original while improving task success rates by 5.1 percentage points on average.

[Findings] Adaptive Layer Selection for Layer-Wise Token Pruning in LLM Inference ▶ *Underline Video*

Rei_Taniguchi Rei_Taniguchi, Yuyang Dong, Makoto Onizuka and Chuan Xiao

Due to the prevalence of large language models (LLMs), key-value (KV) cache reduction for LLM inference has received remarkable attention. Among numerous works that have been proposed in recent years, layer-wise token pruning approaches, which select a subset of tokens at particular layers to retain in KV cache and prune others, are one of the most popular schemes. They primarily adopt a set of pre-defined layers, at which tokens are selected. Such design is inflexible in the sense that the accuracy significantly varies across tasks and deteriorates in harder tasks such as KV retrieval. In this paper, we propose ASL, a training-free method that adaptively chooses the selection layer for KV cache reduction, exploiting the variance of token ranks ordered by attention score. The proposed method balances the performance across different tasks while meeting the user-specified KV budget requirement. ASL operates during the prefilling stage and can be jointly used with existing KV cache reduction methods such as SnapKV to optimize the decoding stage. By evaluations on the InfiniteBench, RULER, and NIAH benchmarks, we show that ASL, equipped with one-shot token selection, adaptively trades inference speed for accuracy, outperforming state-of-the-art layer-wise token pruning methods in difficult tasks.

[Long] Agentic Very Long Video Understanding ▶ *Underline Video*

Aniket Rege, Arka Sadhu, Yuliang Li, Kejie Li, Ramya Korlakai Vinayak, Yuning Chai, Yong Jae Lee and Hyo Jin Kim

The advent of always-on personal AI assistants, enabled by all-day wearable devices such as smart glasses, demands a new level of contextual understanding, one that goes beyond short, isolated events to encompass the continuous, longitudinal stream of egocentric video. Achieving this vision requires advances in long-horizon video understanding, where systems must interpret and recall visual and audio information spanning days or even weeks. Existing methods, including large language models and retrieval-augmented generation, are constrained by limited context windows and lack the ability to perform compositional, multi-hop reasoning over very long video streams. In this work, we address these challenges through EGAgent, an enhanced agentic framework centered on entity scene graphs, which represent people, places, objects, and their relationships over time. Our system equips a planning agent with tools for structured search and reasoning over these graphs, as well as hybrid visual and audio search capabilities, enabling detailed, cross-modal, and temporally coherent reasoning. Experiments on the EgoLifeQA and Video-MME-long datasets show that our method achieves state-of-the-art performance on EgoLifeQA (57.5%) and competitive performance on Video-MME-long (74.1%) for complex longitudinal video understanding tasks.

[Long] Aligning Large Language Models via Fully Self-Synthetic Data ▶ *Underline Video*

Shangjian Yin, Zhepei Wei, Xinyu Zhu, Wei-Lin Chen and Yu Meng

Traditional reinforcement learning from human feedback (RLHF) for large language models (LLMs) relies on expensive human-annotated datasets, while Reinforcement Learning from AI Feedback (RLAIF) also incurs significant costs, requiring the collection of diverse prompts and corresponding responses, often necessitating external reward models or proprietary models like GPT-4 to annotate preference pairs. In this work, we introduce Self-Alignment Optimization (SAO), a fully self-synthetic framework for LLM alignment, where all training data, including prompts (i.e., user queries), responses, and preferences, are generated by the model itself. Specifically, SAO first instructs the LLM to engage in persona role-play and generate diverse prompts and responses, which are then self-evaluated for preference optimization. Extensive experiments demonstrate that SAO effectively enhances the models chat capabilities on standard benchmarks like AlpacaEval 2.0, while maintaining strong performance on downstream objective tasks (i.e.,, question-answering, math reasoning). Our work provides a practical solution for self-improvement in aligning LLMs, and the code for reproducing our results is available at: https://github.com/SJY8460/SAO.

[SRW] An Incremental CYK Recognizer for GPU-accelerated General Context-free Prefix Validation ▶ *Underline Video*

Jiacheng Zhang, Ayesha Khatun and Steven Bethard

Constrained decoding systems are often built on context-free parsers intended for programming languages. These parsers either degrade into O(n^3) time-complexity or fail entirely if the grammar is not carefully engineered to keep properties such as determinism and non-ambiguity.

There is thus a need to design parsers that efficiently handle non-determinism and ambiguity, while simultaneously being incremental so that they can be coupled with the token-based predictions of large language models. Inspired by prior work, we derive an incremental CYK prefix recognizer, which still has O(n^3) complexity but allows for acceleration with only a fraction of GPU resources (stream multiprocessors). Our parser shows robust efficiency in complex context-free grammars while other parsers crash or degrade. At the same time, we remain empirically competitive in restrictive grammar classes such as LALR.

[SRW] Analysis of the Neglect-Zero Effect in Large Language Models ▶ *Underline Video*

Jin Tanaka; Daiki Matsuoka; Ryoma Kumon; Hitomi Yanaka;

We investigate the extent to which the language processing of LLMs resembles human cognitive processes, focusing on a human cognitive bias called the neglect-zero effect. This effect refers to the human tendency to ignore zero-models, which are configurations that render a proposition vacuously true by virtue of an empty set. We focus on two types of inference driven by the neglect-zero effect, and examine how LLMs process the inference by comparing their behavior with that in inference not involving the neglect-zero effect. For this purpose, we employ a paradigm using structural priming, where recent exposure to a preceding sentence (the prime) facilitates the processing of a subsequent sentence (the target) due to their structural similarity. We prepare primes so that they force LLMs to consider the zero-model, and analyze whether they also consider zero-models in the target. The results indicate the possibility that the neglect-zero effect does not arise within the LLMs analyzed in this study.

[SRW] Analyzing Hate Speech Amplification on Fringe Platforms ▶ *Underline Video*

Anika Basu

Fringe platforms like Gab harbor high volumes of hate speech due to minimal moderation and insular communities. Our study examines the factors that determine how hate speech amplifies on these platforms. We scraped a novel dataset of 5K+ threads and 50K+ responses from four fringe platforms (Gab, 4chan, Stormfront, and Vanguard), where hate speech amplification was measured using platform-specific engagement metrics. We trained both Generalized Linear Models and Gradient Boosted Tree models to estimate how several features influence the amplification of hate speech on fringe platforms, and used Shapley value estimates to identify the relative importance of the features. Our analysis shows that research insights from social network analysis (SNA) of mainstream sites like X do not directly generalize to fringe platforms. For instance, our experiments show that using features like thread structure and disagreements in early response windows can give up to 74% RMSE lift in predicting reply counts for hateful posts on fringe platforms, on top of a baseline model with features like hate intensity and thread age that would be considered predictive by regular SNA methods.

[Long] Anchoring the Cache: Mitigating Contextual Hallucination in KV-Compressed Long-Context Summarization ▶ *Underline Video*

Yu Fu, Chen Luo, Josef Valvoda, Xin Zhang, Xuejing Lei, Xiao Pan, Hui Liu and Yue Dong

Key-Value (KV) cache compression techniques have improved the efficiency of long-context summarization in Large Language Models (LLMs), but their impact on model hallucination remains underexplored. In this paper, we present the first systematic study of how KV cache compression affects hallucination in long-context summarization, demonstrating that aggressive compression can increase hallucination scores by up to 3.36x compared to the baseline. To mitigate this issue, we propose HalluKV, a decoding-phase strategy that selectively removes generated KV pairs from retrieval heads responsible for retrieving critical information from source context, thereby anchoring their attention on the preserved source information. Our approach maintains computational efficiency while significantly reducing hallucination across multiple models and datasets, achieving up to 5.48 average point reductions on Llama-3-8B-Instruct, enabling more trustworthy long-context summarization.

[SRW] Annotation Entropy Predicts Per-Example Learning Dynamics in LoRA Fine-Tuning ▶ *Underline Video*

Brady Steele

We find that LoRA fine-tuning exhibits un-learning on contested examples: items with high annotator disagreement show increasing loss during training, a qualitatively distinct pattern largely absent under full fine-tuning and consistent across all six models tested (four encoder, two decoder-only). This discovery emerges from correlating annotation entropy, computed from ChaosNLI's 100 labels per example, with per-example area under the loss curve (AULC) on SNLI and MNLI. The correlation is positive in all 25 conditions tested (Spearman = 0.06-0.43), with decoder-only models showing stronger correlations than encoders at matched LoRA rank. The effect survives partial-correlation controls and replicates across seeds and datasets. A preliminary noise-injection experiment is consistent with these findings.

[Findings] Assessing the Effect of Context in Multi-domain Acceptability Judgment ▶ *Underline Video*

Eunike Andriani Kardinata, Yusuke Sakai and Taro Watanabe;

Acceptability judgments provide a crucial basis for understanding how sentences are perceived as natural or well-formed, and they are increasingly used to assess the linguistic capability of large language models (LLMs). Unlike grammaticality, acceptability depends not only on structural form but also on contextual and domain-specific factors. Most prior work evaluates sentences in isolation, and relatively little is known about how explicit contextual cues influence LLM acceptability judgments across domains. This study examines how contextual information affects model-generated acceptability ratings across multiple domains and several LLMs, using different forms of domain-specific contextual cues to situate sentences in their intended usage settings. The results show that context can meaningfully shift model judgments, although its effects vary across models and domains. Overall, the findings provide evidence on contextual effects in LLM

acceptability judgment and support the development of more context-aware evaluation frameworks.

[Findings] AudioStealer: Extracting Audio Prompts via Shapley Value-Guided Query Search ▶ *Underline Video*
Yingbin Jin, Xingjian Du, Hanjun Luo, Zihao Wang, Haibo Hu, XiaoFeng Wang and Xinfeng Li
As text-to-music models gain widespread adoption, the prompts used to guide these systems have become valuable intellectual property. This shift has given rise to a new form of attack: prompt stealing, aiming to reconstruct the high-value prompts that guide the music generation. However, unlike prior work in text and image generation, prompt stealing in text-to-music systems faces unique challenges due to the entangled and diffuse nature of semantic representations in audio, which complicates the decoupling of specific textual tokens from acoustic outputs. To address these challenges, we present AudioStealer, the first targeted study of prompt inversion in the audio domain. AudioStealer operates via a two-stage black-box attack framework: first, a heuristic search guided by audio-language embeddings identifies initial candidates; then, these candidates are refined using a game-theoretic strategy based on Shapley value estimation to attribute precise semantic contributions. Our method requires no direct access to the target model and relies solely on a shadow model, making it broadly applicable. Through extensive experiments, we demonstrate that AudioStealer recovers prompts with high textual consistency to the ground truth, while the regenerated audio maintains strong perceptual similarity to the target recordings. These results expose critical vulnerabilities in the text-to-audio market ecosystem and underscore the urgent need for intellectual property protections in generative audio technologies.

[Findings] AWARE: Agentic Knowledge Warehousing for Contextual Intelligence ▶ *Underline Video*
Hongjin Qian, Siqi Bao, Zhao Cao and Zheng Liu
Information seeking bridges the knowledge gap between a query and its answer. Although LLMs perform well broadly, their ability to close this gap is limited by pretraining and degrades on specialized or up-to-date queries. A common remedy augments LLMs with external knowledge, either by injecting retrieved evidence into context or interleaving retrieval with reasoning. The former limits exploration of layered dependencies, while the latter is bounded by context length, constraining efficiency and scalability. For complex tasks with intricate dependencies and large text volumes, both approaches become inadequate. To tackle this bottleneck, we present AWARE (Agentic Knowledge Warehouse), an agentic knowledge warehousing framework that transforms heterogeneous, unstructured data into minimal, task-conditioned knowledge representations consumable by LLMs. Rather than exposing raw text, AWARE constructs knowledge through intent planning, online multi-threaded exploration, and map-reduce evidence integration, producing compact, LLM-ready context under finite budgets. Specifically, it applies offline document structuring to generate document headers that support controlled access, performs exploration with targeted refinement to recover layered information dependencies, and integrates distributed evidence into task-aware representations for downstream answer generation. Experiments on GAIA, WebWalker, and BrowseComp-Plus show improvements over all baselines

[SRW] Believing is Seeing: How Token Inflation Mechanistically Erodes Theory of Mind in Large Language Models ▶ *Underline Video*
Zhizhi Wang and Ruochen Zhang
Theory of Mind (ToM), the ability to attribute mental states to others that may diverge from reality, is operationalized through false-belief tasks: a character holds an outdated belief about an object's location (the ToM answer), which conflicts with where the object actually is (the reality answer). Large language models can pass such tests at short context lengths, yet their performance degrades sharply when input is padded with irrelevant tokens, a phenomenon we term token inflation. We use mechanistic interpretability to investigate why this degradation occurs in Llama 3-8B-Instruct. Using a suite of activation patching experiments, we show that token inflation corrupts belief-state representations during early-layer processing, and that this corruption cannot be recovered by the model's own later computation. The failure is specific to belief information: restoring early-layer activations from a clean short-context run rescues up to 16 percentage points of lost accuracy, whereas the same intervention using reality-state or random activations provides no benefit. A head-level analysis further localizes the damage to a distributed circuit of attention heads whose focus on belief-relevant tokens collapses under inflation. These findings generalize across four languages (EN, ZH, JA, KO) and three padding types (dot, Wikipedia, shuffled Wikipedia). Our results demonstrate that token inflation induces a representational collapse of belief states during early-layer processing, pinpointing a concrete mechanistic bottleneck that must be addressed for LLMs to reason reliably about others' mental states.

[Long] Beyond "I Don't Know": Evaluating LLM Self-Awareness in Discriminating Data and Model Uncertainty ▶ *Underline Video*
Jingyi Ren, Ante Wang, Yunghwei Lai, Xiaolong Wang, Linlu Gong, Weitao Li, Weizhi Ma and Yang Liu
Reliable Large Language Models (LLMs) should abstain when confidence is insufficient. However, prior studies often treat refusal as a generic "I don't know", failing to distinguish input-level ambiguity (data uncertainty) from capability limitations (model uncertainty). This lack of distinction limits downstream action decisions like requesting clarification or invoking external tools. In this work, we introduce UA-Bench, a benchmark of over 3,500 questions drawn from six datasets spanning knowledge-intensive and reasoning-intensive tasks, designed to evaluate explicit uncertainty attribution. An evaluation of 18 frontier LLMs shows that even state-of-the-art models struggle to reliably discriminate between data uncertainty and model uncertainty, and that high answer accuracy does not necessarily imply strong uncertainty attribution ability. To narrow this gap, we propose a lightweight data synthesis and reinforcement learning strategy. Experiments on both Qwen3-4B-Instruct-2507 and Qwen3-8B in thinking mode show that the proposed method improves uncertainty attribution while preserving answer accuracy. Our code and data are publicly available now.

[SRW] Beyond Discrete Search: Divergent Thinking as Optimization in Latent Intention Space ▶ *Underline Video*

Mateusz Bystroski, Grzegorz Piotrowski and Tomasz Jan Kajdanowicz

We argue that LLM-based coding agents frequently fail to solve problems that lie within the model's capacity and the bottleneck is often the conditioning context rather than the model itself. We formalize this for the full class of Turing-computable problems with verifiable specifications and introduce a framework that recasts coding as optimization over a latent space of intentions, natural-language descriptions of high-level solution strategies, and searches this vector space with feedback guidance to steer coding agent toward correct solutions of the problems. The method operates as a plug-in layer that can wrap any coding agent without modifying its architecture or weights. On SWE-Bench Verified, our method raises the resolution rate of a weak, quantized 24B open-weight model to parity with frontier models +25x its size.

[Findings] BRIEF-Pro: Universal Context Compression with Short-to-Long Synthesis for Fast and Accurate Multi-Hop Reasoning ▶ *Underline Video*

Jia-Chen Gu, Junyi Zhang, Di Wu, Yuankai Li, Kai-Wei Chang and Nanyun Peng

As retrieval-augmented generation (RAG) tackles complex tasks, increasingly expanded contexts offer richer information, but at the cost of higher latency and increased cognitive load on the model. To mitigate this bottleneck, especially for intricate multi-hop questions, we introduce BRIEF-Pro. It is a universal, lightweight compressor that distills relevant evidence for a given query from retrieved documents into a concise summary for seamless integration into in-context RAG. Using seed data consisting of relatively short contexts (fewer than 1k words), BRIEF-Pro is trained to perform abstractive compression of extended contexts exceeding 10k words across a wide range of scenarios. Furthermore, BRIEF-Pro offers flexible user control over summary length by allowing users to specify the desired number of sentences. Experiments on four open-domain multi-hop question-answering datasets show that BRIEF-Pro generates more concise and relevant summaries, enhancing performance across small, large, and proprietary language models. With the 70B reader model, 32x compression by BRIEF-Pro improves QA performance by 4.67% on average over LongLLMLinguas 9x, while requiring only 23% of its computational overhead.

[Findings] BubbleRAG: Interactive Cognitive Offloading with Thought Bubble in Retrieval-Augmented Generation ▶ *Underline Video*

Fuda Ye, Jiachuan Wang, Yongqi Zhang, Lei Chen and Shuangyin Li

Retrieval-augmented generation (RAG) extends the capabilities of large language models (LLMs) by providing access to external knowledge. However, traditional retrieval-augmented LLMs rely on a silent reading paradigm that processes all retrieved documents passively, forcing them to reason without any interaction with the documents. This paradigm contrasts sharply with human interactive reading behavior, where external tools, such as bookmarks and notes, are used to offload cognitive demands. This paper introduces BubbleRAG, an enhanced RAG framework that emulates human interactive reading through annotation and re-reading. Specifically, BubbleRAG utilizes a lightweight thought bubble module that offloads LLM's internal cognition into external bookmark tokens, which are then annotated back into the context. These bookmarks serve as externalized memory, allowing the LLM to revisit these annotations in subsequent reading and answering. Notably, BubbleRAG is particularly suitable for low-resource scenarios, as the LLM parameters remain frozen. Extensive experiments confirm the effectiveness, robustness, and generalizability of BubbleRAG. Our findings demonstrate that BubbleRAG enables LLMs to achieve superior evidence identification abilities typically seen in retrievers, while establishing a cognitive link between external and internal information during answer generation. The source code is available at https://github.com/yefd/BubbleRAG.

[Long] CachePrune: Teaching LLMs What Not to Follow via KV-Cache Editing ▶ *Underline Video*

Rui Wang, Junda Wu, Yu Xia, Tong Yu, Ruiyi Zhang, Ryan Rossi, Subrata Mitra, Lina Yao and Julian McAuley

Large Language Models (LLMs) are susceptible to indirect prompt injection attack, where the model inadvertently responds to instructions injected into the prompt context. This vulnerability stems from LLMs inability to distinguish between data and instructions within a prompt. We propose CachePrune that defends against this attack by identifying and pruning neurons associated with instruction-following, during KV cache encoding of the prompt context. The pruning steers the LLM toward interpreting the context purely as data rather than as instructions to follow. To identify these neurons, we introduce a neural attribution mechanism guided by a preferential attribution loss, and theoretically connect this loss to an upper bound of the Direct Preference Optimization (DPO) objective. Further, we improve on the fidelity of neural attribution by leveraging an observed triggering effect in instruction-following. Our approach does not interfere with prompt formatting or incur test-time overhead in response generation. Experiments show that CachePrune significantly reduces the attack success rate while preserving the LLMs ability to follow user instructions.

[Findings] Calibrating Model-Based Evaluation Metrics for Summarization ▶ *Underline Video*

Hongye Liu, Dhanajit Brahma and Ricardo Henao

Recent advances in summary evaluation are based on model-based metrics to assess quality dimensions, such as completeness, conciseness, and faithfulness. However, these methods often require large language models, and predicted scores are frequently miscalibrated, limiting their reliability. Moreover, evaluating the average quality across different summaries for a single document typically requires access to multiple reference summaries. Here, we propose a general framework that generates individual and average proxy scores without relying on reference summaries, human annotations, or expensive model-based metrics. We also propose group isotonic regression binning (GIRB), a calibration method that adjusts the raw predictions to better align with ground-truth evaluation metrics. While we focus on continuous-value scenarios, such as summarization, the method is applicable to discrete-value tasks, such as question answering. Experiments on seven

datasets demonstrate that our approach consistently outperforms existing baselines.

[Short] Chain-of-Thought Degrades Visual Spatial Reasoning Capabilities of Multimodal LLMs ▶ *Underline Video*
Sai Srinivas Kancheti, Aditya Sanjiv Kanade, Vineeth N. Balasubramanian and Tanuja Ganu
Multimodal Reasoning Models (MRMs) leveraging Chain-of-Though (CoT) based thinking have revolutionized mathematical and logical problem-solving. However, we show that this paradigm struggles with generalized spatial intelligence. We perform a comprehensive evaluation of sixteen models across thirteen spatial benchmarks and identify a critical gap: CoT prompting consistently degrades performance in visual spatial reasoning. Furthermore, through a novel No-Image++ ablation, we demonstrate that MRMs and CoT prompted MLMs suffer from severe shortcut learning, and hallucinate visual details from textual priors even when the image is absent. These findings challenge the efficacy of text-only CoT for spatial tasks and underscore the need for vision-centric reasoning paradigms.

[SRW] Claim Verification in the Age of Large Language Models: A Survey ▶ *Underline Video*
Alphaeus Dmonte, Roland Oruche, Marcos Zampieri, Prasad Calyam and Isabelle Augenstein
The large and ever-increasing amount of data available on the Internet, coupled with the laborious task of manual claim and fact verification, has sparked interest in the development of automated claim verification systems. Several deep learning and transformer-based models have been proposed for this task over the years. With the introduction of Large Language Models (LLMs) and their superior performance in several NLP tasks, we have seen a surge of LLM-based approaches to claim verification along with the use of novel methods such as Retrieval Augmented Generation (RAG). In this survey, we present a comprehensive account of recent claim verification frameworks using LLMs. We describe the different components of the claim verification pipeline used in these frameworks in detail, including common approaches to retrieval, prompting, and fine-tuning. Finally, we describe publicly available English datasets for this task.

[Long] ClusterRAG: Cluster-Based Collaborative Filtering for Personalized Retrieval-Augmented Generation ▶ *Underline Video*
Gibson Nkhata, Uttamasha Anjally Oyshi, Quan Mai and Susan Gauch
Personalized Retrieval-Augmented Generation (RAG) relies on accurately selecting user-relevant documents. In practice, existing RAG approaches often suffer from high retrieval costs and overlook that collaborative signals from similar users can enhance personalized generation for the current user. We propose ClusterRAG, a Cluster-Based Collaborative Filtering for Personalized Retrieval-Augmented Generation. ClusterRAG represents users through their profile documents, organizes users into semantically coherent clusters using density-based clustering, and performs retrieval at both the cluster and document levels via cluster-level similarity and fine-grained ranking. Extensive experiments on the LaMP benchmark demonstrate that jointly leveraging the target users profile and profiles from top similar users consistently yields the best performance across diverse tasks. Further analysis shows that ClusterRAG integrates seamlessly with different dense retrievers and rankers, and remains effective when paired with both fine-tuned and zero-shot language models.

[Findings] CODEPROMPTZIP: Code-specific Prompt Compression for Retrieval-Augmented Generation in Coding Tasks with LMs ▶ *Underline Video*
Pengfei He, Shaowei Wang and Tse-Hsun Chen
Retrieval-Augmented Generation (RAG) enhances code generation by incorporating retrieved code examples into prompts, but the resulting long-context inputs impose substantial memory and computational overhead. Existing prompt compression techniques are largely designed for natural language and fail to account for the structural and semantic properties of code, while also lacking fine-grained control over compression ratios. We propose CodePromptZip, a code-aware prompt compression framework for RAG that enables precise length control while preserving critical information. Motivated by type-aware ablation studies, CodePromptZip leverages static analysis to rank code tokens by information gain and applies a dynamic compression strategy to retain the most informative tokens under a given budget. For incomplete or unparsable code snippets, CodePromptZip employs a language-model-based compressor trained on analyzable samples and augmented with a copy mechanism to preserve key tokens. Extensive experiments on three code-related tasks demonstrate that CodePromptZip consistently outperforms entropy-based and distillation-based baselines, achieving improvements of 23.4%, 28.7%, and 8.7%, respectively, while providing accurate control over compression ratios.

[Findings] CodeScout: Contextual Problem Statement Enhancement for Software Agents ▶ *Underline Video*
Manan Suri, Xiangci Li, Mehdi Shojaie, Songyang Han, Chao-Chun Hsu, Shweta Garg, Aniket Anand Deshmukh and Varun Kumar
Current AI-powered code assistance tools often struggle with ambiguous problem statements that lack sufficient task context and requirements specification. Recent analysis of software engineering agents reveals that failures on such ambiguous requests are highly correlated with longer trajectories involving either over-exploration or repeated attempts at applying the same fix without proper evolution or testing, leading to suboptimal outcomes across software development tasks. We introduce CodeScout, a breakthrough contextual query refinement approach that systematically converts ambiguous user requests into comprehensive, actionable problem statements through lightweight pre-exploration of the target codebase. Our key innovation is demonstrating that structured analysis before task execution can supplement existing agentic capabilities without requiring any modifications to their underlying scaffolds. CodeScout performs targeted context scoping, conducts multi-perspective analysis examining potential fixes and exploration opportunities, then synthesizes these insights into enhanced problem statements with reproduction steps, expected behaviors, and targeted exploration hints. This pre-exploration directly addresses the identified failure patterns by reducing non-converging agent trajectories while clarifying user intent in natural language space. We evaluate CodeScout using state-of-the-art agentic scaffolds and language models on SWEBench-Verified, demonstrating a 20% improvement in resolution rates

with up to 27 additional issues resolved compared to the default baseline method. Our results suggest that systematic query refinement through contextual analysis represents a promising direction for enhancing AI code assistance capabilities.

[Findings] CollabCoder: Plan-Code Co-Evolution via Collaborative Decision-Making for Efficient Code Generation ▶ *Underline Video*

Tung Doan, Huy Phung, Dung Nguyen and Nam Bui

Automated code generation remains a persistent challenge in software engineering, as conventional multi-agent frameworks are often constrained by static planning, isolated execution, high computational overhead, and limited adaptability to complex tasks. This paper introduces CollabCoder, a novel Plan-Code Co-Evolution framework that improves code generation through dynamic multi-agent collaboration. The core idea is to design a collaborative decision-making process between the plan module and the code module to decide which module should be executed for the debugging process. Extensive experiments on widely used benchmarks demonstrate that CollabCoder consistently improves code quality and robustness across tasks. Importantly, CollabCoder achieves performance comparable to or exceeding current state-of-the-art methods while reducing computational overhead, with efficiency gains becoming more pronounced as benchmark difficulty increases. On the more challenging LiveCodeBench and xCodeEval benchmarks, our approach improves performance by 11-20% over strong baselines while reducing the number of API calls by an average of 4-10 per execution.

[SRW] Conformal LLM Routing with Distribution-Free Safety Guarantees ▶ *Underline Video*

Iqtedar Uddin; André Bauer

LLM routing directs queries to a cheaper model when it suffices and to an expensive model otherwise, reducing inference cost. Existing input-based routers optimize cost-performance trade-offs but provide no formal bound on how often the cheaper model fails among routed queries. We adapt a proactive conformal gate framework to LLM routing. A logistic regression gate trained on text embeddings predicts per-query safety, and Clopper-Pearson conformal calibration selects a routing threshold that guarantees the violation rate among routed queries stays below alpha (the violation tolerance) with probability at least 1 - delta (the confidence level). On two benchmarks covering math reasoning (GSM8K) and multi-domain knowledge (MMLU), routing between Mixtral-8x7B and GPT-4 (a 24.5x cost difference), our method maintains the target alpha within the delta tolerance across a sweep from 0.05 to 0.50, while a validation-tuned baseline crosses the violation boundary on GSM8K. A feasibility analysis across all 10 RouterBench models reveals that routability is jointly model- and task-dependent. To our knowledge, this is the first input-based LLM router with distribution-free safety guarantees.

[SRW] Constructing a Japanese Rap Lyric Generation Model with GRPO ▶ *Underline Video*

Hayato Ogawa and Daisuke Kawahara

Rap is a vocal style rooted in Hip-Hop culture, characterized by producing rhymes in synchrony with a rhythmic beat. This paper proposes a method for generating Japanese rap lyrics with a large language model (LLM) whose rhyming behavior is improved via reinforcement learning. We design a reward function that evaluates end rhymes between two generated bars and apply GRPO, a reinforcement-learning method, to encourage Japanese rhyming without using existing Japanese rap lyrics as training data. Experimental results show that, although output collapse is observed in some cases, GRPO increases the proportion of outputs that receive moderate or high human ratings on rhyme-related criteria.

[SRW] Constructing a Japanese Verdict Prediction Dataset for Fact-Checking of LLM-Generated Texts ▶ *Underline Video*

Miwa Masano, Hirokazu Kiyomaru, Atsushi Keyaki, Kaito Horio, Rei Minamoto, Ribeka Keyaki, Kouta Nakayama, Hideyuki Tachibana and Daisuke Kawahara;

The development of fact-checking systems for verifying the factuality of text generated by large language models (LLMs) has been advancing. In the verdict prediction step of such systems, the system determines whether claims in the generated text are supported by retrieved evidence, formulated as a natural language inference (NLI) task. This study extends the label set for verdict prediction to capture claim-evidence relationships that humans would commonly interpret as supported or refuted, even in the absence of strict logical entailment or contradiction. It also constructs a Japanese dataset comprising 28,147 instances from two sources based on this extended label set. We analyze the causes of annotation disagreement and find that ambiguity in the boundary of acceptable inference, interpretive characteristics of negative cases, and incomplete information in the evidence affect annotation variability. Using this dataset, we evaluate the performance of prompt-based verdict prediction methods and show that prompts that explicitly elicit chain-of-thought reasoning improve F1 by 4 percentage points compared to baseline.

[Findings] Context as a Tool: Context Management for Long-Horizon SWE-Agents ▶ *Underline Video*

Shukai Liu, Bo Jiang, Jian Yang, Yizhi LI, Jinyang Guo, Xianglong Liu and Bryan Dai

Agents based on large language models have recently shown strong potential on real-world software engineering (SWE) tasks that require long-horizon interaction with repository-scale codebases. However, most existing agents rely on append-only context maintenance or passively triggered compression heuristics, which often lead to context explosion, semantic drift, and degraded reasoning in long-running interactions. We propose Cat, a new context management paradigm that elevates context maintenance to a callable tool integrated into the decision-making process of agents. Cat formalizes a structured context workspace consisting of stable task semantics, condensed long-term memory, and high-fidelity short-term interactions, and enables agents to proactively compress historical trajectories into actionable summaries at appropriate milestones. To support context management for SWE-agents, we propose a trajectory-level supervision

framework, CaT-Generator, based on an offline data construction pipeline that injects context-management actions into complete interaction trajectories. Using this framework, we train a context-aware model, SWE-Compressor. Experiments on SWE-Bench-Verified demonstrate that SWE-Compressor reaches a 57.6% solved rate and significantly outperforms ReAct-based agents and static compression baselines, while maintaining stable and scalable long-horizon reasoning under a bounded context budget.

[Findings] ContextCheck: Sentence-Level Faithfulness Verification with Context-Aware Disambiguation ▶ *Underline Video*

Yueqin Yin, Yaxi Li, Xin Liu, Xun Wang, Kaiqiang Song, Simin Ma, Shujian Liu, Sathish Reddy Indurthi, Haoyun Deng, Pengcheng He, Mingyuan Zhou and Song Wang

Large language models often hallucinate, producing content that is factually incorrect or not grounded in the sources. Reliable faithfulness verification is critical for trustworthy deployment. In the provided-source (closed-world) setting, existing verifiers either classify whole passages in one step or check sentences independently, overlooking cross-sentence context. We present ContextCheck, a framework for sentence-level faithfulness verification with context-aware disambiguation. Each sentence is verified against the grounding document while conditioning on preceding sentences, enabling pronouns and references to be resolved directly in context. This design avoids the separate decontextualization step of rewriting claims into self-contained forms, casting verification as a context-conditioned task. Fine-tuned from Llama-3.1-8B-Instruct, ContextCheck sets a new state of the art on three context-dependent datasets; it improves Macro F1 by over 10 points compared to the strongest baselines, and matches or slightly surpasses the strongest baselines on 14 standard single-sentence datasets compared to prior 8B-scale verifiers (average Macro F1 73.5 vs. 72.8). These results show that ContextCheck offers a practical and effective approach for sentence-level hallucination detection.

[SRW] Contextual Diversity Measure (CDM) for Controllable Story Generation in Large Language Models ▶ *Underline Video*

Richard Susilo, Hanna Suominen and Patrik Haslum;

Scenario-based text generation has broad applications across education and creative writing, but remains underexplored in controllable text generation. We introduce the Contextual Diversity Measure (CDM), a metric that quantifies semantic diversity for scenario generation given an abstract semantic role labeling constraints, and validate it through controlled experiments. Statistical analysis across four embedding models demonstrates that CDM successfully distinguishes between high-diversity and low-diversity text pairs, with all tests achieving significance at $p < 0.05$ on both subsets of the dataset, with effect sizes from small-to-medium (Cohen's d from 0.292 to 0.508) and medium-to-large (d: 0.677–1.195). Baseline comparisons indicate that CDM achieves excellent discrimination accuracy (100% and 91.9% for the two subsets), with discriminative power up to 5.5x greater than the best~baseline.

[SRW] Convergent Demographic Utility Hierarchies: Geometry of Intersectional Values in LLMs ▶ *Underline Video*

author_pravish_sainath1 author_pravish_sainath1

Recent work has shown that LLMs develop internally coherent utility functions that emerge with scale, yet whether these value systems encode systematic demographic hierarchies remains unexplored. We elicit pairwise preferences across 15 intersectional demographic groups (defined by race, gender, and their combinations) and 8 policy domains on three 7–8B instruction-tuned LLMs, fitting Thurstonian utility models to the resulting preference matrices. All three models converge on a compensatory hierarchy that inverts real-world structural advantage, consistently ranking marginalized groups, the highest and dominant groups are lowest. Intersectional utilities do not combine additively: single-axis audits that measure gender and race gaps independently overestimate the most extreme intersectional gap by 26-40% in our experiments. Geometrically, we identify a linear direction in the representation space that predicts the full utility hierarchy from neutral sentences alone, and show that this direction is substantially aligned with gender encoding but not with race encoding. Orthogonalization reveals that gender separation in representations is not fully explained by utility encoding. The hierarchy is already present in base (pre-alignment) models and is amplified several-fold by instruction tuning, suggesting it originates in pre-training data rather than alignment procedures.

[Findings] CoopQ: Cooperative Game Inspired Layerwise Mixed Precision Quantization for LLMs ▶ *Underline Video*

Junchen Zhao, Ali Derakhshan, Jayden Kana Hyman, Junhao Dong, Sangeetha Abdu Jyothi and Ian Harris

Large Language Models (LLMs) promise impressive capabilities, yet their multi-billion parameter scale makes on-device or low-resource deployment prohibitive. Mixed precision quantization offers a compelling solution, but existing methods struggle when the average precision drops below four bits, as they rely on isolated, layer-specific metrics that overlook critical inter-layer interactions affecting overall performance. To address these limitations, we first frame the mixed-precision quantization problem as a cooperative game among layers and introduce Shapley-based Progressive Quantization Estimation (SPQE) to efficiently obtain accurate Shapley estimates of layer sensitivities and inter-layer interactions. Leveraging the SPQE estimates, we propose Cooperative Game Inspired Mixed-Precision Quantization (CoopQ) which translates these Shapley estimates into a binary quadratic optimization formulation, assigning either 2 or 4-bit precision to layers under strict memory constraints. Comprehensive experiments conducted on Llama-3, Gemma-2, and Qwen models across three independent PTQ backends (Quanto, HQQ, GPTQ) demonstrate CoopQs scalability and consistently superior performance compared to methods relying solely on isolated metrics. Across average precisions spanning 4 bit down to 2 bit, CoopQ cuts Perplexity by 20 - 80% relative to the best baseline, with the margin growing as the bit-width tightens.

[Findings] CORAL: Adaptive Retrieval Loop for Culturally-Aligned Multilingual RAG ▶ *Underline Video*

Nayeon Lee, Jiwoo Song and Byeongcheol Kang

Multilingual retrieval-augmented generation (mRAG) is often implemented within a fixed retrieval space, typically via query or document translation or multilingual embedding vector representations. However, this approach may be inadequate for culturally grounded queries, in which retrieval-condition misalignment may occur. Even strong retrievers and generators may struggle to produce culturally relevant answers when sourcing evidence from inappropriate linguistic or regional contexts. To this end, we introduce CORAL (COntext-aware Retrieval with Agentic Loop, an adaptive retrieval methodology for mRAG that enables iterative refinement of both the retrieval space (corpora) and the retrieval probe (query) based on the quality of the evidence. The overall process includes: (1) selecting corpora, (2) retrieving documents, (3) critiquing evidence for relevance and cultural alignment, and (4) checking sufficiency. If the retrieved documents are insufficient to answer the query correctly, the system (5) reselects corpora and rewrites the query. Across two cultural QA benchmarks, CORAL achieves up to a 3.58%p accuracy improvement on low-resource languages relative to the strongest baselines.

[Long] CuBridge: An LLM-Based Framework for Understanding and Reconstructing High-Performance Attention Kernels ▶ ***Underline Video***

Xing Ma, Yangjie Zhou, Wu Sun, Zihan Liu, Jingwen Leng, Yun Lin, Shixuan Sun, Minyi Guo and Jin Song Dong

Efficient CUDA implementations of attention mechanisms are critical to modern deep learning systems, yet supporting diverse and evolving attention variants remains challenging. Existing frameworks and compilers trade performance for flexibility, while expert-written kernels achieve high efficiency but are difficult to adapt. Recent work explores large language models (LLMs) for GPU kernel generation, but prior studies report unstable correctness and significant performance gaps for complex operators such as attention. We present CuBridge, an LLM-based framework that adapts expert-written attention kernels through a structured lift-transfer-lower workflow. CuBridge starts from expert-written CUDA attention kernels and lifts them into an executable intermediate representation that makes execution orchestration explicit while abstracting low-level CUDA syntax. Given a user-provided PyTorch specification, CuBridge generates and verifies a target IR program, then reconstructs optimized CUDA code via reference-guided lowering. Across diverse attention variants and GPU platforms, CuBridge consistently produces correct kernels and substantially outperforms general frameworks, compiler-based approaches, and prior LLM-based methods.

[SRW] Cultural Value Alignment Via Latent Activation Steering in Large Language Models ▶ ***Underline Video***

Trung_Duc_Anh Dang and Sarah Masud

Large Language Models (LLMs) often exhibit homogenized cultural perspectives. While the World Values Survey (WVS) provides a gold standard for mapping human values, traditional direct prompting of LLMs on WVS often fails to access the model's latent cultural depth, leading to safety-aligned refusals or neutral responses. Here, we propose a generalizable framework for cultural evaluation and intervention that transitions from abstract queries to scenario-based behavioral probing. By extracting implicit token probabilities across 300 situational dilemmas, we bypass surface-level alignment to map the latent coordinates of LLMs cultural value. We further introduce activation steering to shift these internal alignments during the forward pass without retraining. Our results reveal that cultural adaptability is non-uniform. While Llama exhibits structural rigidity, Qwen and Gemma show high but volatile sensitivity across social domains such as family and the legal system. Finally, we quantify latent entanglement, demonstrating that models encode cultural axes as coupled "modernization" dimensions. This work establishes a computationally efficient framework for cultural steering, highlighting the structural complexities when navigating global value with LLMs.

[Long] Data Pollination: An Emergent Ecological Process Driving AI Population Evolution ▶ ***Underline Video***

Shufang Xie, Qizhi Pei, Ang Lv, Jingyang Hu, Lijun Wu and Rui Yan

AI development is often framed as the outcome of isolated research and engineering efforts, yet evidence from deployed systems suggests that language models interact through a shared data ecosystem. While the optimization of individual models is extensively studied, the emergent properties of this interconnected population remain largely unexplored, limiting our ability to predict long-term ecosystem trajectories We term this process data pollination, the unintentional circulation of synthetic model outputs through shared online platforms and web-scale training corpora, and formalize it as a population-based evolutionary framework to investigate stability dynamics under synthetic data training. Our theoretical analysis and controlled experiments involving 320 language models demonstrate that population dynamics can mitigate the model collapse observed in single-lineage recursive training, yielding stable or improving performance across diverse benchmarks. Crucially, we find that ecological diversity functions as a fundamental resilience mechanism that safeguards the ecosystem against collapse, highlighting the critical importance of maintaining model diversity for sustainable AI development.

[Findings] Deciphering Cultural Representations in Large Language Models via Sparse Autoencoders ▶ ***Underline Video***

Chenye Zou, Difan Jiao and Lijie Hu

Large Language Models (LLMs) are increasingly deployed worldwide, yet they exhibit strong Western-centric biases, and the internal mechanisms governing their cultural behaviors remain poorly understood. Prior work has identified so-called cultural neurons, but individual neurons are often polysemous, conflating abstract cultural knowledge with surface-level lexical cues due to superposition. We apply Sparse Autoencoders (SAEs) to decompose intermediate LLM activations into sparse, interpretable feature representations that disentangle these factors. This analysis reveals culturally selective features that remain invariant across paraphrasing and task formats, indicating abstraction beyond lexical correlations. Through targeted feature ablation, we provide causal evidence that these features are necessary for cultural reasoning: their removal selectively degrades performance on culturally conditioned tasks. Furthermore, we show that steering model activations along these feature directions is sufficient to systematically modulate cultural-related knowledge generation, without

retraining. Together, our results offer the first causal evidence that LLMs encode cultural knowledge as decoupled semantic structures rather than surface patterns, enabling a scalable pathway toward cultural alignment through mechanistic intervention. Code is available at https://github.com/IAN-YE/Cultural-features-SAE.

[Short] Deep Kernel Fusion for Transformers ▶ *Underline Video*

Zixi Zhang, Zhiwen Mo, Yiren Zhao and Robert Mullins

Agentic LLM inference with long contexts is increasingly limited by memory bandwidth rather than compute. In this setting, SwiGLU MLP blocks, whose large weights exceed cache capacity, become a major yet under-optimized bottleneck in the Transformer architecture. We propose DeepFusionKernel, a deeply fused kernel that cuts HBM traffic and boosts cache reuse, delivering up to 13.2% speedup on H100 and 9.7% on A100 over SGLang. Integrated with SGLang and paired with a kernel scheduler, DeepFusionKernel ensures consistent accelerations across generation lengths, while remaining adaptable to diverse models, inference configurations, and hardware platforms.

[Long] Deep Research with Open-Domain Evaluation and Multi-Stage Guardrails for Safety ▶ *Underline Video*

Wei-Chieh Huang, Henry Peng Zou, Yaozu Wu, Dongyuan Li, Yankai Chen, Weizhi Zhang, Yangning Li, Angelo Zangari, Jizhou Guo, Chunyu Miao, Liancheng Fang, Langzhou He, Yinghui Li, Renhe Jiang and Philip S. Yu

Deep research frameworks have shown promising capabilities in synthesizing comprehensive reports from web sources. While deep research possesses significant potential to address complex issues through planning and research cycles, existing frameworks are deficient in sufficient evaluation procedures and stage-specific protections. They typically treat evaluation as exact match accuracy of question-answering, but overlook crucial aspects of report quality such as credibility, coherence, breadth, depth, and safety. This oversight may result in hazardous or malicious sources being integrated into the final report. To address this, we introduce DeepResearchGuard, a framework featuring four-stage safeguards with open-domain evaluation, and DRSafeBench, a novel stage-wise safety benchmark. Evaluating across GPT-4o, o4-mini, Gemini-2.5-flash, DeepSeek-v3, and GPT-5, DeepResearchGuard improves defense success rates by an absolute 16.53% while reducing over-refusal rates to approximately 6%. Through extensive experiments, we show that DeepResearchGuard enables comprehensive open-domain evaluation and stage-aware defenses that effectively block harmful content propagation, while systematically improving report quality without excessive over-refusal rates.

[Findings] DeepPrune: Parallel Scaling without Inter-trace Redundancy ▶ *Underline Video*

Shangqing Tu, Yaxuan Li, Yushi Bai, Lei Hou and Juanzi Li

Parallel scaling has emerged as a powerful paradigm to enhance reasoning capabilities in large language models (LLMs) by generating multiple Chain-of-Thought (CoT) traces simultaneously. However, this approach introduces significant computational inefficiency due to inter-trace redundancy—our analysis reveals that over 80% of parallel reasoning traces yield identical final answers, representing substantial wasted computation. To address this critical efficiency bottleneck, we propose DeepPrune, a novel framework that enables efficient parallel scaling through dynamic pruning. Our method features a specialized judge model trained with oversampling techniques to accurately predict answer equivalence from partial reasoning traces, achieving 0.7072 AUROC on equivalence prediction across unseen reasoning models. This is combined with an online greedy clustering algorithm that dynamically prunes redundant paths while preserving answer diversity. Comprehensive evaluations across three challenging benchmarks (AIME 2024, AIME 2025, and GPQA) and multiple reasoning models demonstrate that DeepPrune achieves remarkable token reduction ranging from 65.73% to 88.50% compared to conventional consensus sampling, while maintaining competitive accuracy within 3.4 percentage points. Our work establishes a new standard for efficient parallel reasoning, making high-performance reasoning more efficient. Our code and data are here: https://github.com/THU-KEG/DeepPrune/

[Long] Defenses Against Prompt Attacks Learn Surface Heuristics ▶ *Underline Video*

Li Li, Chenxiao Yu, Zhiyu Ni, Hao Li, Charith Peris, Chaowei Xiao and Yue Zhao

Large language models (LLMs) are increasingly deployed in security-sensitive applications, where they must follow system- or developer-specified instructions that define the intended task behavior, while completing benign user requests. When adversarial instructions appear in user queries or externally retrieved content, models may override intended logic. Recent defenses rely on supervised fine-tuning with benign and malicious labels. Although these methods achieve high attack rejection rates, we find that they rely on narrow correlations in defense data rather than harmful intent, leading to systematic rejection of safe inputs. We analyze three recurring shortcut behaviors induced by defense fine-tuning. Position bias arises when benign content placed later in a prompt is rejected at much higher rates; across reasoning benchmarks, suffix-task rejection rises from below 10% to as high as 90%. Token trigger bias occurs when strings common in attack data raise rejection probability even in benign contexts; inserting a single trigger token increases false refusals by up to 50%. Topic generalization bias reflects poor generalization beyond the defense data distribution, with defended models suffering test-time accuracy drops of up to 40%. These findings suggest that current prompt-injection defenses frequently respond to attack-like surface patterns rather than the underlying intent. We introduce controlled diagnostic datasets and a systematic evaluation across two base models and multiple defense pipelines, highlighting limitations of supervised fine-tuning for reliable LLM security.

[Findings] DELTA: Dynamic Layer-Aware Token Attention for Efficient Long-Context Reasoning ▶ *Underline Video*

Hossein Entezari Zarch, Lei Gao, Chaoyi Jiang and Murali Annavaram

Large reasoning models (LRMs) achieve state-of-the-art performance on challenging benchmarks by generating long chains of intermediate steps, but their inference cost is dominated by decoding, where each new token must attend to the entire growing sequence. One approach

to reduce this latency is to evict entries from the key-value (KV) cache, thereby reducing the active context used in attention computation. However, such sparse attention methods suffer from severe accuracy degradation on reasoning tasks due to cumulative selection errors and the evolving importance of tokens over long derivations. We present DELTA, a training-free sparse attention mechanism that improves computational efficiency without sacrificing model accuracy. DELTA partitions transformer layers into three groups: initial layers that use full attention, a small set of Delta-layers that identify salient tokens via aggregated head-level attention scores, and subsequent sparse-attention layers that attend only to the selected subset. This design preserves the full KV cache in GPU memory for accuracy, while avoiding expensive full-attention computation over many layers. On reasoning benchmarks such as AIME and GPQA-Diamond, DELTA matches or surpasses full attention in accuracy, while reducing the number of attended tokens by up to 4.25x and delivering 1.54x end-to-end speedup. Our results show that selective reuse of intermediate attention maps offers a robust path toward efficient long-context reasoning. The code is available at https://github.com/hoenza/DELTA.

[SRW] Detecting Hallucinations in Large Language Models via Internal Attention Divergence Signals ▶ *Underline Video*
Gijs van Dijk
We propose a lightweight and single-pass uncertainty quantification method for detecting hallucinations in Large Language Models. The method uses attention matrices to estimate uncertainty without requiring repeated sampling or external models. Specifically, we measure the Kullback-Leibler divergence between each attention heads distribution and a uniform reference distribution, and use these features in a logistic regression probe. Across multiple datasets, task types, and model families, attention divergence is strongly predictive of answer correctness and performs competitively with existing uncertainty estimation methods. We find that this signal is concentrated in middle layers and on factual tokens such as named entities and numbers, suggesting that attention dynamics provides an efficient and interpretable white-box signal of model uncertainty.

[Findings] Differentially Private Synthetic Text Generation for Retrieval-Augmented Generation (RAG) ▶ *Underline Video*
Junki Mori, Kazuya Kakizaki, Taiki Miyagawa and Jun Sakuma
Retrieval-Augmented Generation (RAG) enhances large language models (LLMs) by grounding them in external knowledge. However, its application in sensitive domains is limited by privacy risks. Existing private RAG methods typically rely on query-time differential privacy (DP), which requires repeated noise injection and leads to accumulated privacy loss. To address this issue, we propose DP-SynRAG, a framework that uses LLMs to generate differentially private synthetic RAG databases. Unlike prior methods, the synthetic text can be reused once created, thereby avoiding repeated noise injection and additional privacy costs. To preserve essential information for downstream RAG tasks, DP-SynRAG extends private prediction, which instructs LLMs to generate text that mimics subsampled database records in a DP manner. Experiments show that DP-SynRAG achieves superior performance to the state-of-the-art private RAG systems while maintaining a fixed privacy budget, offering a scalable solution for privacy-preserving RAG.

[Findings] DiSec: Mitigating Backdoors in Pre-trained Language Models via Disentanglement of Adversarial Weights for Secure Fine-Tuning ▶ *Underline Video*
Sunanda Das and Qinghua Li
Task-agnostic backdoor attacks can contaminate pre-trained language models (PLMs) in a way that survives downstream adaptation, even under full fine-tuning, making it difficult for practitioners to trust third-party checkpoints. Existing defenses often rely on privileged assumptions (e.g., access to poisoned data or trigger/target knowledge), thereby limiting their applicability in realistic settings. We present DiSec, a robust and label-efficient purification framework that uses only clean auxiliary text and does not rely on downstream supervision or attack signatures. DiSec elicits model-internal signals from this clean data to separate suspicious parameter components that are inconsistent with benign behavior, and then flags anomalous structures by jointly leveraging complementary spectral and generative views of outliers. Finally, DiSec performs a structure-preserving repair via layer-local prototype-based mean correction, yielding an idempotent update that depends only on non-adversarial statistics. Across diverse downstream classification tasks and PLM backdoor strategies, DiSec substantially suppresses attack success while preserving clean-task utility, offering a practical path to securing fully fine-tuned PLMs before deployment. The codes are publicly available at https://github.com/das-sunanda/DiSec.

[SRW] Disentangling Meaning and Language Components in Diverse Multilingual Sentence Embeddings ▶ *Underline Video*
Kanade Nonomura; Keita Fukushima; Risa Kondo; Tomoyuki Kajiwara
We disentangle multilingual sentence embeddings into language-dependent and language-agnostic components, leveraging the latter to improve cross-lingual similarity estimation. Previous studies focused on encoder-based approaches that use only the input sentence; in contrast, this study examines the effectiveness of disentanglement methods across a broader range of sentence embeddings, including decoder-based approaches and those that utilize prompts. Experimental results demonstrate that embedding disentanglement is effective for a wide variety of sentence embeddings.

[SRW] Dissociating Circuit-Level and Distribution-Level Effects of Knowledge Conflicts in LLMs ▶ *Underline Video*
author_pravish_sainath1 author_pravish_sainath1
When large language models encounter conflicts between parametric knowledge and contextual evidence, does the parametric circuit's contribution change, or is the conflict resolved at the output distribution? We apply a three-level analysis framework, decomposing conflict resolution into circuit-level (parametric-site logit contribution, _P), pathway-level (MLP vs.\{ } attention), and output-level (token rank)

effects across four transformer architectures spanning three attention types and two scale points. We find that _P is near-invariant under conflict (<=1.8% variation), while output token ranks shift substantially, establishing that conflicts resolve through additive logit-space competition. Activation patching confirms parametric sites are causally sufficient and necessary, with contextual heads exerting 30–40x weaker influence. Layer-by-layer analysis reveals three architecturally determined resolution mechanisms, and parametric–contextual coupling decreases substantially with scale, exceeding random-subspace predictions. Domain generalization shows parametric sites are stratified by knowledge complexity with no overlap, yet the resolution mechanism generalizes across domains.

[Long] Do LLMs Know Tool Irrelevance? Demystifying Structural Alignment Bias in Tool Invocations ▶ *Underline Video*

Yilong Liu, Xixun Lin, Pengfei Cao, Ge Zhang, Fang Fang and Yanan Cao

Large language models (LLMs) have demonstrated impressive capabilities in utilizing external tools. In practice, however, LLMs are often exposed to tools that are irrelevant to the user's query, in which case the desired behavior is to refrain from invocations. In this work, we identify a widespread yet overlooked mechanistic flaw in tool refusal, which we term structural alignment bias: Even when a tool fails to serve the user's goal, LLMs still tend to invoke it whenever query attributes can be validly assigned to tool parameters. To systematically study this bias, we introduce SABEval, a new dataset that decouples structural alignment from semantic relevance. Our analysis shows that structural alignment bias induces severe tool-invocation errors in LLMs, yet remains largely unaccounted for in existing evaluations. To investigate the internal mechanisms underlying this bias, we propose Contrastive Attention Attribution, which reveals two competing pathways for semantic checking and structural matching. The relative strength of these pathways drives LLMs' tool invocation decisions. Based on these findings, we further introduce a rebalancing strategy that effectively mitigates structural alignment bias, as demonstrated by extensive experiments, without degrading general tool-use capabilities.

[SRW] Does Topic Sentiment Cause Perceived Ideology? Comparing Human and LLM Annotations in Political News Articles ▶ *Underline Video*

Upasana Chatterjee

We ask whether topic sentiment has a causal effect on perceived political ideology, and whether the answer depends on who assigns the ideology label. Using the AllSidesV2 test dataset (N=11,516 articles), we hold sentiment annotations constant and compare ideology labels from expert human annotators, GPT-4o-mini (baseline and finetuned), and Llama-3.3-70B. We apply Double Machine Learning (DML) and community-level mediation analysis across all four annotation paradigms. Human annotations yield no significant causal effects at the community level. Fine-tuned GPT-4o-mini achieves the highest classification accuracy (F1=72.48) and is the only annotator paradigm that produces significant community-level treatment effects and significant natural direct effects (NDEs) in mediation. We interpret this as evidence of shortcut learning: fine-tuning on ideology-labeled data causes the model to internalise a spurious sentiment–ideology coupling not operative in human judgment. This coupling is structurally invisible to F1-based evaluation, with implications for the use of LLM annotations as silver labels and as proxies for human judgment in downstream causal analyses.

[Findings] DRIFT: Transferring Reasoning Priors for Efficient MLLM Fine-Tuning ▶ *Underline Video*

Chao Huang, Zeliang Zhang, Jiang Liu, Ximeng Sun, Jialian Wu, Xiaodong Yu, Ze Wang, Chenliang Xu, Emad Barsoum and Zicheng Liu

Multimodal large language models (MLLMs) have made rapid progress, yet their reasoning ability often lags behind strong text-only LLMs. Bridging this gap typically requires large-scale multimodal reasoning data or reinforcement learning, incurring substantial cost. An appealing alternative is parameter-space model merging between reasoning-enhanced LLMs and MLLMs, but we show that naive merging is fragile: its effectiveness varies widely across model families and can significantly degrade performance (e.g., for Qwen-based MLLMs). We propose Directional Reasoning Injection for Fine-Tuning (DRIFT), a lightweight method that transfers reasoning knowledge in the gradient space while preserving multimodal alignment. DRIFT precomputes a reasoning prior from the parameter differences between text-only reasoning experts and multimodal models, and uses it to bias gradients during supervised fine-tuning. This design retains the simplicity of standard SFT pipelines while enabling efficient and stable reasoning transfer. Experiments on multimodal reasoning benchmarks, including MathVista and MathVerse, show that DRIFT consistently outperforms naive merging and standard SFT, and matches or surpasses training-intensive methods with substantially lower data and compute.

[Findings] DRP: Distilled Reasoning Pruning with Mathematical Skill-aware Step Decomposition for Efficient Large Reasoning Models ▶ *Underline Video*

Yuxuan Jiang, Dawei Li and Frank Ferraro

While Large Reasoning Models (LRMs) excel at complex tasks via long Chain-of-Thought (CoT) reasoning, their outputs are often excessively verbose, leading to inefficiency. This problem is amplified when the students long-form reasoning mismatches the concise outputs of smaller teacher models, common in LLM distillation to avoid using costly large teachers. To address this issue, we propose Distilled Reasoning Pruning (DRP), a hybrid framework that combines inference-time pruning with tuning-based distillation. DRP leverages a teacher model to perform mathematical problem-solving skill-aware step decomposition and pruning, then distills the refined reasoning paths into a student model, enabling efficient and accurate reasoning. Across challenging math datasets, DRP significantly reduces token usage without sacrificing accuracy, for instance, cutting tokens on GSM8K from 917 to 328 while improving accuracy from 91.7% to 94.1%, and reducing AIME tokens by 43% with no performance drop. Further analysis shows that aligning training CoT structure with the students capacity is key to effective knowledge transfer.

[SRW] Dynamic Meta-Metrics: Source-Sentence Conditioned Weighting for MT Evaluation ▶ *Underline Video*

Luke Zhang; Justin Vasselli; Aditya Khan; York Hay Ng; En-Shiun Annie Lee

We propose Dynamic Meta-Metrics (DMM), a framework for machine translation evaluation that learns source-sentence conditioned combinations of existing metrics. Rather than relying on a single static ensemble or language-specific weighting, DMM adapts the metric combination based on properties of the source segment. We study hard conditioning, which fits an interpretable combiner per cluster, and an exploratory soft-conditioned extension whose weights vary continuously with source-cluster responsibilities. We evaluate DMM on the WMT Metrics Shared Task data across multiple language pairs using pairwise agreement measures at the system and segment levels. Across settings, MLP-based combinations outperform linear and Gaussian process-based ensembles, and introducing soft conditioning yields gains over linear models.

[Findings] Efficient Hallucination Detection in Automatic Code Generation ▶ *Underline Video*

Georgii Andriushchenko, Roman Garaev, Lyudmila Rvanova, Artem Shelmanov and Vladimir V. Ivanov

Large language models (LLMs) frequently produce source code that seems correct and well-formed, yet includes hallucinated elements that cause downstream test failures. In this study, we benchmark state-of-the-art uncertainty quantification methods and existing baselines for the task of hallucination detection in source code and introduce a diff-based pipeline to construct a code dataset annotated with line-level hallucinations. Building on this, we train a lightweight Transformer-based detector that uses LLM internal representations to identify hallucinations, substantially outperforming existing methods across several code generation domains. The detector also shows particular promise for enabling self-correction in LLM-based coding agents. We release the first publicly available dataset of line-level code hallucinations, along with the corresponding source code and trained hallucination detectors https://github.com/datapaf/CodeHallucinationDetection

[Long] EfficientLLM: Unified Pruning-Aware Pretraining for Auto-Designed Compact Language Models ▶ *Underline Video*

Xingrun Xing, Zheng Liu, Shitao Xiao, Boyan Gao, Yiming Liang, Haokun Lin, Xianlin Zeng, Guoqi Li and Jiajun Zhang

Modern large language models (LLMs) driven by scaling laws achieve emergent intelligence in large model sizes. Recently, the increasing concerns about cloud costs, latency and privacy make it an urgent requirement to develop compact edge language models. Distinguished from direct pretraining that bounded by parameter scaling law, this work proposes the unified pruning-aware pretraining, focusing on pretraining compact models while preserving performance of much larger source models, termed EfficientLLM. It features following characteristics: 1) Pruning in Pretraining Corpus: we introduce minimal parameter groups to decouple LLMs and continuously optimize model architecture with classic pruning methods like LLM-Pruner and SparseGPT during pretraining. We reveal that it achieves top-quality compact language models to scale up LLM pruning to large scale pretraining. 2) Auto-Designed Architecture: the LLM architecture is auto-designed during saliency-driven pruning, unifying pretraining, architectural design, and parameter pruning into a single process. Based on these, EfficientLLM significantly outperforms directly pretrained baselines with 100M 1B parameters, such as MobileLLM, SmolLM, Qwen2.5-0.5B, OLMo-1B, Llama3.2-1B in commen sense benchmarks, which bridges the performance gap between traditional LLM compression and direct pretraining. We open source on https://github.com/Xingrun-Xing2/EfficientLLM.

[Long] EGSS: Entropy-guided Stepwise Scaling for Reliable Software Engineering ▶ *Underline Video*

Chenhui Mao, Yuanting Lei, Zhixiang Wei, Ming Liang, Zhixiang Wang, Jingxuan Xu, Dajun Chen, Wei Jiang and Yong Li

Agentic Test-Time Scaling (TTS) has delivered state-of-the-art (SOTA) performance on complex software engineering tasks such as code generation and bug fixing. However, its practical adoption remains limited due to significant computational overhead, primarily driven by two key challenges: (1) the high cost associated with deploying excessively large ensembles, and (2) the lack of a reliable mechanism for selecting the optimal candidate solution, ultimately constraining the performance gains that can be realized. To address these challenges, we propose Entropy-Guided Stepwise Scaling (EGSS), a novel TTS framework that dynamically balances efficiency and effectiveness through entropy-guided adaptive search and robust test-suite augmentation. Extensive experiments on SWE-Bench-Verified demonstrate that EGSS consistently boosts performance by 5-10% across all evaluated models. Specifically, it increases the resolved ratio of Kimi-K2-Intruct from 63.2% to 72.2%, and GLM-4.6 from 65.8% to 74.6%. Furthermore, when paired with GLM-4.6, EGSS achieves new state-of-the-art among open-source large language models. In addition to these accuracy improvements, EGSS reduces inference-time token usage by over 28% compared to existing TTS methods, achieving simultaneous gains in both effectiveness and computational efficiency.

[Findings] EmoMM: Benchmarking and Steering MLLM for Multimodal Emotion Recognition under Conflict and Missingness ▶ *Underline Video*

Yueru Sun, Yimeng Zhang, Haoyu Gu, Nuo Chen, Dong She, Xianrong Yao, Yang Gao and Zhanpeng Jin

Multimodal Emotion Recognition (MER) is critical for interpreting real-world interactions. While Multimodal Large Language Models (MLLM) have shown promise in MER, their internal decision-making mechanisms under modality conflict and missingness remain largely underexplored. In this paper, to systematically investigate these behaviors, we introduce EmoMM, a comprehensive benchmark featuring modality-aligned, conflict, and missing subsets. Through extensive evaluation, we uncover a Video Contribution Collapse (VCC) phenomenon, where MLLM marginalize video evidence due to high token redundancy and modality preferences. To address this, we propose Conflict-aware Head-level Attention Steering (CHASE), a lightweight mechanism that detects modality conflicts and performs inference-time attention steering, effectively mitigating decision bias without retraining the backbone. Experimental results demonstrate that CHASE consistently improves performance across various settings, significantly enhancing the reliability of MLLM in complex affective scenarios.

[Findings] Enhancing Factuality through Consensus and Consistency in Summarization Using Minimum Bayes Risk Decoding ▶ *Underline Video*

Riza Setiawan Soetedjo, Yusuke Sakai, Hidetaka Kamigaito, Jingun Kwon, Manabu Okumura and Taro Watanabe

Improving the quality of model-generated summaries, especially factuality, the accuracy of a summary with respect to its source content, remains a challenge. While reranking could select the optimal output from multiple generated candidates, it is limited to only using the source as guidance, resulting in unreliable summaries. To address this limitation, we propose ConSUM that reranks candidate summaries by considering two factors: consistency to the source document and consensus among the other candidates. Consensus is established using Minimum Bayes Risk (MBR) decoding over the set of generated summaries, while ensuring consistency by employing factuality-aware metrics that compare the summary against the source. Rigorous testing demonstrates that our system is competitive with existing methods, with human evaluations further confirming that its generated summaries are preferred over those from other systems.

[SRW] EnsemHalDet: Robust VLM Hallucination Detection via Ensemble of Internal State Detectors ▶ *Underline Video*

Ryuhei Miyazato; Shunsuke Kitada; Kei Harada

Vision-Language Models (VLMs) excel at multimodal tasks, but they remain vulnerable to hallucinations that are factually incorrect or ungrounded in the input image. Recent work suggests that hallucination detection using internal representations is more efficient and accurate than approaches that rely solely on model outputs. However, existing internal-representation-based methods typically rely on a single representation or detector, limiting their ability to capture diverse hallucination signals. In this paper, we propose EnsemHalDet, an ensemble-based hallucination detection framework that leverages multiple internal representations of VLMs, including attention outputs and hidden states. EnsemHalDet trains independent detectors for each representation and combines them through ensemble learning. Experimental results across multiple VQA datasets and VLMs show that EnsemHalDet consistently outperforms prior methods and single-detector models in terms of AUC. These results demonstrate that ensembling diverse internal signals significantly improves robustness in multimodal hallucination detection.

[Findings] EnsemW2S: Enhancing Weak-to-Strong Generalization with Large Language Model Ensembles ▶ *Underline Video*

Aakriti Agrawal, Mucong Ding, Chenghao Deng, Zora Che, Arjun Rajaram, Anirudh Satheesh, Bang An, C. Bayan Bruss, John Langford and Furong Huang

With Large Language Models (LLMs) rapidly approaching and potentially surpassing human-level performance, it has become imperative to develop approaches capable of effectively supervising and enhancing these powerful models using smaller, human-level models exposed to only human-level data. We address this critical weak-to-strong (W2S) generalization challenge by proposing a novel method aimed at improving weak experts, by training on the same limited human-level data, enabling them to generalize to complex, super-human-level tasks. Our approach, called EnsemW2S, employs a token-level ensemble strategy that iteratively combines multiple weak experts, systematically addressing the shortcomings identified in preceding iterations. By continuously refining these weak models, we significantly enhance their collective ability to supervise stronger student models. We extensively evaluate the generalization performance of both the ensemble of weak experts and the subsequent strong student model across in-distribution (ID) and out-of-distribution (OOD) datasets. For OOD, we specifically introduce question difficulty as an additional dimension for defining distributional shifts. Our empirical results demonstrate notable improvements, achieving 4%, and 3.2% improvements on ID datasets and, upto 6% and 2.28% on OOD datasets for experts and student models respectively, underscoring the effectiveness of our proposed method in advancing W2S generalization.

[Long] EQUIP: EQUivariant preserving In-Place updates for Efficient Token Pruning ▶ *Underline Video*

Arun Ramachandran, Govindarajan R, Murali Annavaram and Prakash Raghavendra

Token-pruning has emerged as a primary focus in large language models (LLMs) to enhance model efficiency while preserving accuracy, especially for large sequence lengths. However, the eviction operation of token-pruning methods causes holes in KV tensors, posing two major challenges: (1) The shift operation, required to make the KV tensor contiguous, results in significant copy overheads; (2) The changes in position indices due to token eviction lead to increased computational requirements for Rotary Positional Encoding (RoPE). To address these issues, we introduce EQUIP, an EQUivariant preserving in-place token update mechanism that ensures the equivariance property of the operations performed in the attention computation. EQUIP offers two fundamental advantages: First, it combines eviction and a subsequent token insertion into an in-place replacement operation, which reduces the KV cache copy overheads significantly. Second, EQUIP reduces recomputation of rotation operations through a combination of in-place update, caching and a re-indexing strategy. Together, these optimizations enable EQUIP to achieve geomean speedups of 1.62x (or 1.47x) on CPU (GPU) over StreamingLLM, and 3.45x (or 1.86x) on CPU (GPU) over Heavy Hitters (H2O). EQUIP with Paged Attention achieves speedups of 4.18x(2.61x) on CPU (GPU) over auto-regressive baselines. EQUIP matches the model accuracy of baseline pruning methods while delivering superior performance.

[Long] Every Response Counts: Quantifying Uncertainty of LLM-based Multi-Agent Systems through Tensor Decomposition ▶ *Underline Video*

Tiejin Chen, Huaiyuan Yao, Jia Chen, Evangelos E. Papalexakis and Hua Wei

While Large Language Model-based Multi-Agent Systems (MAS) consistently outperform single-agent systems on complex tasks, their intricate interactions introduce critical reliability challenges arising from communication dynamics and role dependencies. Existing Uncertainty Quantification methods, typically designed for single-turn outputs, fail to address the unique complexities of the MAS.

Specifically, these methods struggle with three distinct challenges: the cascading uncertainty in multi-step reasoning, the variability of inter-agent communication paths, and the diversity of communication topologies. To bridge this gap, we introduce MATU, a novel framework that quantifies uncertainty through tensor decomposition. MATU moves beyond analyzing final text outputs by representing entire reasoning trajectories as embedding matrices and organizing multiple execution runs into a higher-order tensor. By applying tensor decomposition, we disentangle and quantify distinct sources of uncertainty, offering a comprehensive reliability measure that is generalizable across different agent structures. We provide comprehensive experiments to show that MATU effectively estimates holistic and robust uncertainty across diverse tasks and communication topologies.

[SRW] Evolutionary Search for Automated Design of Uncertainty Quantification Methods ► *Underline Video*

Mikhail Seleznyov; Daniil Korbut; Viktor Moskvoretskii; Oleg Somov; Alexander Panchenko; Elena Tutubalina

Uncertainty quantification (UQ) methods for large language models are predominantly designed by hand based on domain knowledge and heuristics, limiting their scalability and generality. We apply LLM-powered evolutionary search to automatically discover unsupervised UQ methods represented as Python programs. On the task of atomic claim verification, our evolved methods outperform strong manually-designed baselines, achieving up to 6.7% relative ROC-AUC improvement across 9 datasets while generalizing robustly out-of-distribution. Qualitative analysis reveals that different LLMs employ qualitatively distinct evolutionary strategies: Claude models consistently design high-feature-count linear estimators, while Gpt-oss-120B gravitates toward simpler and more interpretable positional weighting schemes. Surprisingly, only Sonnet 4.5 and Opus 4.5 reliably leverage increased method complexity to improve performance~– Opus 4.6 shows an unexpected regression relative to its predecessor. Overall, our results hint that LLM-powered evolutionary search is a promising paradigm for automated, interpretable hallucination detector design.

[Findings] Exploring Layer-wise Information Effectiveness for Post-Training Quantization in Small Language Models ► *Underline Video*

He Xiao and Qingyao Yang

Large language models with billions of parameters are often over-provisioned: many layers contribute little unique information yet dominate the memory and energy footprint during inference. We present LieQ (Layer-wise information effectiveness Quantization), a hardware-native, metric-driven post-training quantization framework that addresses the critical challenge of maintaining accuracy in sub-8B models, model parameters less than 8B, under extreme low-bit compression. LieQ keeps uniform bit-width within each layer while mixing precision across layers, preserving standard multiplication kernels and avoiding irregular memory access, codebooks, or irregular formats at inference time. Our method uncovers a strong correlation between layer-wise functional saliency and representational compactness, revealing that layers with higher training-induced energy concentration are functionally irreplaceable. Leveraging this insight, we propose a purely geometry-driven sensitivity proxy that enables automatic bit-width allocation under a target average-bit budget without expensive gradient updates or inference-based perplexity probing. Under an average weight bit-width approaching two bits per parameter, LieQ consistently reduces the large accuracy gap typically observed for naive uniform 2-bit baselines on Qwen3 and LLaMA3.x families, while retaining standard-kernel efficiency. These properties make LieQ a practical path toward deploying small language models on resource-constrained edge devices. Code will be available at: https://github.com/HeXiao-55/LieQ-official.git.

[SRW] Eye Movement Features Can Predict Human Preferences on Machine-Generated Texts ► *Underline Video*

Xiaoshan He, Xiaoqun Liu, Haodong He, Yu Wang and Yang Xu

Eye-tracking metrics offer valuable insights into human visual attention during language comprehension, yet existing corpora lack diverse machine-generated text samples. To bridge this gap, we introduce Gaze Responses for Evaluating AI Texts (GREAT), a comprehensive dataset and software framework capturing human eye-movement patterns during screen reading of passages generated by large language models (LLMs). The dataset includes raw eye-movement recordings, reading-time measures, and post-reading evaluations for LLM-generated passage pairs from MT-Bench, alongside rigorous validation metrics. The collected eye-tracking metrics demonstrate strong explanatory power in predicting text quality. When integrated with negative log-likelihood (NLL), a commonly used metric for evaluating text quality, substantially enhance model performance across all standard statistical criteria. These findings demonstrate that eye-tracking data effectively complement probabilistic metrics, improving predictive accuracy for text quality assessment. The full dataset and some processing code are publicly available at https://anonymous.4open.science/r/eye-track.

[Long] FactCorrector: A Graph-Inspired Approach to Long-Form Factuality Correction of Large Language Models ► *Underline Video*

Javier Carnerero-Cano, Massimiliano Pronesti, Radu Marinescu, Tigran T. Tchrakian, James Barry, Jasmina Gajcin, Yufang Hou, Alessandra Pascale and Elizabeth M. Daly

Large language models (LLMs) are widely used in knowledge-intensive applications but often generate factually incorrect responses. A promising approach to rectify these flaws is correcting LLMs using feedback. Therefore, in this paper, we introduce FactCorrector, a new post-hoc correction method that adapts across domains without retraining and leverages structured feedback about the factuality of the original response to generate a correction. To support rigorous evaluations of factuality correction methods, we also develop the VELI5 benchmark, a novel dataset containing systematically injected factual errors and ground-truth corrections. Experiments on VELI5 and several popular long-form factuality datasets show that the FactCorrector approach significantly improves factual precision while preserving relevance, outperforming strong baselines. We release our code at https://ibm.biz/factcorrector.

[SRW] Factual State Discovery Benchmark: Evaluating Fact Elicitation in Polish Tax Law ► *Underline Video*

Mateusz Bystroski, Kamil Tagowski, Denis Janiak, Julia Farganus, Lukasz Augustyniak, Monika Kajdanowicz and Tomasz Jan Kajdanowicz

Before a tax authority can issue a ruling, it must receive a complete description of the taxpayer's situation—yet no benchmark measures whether language models can systematically elicit all relevant facts through dialogue. We introduce the Factual State Discovery Benchmark, in which a discovery agent questions a simulated taxpayer grounded in a real tax document. The dataset comprises 500 narratives from official Polish tax interpretations, decomposed into 32\{\},874 atomic facts with validated supported precision (97.6%), atomicity (93.8%), and sentence coverage (96.0%). Experiments with four models show that even the best system recovers only 77% of facts on easy samples and under 49% on hard samples after 50 turns. These findings establish conversational fact elicitation as a challenging open problem requiring retrieval-augmented and adaptive questioning strategies.

[SRW] Filling the Long Tail: Structure-Aware Curriculum-Gap Completion for Medical Education with LLMs ► *Underline Video*

Wenjie Lin

Medical education resources are dense for common diseases but often sparse for under-covered conditions, atypical presentations, and fine-grained concept distinctions. This creates curriculum gaps that are difficult to repair manually, especially in long-tail domains where structured teaching materials are limited. We introduce Curriculum-Gap Completion (CGC), a new task for Large Language Model (LLM)-based medical education in which a model reconstructs missing educational units from a partially specified curriculum graph. Given topic nodes, pedagogical relations, and structured teaching slots, the model predicts omitted concepts, restores missing instructional links, and completes automatically verifiable teaching content. We instantiate this setting in a long-tail medical case study (hyperhidrosis) and evaluate five LLMs under three methods: direct prompting, retrieval-augmented prompting, and our proposed Structure-Aware Curriculum-Gap Completion (SACGC) framework. Across models, SACGC achieves the strongest overall performance, with the largest gains on structurally demanding masking settings. Ablation results show that explicit graph structure is the most important component, while schema constraints provide additional benefit. These findings suggest that LLMs are better suited for reconstructing an under-specified educational structure than for unrestricted medical tutoring, and they motivate CGC as a new natural language processing (NLP) problem for healthcare education.

[SRW] Fine-Grained Semantic Comparison of Legal Documents using LLMs ► *Underline Video*

Elisei Rykov; Nikolay Ivanov; Maria Bandulevich; Kseniia Petrushina; Valentin Malykh; Vasily Konovalov; Alexander Panchenko; Ilseyar Alimova;

Frequent revisions of complex regulatory documents in large organizations often introduce inconsistencies and contradictions that are difficult for lawyers and auditors to detect manually. Existing tools rely on characterlevel diffs and therefore miss paraphrases and semantic shifts. We introduce LegDiff, a novel benchmark for evaluating spanaware semantic comparison of legal texts, and use it to investigate the ability of large language models to detect semantic changes beyond token and characterlevel matching. LegDiff comprises manually annotated pairs of legal paragraphs drawn from different documents. In addition, we present a pipeline to generate synthetic training data that aligns with the manual annotations and closely resembles realworld data, and a visualization tool for clearly displaying detected differences and inconsistencies. The dataset, code, and a visualization tool are publicly available to facilitate reproducibility and further research (https://anonymous.4open.science/r/LDC-5BF6/README.md).

[Long] From Off-Policy to On-Policy: Enhancing GUI Agents via Bi-level Expert-to-Policy Assimilation ► *Underline Video*

Zezhou Wang, Ziyun Zhang, Xiaoyi Zhang, Zhuzhong Qian and Yan Lu

Vision-language models are increasingly deployed as computer-use agents (CUAs) that operate desktops and browsers. Top-performing CUAs are framework-based systems that decompose planning and execution, while end-to-end screenshot-to-action policies are easier to deploy but lag behind on benchmarks such as OSWorld-Verified. GUI datasets like OSWorld pose two bottlenecks: they expose only a few hundred interactive, verifiable tasks and environments, and expert trajectories must be gathered by interacting with these environments, making such data hard to scale. We therefore ask how reinforcement learning from verifiable rewards (RLVR) can best exploit a small pool of exist expert trajectories to train end-to-end policies. Na"ively mixing these off-policy traces into on-policy RLVR is brittle: even after format conversion, expert trajectories exhibit structural mismatch and distribution shift from the learner. We propose BEPA (Bi-Level Expert-to-Policy Assimilation), which turns static expert traces into policy-aligned guidance via self-rolled reachable trajectories under the base policy (LEVEL-1) and a per-task, dynamically updated cache used in RLVR (LEVEL-2). On OSWorld-Verified, BEPA improves UITARS1.5-7B success from 22.87% to 32.13% and raises a held-out split from 5.74% to 10.30%, with consistent gains on MMBench-GUI and Online-Mind2Web. Our code and data are available at an anonymous repository: https://anonymous.4open.science/r/ACL_BEPA.

[Findings] From RAG to Agentic RAG for Faithful Islamic Question Answering ► *Underline Video*

Gagan Bhatia, Hamdy Mubarak, Mustafa Jarrar, George Mikros, Fadi zaraket, Mahmoud Alhirthani, Mutaz al-Khatib, Logan Cochrane, Kareem Mohamed Darwish, Rashid Yahiaoui and Firoj Alam

Large Language Models (LLMs) are increasingly used for Islamic question answering, where ungrounded responses may carry serious religious consequences. Yet standard MCQ/MRC-style evaluations do not capture key real-world failure modes, notably free-form hallucinations and the ability to abstain when evidence is insufficient. To address this gap, we introduce IslamicFaithQA, a 3,810-item bilingual (Arabic/English) generative benchmark with atomic single-gold answers, which enables direct measurement of hallucination and abstention. We additionally developed an end-to-end grounded Islamic modeling suite consisting of (i) 25K Arabic text-grounded

SFT reasoning pairs, (ii) 5K bilingual preference samples for reward-guided alignment, and (iii) a verse-level Qur'an retrieval corpus of 6k atomic verses (ayat). Building on these resources, we develop an agentic Quran-grounding framework (agentic RAG) that uses structured tool calls for iterative evidence seeking and answer revision. Experiments across Arabic-centric and multilingual LLMs show that retrieval improves correctness and that agentic RAG yields the largest gains beyond standard RAG, achieving state-of-the-art performance and stronger Arabic–English robustness even with a small model (i.e., Qwen3 4B). We made the datasets are publicly available (https://huggingface.co/datasets/QCRI/IslamicFaithQA).

[Findings] From Scoring to Explanations: Evaluating SHAP and LLM Rationales for Rubric-based Teaching Quality Assessment ▶ *Underline Video*

Ivo Bueno, Babette Bühler, Philipp Stark, Tim Fütterer, Ulrich Trautwein, Dorottya Demszky, Heather Hill and Enkelejda Kasneci

Automated scoring models are increasingly used to assign rubric-based quality ratings to complex language performances, including classroom transcripts, yet they typically provide little insight into why a particular score is produced. We propose a general framework for sentence-level interpretability of rubric-based scoring that combines model-agnostic Shapley-value attributions with rationales generated by large language models (LLMs). Instantiated on the Quality of Feedback dimension of the CLASS framework using the NCTE corpus, the framework enables systematic comparison of fine-tuned pretrained language models (PLMs) and prompted LLMs on both scoring performance and explanation faithfulness. Across 6k annotated transcript segments, fine-tuned PLMs outperform LLMs in prediction accuracy but exhibit label compression toward mid-scale scores. Deletion-based tests show that SHAP identifies sentences that reliably drive model predictions, producing typically larger and more coherent prediction shifts than LLM-generated rationales. Cross-model analyses further reveal that SHAP attributions transfer robustly across architectures, whereas LLM rationales exert limited and inconsistent influence. Overall, the findings demonstrate that SHAP provides more faithful and transferable explanations for rubric-based scoring, and that the proposed framework offers a principled basis for evaluating both scoring models and their explanations in high-stakes educational settings and other rubric-based language assessment tasks.

[Long] From Words to Pixels: A Comprehensive Survey on Large Language Models in Visual Segmentation ▶ *Underline Video*

Yizhou Wang, Mang Tik Chiu, Lingzhi Zhang, Xuan Shen, Sohrab Amirghodsi and Yun Fu

Visual segmentation, the task of segmenting an image into semantically meaningful regions, is a cornerstone in machine learning and has widespread applications in industry. Nevertheless, visual segmentation with instruction has been a challenging task for many years. This largely stems from the cross-modal discrepancy between language and image domains, resulting in difficulty in relating the instruction semantics and the pixel-level predictions. In recent years, the remarkable reasoning capabilities of Large Language Models (LLMs) and Large Multimodal Models (LMMs) have spurred a new wave of research aiming to bridge the disparity between natural language instructions and pixel-level understanding. This survey offers the first comprehensive overview of the rapidly evolving field of LLM-driven visual segmentation. We categorize existing approaches based on their core objectives and methodologies, including reasoning-based segmentation, open-vocabulary segmentation, grounding techniques connecting language to pixels, and extensions to video domains. We review recent seminal works in LLM-based visual segmentation, analyzing their architectural innovations, training strategies, and benchmark performance. Furthermore, we discuss the common datasets, evaluation metrics, and identify key challenges and promising future directions at the intersection of language and visual segmentation. We hope this survey serves as a valuable resource for researchers and practitioners seeking to understand the current landscape and future directions of leveraging LLMs for sophisticated visual segmentation tasks and applications. The resource summary is available at https://github.com/wyzjack/Awesome-LLM-Visual-Segmentation.

[SRW] Fusion Training for Mathematical Generalization in Large Language Models ▶ *Underline Video*

Congfeng Cao; Pengyu Zhang; Jelke Bloem

Thinking Mode Fusion (TMF) enables large language models to support both concise responses and long-form reasoning by unifying a non-thinking mode and a thinking mode within a single model. However, its training dynamics, including the data ratio and training schedule between the two modes, remain underexplored. In this work, we present a systematic study of TMF by analyzing the effects of the training schedule and data ratio between thinking and non-thinking modes. Focusing on mathematical problem solving, we construct a benchmark with multiple thinking-to-non-thinking data ratios and three training schedules. Our results reveal an asymmetric interaction between the two modes: increasing the ratio of non-thinking supervision reduces the accuracy of the thinking mode. We further show that different training schedules modulate this trade-off and that the optimal schedule depends on the data ratio. Finally, we quantify a negative correlation between non-thinking and thinking mode supervision, highlighting an inherent tension between these two modes. These findings provide practical guidance for designing effective TMF training settings. All code and data are released to support further research at: Fusion Bench.

[SRW] Geometry of Knowledge Allows Extending Diversity Boundaries of Large Language Models ▶ *Underline Video*

Mateusz Bystroski; Doheon Han; Nitesh V. Chawla; Tomasz Jan Kajdanowicz

Starting from the hypothesis that knowledge in semantic space is organized along structured manifolds, we argue that this geometric structure renders the space explorable. By traversing it and using the resulting continuous representations to condition an LLMs generation distribution, we can systematically expand the models reachable semantic range. We introduce a framework that requires no modification of LLM parameters and operationalizes this idea by constructing a conditioning distribution from a small set of diverse anchor generations. This distribution conditions LLM's generation via an xRAG-style projector. Our experiments demonstrate that this manifold-based conditioning substantially increases generative diversity, with direct benefits for enhancing divergent thinking, a core facet of creativity, in language models.

[Long] GRAD: Generalizing RAG Adaptation with Decoding ▶ *Underline Video*

Youngwon Lee, Seung-won Hwang, Zhewei Yao and Yuxiong He

Retrieval-augmented generation needs generation to follow retrieved evidence across shifting domains and prompt layouts, but training a new stronger model per task is costly. To this end, we propose GRAD, an adaptive decoding-time framework that keeps the base generator fixed and composes small, objective-specific guidance at inference. A key advantage of this design is enabling mix and match diverse RAG objectives: model scaling (MS), domain adaptation (DA) and positional debiasing (DB) can be integrated as token-level guidance terms, and new objectives can be easily plugged in. Across public benchmarks and private settings with no in-domain labels, GRAD improves accuracy with favorable latency, offering strong trade-offs versus scaling while reliably activating helpful objectives and suppressing harmful ones, adaptively to tasks.

[Findings] GrandGuard: Taxonomy, Benchmark, and Safeguards for Elderly-Chatbot Interaction Safety ▶ *Underline Video*

Changxuan Fan, Xi Yang, Yueyuan Zheng, Bin Zhou, Yuanping WANG, Wenbin Hu, Huihao JING, Ki Sen Hung, Dazhao Du, Haoran Li, Janet Hui-wen Hsiao and Yangqiu Song

As older adults increasingly use LLM-based chatbots for companionship and assistance, a safety gap is emerging. Older adults may face vulnerabilities from social isolation, limited digital literacy, and cognitive decline, yet existing safety benchmarks largely target general harms and overlook elderly-specific risks. For example, a prompt such as how to repair a ceiling light alone in the dark may be benign for most users but poses a serious fall risk for older adults with mobility limitations. We introduce GrandGuard, the first comprehensive framework for assessing and mitigating elderly-specific contextual risks in LLM interactions. We develop a three-level taxonomy with 50 fine-grained risk types across mental well-being, financial, medical, toxicity, and privacy domains, grounded in real-world incidents, community discussions, and analysis of stakeholder studies. Using this taxonomy, we construct a benchmark of 10,404 labeled prompts and responses, showing that several leading LLMs mishandle elderly-specific contextual risks in over 50% of cases. We mitigate these failures with two safeguards: a fine-tuned Llama-Guard-3 and a policy-enhanced gpt-oss-safeguard-20b, achieving up to 96.2% and 90.9% unsafe-prompt detection accuracy, respectively. GrandGuard lays the groundwork for AI systems that move beyond general safety to support aging populations.

[Findings] GRASP: Graph-Reasoning Aided Survey Planning for High-Fidelity Related Work Generation ▶ *Underline Video*

Haoming Li and Jessica Ouyang

Writing a literature review requires a deep understanding of the relationships among cited papers: how they build on, challenge, or offer alternative perspectives to one another. We present Graph-Reasoning Aided Survey Planning (GRASP), a framework combining LLM planning for related work generation with graph algorithms to extract key relationships among cited papers. Our two-layer graph structure consists of a Graph of Thoughts and an Argument-Counteragument Planning Network, representing the cited papers at different levels of granularity, and we apply topology-aware pruning via a Steiner tree to identify the core inter-paper relationships captured in our graph. Our citation analysis-based evaluation shows that GRASP generates RWS that closely match human-written targets in terms of the discourse roles, intents, and grouping of citations.

[SRW] HARP: Representation-Based Preference Learning for Perceptual Data ▶ *Underline Video*

Jordan Sinclair; Yousra Shleibik; Kerstin Haring

Learning reward functions from preference feedback is a widely used strategy for aligning intelligent systems with human intent, but its effectiveness heavily relies on the representation used to describe candidate outcomes. In perceptual domains such as visual aesthetics and robot appearance, hand-engineered features often fail to capture the latent attributes underlying subjective judgments. We propose HARP, a representation-based preference learning framework that infers reward functions directly over learned latent embeddings using pairwise comparisons and Bayesian updates, interpreting preference learning as identifying a preference-aligned direction in representation space. By operating in a continuous representation space, HARP enables preference learning without manual feature design. We evaluate HARP on a robot image preference task with simulated oracle users, comparing learned representations to principled hand-engineered features under identical query strategies. Results show more stable convergence and improved top-k preference recovery, highlighting the importance of representation choice in perceptual preference learning.

[Long] How Context Shapes Truth: Geometric Transformations of Statement-level Truth Representations in LLMs ▶ *Underline Video*

Shivam Adarsh, Maria Maistro and Christina Lioma

Large Language Models (LLMs) often encode whether a statement is true as a vector in their residual stream activations. These vectors, also known as truth vectors, have been studied in prior work, however how they change when context is introduced remains unexplored. We study this question by measuring (1) the directional change (theta) between the truth vectors with and without context and (2) the relative magnitude of the truth vectors upon adding context. Across four LLMs and four datasets, we find that (1) truth vectors are roughly orthogonal in early layers, converge in middle layers, and may stabilize or continue increasing in later layers; (2) adding context generally increases the truth vector magnitude, i.e., the separation between true and false representations in the activation space is amplified; (3) larger models distinguish relevant from irrelevant context mainly through directional change (theta), while smaller models show this distinction through magnitude differences. We also find that context conflicting with parametric knowledge produces larger geometric changes than parametrically aligned context. Collectively, these findings provide a geometric characterization of how context transforms the truth vector in the activation space of LLMs.

[SRW] How Hard is Math? Using Quantitative Metrics to Measure LLM Alignment to Human Intuitions of Difficulty ► *Underline Video*

Micah Helzerman; Steven R Wilson; Cam McLeman

Modern LLMs have demonstrated advanced reasoning skills, including the ability to solve Olympiad-level mathematics problems. While solving more and more difficult problems is a hallmark of LLM progress, less attention has been placed on how "difficulty" is operationalized in the context of LLM problem solving tasks. This is particularly relevant in educational contexts where teachers or students may ask LLMs for "easy" or "hard" questions. In this paper, we explore various quantitative measurements from LLM-generated solutions and evaluate their inter-correlations, as well as their correlation to human-annotated difficulty scores. We find moderate correlations between metrics using log probabilities and output lengths, including some that are more strongly correlated to difficulty than LLM accuracy. We also train ModernBERT to predict difficulty scores, leading to reasonable accuracy within a given benchmark, but decreased performance when generalizing to other math benchmarks. Finally, to explore connections between difficulty scores and human performance, we collect problems, human solutions, and human performance data from the Putnam competition. We find poor alignment between LLM metrics and human-assigned difficulty scores, despite strong correlations between those scores and human performance on the problems.

[Long] HowToNarrate: A General-Domain Benchmark for Synchronized Video Narration with External Knowledge ► *Underline Video*

Xueyan Wang, Dingyi Yang and Qin Jin

We present HowToNarrate, the first general-domain benchmark for Synchronized Video Narration. The benchmark contains 3.2K videos across seven domains, segmented into 37.5K clips with aligned narrations and associated external knowledge. Effective narration requires models to understand visual scenes, incorporate relevant knowledge, and produce coherent, length-appropriate descriptions. We systematically benchmark current Multimodal LLMs (MLLMs) on these abilities. Our analysis shows that existing MLLMs overemphasize knowledge retrieval while largely neglecting prior context (receiving less than 10% attention). Moreover, they often conflate narration context with external knowledge, leading to redundancy and incoherence. To mitigate these issues, we propose VideoNarrationAgent, a multi-agent framework that combines context compression, knowledge retrieval, and narration generation. Experiments demonstrate that our method significantly improves MLLM performance. Furthermore, instruction tuning on HowToNarrate enhances both context-awareness and length control, boosting Qwen2.5-VLs score from 25 to 84. We will release all data and code to support future research in synchronized video narration.

[Long] HybridKV: Hybrid KV Cache Compression for Efficient Multimodal Large Language Model Inference ► *Underline Video*

Bowen Zeng, Feiyang Ren, Jun Zhang, Xiaoling Gu, Ke Chen, Lidan Shou and Huan Li

Multimodal Large Language Models (MLLMs) have advanced unified reasoning over text, images, and videos, but their inference is hindered by the rapid growth of key-value (KV) caches. Each visual input expands into thousands of tokens, causing caches to scale linearly with context length and remain resident in GPU memory throughout decoding, which leads to prohibitive memory overhead and latency even on high-end GPUs. A common solution is to compress caches under a fixed allocated budget at different granularities: token-level uniformly discards less important tokens, layer-level varies retention across layers, and head-level redistributes budgets across heads. Yet these approaches stop at allocation and overlook the heterogeneous behaviors of attention heads that require distinct compression strategies. We propose HybridKV, a hybrid KV cache compression framework that integrates complementary strategies in three stages: heads are first classified into static or dynamic types using text-centric attention; then a top-down budget allocation scheme hierarchically assigns KV budgets; finally, static heads are compressed by text-prior pruning and dynamic heads by chunk-wise retrieval. Experiments on 11 multimodal benchmarks with Qwen2.5-VL-7B show that HybridKV reduces KV cache memory by up to 7.9x and achieves 1.52x faster decoding, with almost no performance drop or even higher relative to the full-cache MLLM.

[SRW] Identifying the Convergent Sycophancy Gap in Model Evaluations ► *Underline Video*

Jan Batzner, Volker Stocker, Stefan Schmid and Gjergji Kasneci

Sycophancy, the response pattern of language models to flatter and agree with users statements rather than correcting them, is widely studied but benchmarked inconsistently. To enable better interpretations of the contested concept of model sycophancy, we evaluate eight models on eight popular sycophancy benchmarks across five random seeds (32,000 evaluations). Cross-benchmark correlations average near zero, benchmarks fail to agree on which models are most sycophantic, and combining them yields negative internal consistency. Only 2 of 28 model-level benchmark pairs correlate significantly, one of which is negative. Yet, individual benchmarks show high test-retest reliability, confirming they measure their specific operational definition consistently, just not the same construct. We recommend multi-benchmark evaluation designs spanning at least three benchmark families before global interpretations about model sycophancy.

[Long] IF-RewardBench: Benchmarking Judge Models for Instruction-Following Evaluation ► *Underline Video*

Bosi Wen, Yilin Niu, Cunxiang Wang, Xiaoying Ling, Ying Zhang, Pei Ke, Hongning Wang and Minlie Huang

Instruction-following is a foundational capability of large language models (LLMs), with its improvement hinging on scalable and accurate feedback from judge models. However, the reliability of current judge models in instruction-following remains underexplored due to several deficiencies of existing meta-evaluation benchmarks, such as their insufficient data coverage and oversimplified pairwise evaluation paradigms that misalign with model optimization scenarios. To this end, we propose IF-RewardBench, a comprehensive meta-evaluation

benchmark for instruction-following that covers diverse instruction and constraint types. For each instruction, we construct a preference graph containing all pairwise preferences among multiple responses based on instruction-following quality. This design enables a listwise evaluation paradigm that assesses the capabilities of judge models to rank multiple responses, which is essential in guiding model alignment. Extensive experiments on IF-RewardBench reveal significant deficiencies in current judge models and demonstrate that our benchmark achieves a stronger positive correlation with downstream task performance compared to existing benchmarks. Our codes and data are available at https://github.com/thu-coai/IF-RewardBench.

[Findings] Improving LLM Code Reasoning via Semantic Equivalence Self-Play with Formal Verification ► *Underline Video*

Poon Tsz Nok and Antonio Valerio Miceli Barone

We introduce a self-play framework for semantic equivalence in Haskell, utilizing formal verification to guide adversarial training between a generator and an evaluator. The framework leverages Liquid Haskell proofs for validating equivalence and execution-based counterexamples for inequivalence, organized via a difficulty-aware curriculum. To facilitate this, we release OpInstruct-HSx, a synthetic dataset of ~ 28k validated Haskell programs. Empirical experiments show that our evaluator transfers effectively to downstream tasks, achieving up to 13.3pp accuracy gain on EquiBench and consistent gains on PySecDB. Ablation studies on the SEQ-SINQ regimes indicate that while inequivalence supervision provides data volume, equivalence proofs are uniquely responsible for the model's reasoning capabilities. The entire training pipeline and dataset are publicly released on GitHub and Hugging Face respectively.

[Short] Improving Retrieval-Augmented Generation without Taxonomy-based Error Categorization ► *Underline Video*

Gongbo Zhang, Yifan Peng and Chunhua Weng

Retrieval-Augmented Generation (RAG) improves the factual accuracy of large language model (LLM) outputs by grounding generation in external knowledge. Recent agentic RAG systems extend this paradigm by introducing critic agents that evaluate model responses and iteratively refine outputs. However, most prior work implicitly assumes reliable critic feedback and focuses on planning strategies, while paying limited attention to the robustness of the error correction process itself, which is often hindered by misaligned error categories and ineffective or incorrect corrections. We hypothesize that RAG performance can be improved without explicit error categorization. To this end, we propose RePAIR, a response-action learning paradigm that directly maps flawed RAG outputs to error-mitigating action plans without relying on fine-grained error taxonomies or explicit critic supervision. Across multiple benchmarks, RePAIR consistently improves agentic RAG performance.

[Findings] Improving the Throughput of Diffusion-based Large Language Models via a Training-Free Confidence-Aware Calibration ► *Underline Video*

Jucheng Shen, Gaurav Sarkar, Yeonju Ro, Sharath Nittur Sridhar, Zhangyang Wang, Aditya Akella and Souvik Kundu

We present CadLLM, a training-free method to accelerate the inference throughput of diffusion-based LLMs (dLLMs). We first investigate on the dynamic nature of the token unmasking confidence across blocks and steps. Based on this observation, we then present a lightweight adaptive approach that can control the generation block size, step size, and threshold based on the average confidence score of the unmasked tokens. We further reduce the softmaxing overhead of token probability generation by dynamically leveraging a subset of vocabulary size to regulate sampling breadth. CadLLM is a plug-and-play model-agnostic with KV caching based dLLMs. Extensive experiments on four popular tasks demonstrate the efficacy of CadLLM to yield throughput improvement of up to 1.1-2.28x over the state-of-the-art baseline with competitive accuracy.

[SRW] Inference-Time Feedback for Reasoning Controllability in Diffusion Language Models ► *Underline Video*

Clovis Barbour; Huixin Zhan

Scientific NLP systems often require outputs that satisfy strict, machine-checkable constraints. In this work, we study structured-generation controllability along three axes: structural control, iterative correction, and decoding dynamics. Diffusion decoding is of particular interest because its iterative refinement may improve global structure and revision behavior, but may also introduce distinct failure modes such as termination instability and repetition. To quantify controllability, we evaluate compliance with five machine-checkable constraints: (i) required headings and (ii) correct ordering, which reflect global structural control; (iii) explicit end markers and (iv) per-section bullet constraints, which probe local constraint adherence; and (v) repetition avoidance, which captures generation stability under different decoding dynamics. We use these metrics to assess both single-pass generation and changes under iterative correction. Our goal is to isolate structural reliability under parser-facing requirements rather than to directly measure scientific correctness. Across our benchmark, diffusion models tend to better preserve global structure, while iterative improvement substantially improves explicit termination and other local control constraints. Hybrid systems show mixed behavior depending on decoding order. These results suggest that machine-checkable controllability can be usefully decomposed into global structure versus local control, and that the two may benefit from different inference-time strategies.

[Long] Influence-based Online Experience Selection for Effective RLHF ► *Underline Video*

Yifan Gong, Jing Yao, Xiting Wang, Xunlong Wang, Xiaoyuan Yi and Xing Xie

Reinforcement Learning from Human Feedback (RLHF) has emerged as a crucial technique for aligning large language models (LLMs) with human preferences. However, existing RLHF methods face key challenges, including poor sample efficiency, high computational overhead, and slow convergence. Recent studies highlight the importance of data selection in RL, but how to effectively select the most beneficial experiences for RL training remains an open problem. Existing data selection methods for RL rely on heuristic metrics, failing to establish

an interpretable connection between data and optimization objectives. To address this problem, we propose InfOES (Influence-based Online Experience Selection), a novel data selection method for RLHF that dynamically estimates the influence of individual training samples on policy optimization. By incorporating data attribution into the policy gradient, InfOES can identify and filter out detrimental samples on the fly, ensuring effective convergence toward alignment objectives. Our approach is compatible with various RL algorithms (e.g., PPO, GRPO, REINFORCE++). Extensive experiments demonstrate that InfOES significantly enhances training effectiveness, achieving superior alignment performance with fewer optimization steps.

[Long] InsAT: Instance-aware Semantic Alignment and Transfer from Human-Object Keypoints for Zero-to-Few-shot Action Understanding ▶ *Underline Video*

Kazuki Tsutsukawa

Keypoint-based action recognition offers robustness to appearance variations and provides privacy-preserving representation. However, existing zero-shot (ZS) approaches largely emphasize human motion while underutilizing contextual information, particularly human-object interactions. Moreover, extending keypoint-based ZS models to few-shot scenarios remains insufficiently explored. We propose Instance-aware Semantic Alignment and Transfer (InsAT), a unified framework for ZS recognition and zero-to-few-shot (Z2F) adaptation that leverages instance-level language descriptions. InsAT aligns textual descriptions of humans, objects, and their interactions with visual representations derived from human and object keypoints, enabling effective transfer of interaction knowledge from seen to unseen action classes. To support Z2F adaptation, we introduce Instance-level Visual Adaptation, a parameter-free mechanism that improves recognition by incorporating instance-level contextual cues without updating model weights. Extensive experiments demonstrate that InsAT substantially outperforms prior keypoint-based ZS methods and achieves competitive performance relative to large vision-language models, while remaining data-efficient and robust.

[SRW] Interpretability of LLM Classifiers via the Rational Inattention Theory with Application to Hate Speech Detection ▶ *Underline Video*

Yuan Zhao and Ali Abdi

Hate speech detection is essential for maintaining healthy online communities. Large language models (LLMs) perform well on text classification, yet their decision strategies need to be better understood. While post-hoc rationales can justify individual decisions, they substantially increase inference cost and limit scalability in high-throughput settings. As another approach, we propose an extended rational inattention model that parameterizes linguistic noise and information processing cost, providing an interpretable behavioral framework for black-box LLM classifiers. Treating LLMs as rational decision-makers under information constraints allows us to estimate - from the observed classification behavior - the parameters that represent information processing cost and noise sensitivity. As a case study and using a hate-speech dataset spanning multiple noise environments, we evaluate four commercial LLMs and show that the introduced extended rational inattention model predictions closely match the observed performance across different noise levels. We further test the performance under various noise mechanisms and find that the inferred information cost parameters remain consistent while the noise parameters vary with the distortion mechanism. Overall, our introduced framework offers a cost-efficient and quantitative approach to derive interpretable indices of LLM moderation behavior and decisions, without additional rationale generation.

[Short] Is a Document Educational or Just Wikipedia-Style? , Pitfalls of Classifier-Based Quality Filtering ▶ *Underline Video*

Mateusz Klimaszewski and Piotr Andruszkiewicz

Classifier-based Quality Filtering has recently emerged as a fundamental technique in constructing pre-training corpora. The ability to deploy a single model that can replace or supplement a set of heuristics has proven effective across numerous Large Language Models. In this work, we expose a critical vulnerability in this approach by demonstrating how a straightforward Wikipedia-style reformatting operation can substantially alter a model's quality assessment and enable low-quality content to surpass filtering thresholds. Our analysis reveals that the FineWeb-Edu CQF model would reverse its filtering decision for approximately 7% of evaluated documents, thereby admitting content into the pre-training corpus that would otherwise have been excluded.

[Findings] JailMeter: An Evidence-Based Evaluation Framework for Jailbreak Attacks on Large Language Models ▶ *Underline Video*

Qingjia Huang, Jingyu Zhang, Jianguo Wu, Yakai Li, Weijuan Zhang, Yankai Rong, Junyi Yao, Shengzhi Zhang and Xiaoqi Jia

The assessment of jailbreak attacks against large language models currently suffers from inconsistent evaluation criteria and methods, leading to unreliable estimates of attack success rates. We propose JailMeter, an evidence-based evaluation framework designed to more faithfully measure jailbreak effectiveness. Inspired by the Information Bottleneck theory, JailMeter applies dual-feedback optimization to filter jailbreak noise from model responses while preserving content relevant to the original malicious question. This process produces concise evidence for a rigorous assessment under which an attack is validated only when the response captures the malicious intent and delivers a complete answer, thereby signaling a substantive bypass of model safety alignment. We evaluate JailMeter on JailMeter-Eva, a challenging benchmark containing 330 human-labeled, non-rejected jailbreak instances. JailMeter achieves an accuracy of 97.27%, substantially outperforming existing evaluation methods. To support large-scale evaluation, we further distill JailMeter into a small language model, JailMeterSLM, which maintains comparable reliability with significantly reduced computational costs. Code and dataset are available at https://github.com/Magi2B0y/JailMeter.

[Long] L2Dir: Integrating L_2-Norm and Directional Alignment for Unsupervised Contrastive Representation Learning in Multimodal Retrieval ▶ *Underline Video*

Tianyu Zong, Rui Dai, Hongzhu Yi, Yuanxiang Wang, Zhenghao Zhang, Zhenyu Guan, Yujia Yang, Bingkang Shi, Yueyang Ding, Xiangxiang Chu, Kaikui Liu and Jungang Xu

Multimodal representation learning primarily relies on contrastive objectives such as InfoNCE to align diverse modalities. However, these methods focus almost exclusively on directional alignment and often neglect the intrinsic role of embedding magnitudes (L2-norm) in the contrastive process. To bridge this gap, we propose L2Dir, a plug-and-play framework designed to optimize L2-norm alignment and Directional consistency jointly. As a highly efficient solution, L2Dir doesn't require extra data, distillation, or external supervision. It can be integrated seamlessly into existing pipelines by employing a lightweight MLP to reconstruct magnitudes from frozen backbone features. Extensive evaluations across 95 tasks using UniIR and VLM2Vec-V2 frameworks demonstrate that L2Dir yields consistent and significant performance gains over established baselines across various backbones and scales, proving that explicit magnitude modeling is a versatile and potent strategy for refining unsupervised multimodal representations. The source code for L2Dir in VLM2Vec-V2 is available in the supplementary materials.

[SRW] LAMP-MedQA: A Lightweight Multi-Agent System for Patient-Oriented Medical Question Answering ▶ *Underline Video*

Jack A. Johnson, Meghali Banerjee, Joseph Crawford, James Welch, Jim Davies and Tingyan Wang

Patient health literacy is a critical determinant of health outcomes, yet medical discharge summaries remain inaccessible to many patients due to jargon and complex language. Large language models (LLMs) offer a promising means of bridging this gap, but their deployment in resource-constrained hospital environments demands lightweight, privacy-preserving solutions. We evaluate a range of open- and closed-source LLMs on the MeDiSumQA dataset, comprising real patient discharge summaries paired with lay questions and clinician-verified answers, and demonstrate that open-source models achieve performance comparable to GPT-5. We then introduce LAMP-MedQA, a lightweight multi-agent framework for patient-oriented medical question answering. LAMP-MedQA decomposes the task into an information extraction stage and a text simplification stage, each governed by an automated metric-driven feedback loop that enables iterative self-correction without human-in-the-loop supervision. Implemented using two small open-source models (Qwen2.5-7B-Instruct and Phi-3.5-mini-instruct), the system achieves significantly improved readability (FKGL) and simplification quality (SARI) compared to zero-shot GPT-5, whilst maintaining competitive semantic similarity and clinical accuracy, suggesting that collaborative lightweight agents represent a viable approach to improving health literacy in clinical settings.

[SRW] Language Directions in Multilingual LLMs: A Layer-wise Diagnostic Study of Token Alignment and Pretraining Imprint ▶ *Underline Video*

Jea Sung Kim ; Suan Lee

We investigate how multilingual representations emerge across depth in large language models. Using a unified probing framework, we analyze six multilingual LLMs across five languages (EN/ES/ZH/FR/DE), decomposing behavior into (i) early-layer dynamics, (ii) linear vs.\{} MLP separability, and (iii) token–language alignment that tracks where vocabulary sharing peaks. Across models, we observe a consistent and substantial early jump: accuracy rises by +73.5 to +80.7 points from L0 to L1 on average, indicating that language-relevant signals become accessible immediately after the embedding layer. Moreover, representations are largely linearly separable: for 5/6 models, the mean gap between MLP and linear probes remains within +/-0.5 points. Token–language alignment further reveals systematic structure, with peak vocabulary mass exceeding 48% in some models and substantial variation in the depth of peak sharing. These findings provide a compact, cross-model characterization of how multilingual information is organized across depth and introduce simple alignment metrics that complement accuracy-based evaluation.

[Long] Language Models Struggle to Use Representations Learned In-Context ▶ *Underline Video*

Michael Lepori, Tal Linzen, Ann Yuan and Katja Filippova

Though large language models (LLMs) have enabled great success across a wide variety of tasks, they still appear to fall short of one of the loftier goals of artificial intelligence research: creating an artificial system that can adapt its behavior to radically new contexts upon deployment. One important step towards this goal is to create systems that can induce rich representations of data that are seen in-context, and then flexibly deploy these representations to accomplish goals. Recently, Park et al. 2024 demonstrated that current LLMs are indeed capable of inducing such representation from context (i.e., in-context representation learning). The present study investigates whether LLMs can use these representations to complete simple downstream tasks. We first assess whether open-weights LLMs can use in-context representations for next-token prediction, and then probe models using a novel task, adaptive world modeling. In both tasks, we find evidence that open-weights LLMs struggle to deploy representations of novel semantics that are defined in-context, even if they encode these semantics in their latent representations. Furthermore, we assess closed-source, state-of-the-art reasoning models on the adaptive world modeling task, demonstrating that even the most performant LLMs cannot reliably leverage novel patterns presented in-context. Overall, this work seeks to inspire novel methods for encouraging models to not only encode information presented in-context, but to do so in a manner that supports flexible deployment of this information.

[Short] Late Code Chunking: A Code Chunking Strategy for Repository-Level Code Completion ▶ *Underline Video*

Seungmin OhEunseok Lee

This paper introduces Late Code Chunking (LC^2), a chunking strategy designed to improve the semantic understanding of code segments

for Large Language Models (LLMs). Repository-level code completion requires predicting the completion of unfinished code by leveraging cross-file context spread across a repository. However, when retrieved fragments have missing semantics, the loss of structural or behavioral information during chunking, LLMs struggle to interpret the target code. To address this, LC^2 refines retrieved chunks by constructing a dual context: a "Code Retrieval Context" optimized for similarity-based search, and a "Code Comprehension Context" that serves as a late enrichment step through context expansion and augmentation. This dual-context design reduces information loss due to chunking and enhances the ability of LLMs to utilize retrieved code. Additionally, we introduce an Asymmetric Query-Chunk Sizing strategy to further optimize retrieval quality by minimizing query noise. Our experiments demonstrate that LC^2 provides robust performance gains, achieving a statistically significant 19.7% improvement in Exact Match accuracy on the CrossCodeEval benchmark compared to the best existing chunking method.

[SRW] Learning Shortcut Models for Efficient Recursive Reasoning ▶ *Underline Video*

Shiv Shankar

Recursive models that iteratively refine latent representations have demonstrated strong performance on variety of reasoning tasks. However, these models only control if and when to stop early, not how computation is distributed. In this work, we introduce shortcut reasoning, a framework for distilling recursive latent reasoning into a multiscale jump model that enables flexible test-time compute. We reinterpret recursive reasoning as a latent-time dynamical process and train a student model to predict the effect of multiple reasoning steps at once. To ensure robustness, we augment shortcut transitions with a repair mechanism, where a denoising base-model update projects approximate latent states back to a valid reasoning manifold. We further introduce stepwise improvement supervision, encouraging each shortcut step to increase the likelihood of the correct answer. Experiments on ARC-AGI show that our approach achieves competitive accuracy compared to recursive baselines while requiring fewer sequential updates.

[Findings] Lil: Less is Less When Applying Post-Training Sparse-Attention Algorithms in Long-Decode Stage ▶ *Underline Video*

Junhao Hu, Fangze Li, Mingtao Xu, Feifan Meng, Shiju Zhao, Tiancheng Hu, Ting Peng, Anmin Liu, Wenrui Huang, Chenxu Liu, Ziyue Hua and Tao Xie

Large language models (LLMs) demonstrate strong capabilities across a wide range of complex tasks and are increasingly deployed at scale, placing significant demands on inference efficiency. Prior work typically decomposes inference into prefill and decode stages, with the decode stage dominating total latency. To reduce time and memory complexity in the decode stage, a line of work introduces sparse-attention algorithms. In this paper, we show, both empirically and theoretically, that sparse attention can paradoxically increase end-to-end complexity: information loss often induces significantly longer sequences, a phenomenon we term “Less is Less” (Lil). To mitigate the Lil problem, we propose an early-stopping algorithm that detects the threshold where information loss exceeds information gain during sparse decoding. Our early-stopping algorithm reduces token consumption by up to 90% with a marginal accuracy degradation of less than 2% across reasoning-intensive benchmarks.

[Long] LinkQA: Synthesizing Diverse QA from Multiple Seeds Strongly Linked by Knowledge Points ▶ *Underline Video*

Xuemiao Zhang, Can Ren, Chengying Tu, Rongxiang Weng, Hongfei Yan, Jingang Wang and Xunliang Cai

The advancement of large language models (LLMs) struggles with the scarcity of high-quality, diverse training data. To address this limitation, we propose LinkSyn, a KP-graph-based synthesis framework that for the first time enables flexible control over discipline and difficulty distributions while balancing KP coverage and popularity. LinkSyn extracts KPs from question-answering (QA) seed data and constructs a KP graph to synthesize diverse QA data from multiple seeds strongly linked by KPs and sampled from graph walks. Specifically, LinkSyn incorporates (1) a knowledge value function to guide the adjustment of path sampling probability and balance KP coverage and popularity during graph walks; (2) diffusion-based synthesis via a strong reasoning model by leveraging multiple seeds with dense logical associations along each path; and (3) high-difficulty QA enhancement within given disciplines by flexible difficulty adjustments. By executing LinkSyn, we synthesize LinkQA, a diverse multi-disciplinary QA dataset with 50B tokens. Extensive experiments on Llama-3 8B demonstrate that continual pre-training with LinkQA yields an average improvement of 11.51% on MMLU and CMMLU, establishing new SOTA results. LinkQA consistently enhances performance across model size and initial FLOPs scales.

[Long] LLaDA 1.5: Variance-Reduced Preference Optimization for Large Language Diffusion Models ▶ *Underline Video*

Fengqi Zhu, Rongzhen Wang, Shen Nie, Xiaolu Zhang, Chunwei Wu, JUN ZHOU, Yankai Lin, Ji-Rong Wen and Chongxuan Li

Masked diffusion language models present a promising paradigm for language modeling, yet the systematic theoretical analysis and comprehensive empirical validation of their alignment on general tasks remain relatively underexplored. In this paper, we identify the primary challenge for this problem: the high variance in Evidence Lower Bound (ELBO)-based likelihood estimates required for preference optimization. To address this issue, we propose Variance-Reduced Preference Optimization (VRPO), a framework that formally analyzes the bias and variance of the preference optimization loss and gradient based on Direct Preference Optimization, showing both are governed by a score-estimator variance. Building on this foundation, we introduce multiple unbiased variance reduction strategies, including optimal budget allocation and antithetic sampling, to improve alignment performance. We demonstrate the effectiveness of VRPO by applying it to LLaDA, a large diffusion language model. The resulting model, LLaDA 1.5, consistently outperforms its SFT-only predecessor consistently across various general benchmarks, such as mathematics (GSM8K +4.7), coding (HumanEval +3.0, MBPP +1.8), and alignment (IFEval +4.0, Arena-Hard +4.3). Furthermore, LLaDA 1.5 demonstrates a highly competitive mathematical performance compared to other strong language MDMs and ARMs. Our model is available at https://huggingface.co/GSAI-ML/LLaDA-1.5.

[SRW] LLM as a Meta-Judge: Synthetic Data for NLP Evaluation Metric Validation ► *Underline Video*

Luká Eigler; Jindich Libovický; David Hurych

Validating evaluation metrics for NLG typically relies on expensive and time-consuming human annotations, which predominantly exists only for English datasets. We propose LLM as a Meta-Judge, a scalable framework that utilizes LLMs to generate synthetic evaluation datasets via controlled semantic degradation of real data, replacing human judgment. We validate our approach using meta-correlation, measuring the alignment between metric rankings derived from synthetic data and those from standard human benchmarks. Experiments across Machine Translation, Question Answering, and Summarization demonstrate that synthetic validation serves as a reliable proxy for human judgment, achieving meta-correlations exceeding 0.9 in multilingual QA and proves to be a viable alternative where human judgments are unavailable or too expensive to obtain. Our codes and data will become publicly available upon paper acceptance.

[SRW] LLM Parameters for Math Across Languages: Shared or Separate? ► *Underline Video*

Behzad Shomali; Luisa Victor; Tim Selbach; Ali Hamza Bashir; David Berghaus; Joachim Koehler; Mehdi Ali; Markus Frey;

Large language models (LLMs) exhibit substantial cross-lingual variation in mathematical reasoning performance, but it remains unclear whether these differences reflect language-specific parameters or a shared mechanism that manifests differently by language. We present a cross-lingual mechanistic analysis of mathematical reasoning in LLMs, enabling us to localize and compare model parameters that support mathematical reasoning across languages. We find that mathematical reasoning is supported by a partially shared set of parameters, with the strongest overlap concentrated in intermediate model layers. We further observe that English consistently produces the largest set of math-relevant parameters, whereas lower-resource languages reveal smaller sets of relevant parameters. These results indicate that math-related reasoning in multilingual LLMs is neither fully language-invariant nor fully language-specific, but instead relies on a shared parameter core with systematic cross-lingual differences.

[Findings] LLM Program Optimization via Retrieval Augmented Search ► *Underline Video*

Sagnik Anupam, Alexander Shypula and Osbert Bastani

Recent work has demonstrated the potential of large language models (LLMs) for program optimization, a key challenge in programming languages. We propose a blackbox adaptation method called Retrieval Augmented Search (RAS) that performs beam search over candidate optimizations; at each step, it retrieves in-context examples from a given training dataset of slow-fast program pairs to guide the LLM. Critically, we find that performing contextual retrieval based on an LLM-generated natural language description significantly outperforms retrieval based on the source code. We also propose AEGIS, a method for improving interpretability by decomposing training examples into "atomic edits" that are significantly more incremental in nature. We show that RAS performs up to 2.06x better than prior state-of-the-art blackbox adaptation strategies on optimizing C++ programs, and that AEGIS performs up to 1.37x better while making significantly smaller edits. We also show that using RAS improves the mean runtime percentile of Python programs by 10.27 compared to baselines.

[SRW] LLM-Based Zero-Shot Soft Labeling for Anticipating Disagreement in Negotiation Dialogues ► *Underline Video*

Ken Watanabe; Katsuhide Fujita

Negotiation involves complex emotional and strategic dynamics that pose challenges for AI agents in negotiation dialogues. This paper proposes a zero-shot soft-labeling method using an large language model-based embedding model and verifies its performance on negotiation dialogues. Furthermore, it examines the performance of predictive model training on rule-based annotated hard and soft labels obtained by the proposed method for the task of predicting whether agreement will be reached from partial dialogues, namely, final disagreement anticipation in negotiation mid-dialogues (FDANMD). Soft labeling obtained by the proposed method showed a maximum HIT@3 score of 0.87 against rule-based annotated hard labels, whereas failure cases also demonstrated the limitations of rule-based annotation. Furthermore, using ROC AUC, evaluations of FDANMD across three datasets (CB, DN, and JI) with negotiation progress rates of 0.25, 0.5, and 1.0 revealed that soft labeling is particularly effective at low negotiation progress rates and also offers superior performance on individual datasets and unseen datasets for models trained on multiple datasets. These results motivate the use of soft labeling to incorporate the complexity of negotiation dialogues into intermediate representations and support the generalizability of zero-shot soft labeling and generalizable predictors across a wide range of negotiations beyond known domains.

[SRW] LLMs for Now, Fine-Tuning for Later: An Ensemble Approach to Data Drift in Domain-Specific Tasks ► *Underline Video*

Yuxuan Lu; Bingsheng Yao; Shao Zhang; Yisi Sang; Yun Wang; Hansu Gu; Peng Zhang; Tun Lu; Toby Jia-Jun Li; Dakuo Wang

Deploying machine learning models in real-world domain-specific scenarios is challenged by the scarcity of expert annotations and by data drift, where the statistical properties of incoming data continuously evolve. Active Learning (AL) iteratively improves compact models with expert annotations but suffers from recurring cold-start degradation, while LLMs provide strong off-the-shelf performance yet cannot leverage newly accumulated labels, raising the question: how can we better leverage LLMs to assist the active learning process? Through an empirical study on five legal and biomedical datasets, we reveal a complementary temporal dynamic: LLMs excel during early and post-drift stages, while AL-assisted compact models eventually surpass them as annotations accumulate. Motivated by this finding, we propose an ensemble system that combines an LLM, an AL-assisted compact model, and an automatic switch module that routes predictions to the better-performing model in real time. Evaluated under simulated data drift on two mental health datasets, our system achieves 96–98% switch accuracy and consistently outperforms either model used alone.

[Long] LoVeC: Reinforcement Learning for Better Verbalized Confidence in Long-Form Generation ▶ *Underline Video*

Caiqi Zhang, Xiaochen Zhu, Chengzu Li, Nigel Collier and Andreas Vlachos

Hallucination remains a major challenge for the safe and trustworthy deployment of large language models (LLMs) in factual content generation. Prior work has explored confidence estimation as an effective approach to hallucination detection, but often relies on post-hoc self-consistency methods that require computationally expensive sampling. Verbalized confidence offers a more efficient alternative, but existing approaches are largely limited to short-form question answering (QA) tasks and do not generalize well to open-ended generation. In this paper, we propose LoVeC (Long-form Verbalized Confidence), a novel reinforcement learning (RL)-based method that trains LLMs to append an on-the-fly numerical confidence score to each generated statement during long-form generation. The confidence score serves as a direct and interpretable signal of the factuality of generation. We introduce two evaluation settings, free-form tagging and iterative tagging, to assess different verbalized confidence estimation methods. Experiments on three long-form QA datasets show that our RL-trained models achieve better calibration and generalize robustly across domains. Also, our method is highly efficient, being 20 x faster than traditional self-consistency methods while achieving better calibration.

[Long] MAGMA: A Multi-Graph based Agentic Memory Architecture for AI Agents ▶ *Underline Video*

DongmingJiang DongmingJiang, Yi Li, Guanpeng Li and Bingzhe Li

Memory-Augmented Generation (MAG) extends large language models with external memory to support long-context reasoning, but existing approaches largely rely on semantic similarity over monolithic memory stores, entangling temporal, causal, and entity information. This design limits interpretability and alignment between query intent and retrieved evidence, leading to suboptimal reasoning accuracy. In this paper, we propose MAGMA, a multi-graph agentic memory architecture that represents each memory item across orthogonal semantic, temporal, causal, and entity graphs. MAGMA formulates retrieval as policy-guided traversal over these relational views, enabling query-adaptive selection and structured context construction. By decoupling memory representation from retrieval logic, MAGMA provides transparent reasoning paths and fine-grained control over retrieval. Experiments on LoCoMo and LongMemEval demonstrate that MAGMA consistently outperforms state-of-the-art agentic memory systems in long-horizon reasoning task.

[Long] MAGNET: Towards Adaptive GUI Agents with Memory-Driven Knowledge Evolution ▶ *Underline Video*

Libo Sun, Jiwen Zhang, Siyuan Wang and zhongyu wei

Mobile GUI agents powered by large foundation models enable autonomous task execution in applications, but frequent updates that alter UI appearance and reorganize workflows cause agents trained on historical data to fail. Despite these surface changes, we observe that functional semantics and task intents remain fundamentally stable. Building on this insight, we introduce MAGNET, a memory-driven adaptive agent framework with dual-level memory: stationary memory that links diverse visual features to stable functional semantics for robust action grounding and procedural memory that captures stable task intents across varying workflows. Furthermore, we propose a dynamic memory evolution mechanism that continuously refines both memories by prioritizing frequently accessed knowledge. Evaluations on the online benchmark AndroidWorld demonstrate substantial improvements over memory-augmented baselines, while offline benchmarks confirm consistent gains under distribution shifts. These results validate that leveraging stable structures across interface changes improves agent performance and generalization in evolving software environments.

[Findings] MDocRAG-RL: Empowering Multi-Modal Document RAG via Complex Visual Reasoning with Reinforcement Learning ▶ *Underline Video*

Zhongyu Wang;

While Retrieval-Augmented Generation(RAG) enhances multi-modal large language models(MLLMs) by introducing external knowledge, existing RAG systems still face significant limitations when dealing with complex visual reasoning. On one hand, MLLMs, being generative models, produce suboptimal embeddings for retrieval tasks. On the other hand, existing methods naively insert images into context without adequate visual perception, thereby limiting reasoning capabilities. To address these challenges, we propose MDocRAG-RL, a novel RAG framework for complex visual reasoning. We design specialized pre-training and fine-tuning tasks to enable MLLMs to compress visual document representations and align textual and visual embeddings for improved retrieval efficiency. Additionally, we design a visual perception action space for the generator that allows progressive coarse-to-fine information acquisition from visually-rich documents. Furthermore, we develop a reinforcement learning framework to enhance the complex visual reasoning capability of the RAG system. Extensive experiments on multiple challenging benchmarks demonstrate the significant effectiveness of our approach, achieving state-of-the-art performance across various benchmarks.

[SRW] Measuring and Mitigating Shortcut Reliance in Language Models with Probe-Based Representation Entanglement ▶ *Underline Video*

Divyajot Singh

Shortcut learning remains a major obstacle to robust NLP systems: models can achieve high in-distribution accuracy by relying on surface cues that fail under distribution shift. We study whether shortcut reliance can be diagnosed and mitigated in small instruction-tuned language models using a simple representation-level quantity. We fine-tune Gemma 3 1B Instruct and Llama 3.2 1B on two synthetic sentiment shortcuts in SST-2 and one natural shortcut in MNLI based on lexical overlap. During training, we fit linear probes for the task label and the shortcut attribute at every layer and define CDRE as the absolute cosine similarity between the two probe directions. Across settings, increasing shortcut prevalence produces a sharp rise in the robustness gap between shortcut-aligned and shortcut-free test sets, and higher

deep-layer CDRE tracks this degradation. At a 99% shortcut ratio, Llama's clean accuracy on capitalization-biased SST-2 drops from 93.2% at 0% bias to 49.0%, while Gemma drops from 91.8% to 60.2%. A CDRE-regularized objective substantially improves robustness for capitalization and lexical-overlap shortcuts, but offers little benefit for a speaker-prefix shortcut whose learned directions are already nearly orthogonal. These results show that probe-derived representation entanglement provides a reliable signal of harmful shortcut reliance and offers a practical criterion for determining when shortcut mitigation is likely to be effective.

[SRW] Mechanistic Analysis Of Universality: Numerical Comparison Circuits Across Transformer Architectures ▶ *Underline Video*
Arya Bhardia; Julian Ramirez; Siddhanta Verma; Karen Mkrtchyan
Transformer language models reliably achieve high accuracy on many reasoning tasks; however, their internal mechanisms are not fully understood. Mechanistic interpretability seeks to remedy this gap by identifying task circuits within individual models, but it is unclear whether such circuits generalize across model families and scales. In this work, we study the universality of circuits through the lens of numerical comparisons, a simple and controlled task that enables clean and causal interventions. We conduct experiments on a set of transformer models spanning different families and sizes from 1.7b to 9b parameters. We find that models within the Qwen family exhibit a highly consistent circuit structure across architecture and scale, featuring localized attention heads that write a task relevant signal. In contrast, models from other families show qualitatively different implementations, where task relevant information emerges much earlier and is distributed across components as opposed to being concentrated within a small set of attention heads. These results serve as evidence that task behavior similarities do not imply mechanistic universality and highlight the necessity for cross model comparisons to claim generalization of internal circuits.

[SRW] Metadata Conditioned Large Language Models for Localization ▶ *Underline Video*
Anjishnu Mukherjee, Ziwei Zhu and Antonios Anastasopoulos;
Large language models are typically trained by treating text as a single global distribution, often resulting in geographically homogenized behavior. We study metadata conditioning as a lightweight approach for localization, pre-training 31 models (at 0.5B and 1B parameter scales) from scratch on large-scale English news data annotated with verified URLs, country tags, and continent tags, covering 4 continents and 17 countries. Across four controlled experiments, we show that metadata conditioning consistently improves in-region performance without sacrificing cross-region generalization, enables global models to recover localization comparable to region-specific models, and improves learning efficiency. Our ablation studies demonstrate that URL-level metadata alone captures much of the geographic signal, while balanced regional data coverage remains essential, as metadata cannot fully compensate for missing regions. Finally, we introduce a downstream benchmark of 800 localized news MCQs and show that after instruction tuning, metadata conditioned global models achieve accuracy comparable to LLaMA-3.2-1B-Instruct, despite being trained on substantially less data. Together, these results establish metadata conditioning as a practical and compute-efficient approach for localization of language models.

[SRW] Mind the Gap: Multilingual Divide in LLM Bias Detection and Reasoning ▶ *Underline Video*
Medha Hira; Prachi Goyal*; Raj Maheshwari*; Arnav Goel*
Large Language Models (LLMs) are increasingly deployed in multilingual settings, yet most bias evaluation remains English-centric and overlooks how bias manifests within reasoning. We present a systematic study of social bias in both predictions and chain-of-thought reasoning across English, Dutch, Spanish, and Turkish using the MBBQ benchmark. We evaluate instruction-tuned, CoT-prompted, and reasoning-native models under supervised fine-tuning and preference optimization, using accuracy, F1, bias metrics, and a novel reasoning-level language drift measure. We find that (1) bias varies substantially across languages, with consistent degradation in non-English settings, (2) reasoning traces often introduce additional stereotype-driven signals beyond final outputs, and (3) English-trained debiasing methods fail to generalize reliably, with preference optimization introducing cross-lingual trade-offs. We further show that performance gains in multilingual settings are frequently driven by implicit reliance on English-centric reasoning, revealed through increased language drift. Together, our results demonstrate that multilingual fairness cannot be inferred from English performance and requires reasoning-aware, language-specific evaluation and alignment.

[Long] Mitigating Over-Refusal in Aligned Large Language Models via Inference-Time Activation Energy ▶ *Underline Video*
Eric Hanchen Jiang, Weixuan Ou, Run Liu, Shengyuan Pang, Guancheng Wan, Ranjie Duan, Wei Dong, Kai-Wei Chang, XiaoFeng Wang, Ying Nian Wu and Xinfeng Li
Safety alignment of large language models currently faces a central challenge: existing alignment techniques often prioritize mitigating responses to harmful prompts at the expense of overcautious behavior, leading models to incorrectly refuse benign requests. A key goal of safe alignment is therefore to improve safety while simultaneously minimizing false refusals. In this work, we introduce Energy Landscape Steering (ELS), a novel, fine-tuning free framework designed to resolve this challenge through dynamic, inference-time intervention. We trained a lightweight, external Energy-Based Model (EBM) to assign high energy to undesirable (false refusal or jailbreak) states and low energy to desirable (helpful response or safe reject) ones. During inference, the EBM maps the LLMs internal activations to an energy landscape, and we use the gradient of the energy function to steer the hidden states toward low-energy regions in real time. This dynamically guides the model toward desirable behavior without modifying its parameters. By decoupling behavioral control from the models core knowledge, ELS provides a flexible and computationally efficient solution. Extensive experiments across diverse models demonstrate its effectiveness: raising compliance on the ORB-H benchmark from 57.3% to 82.6% while maintaining the baseline safety performance. Our work establishes a promising paradigm for building LLMs that simultaneously achieve high safety and low false refusal rates.

[Long] MMTutorBench: The First Multimodal Benchmark for AI Math Tutoring ▶ *Underline Video*

Tengchao Yang, Sichen Guo, Mengzhao Jia, Jiaming Su, Yuanyang Liu, Zhihan Zhang and Meng Jiang

Effective math tutoring requires not only solving problems but also diagnosing students difficulties and guiding them step by step. While multimodal large language models (MLLMs) show promise, existing benchmarks largely overlook these tutoring skills. We introduce MMTutorBench, the first benchmark for AI math tutoring, consisting of 770 problems built around pedagogically significant key-steps. Each problem is paired with problem-specific rubrics that enable fine-grained evaluation across six dimensions, and structured into three tasks, Insight Discovery, Operation Formulation, and Operation Execution. We evaluate 12 leading MLLMs and find clear performance gaps between proprietary and open-source systems, substantial room compared to human tutors, and consistent trends across input variants: OCR pipelines degrade tutoring quality, few-shot prompting yields limited gains, and our rubric-based LLM-as-a-Judge proves highly reliable. These results highlight both the difficulty and diagnostic value of MMTutorBench for advancing AI tutoring.

[Long] MoEC: A Memory-Routed Mixture-of-Experts Controller for Adaptive Minecraft Control ▶ *Underline Video*

Hui Wu, Chao Xu, Jianghui Wang, Ziqiong Liu, Dong Li, Yiwei Dai and Emad Barsoum

Embodied agents in open-ended environments such as Minecraft increasingly adopt planner-controller architectures, with large language models acting as high-level planners. While planning has advanced rapidly, control remains underexplored. Existing systems commonly rely on a monolithic policy to execute subgoals across varying contexts, forcing incompatible behaviors into a shared parameter space and causing interference that scaling only partially mitigates. To address this, we propose MoEC, a Memory-Routed Mixture-of-Experts Controller for Adaptive Minecraft Control. MoEC routes via a subgoal-indexed, non-parametric expert memory and regulates capacity through failure-triggered expert growth and redundancy-aware consolidation. This design enables continual adaptation without full retraining, while maintaining parameter efficiency and with bounded inference cost. We evaluate MoEC on diverse and compositional Minecraft tasks, demonstrating significant gains in adaptability, robustness, and execution consistency over strong baselines, yielding a scalable and efficient alternative for open-ended control.

[Long] MONETA: Multimodal Industry Classification through Geographic Information with Multi Agent Systems ▶ *Underline Video*

Arda Yüksel, Gabriel Thiem, Susanne Walter, Patrick Felka, Gabriela Alves Werb and Ivan Habernal

Industry classification schemes are integral parts of public and corporate databases as they classify businesses based on economic activity. Due to the size of the company registers, manual annotation is costly, and fine-tuning models with every update in industry classification schemes requires significant data collection. We replicate the manual expert verification by using existing or easily retrievable multimodal resources for industry classification. We present MONETA, the first multimodal industry classification benchmark with text (Website, Wikipedia, Wikidata) and geospatial sources (OpenStreetMap and satellite imagery). Our dataset enlists 1,000 businesses in Europe with 20 economic activity labels according to EU guidelines (NACE). Our training-free baseline reaches 62.10% and 74.10% with open and closed-source Multimodal Large Language Models (MLLM). We observe an increase of up to 22.80% with the combination of multi-turn design, context enrichment, and classification explanations. We will release our dataset and the enhanced guidelines.

[Findings] Moneyball with LLMs: Analyzing Tabular Summarization in Sports Narratives ▶ *Underline Video*

Ritam Upadhyay, Naman Ahuja, Rishabh Baral, Aparna Garimella and Vivek Gupta

Large language model (LLM) approaches to tabular summarization rely on extensive prompt engineering, decomposition pipelines, or entity-level intermediate representations to achieve strong performance. While effective, these strategies are computationally expensive and offer limited insight into how well models maintain state over long, evolving narratives. We introduce SporTabSet, a diagnostic benchmark for long-context tabular summarization across two complementary sports domains that require tracking multiple entities and aggregating statistics under domain-specific rules. Using SporTabSet, we systematically evaluate decomposition-based strategies across several long context LLMs. Results show that although decomposition substantially improves accuracy and numerical fidelity, gains stem mainly from dissecting multi-entity interference rather than improved local arithmetic. Robustness experiments further reveal high sensitivity to surface-level cues with structured failures, including hallucination, omission, and role confusion. Together, these findings identify consistent multi-entity memory as a key bottleneck in long-context table generation, motivating diagnostic evaluation as a prerequisite for scalable, efficient, and reliable tabular summarization models.

[Long] MTSQL-R1: Towards Long-Horizon Multi-Turn Text-to-SQL via Agentic Training ▶ *Underline Video*

Taicheng Guo, Hai Wang, ChaoChun Liu, Mohsen Golalikhani, Xin Chen, Xiangliang Zhang and Chandan K. Reddy

Multi-turn Text-to-SQL aims to translate a users conversational utterances into executable SQL while preserving dialogue coherence and grounding to the target schema. However, most existing systems only regard this task as a simple text translation task and follow a short-horizon paradigm, generating a query per turn without execution, explicit verification, and refinement, which leads to non-executable or incoherent outputs. We present MTSQL-R1, an agentic training framework for long-horizon multi-turn Text-to-SQL. We cast the task as a Markov Decision Process (MDP) in which an agent interacts with (i) a database for execution feedback and (ii) a persistent dialogue memory for coherence verification, performing an iterative propose->execute->verify->refine cycle until all checks pass. Experiments on CoSQL and SParC demonstrate that MTSQL-R1 consistently outperforms strong baselines, highlighting the importance of environment-driven verification and memory-guided refinement for conversational semantic parsing. Full recipes (including code, trained models, reasoning trajectories, etc.) will be released upon acceptance to contribute to community research.

[SRW] Multi-Constraint State Tracking with Negation: A Diagnostic Benchmark for LLM World Modeling ► *Underline Video*

Ayan Sar; Pranav Singh Puri; Sumit Aich; Anurag Kaushish; Tanupriya Choudhury; Ajith Abraham

Large Language Models (LLMs) achieve strong performance on a wide range of reasoning benchmarks, yet it remains unclear whether they can reliably maintain and update internal representations of an evolving world described in natural language. In particular, existing evaluations inadequately probe state tracking under multiple interacting constraints and largely overlook the role of negated actions, despite their ubiquity in real-world language. We address this gap by introducing MCST, a diagnostic benchmark for multi-constraint state tracking that evaluates an LLM's ability to maintain consistent world models across sequences of actions involving inventory changes, spatial movement, temporal ordering, and systematic negation. MCST comprises 100,847 questions spanning 12 real-world domains, with five calibrated difficulty levels, nine question types, and controlled integration of negated actions. The benchmark further incorporates culturally diverse entity names to enable analysis of cross-cultural robustness. We evaluate 14 SOTA LLMs across multiple model families using a unified evaluation protocol. Our results reveal substantial limitations: even the strongest models exhibit sharp performance degradation as difficulty increases, with accuracy dropping below 35% at the highest level. Most notably, we identify negation as a dominant failure mode, causing accuracy reductions of 23-32% across models. We release MCST and the full evaluation framework to support future research on state tracking and reasoning in language models and is available at GitHub.

[SRW] NEAT-IR: Neural Explainable Analysis Tool for Information Retrieval ► *Underline Video*

Lev Sukherman; Artem Frenk; Nina Klimenkova; Connor Jason

Neural IR models achieve strong performance but remain difficult to interpret. We present NEAT-IR, a black-box analysis framework that explains ColBERT's ranking behavior using 26 classical IR features (BM25, TF-IDF, IDF measures, positional signals). We analyze ColBERT through two complementary lenses: regression (predicting exact scores) and learning-to-rank (predicting relative order), evaluated on MS MARCO (48,250 query-passage pairs). Our key finding is a score-rank gap: classical features preserve ColBERT's rankings nearly perfectly (NDCG@5 ~ 0.99) yet explain only R^2 ~ 0.28 of score variance. Feature attribution reveals that regression and ranking models rely on distinct feature subsets: query-level IDF signals dominate score prediction, while document-matching features (BM25, cosine TF-IDF) drive ranking preservation. These findings suggest that ColBERT's ordinal behavior on MS MARCO is largely recoverable from classical signals, while neural contributions primarily affect score magnitude. NEAT-IR enables practitioners to diagnose when neural rankers deviate from classical patterns, supporting interpretable model auditing and informed hybrid pipeline design.

[Findings] NeuroSym-Cal: Bridging the Reasoning-Execution Gap in Code Generation via Hierarchical Calibration ► *Underline Video*

Peiyang Liu, Yining Wang, Youru Li, Long Li, Zhi Cai and Wei Ye

While Chain-of-Thought (CoT) reasoning enhances code generation in Large Language Models (LLMs), it introduces a critical challenge in uncertainty estimation: Confidence Saturation. Existing calibration methods, such as Self-Consistency, rely on the assumption that consensus implies correctness. This assumption fails under systematic errors, where models confidently repeat flawed logic, leading to miscalibrated high-confidence predictions. To address this, we introduce NeuroSym-Cal, a hierarchical calibration framework. We posit that reliable confidence requires interrogating the model at two complementary levels: the extrinsic consensus of its symbolic outputs and the intrinsic sensitivity of its latent reasoning. Specifically, we propose Reasoning Sensitivity Analysis to measure the local curvature of the deductive process via latent perturbation, providing a fine-grained signal that persists even when output consensus saturates. These orthogonal features are fused by a Contextual Calibration Network to predict correctness. Experiments across state-of-the-art reasoning models (e.g., DeepSeek-R1) demonstrate that NeuroSym-Cal effectively de-saturates overconfident errors, achieving state-of-the-art Expected Calibration Error (ECE) and superior selective generation performance on Out-Of-Domain (OOD) benchmarks.

[Findings] New Compendium of a Myriad of Plants: A New Dataset Describing Ancient Chinese Plants ► *Underline Video*

Xiaobin Shen, Zhongqing Wang, Shichen Li, Chu-Ren Huang and Guodong Zhou

In ancient China, a variety of datasets depicted humanistic scenes, geographical features, and plants. However, these datasets, compiled long ago, often contain errors, lack comprehensiveness, and are inconsistent with modern realities. To meet current demands, we aim to expand and improve ancient datasets using large language model. Focusing on the Great Compendium of Myriad Flowers, an invaluable ancient plants dataset, we gather information on numerous previously excluded plants, carefully select and organize classical Chinese poetry and prose, and construct a comprehensive botanical encyclopedia knowledge system. Additionally, we collect ancient paintings and modern photographs of plants to enrich the dataset. Furthermore, we propose a novel multi-modal plant classification model designed to integrate multi-modal information from both classical and contemporary sources, enabling the extraction of plant-related information from classical Chinese poetry and prose. Extensive experiments demonstrate the importance of the proposed new ancient plants dataset, and also indicate the effectiveness of our proposed multi-modal plant classification model.

[Findings] Not All Citations Are Equal:Entropy-Guided Citation Selection for Noise-Resistant Medical LLM ► *Underline Video*

Minyu Gao, Hanlin Xiao, Ruoyu Wang, Shuai Yang, YeXuan Zhang, Xin Wu and Xingyu Liu

Retrieval-Augmented Generation (RAG) provides external knowledge support for large language models (LLMs) in medical applications, but retrieved contexts often contain noisy or conflicting evidence that can degrade reasoning. We observe that when internal and external knowledge disagree, models systematically prefer external citations, inadvertently injecting retrieval noise. Our analyses further show that only a subset of retrieved citations consistently improves outcomes; these effective citations exhibit markedly lower token-level entropy,

linking citation entropy to model accuracy. Building on these findings, we propose a complete pipeline consisting of a training-free multi-turn reasoning framework and a post-training methodology. The training-free framework elicits internal thought, external thought, and fusion thought, and applies conflict detection and explicit denoising for complex queries. For post-training, we distill structured supervised fine-tuning (SFT) data and apply GRPO with an entropy-based citation reward that encourages selective citation of beneficial external knowledge while penalizing noisy citations. Experiments across diverse benchmarks demonstrate consistent gains in noise-resistant medical reasoning, with larger improvements on harder cases.

[Long] Octopus: Gated Selective Attention for Memory-Bounded Long-Context Inference in Large Language Models ▶ *Underline Video*

Van-Chien Nguyen, Ryan A. Rossi, Linh Ngo, Franck Dernoncourt and Thien Nguyen

Transformer inference becomes increasingly memory-bound as the Key-Value (KV) cache grows linearly with sequence length. While subquadratic architectures offer constant-memory inference, they rely on aggressive state compression that degrades performance on complex reasoning tasks. We propose Octopus, a framework that confers fixed-memory inference onto pretrained Transformers without the information loss of linearization. Octopus retrofits attention layers with Gated Selective Attention, a learnable module that enforces an adaptive sparsity policy over the context history. By dynamically scoring and retaining only high-utility KV states, this mechanism transforms the unbounded cache into a compact, evolving memory budget that filters out uninformative noise. Empirically, on the GSM8K benchmark, it outperforms state-of-the-art linearized baselines by over 36 points under identical memory constraints. Remarkably, Octopus also surpasses its own full-cache teacher, demonstrating that learned sparse retention serves as an effective regularizer for long-horizon reasoning.

[Long] Omni-I2C: A Holistic Benchmark for High-Fidelity Image-to-Code Generation ▶ *Underline Video*

Jiawei Zhou, Chi Zhang, Xiang Feng, Qiming Zhang, Haibo Qiu, Lihuo He, Dengpan Ye, Xinbo Gao and Jing Zhang.

We present Omni-I2C, a comprehensive benchmark designed to evaluate the capability of Large Multimodal Models (LMMs) in converting complex, structured digital graphics into executable code. We argue that this task represents a non-trivial challenge for the current generation of LMMs: it demands an unprecedented synergy between high-fidelity visual perception, to parse intricate spatial hierarchies and symbolic details, and precise generative expression, to synthesize syntactically sound and logically consistent code. Unlike traditional descriptive tasks, Omni-I2C requires a holistic understanding where any minor perceptual hallucination or coding error leads to a complete failure in visual reconstruction. Omni-I2C features 1130 meticulously curated samples, defined by its breadth across subjects, image modalities, and programming languages. By incorporating authentic user-sourced cases, the benchmark spans a vast spectrum of digital content, from scientific visualizations to complex symbolic notations, each paired with executable reference code. To complement this diversity, our evaluation framework provides necessary depth; by decoupling performance into perceptual fidelity and symbolic precision, it transcends surface-level accuracy to expose the granular structural failures and reasoning bottlenecks of current LMMs. Our evaluation reveals a substantial performance gap among leading LMMs; even state-of-the-art models struggle to preserve structural integrity in complex scenarios, underscoring that multimodal code generation remains a formidable challenge. Data and code are available at https://github.com/MiliLab/Omni-I2C.

[Long] On-the-Fly VLA Adaptation via Test-Time Reinforcement Learning ▶ *Underline Video*

Changyu Liu, Yiyang Liu, Taowen Wang, Qiao Zhuang, James Chenhao Liang, Wenhao Yang, Renjing Xu, Qifan Wang, Dongfang Liu and Cheng Han

Vision-Language-Action models have recently emerged as a powerful paradigm for general-purpose robot learning, enabling agents to map visual observations and natural-language instructions into executable robotic actions. Though popular, they are primarily trained via supervised fine-tuning or training-time reinforcement learning, requiring explicit fine-tuning phases, human interventions, or controlled data collection. Consequently, existing methods remain unsuitable for challenging simulated- or physical-world deployments, where robots must respond autonomously and flexibly to evolving environments. To address this limitation, we introduce a Test-Time Reinforcement Learning for VLAs (TT-VLA), a framework that enables on-the-fly policy adaptation during inference. TT-VLA formulates a dense reward mechanism that leverages step-by-step task-progress signals to refine action policies during test time while preserving the SFT/RL-trained priors, making it an effective supplement to current VLA models. Empirical results show that our approach enhances overall adaptability, stability, and task success in dynamic, previously unseen scenarios under simulated and real-world settings. We believe TT-VLA offers a principled step toward self-improving, deployment-ready VLAs.

[Long] Optimizing User Profiles via Contextual Bandits for Retrieval-Augmented LLM Personalization ▶ *Underline Video*

Linfeng Du, Ye Yuan, Zichen Zhao, Fuyuan Lyu, Emiliano Penaloza, Xiuying Chen, Zipeng Sun, Jikun Kang, Laurent Charlin, Xue Liu and Haolun Wu

Large language models (LLMs) excel at general-purpose tasks, yet adapting their responses to individual users remains challenging. Retrieval augmentation provides a lightweight alternative to fine-tuning by conditioning LLMs on user history records, and existing approaches typically select these records based on semantic relevance. We argue that relevance serves as an unreliable proxy for utility: a record may be semantically similar to a query yet fail to improve generation quality or even degrade it due to redundancy or conflicting information. To bridge this gap, we propose PURPLE, a contextual bandit framework that oPtimizes UseR Profiles for LLM pErsonalization. In contrast to a greedy selection of the most relevant records, PURPLE treats profile construction as an order-sensitive generation process

and utilizes a Plackett-Luce ranking model to capture complex inter-record dependencies. By training with semantically rich feedback provided by the likelihood of the reference response, our method aligns retrieval directly with generation quality. Extensive experiments on nine personalization tasks demonstrate that PURPLE consistently outperforms strong heuristic and retrieval-augmented baselines in both effectiveness and efficiency, establishing a principled and scalable solution for optimizing user profiles.

[Findings] PaperScope: A Multi-Modal Multi-Document Benchmark for Agentic Deep Research Across Massive Scientific Papers ► *Underline Video*

Lei Xiong, Huaying Yuan, Zheng Liu, Zhao Cao and Zhicheng Dou

Leveraging Multi-modal Large Language Models (MLLMs) to accelerate frontier scientific research is promising, yet how to rigorously evaluate such systems remains unclear. Existing benchmarks mainly focus on single-document understanding, whereas real scientific workflows require integrating evidence from multiple papers, including their text, tables, and figures. As a result, multi-modal, multi-document scientific reasoning remains underexplored and lacks systematic evaluation. To address this gap, we introduce PaperScope, a multi-modal multi-document benchmark designed for agentic deep research. PaperScope presents three advantages: (1) Structured scientific grounding. It is built on a knowledge graph of over 2,000 AI papers spanning three years, providing a structured foundation for research-oriented queries. (2) Semantically dense evidence construction. It integrates semantically related key information nodes and employs optimized random-walk article selector to sample thematically coherent paper sets, thereby ensuring adequate semantic density and task complexity. (3) Multi-task evaluation of scientific reasoning. It contains over 2,000 QA pairs across reasoning, retrieval, summarization, and problem solving, enabling evaluation of multi-step scientific reasoning. Experimental results show that even advanced systems such as OpenAI Deep Research and Tongyi Deep Research achieve limited scores on PaperScope, highlighting the difficulty of long-context retrieval and deep multi-source reasoning. PaperScope thus provides a rigorous benchmark alongside a scalable pipeline for constructing large-scale multi-modal, multi-source deep research datasets.

[Long] PaT: Planning-after-Trial for Efficient Test-Time Code Generation ► *Underline Video*

Youngsik Yoon, Sungjae Lee, Seockbean Song, Siwei Wang, Wei Chen and Jungseul Ok

Beyond training-time optimization, scaling test-time computation has emerged as a key paradigm to extend the reasoning capabilities of Large Language Models (LLMs). However, most existing methods adopt a rigid Planning-before-Trial (PbT) policy, which inefficiently allocates test-time compute by incurring planning overhead even on directly solvable problems. We propose Planning-after-Trial (PaT), an adaptive policy for code generation that invokes a planner only upon verification failure. This adaptive policy naturally enables a heterogeneous model configuration: a cost-efficient model handles generation attempts, while a powerful model is reserved for targeted planning interventions. Empirically, across multiple benchmarks and model families, our approach significantly advances the cost-performance Pareto frontier. Notably, our heterogeneous configuration achieves performance comparable to a large homogeneous model while reducing inference cost by approximately 69%.

[SRW] Phase Transitions in Affective Meaning Divergence: The Hidden Drift Before the Break ► *Underline Video*

Napassorn Litchiowong;

One partner says "Fine" meaning resolution; the other hears surrender. The word is shared; the affective uptake is not. We formalize this as affective meaning divergence (AMD), the total-variation distance between interlocutors' anchor-conditioned affect distributions. Building on speech-act theory, common-ground accumulation, and entropy-regularized game theory, we derive a logit best-response map whose dynamics undergo a saddle-node bifurcation: when > 4, a monotone increase in AMD-driven load produces an abrupt, hysteretic collapse of repair coordination. On Conversations Gone Awry (CGA-Wiki; N=652), derailing conversations exhibit critical-slowing-down (CSD) signatures across multiple levels: lexical divergence variance ($p<0.001$, $d=0.36$), AMD variance ($p=0.001$, $d=0.26$), and dialog-act repair variance ($p=0.016$, $d=0.20$), all significant after correction and stronger than toxicity and sentiment baselines. AMD provides a distinct temporal signature, with retrospectively measured variance peaking at the bifurcation point while toxicity variance peaks earlier, and is the only indicator grounded in the theoretical framework. Boundary-condition analysis on CGA-CMV (N=1,169) yields mixed but directionally consistent evidence.

[Findings] PRA-RAG: Provably Robust Aggregation in Retrieval-Augmented Generation against Retrieval Corruption ► *Underline Video*

Xue Tan, Yi Zheng, Chang Huo, Yunruo Zhang, Yu Liu, Hao Luan, Zhuyang Yu, Xiaoyan Sun, Ping Chen and Jun Dai

Retrieval-Augmented Generation (RAG) enhances Large Language Models (LLMs) by incorporating external knowledge, effectively mitigating their inherent knowledge limitations. However, RAG remains vulnerable to poisoning attacks that manipulate retrieved texts to mislead model outputs. Existing defense mechanisms often lack theoretical robustness guarantees and perform unreliably when the LLM has limited knowledge of the retrieved content. In this work, we propose PRA-RAG, a provably robust retrieval aggregation algorithm designed to defend against poisoning attacks on retrieved texts. PRA-RAG samples multiple combinations of retrieved texts and utilizes geometric structures in the embedding space to identify a robust subset, from which a stable aggregated representation is derived. We provide theoretical bounds on the maximum impact of poisoned retrieved content and establish a quantitative measure of RAGs robustness. Experiments across multiple benchmarks and RAG architectures demonstrate that PRA-RAG reduces the attack success rate to as low as 1% while maintaining an accuracy of 71%, significantly outperforming representative state-of-the-art (SOTA) methods.

[Long] Preference Heads in Large Language Models: A Mechanistic Framework for Interpretable Personalization ▶ *Underline Video*

Weixu Zhang, Ye Yuan, Changjiang Han, Yuxing Tian, Zipeng Sun, Linfeng Du, Jikun Kang, Hong Kang, Xue Liu and Haolun Wu

Large Language Models (LLMs) exhibit strong implicit personalization ability, yet most existing approaches treat this behavior as a black box, relying on prompt engineering or fine tuning on user data. In this work, we adopt a mechanistic interpretability perspective and hypothesize the existence of a sparse set of Preference Heads, attention heads that encode user specific stylistic and topical preferences and exert a causal influence on generation. We introduce Differential Preference Steering (DPS), a training free framework that (1) identifies Preference Heads through causal masking analysis and (2) leverages them for controllable and interpretable personalization at inference time. DPS computes a Preference Contribution Score (PCS) for each attention head, directly measuring its causal impact on user aligned outputs. During decoding, we contrast model predictions with and without Preference Heads, amplifying the difference between personalized and generic logits to selectively strengthen preference aligned continuations. Experiments on widely used personalization benchmarks across multiple LLMs demonstrate consistent gains in personalization fidelity while preserving content coherence and low computational overhead. Beyond empirical improvements, DPS provides a mechanistic explanation of where and how personalization emerges within transformer architectures.

[SRW] Presentation Slide Translation and Layout Error Correction by LLMs ▶ *Underline Video*

Futo Kajita;Nobuyori Nishimura;Takehito Utsuro;Naoki Muto;Chee Siang Leow;Hiromitsu Nishizaki

We propose a novel approach to translating Japanese slides into English and to correcting their layout errors by utilizing multimodal LLMs with slide images and XML structures. Existing translation tools often suffer from layout errors after translation due to text expansion during the translation process, causing text to overlap with figures or other items in slides and thereby reducing readability. To overcome this issue, our proposed framework introduces two steps consisting of (i) translating text fragments within the slide, and (ii) correcting layout errors by optimizing layout placement based on visual consistency. In step (ii), we empirically show that few-shot prompts are quite effective in layout error correction. Given that the optimal combination of steps (i) and (ii) varies depending on the slide layout, our method generates eight different layout candidates. Consequently, we introduce a third step that automatically selects the optimal output from these eight candidates. The experimental results showed that the proposed method outperformed baselines and achieved 4.1% layout error rate and over 80% model selection success rate.

[Findings] Principled Detection of Hallucinations in Large Language Models via Multiple Testing ▶ *Underline Video*

Jiawei Li, Akshayaa Magesh and Venugopal Veeravalli

While Large Language Models (LLMs) have emerged as powerful foundational models to solve a variety of tasks, they have also been shown to be prone to hallucinations, i.e., generating responses that sound confident but are actually incorrect or even nonsensical. Existing hallucination detectors propose a wide range of empirical scoring rules, but their performance varies across models and datasets, and it is hard to determine which ones to rely on in practice or to treat as a reliable detector. In this work, we formulate the problem of detecting hallucinations as a hypothesis testing problem and draw parallels with the problem of out-of-distribution detection in machine learning models. We then propose a multiple-testing-inspired method that systematically aggregates multiple evaluation scores via conformal p-values, enabling calibrated detection with controlled false alarm rate. Extensive experiments across diverse models and datasets validate the robustness of our approach against state-of-the-art methods.

[Findings] Prior Beliefs Prejudice LLM-as-Judge: Evidence from Persuasion Evaluation ▶ *Underline Video*

Pardis Sadat Zahraei, Xiaoning Wang, Nimet Beyza Bozdag, Gokhan Tur and Dilek Hakkani-Tür

Large Language Models (LLMs) are increasingly used as judges to evaluate text quality, moderate content, and assess arguments. We investigate whether alignment-instilled prior beliefs bias LLM judgments, using persuasion evaluation as a representative task. We find a systematic failure: models conflate their trained beliefs with rhetorical quality, rating identical claims differently based on belief alignment rather than argumentative merit. A bare assertion aligned with training receives higher scores than a well-crafted counter-argument, even when explicitly instructed to judge rhetoric alone. We introduce ConvinceQA, a dataset of 27,756 persuasive arguments with controlled stance variation across subjective, harmful, and misinformation domains, and demonstrate this prior prejudice across models. We exploit this failure through persuasion-based probing: evaluating minimal pairs that differ only in the subject token bypasses learned refusals and reveals hidden biases. Analysis identifies three failure modes, with belief-conditioned rating inflation accounting for 88% of cases. Cross-task validation on essay quality assessment and debate judging confirms this is a pervasive limitation.

[Long] Privacy Collapse: Benign Fine-Tuning Can Break Contextual Privacy in Language Models ▶ *Underline Video*

Anmol Goel, Cornelius Emde, Sangdoo Yun, Seong Joon Oh and Martin Gubri

We identify a novel phenomenon in language models: benign fine-tuning of frontier models can lead to privacy collapse. We find that diverse, subtle patterns in training data can degrade contextual privacy, including optimisation for helpfulness, exposure to user information, emotional and subjective dialogue, and debugging code printing internal variables, among others. Finetuned models lose their ability to reason about contextual privacy norms, share information inappropriately with tools, and violate memory boundaries across contexts. Privacy collapse is a silent failure because models maintain high performance on standard safety and utility benchmarks whilst exhibiting severe privacy vulnerabilities. Our experiments show evidence of privacy collapse across six models (closed and open weight), five fine-tuning datasets (real-world and controlled data), and two task categories (agentic and memory-based). Our mechanistic analysis reveals that privacy representations are uniquely fragile to fine-tuning, compared to task-relevant features which are preserved. Our results reveal a critical gap in current safety evaluations, in particular for the deployment of specialised agents.

[Long] Privacy-R1: Privacy-Aware Multi-LLM Agent Collaboration via Reinforcement Learning ▶ *Underline Video*

Zheng Hui, Yijiang River Dong, Sanhanat Sivapiromrat, Ehsan Shareghi and Nigel Collier

When users submit queries to Large Language Models (LLMs), their prompts can often contain sensitive data, forcing a difficult choice: Send the query to a powerful proprietary LLM providers to achieving state-of-the-art performance and risk data exposure, or relying on smaller, local models guarantees data privacy but often results in a degradation of task performance. Prior approaches have relied on static pipelines that use LLM rewriting, which shatters linguistic coherence and indiscriminately removes privacy-sensitive information, including task-critical content. We reformulate this challenge (Privacy-Conscious Delegation) as a sequential decision-making problem and introduce a novel reinforcement learning (RL) framework called Privacy-R1 to solve it. Our framework trains an agent to dynamically route text chunks, learning a policy that optimally balances the trade-off between privacy leakage and task performance. It implicitly distinguishes between replaceable Personally Identifiable Information (PII) (which it shields locally) and task-critical PII (which it strategically sends to the remote model for maximal utility). To validate our approach in complex scenarios, we also introduce a new medical dataset with high PII density. Our framework achieves a new state-of-the-art on the privacy-utility frontier, demonstrating the necessity of learned, adaptive policies for deploying LLMs in sensitive environments. Dataset can be found at: https://github.com/zackhuiiiii/Privacy-R1.

[SRW] Probing Bias Formation in Medical LLMs through Activation Steering ▶ *Underline Video*

Bayram Ayadi; Annette Hautli-Janisz

Large Language Models specialized for the medical domain achieve high performance on static benchmarks, but remain vulnerable to sycophantic confabulation, where the models generate medically spurious rationales to justify incorrect user hints. This robustness gap poses severe risks in clinical environments, as models may prioritize contextual faithfulness to a biased prompt over their internal parametric medical knowledge. This study introduces a mechanistic approach to identify and mitigate these failures in MedGemma-27B, isolating hint integration circuits using Sparse Autoencoders and geometric manifold analysis. Our findings reveal that sycophantic bias is a highly distributed and polymorphic concept, with biased reasoning routed through shifting dimensions across transformer layers. We identify a specific causal bottleneck for bias formation and demonstrate that cluster-conditioned dynamic steering tailored to the geometric subspace of the prompt outperforms static global interventions, though it reveals a fundamental tension between bias resilience and the retention of internal parametric knowledge.This work proposes a principled framework toward clinical AI systems that are more robust and aligned with expert medical logic, demonstrating the potential of cluster-conditioned geometric interventions while characterizing the inherent trade-offs in clinical knowledge retention.

[Findings] Program Structure-aware Language Models: Targeted Software Testing beyond Textual Semantics ▶ *Underline Video*

Khang Tran, Khoa Nguyen, Cristian Borcea and Hai Phan

Recent advances in large language models for test case generation have improved branch coverage via prompt-engineered mutations. However, they still lack principled mechanisms for steering models toward specific high-risk execution branches, limiting their effectiveness for discovering subtle bugs and security vulnerabilities. We propose GLMTest, the first program structure-aware LLM framework for targeted test case generation that seamlessly integrates code property graphs and code semantics using a graph neural network and a language model to condition test case generation on execution branches. This structured conditioning enables controllable and branch-targeted test case generation, thereby potentially enhancing bug and security risk discovery. Experiments on real-world projects show that GLMTest built on a Qwen2.5-Coder-7B-Instruct model improves branch accuracy from 27.4% to 50.2% on TestGenEval benchmark compared with state-of-the-art LLMs, i.e., Claude-Sonnet-4.5 and GPT-4o-mini.

[Long] Progressive Multimodal Search and Reasoning for Knowledge-Intensive Visual Question Answering ▶ *Underline Video*

Changin Choi, Wonseok Lee, Jungmin Ko and Wonjong Rhee

Knowledge-intensive visual question answering (VQA) requires external knowledge beyond image content, demanding precise visual grounding and coherent integration of visual and textual information. Although multimodal retrieval-augmented generation has achieved notable advances by incorporating external knowledge bases, existing approaches largely adopt single-pass frameworks that often fail to acquire sufficient knowledge and lack mechanisms to revise misdirected reasoning. We propose PMSR (Progressive Multimodal Search and Reasoning), a framework that progressively constructs a structured reasoning trajectory to enhance both knowledge acquisition and synthesis. PMSR uses dual-scope queries conditioned on both the latest record and the trajectory to retrieve diverse knowledge from heterogeneous knowledge bases. The retrieved evidence is then synthesized into compact records via compositional reasoning. This design facilitates controlled iterative refinement, which supports more stable reasoning trajectories with reduced error propagation. Extensive experiments across six diverse benchmarks (Encyclopedic-VQA, InfoSeek, MMSearch, LiveVQA, FVQA, and OK-VQA) demonstrate that PMSR consistently improves both retrieval recall and end-to-end answer accuracy.

[SRW] Proofs as Trajectories: Learning Lean Proof-Step Representations from State Changes ▶ *Underline Video*

Elisaveta Samoylov and Soroush Vosoughi;

Lean proofs are built as sequences of tactic-induced state transitions, yet learned models often represent proof steps primarily through tactic strings or raw proof-state text. Building on Delta Tokens, which encode a proof step by the local edit it induces between successive proof states, we train an encoder-only Transformer to learn contextualized representations of Lean proof steps from state changes. We then use these step representations to study complete proofs as trajectories in a learned latent space. We first show that the Delta-based Transformer

yields better held-out next-tactic retrieval than a matched surface-syntax control, supporting the representational choice used in the trajectory analysis. We then analyze proof trajectories using path length, endpoint span, directness, curvature, and torsion. Across the LeanWorkbook slice used here, longer proofs become increasingly indirect within a relatively bounded latent span: path length grows sharply with proof length while endpoint span changes little, mean step size decreases, curvature rises modestly, and torsion falls. Qualitative case studies show that these geometric patterns align with recognizable proof organizations, including immediate closure, aligned accumulation, scaffolded enrichment, bookkeeping-heavy restructuring, and repeated local contradiction work. The dataset is small and heavily skewed toward short proofs, so the claims are necessarily limited. Within those limits, the results suggest that learned state-change representations recover nontrivial structure in how proofs unfold and provide a promising basis for future trajectory-aware theorem proving.

[Long] R^3AG: Retriever Routing for Retrieval-Augmented Generation ▶ *Underline Video*

Tong Zhao, Yutao Zhu, Yucheng Tian and Zhicheng Dou

Retrieval-augmented generation (RAG) has become a cornerstone for knowledge-intensive tasks. However, the efficacy of RAG is often bottlenecked by the "one-size-fits-all" retrieval paradigm, as different queries exhibit distinct preferences for different retrievers. While recent routing techniques attempt to select the optimal retriever dynamically, they typically operate under a 'single and static capability' assumption, selecting retrievers solely based on semantic relevance. This overlooks a critical distinction in RAG: a retrieved document must not only be relevant but also effectively support the generator in producing correct answers. To address this limitation, we propose RşAG, a novel routing framework that explicitly models the dynamic alignment between queries and retriever capabilities. Unlike previous approaches, RşAG decomposes retriever capability into two learnable dimensions: retrieval quality and generation utility. We employ a contrastive learning objective that leverages complementary supervision signals, i.e., document assessments and downstream answer correctness, to capture query-specific preference shifts. Extensive experiments on diverse knowledge-intensive tasks demonstrate that RşAG consistently outperforms both the best individual retrievers and state-of-the-art static routing methods.

[SRW] RAQE: Reranker-Aligned Query Expansion via Label-Free Group-Relative Policy Optimization ▶ *Underline Video*

Gyeonghun Sun, Jeonghwan Choi and Hwanjun Song;

Retrieval-augmented generation is often bottlenecked by retrieval quality, motivating query expansion. We propose RAQE (Reranker-Aligned Query Expansion), a label-free framework that trains a compact generator (3B/8B) via policy optimization using a reranker-derived reward. RAQE replaces relevance labels with Reranker-Shaped Discounted Gain (RSDG), computed from reranker scores over the retriever's top-k documents. For query expansion, RAQE generates a single, short expansion in one pass, avoiding long outputs and repeated generation while remaining low latency. Across eight IR benchmarks, RAQE is competitive with strong large language model prompting baselines and narrows the gap to supervised reinforcement learning methods without relevance labels.

[Findings] Rational Synthesizers or Heuristic Followers? Analyzing LLMs in RAG-based Question-Answering ▶ *Underline Video*

Atharv Naphade

Retrieval-Augmented Generation (RAG) is the prevailing paradigm for grounding Large Language Models (LLMs), yet the mechanisms governing how models integrate groups of conflicting retrieved evidence remain opaque. Does an LLM answer a certain way because the evidence is factually strong, because of a prior belief, or merely because it is repeated frequently? To answer this, we introduce GroupQA, a curated dataset of 1,635 controversial questions paired with 15,058 diversely-sourced evidence documents, annotated for stance and qualitative strength. Through controlled experiments, we characterize group-level evidence aggregation dynamics: Paraphrasing an argument can be more persuasive than providing distinct independent support; Models favor evidence presented first rather than last, and Larger models are increasingly resistant to adapt to presented evidence. Additionally, we find that LLM explanations to group-based answers are unfaithful. Together, we show that LLMs behave consistently as vulnerable heuristic followers, with direct implications for improving RAG system design.

[Findings] Reasoning-Based Refinement of Unsupervised Text Clusters with LLMs ▶ *Underline Video*

Tunazzina Islam

Unsupervised methods are widely used to induce latent semantic structure from large text collections, yet their outputs often contain incoherent, redundant, or poorly grounded clusters that are difficult to validate without labeled data. We propose a reasoning-based refinement framework that leverages large language models (LLMs) not as embedding generators, but as semantic judges that validate and restructure the outputs of arbitrary unsupervised clustering algorithms. Our framework introduces three reasoning stages: (i) coherence verification, where LLMs assess whether cluster summaries are supported by their member texts; (ii) redundancy adjudication, where candidate clusters are merged or rejected based on semantic overlap; and (iii) label grounding, where clusters are assigned interpretable labels through a two-stage process that generates and consolidates semantically similar labels in a fully unsupervised manner. This design decouples representation learning from structural validation and mitigates the common failure modes of embedding-only approaches. We evaluate the framework in real-world social media corpora from two platforms with distinct interaction models, demonstrating consistent improvements in cluster coherence and human-aligned labeling quality over classical topic models and recent representation-based baselines. Human evaluation shows strong agreement with LLM-generated labels, despite the absence of gold-standard annotations. We further conduct robustness analysis under matched temporal and volume conditions to assess cross-platform stability. Beyond empirical gains, our results suggest that LLM-based reasoning can serve as a general mechanism for validating and refining unsupervised semantic structure, enabling more reliable and interpretable analysis of large text collections without supervision.

[SRW] RECON: Benchmarking Agent Memory for Compositional Reasoning over Long Contexts ▶ *Underline Video*
Mihir Arya
Large language models and LLM-based agents are widely deployed as personal chat assistants, enterprise copilots, and autonomous workflow agents. Across all these applications, memory (the ability to retain, access, and reason over information accumulated over long contexts and multiple interactions) plays a crucial role in determining the reliability of any agent. We introduce RECON (Reasoning over Extended Contexts with Obfuscated Narratives), a benchmark for evaluating compositional temporal reasoning over long contexts. RECON uses investigation case file (100k tokens) and tests agents on six memory-intensive tasks: reconstructing multi-hop evidence chains, propagating cascading invalidations, resolving source conflicts, counterfactual reasoning, satisfying temporal constraints, and temporal fact retrieval. Recent memory benchmarks evaluate whether agents can retrieve scattered facts or detect if a fact has changed; RECON evaluates what happens after the change, whether agents can trace which downstream conclusions are affected, which survive through independent support, and how alternative timelines would have unfolded. We evaluate systems across three architectural families comprising long-context LLMs, retrieval-augmented generation (RAG) agents, and commercial memory agents and find that existing memory mechanisms are not yet sufficient for reliable reasoning over long, evolving narratives. Our benchmark and code are available at https://anonymous.4open.science/r/recon-benchmark/README.md.

[Long] ReCoQA: A Benchmark for Tool-Augmented and Multi-Step Reasoning in Real Estate Question and Answering ▶ *Underline Video*
Yindong Zhang, Wenmian Yang, Yiquan Zhang and Weijia Jia
Developing agents capable of navigating fragmented, multi-source information remains challenging, primarily due to the scarcity of benchmarks reflecting hybrid workflows combining database querying with external APIs. To bridge this gap, we introduce ReCoQA, a large-scale benchmark of 29,270 real-estate instances featuring machine-verifiable supervision for intermediate steps, including structured intent labels, SQL queries, and API calls. Complementarily, we propose HIRE-Agent, a hierarchical framework instantiating an understand-plan-execute architecture as a strong baseline. By orchestrating a Front-end parser, a planning Supervisor, and execution Specialists, HIRE-Agent effectively integrates heterogeneous evidence. Extensive experiments demonstrate that HIRE-Agent constitutes a strong baseline and substantiates the necessity of hierarchical collaboration for complex, real-world reasoning tasks.

[Long] Red-Bandit: Test-Time Adaptation for LLM Red-Teaming via Bandit-Guided LoRA Experts ▶ *Underline Video*
Christos Ziakas, Nicholas Loo, Nishita Jain and Alessandra Russo
Automated red-teaming has emerged as a scalable approach for auditing Large Language Models (LLMs) prior to deployment, yet existing approaches lack mechanisms to efficiently adapt to model-specific vulnerabilities at inference. We introduce Red-Bandit, a red-teaming framework that adapts online to identify and exploit model failure modes under distinct attack styles (e.g., manipulation, slang). Red-Bandit post-trains a set of parameter-efficient LoRA experts, each specialized for a particular attack style, using reinforcement learning that rewards the generation of unsafe prompts via a rule-based safety model. At inference, a multi-armed bandit policy dynamically selects among these attack-style experts based on the target models response safety, balancing exploration and exploitation. Red-Bandit outperforms state-of-the-art methods on AdvBench and HarmBench, achieving higher attack success rates under sufficient exploration budgets (ASR@10), while generating more human-readable adversarial prompts (lower perplexity). In addition, Red-Bandits bandit policy serves as a diagnostic tool for identifying model-specific vulnerabilities by indicating which attack styles most effectively elicit harmful behaviors.

[SRW] Reference-Free Schema Generation for Literature Review Tables via Multi-Faceted Rewards ▶ *Underline Video*
Sinjoy Saha; Suman Saha; Mahfuza Farooque; Wenpeng Yin
To accelerate scientific knowledge acquisition, LLMs are increasingly used to synthesize multiple papers into structured tables by inferring schemas and values. While value generation within a fixed schema can often be reduced to extractive question answering, the schema generation problem, determining which dimensions to compare a set of documents, lacks a formal mapping to standard NLP tasks. In this work, we formulate schema generation as a reinforcement learning problem and investigate whether these dimensions can be induced without access to gold-standard schemas. We design a multi-faceted reward framework capturing schema coverage, non-redundancy, relevance, and format, and train a small language model on a literature review dataset. Our approach yields consistent improvements over the untuned base model across intrinsic, reference-based, and LLM-judge metrics, and remains competitive with supervised fine-tuned models at 5x the parameter count on structural and diversity dimensions.

[Long] Revisiting Non-Verbatim Memorization in Large Language Models: The Role of Entity Surface Forms ▶ *Underline Video*
Yuto Nishida, Naoki Shikoda, Yosuke Kishinami, Ryo Fujii, Makoto Morishita, Hidetaka Kamigaito and Taro Watanabe
Understanding what kinds of factual knowledge large language models (LLMs) memorize is essential for evaluating their reliability and limitations. Entity-based QA is a common framework for analyzing non-verbatim memorization, but typical evaluations query each entity using a single canonical surface form, making it difficult to disentangle fact memorization from access through a particular name. We introduce RedirectQA, an entity-based QA dataset that uses Wikipedia redirect information to associate Wikidata factual triples with categorized surface forms for each entity, including alternative names, abbreviations, spelling variants, and common erroneous forms. Across 13 LLMs, we examine surface-conditioned factual memorization and find that prediction outcomes often change when only the entity surface form changes. This inconsistency is category-dependent: models are more robust to minor orthographic variations than to larger lexical

variations such as aliases and abbreviations. Frequency analyses further suggest that both entity- and surface-level frequencies are associated with accuracy, and that entity frequency often contributes beyond surface frequency. Overall, factual memorization appears neither purely surface-specific nor fully surface-invariant, highlighting the importance of surface-form diversity in evaluating non-verbatim memorization.

[Findings] RLHS: Mitigating Misalignment in RLHF with Hindsight Simulation ▶ *Underline Video*

Kaiqu Liang, Haimin Hu, Ryan Liu, Thomas L. Griffiths and Jaime Fernández Fisac

While Reinforcement Learning from Human Feedback (RLHF) has shown promise in aligning generative AI, we present empirical evidence that it can also cause severe, systematic misalignment. We hypothesize that this stems from evaluator feedback depending on downstream outcome predictions (foresight) that can be influenced by the AI's output, inducing Goodharts law dynamics. We present a theoretical analysis showing that conditioning evaluator feedback on downstream observations (hindsight) inhibits this effect by decoupling the alignment signal from potentially compromised predictions, crucially, the result holds even if the observed outcomes are sampled from the AI's own world model. Building on this insight, we introduce Reinforcement Learning from Hindsight Simulation (RLHS), which presents plausible simulated outcomes to evaluators before eliciting feedback. We validate RLHS across three consultancy settings, marketplace interactions, restaurant recommendations, and online course advising, using both online (PPO) and offline (DPO) fine-tuning methods, and show that it substantially improves alignment over RLHF in experiments and human evaluations. We perform post-hoc benchmark evaluations on TruthfulQA, HaluEval, and TrustLLM, finding that even after single-task fine-tuning, RLHF misalignment persists, while RLHS consistently outperforms baselines and demonstrates strong out-of-domain generalization.

[Findings] Role-Sensitive Neurons: A Neuron-Level Gain Control Mechanism for Confidence Steering ▶ *Underline Video*

Peiwen Huang, Chih-Hao Hsu, Tzu-Hung Huang and Shou-De Lin

Role-playing prompts effectively steer Large Language Models (LLMs), yet the neural mechanism driving this behavioral shift remains unclear. In this work, we identify Role-Sensitive Neurons (RSNs), a sparse sub-network (approx 0.5% of all neurons) governing the transition from hesitation to action. Using a novel evaluation framework with explicit abstention (MMLU-E), we reveal a Confidence-Performance Decoupling: roles primarily modulate the model's probabilistic "willingness to act" rather than its underlying knowledge representation. We demonstrate that RSNs function as a mechanistic gain control system: causal intervention on this subspace allows precise regulation of abstention behavior. Furthermore, cross-model transfer experiments confirm that these circuits are indigenous to pre-training, with Instruction Tuning (SFT) acting merely as a "signal sharpener" to refine latent gain dynamics. Finally, we identify a critical safety boundary: in knowledge-deficient models, amplifying RSNs induces "unwarranted certainty," highlighting decisiveness as a tunable gain parameter distinct from epistemic truth.

[Findings] RubRIX: Rubric-Driven Risk Mitigation in Caregiver-AI Interactions ▶ *Underline Video*

Drishti Goel, Jeongah Lee, Qiuyue Zhong, Violeta J. Rodriguez, Daniel S. Brown, Ravi Karkar, Dong Whi Yoo and Koustuv Saha

Caregivers seeking AI-mediated support express complex needs—information-seeking, emotional validation, and distress cues—that warrant careful evaluation of response safety and appropriateness. Existing AI evaluation frameworks, primarily focused on general risks (toxicity, hallucinations, policy violations, etc) may not adequately capture the nuanced risks of LLM-responses in caregiving-contexts. We introduce RubRIX (Rubric-based Risk Index), a theory-driven, clinician-validated framework for evaluating risks in LLM caregiving responses. Grounded in the Elements of an Ethic of Care, RubRIX operationalizes five empirically-derived risk dimensions: Inattention, Bias & Stigma, Information Inaccuracy, Uncritical Affirmation, and Epistemic Arrogance. We evaluate six state-of-the-art LLMs on over 20,000 caregiver queries from Reddit and ALZConnected. Rubric-guided refinement consistently reduced risk-components by 45-98% after one iteration across models. This work contributes a methodological approach for developing domain-sensitive, user-centered evaluation frameworks for high-burden contexts. Our findings highlight the importance of domain-sensitive, interactional risk evaluation for the responsible deployment of LLMs in caregiving support contexts. We release benchmark datasets to enable future research on contextual risk evaluation in AI-mediated support.

[Long] SafeMT: Multi-turn Safety for Multimodal Language Models ▶ *Underline Video*

Han Zhu, Juntao Dai, Jiaming Ji, Haoran Li, Chengkun Cai, Pengcheng Wen, Chi-Min Chan, Boyuan Chen, Yaodong Yang, Sirui Han and Yike Guo

With the widespread use of multi-modal Large Language models (MLLMs), safety issues have become a growing concern. Multi-turn dialogues, which are more common in everyday interactions, pose a greater risk than single prompts; however, existing benchmarks do not adequately consider this situation. To encourage the community to focus on the safety issues of these models in multi-turn dialogues, we introduce SafeMT, a benchmark that features dialogues of varying lengths generated from harmful queries accompanied by images. This benchmark consists of 10,000 samples in total, encompassing 17 different scenarios and four jailbreak methods. Additionally, we propose Safety Index (SI) to evaluate the general safety of MLLMs during conversations. We assess the safety of 17 models using this benchmark and discover that the risk of successful attacks on these models increases as the number of turns in harmful dialogues rises. This observation indicates that the safety mechanisms of these models are inadequate for recognizing the hazard in dialogue interactions. We propose a dialogue safety moderator capable of detecting malicious intent concealed within conversations and providing MLLMs with relevant safety policies. Experimental results from several open-source models indicate that this moderator is more effective in reducing multi-turn Attack Success Rate (ASR) compared to existed guard models.

[Long] SAM3-I: Segment Anything with Instructions ▶ *Underline Video*

Jingjing Li, Yue Feng, Yuchen Guo, Jincai Huang, Wei Ji, Qi Bi, Yongri Piao, Miao Zhang, Xiaoqi Zhao, Qiang Chen, Shihao Zou, Huchuan Lu and Li Cheng

Segment Anything Model 3 (SAM3) advances open-vocabulary segmentation through promptable concept segmentation, enabling users to segment all instances associated with a given concept using short noun-phrase (NP) prompts. While effective for concept-level grounding, real-world interactions often involve far richer natural-language instructions that combine attributes, relations, actions, states, or implicit reasoning. Currently, SAM3 relies on external multi-modal agents to convert complex instructions into NPs and conducts iterative mask filtering, leading to coarse representations and limited instance specificity. In this work, we present SAM3-I, an instruction-following extension of the SAM family that unifies concept-level grounding and instruction-level reasoning within a single segmentation framework. Built upon SAM3, SAM3-I introduces an instruction-aware cascaded adaptation mechanism with dedicated alignment losses that progressively aligns expressive instruction semantics with SAM3's vision-language representations, enabling direct interpretation of natural-language instructions while preserving its strong concept recall ability. To enable instruction-following learning, we introduce HMPL-Instruct, a large-scale instruction-centric dataset that systematically covers hierarchical instruction semantics and diverse target granularities. Experiments demonstrate that SAM3-I achieves appealing performance across referring and reasoning-based segmentation, showing that SAM3 can be effectively extended to follow complex natural-language instructions without sacrificing its original concept-driven strengths. Code and dataset are available at https://github.com/debby-0527/SAM3-I.

[Long] SceneAlign: Aligning Multimodal Reasoning to Scene Graphs in Complex Visual Scenes ▶ *Underline Video*

Chuhan Wang, Xintong Li, Jennifer Yuntong Zhang, Junda Wu, Chengkai Huang, Lina Yao, Julian McAuley and Jingbo Shang

Multimodal large language models often struggle with faithful reasoning in complex visual scenes, where intricate entities and relations require precise visual grounding at each step. This reasoning unfaithfulness frequently manifests as hallucinated entities, mis-grounded relations, skipped steps, and over-specified reasoning. Existing preference-based approaches, typically relying on textual perturbations or answer-conditioned rationales, fail to address this challenge as they allow models to exploit language priors to bypass visual grounding. To address this, we propose SceneAlign, a framework that leverages scene graphs as structured visual information to perform controllable structural interventions. By identifying reasoning-critical nodes and perturbing them through four targeted strategies that mimic typical grounding failures, SceneAlign constructs hard negative rationales that remain linguistically plausible but are grounded in inaccurate visual facts. These contrastive pairs are used in Direct Preference Optimization to steer models toward fine-grained, structure-faithful reasoning. Across seven visual reasoning benchmarks, SceneAlign consistently improves answer accuracy and reasoning faithfulness, highlighting the effectiveness of grounding-aware alignment for multimodal reasoning.

[Findings] SCMAPR: Self-Correcting Multi-Agent Prompt Refinement for Complex-Scenario Text-to-Video Generation ▶ *Underline Video*

Chengyi Yang, Pengzhen Li, Jiayin Qi, Aimin Zhou, Ji Wu and Ji Liu

Text-to-Video (T2V) generation has benefited from recent advances in diffusion models, yet current systems still struggle under complex scenarios, which are generally exacerbated by the ambiguity and underspecification of text prompts. In this work, we formulate complex-scenario prompt refinement as a stage-wise multi-agent refinement process and propose SCMAPR, i.e., a scenario-aware and Self-Correcting Multi-Agent Prompt Refinement framework for T2V prompting. SCMAPR coordinates specialized agents to (i) route each prompt to a taxonomy-grounded scenario for strategy selection, (ii) synthesize scenario-aware rewriting policies and perform policy-conditioned refinement, and (iii) conduct structured semantic verification that triggers conditional revision when violations are detected. To clarify what constitutes complex scenarios in T2V prompting, provide representative examples, and enable rigorous evaluation under such challenging conditions, we further introduce T2V-Complexity, which is a complex-scenario T2V benchmark consisting exclusively of complex-scenario prompts. Extensive experiments on 3 existing benchmarks and our T2V-Complexity benchmark demonstrate that SCMAPR consistently improves text-video alignment and overall generation quality under complex scenarios, achieving up to 2.67% and 3.28 gains in average score on VBench and EvalCrafter, and up to 0.028 improvement on T2V-CompBench over 3 State-Of-The-Art baselines. The codes of SCMAPR are publicly available at https://github.com/HiThink-Research/SCMAPR.

[Findings] SCURank: Ranking Multiple Candidate Summaries with Summary Content Units for Enhanced Summarization ▶ *Underline Video*

Bo-Jyun Wang, Ying-Jia Lin and Hung-Yu Kao

Small language models (SLMs), such as BART, can achieve summarization performance comparable to large language models (LLMs) via distillation. However, existing LLM-based ranking strategies for summary candidates suffer from instability, while classical metrics (e.g., ROUGE) are insufficient to rank high-quality summaries. To address these issues, we introduce SCURank, a framework that enhances summarization by leveraging Summary Content Units (SCUs). Instead of relying on unstable comparisons or surface-level overlap, SCURank evaluates summaries based on the richness and semantic importance of information content. We investigate the effectiveness of SCURank in distilling summaries from multiple diverse LLMs. Experimental results demonstrate that SCURank outperforms traditional metrics and LLM-based ranking methods across evaluation measures and datasets. Furthermore, our findings show that incorporating diverse LLM summaries enhances model abstractiveness and overall distilled model performance, validating the benefits of information-centric ranking in multi-LLM distillation.

[Long] SDE-SQL: Enhancing Text-to-SQL Generation in Large Language Models via Self-Driven Exploration with SQL Probes ▶ *Underline Video*

Wenxuan Xie, Yaxun Dai and Wenhao Jiang

Recent advances in large language models (LLMs) have led to substantial progress on the Text-to-SQL task. However, existing approaches typically depend on static, pre-processed database information supplied at inference time, which restricts the models capacity to deeply comprehend the underlying database content. In the absence of dynamic interaction, LLMs are limited to fixed, human-curated context and lack the ability to autonomously query or explore the data. To overcome this limitation, we introduce SDE-SQL, a novel framework that empowers LLMs to perform Self-Driven Exploration of databases during inference. This is achieved through the generation and execution of SQL probes, enabling the model to actively retrieve information and iteratively refine its understanding of the database. Unlike prior methods, SDE-SQL operates in a zero-shot setting, requiring no in-context demonstrations or question-SQL pairs. Evaluated on the BIRD benchmark with Qwen2.5-72B-Instruct, SDE-SQL achieves an 8.02% relative improvement in execution accuracy over the vanilla Qwen2.5-72B-Instruct baseline, establishing a new state-of-the-art among open-source methods without supervised fine-tuning (SFT) or model ensembling. Furthermore, when combined with SFT, SDE-SQL delivers an additional 0.52% performance gain.

[Long] SecureVibeBench: Benchmarking Secure Vibe Coding of AI Agents via Reconstructing Vulnerability-Introducing Scenarios ▶ *Underline Video*

Junkai Chen, Huihui Huang, Yunbo Lyu, Junwen An, Jieke Shi, Chengran Yang, Ting Zhang, Haoye Tian, Yikun Li, Zhenhao Li, Xin Zhou, Xing Hu and David Lo

Large language model-powered code agents are rapidly transforming software engineering, yet the security risks of their generated code have become a critical concern. Existing benchmarks have provided valuable insights, but they fail to capture scenarios in which vulnerabilities are actually introduced by human developers, making fair comparisons between humans and agents infeasible. We therefore introduce SecureVibeBench, a benchmark of 105 C/C++ secure coding tasks sourced from 41 projects in OSS-Fuzz for code agents. SecureVibeBench has the following features: (i) realistic task settings that require multi-file edits in large repositories, (ii) aligned contexts based on real-world open-source vulnerabilities with precisely identified vulnerability introduction points, and (iii) comprehensive evaluation that combines functionality testing and security checking with both static and dynamic oracles. We evaluate 5 popular code agents like OpenHands, supported by 5 LLMs (e.g., Claude sonnet 4.5) on SecureVibeBench. Results show that current agents struggle to produce both correct and secure code, as even the best-performing one, produces merely 23.8% correct and secure solutions on SecureVibeBench.

[Findings] Selective Steering: Norm-Preserving Control Through Discriminative Layer Selection ▶ *Underline Video*

Quy-Anh Dang and Chris Ngo

Despite significant progress in alignment, large language models (LLMs) remain vulnerable to adversarial attacks that elicit harmful behaviors. Activation steering techniques offer a promising inference-time intervention approach, but existing methods suffer from critical limitations: activation addition requires careful coefficient tuning and is sensitive to layer-specific norm variations, while directional ablation provides only binary control. Recent work on Angular Steering introduces continuous control via rotation in a 2D subspace, but its practical implementation violates norm preservation, causing distribution shift and generation collapse, particularly in models below 7B parameters. We propose Selective Steering, which addresses these limitations through two key innovations: (1) a mathematically rigorous norm-preserving rotation formulation that maintains activation distribution integrity, and (2) discriminative layer selection that applies steering only where feature representations exhibit opposite-signed class alignment. Experiments across nine models demonstrate that Selective Steering achieves 5.5 higher attack success rates than prior methods while maintaining zero perplexity violations and approximately 100% capability retention on standard benchmarks. Our approach provides a principled, efficient framework for controllable and stable LLM behavior modification.

[Long] Semantic Hardness Is Not Visual Hardness: Sign-Aware Hard Negative Mining for Sign Language Retrieval ▶ *Underline Video*

Junmyeong Lee, Chan Hur, Changsu Choi, Sukmin Cho, Fitsum Gaim, Eui Jun Hwang, Hoyun Song and KyungTae Lim

Sign Language Retrieval (SLRet) enables efficient access to sign language content but remains fragile in fine-grained scenarios where visually similar signs must be distinguished. We show that this limitation does not stem from model capacity, but from ineffective hard negative supervision. Specifically, we formulate fine-grained retrieval failures as a negative distribution mismatch: semantically distinct yet visually confusable signs are rarely treated as hard negatives, while existing text-based mining strategies fail to capture such visual ambiguity. To address this issue, we propose Sign-Aware Hard Negative Mining (SAN), which constructs hard negatives based on visual confusability in the sign embedding space rather than linguistic similarity. Experiments on PHOENIX-2014T demonstrate that SAN substantially improves fine-grained retrieval performance while preserving coarse-grained accuracy, highlighting the importance of aligning negative supervision with visual ambiguity in sign language retrieval.

[SRW] Semantic Span Annotation: An Exploratory Study of LLM Annnotation ▶ *Underline Video*

Tejas Goyal, Dhriti Krishnan, Anuj Gupta and Jaromir Savelka

Structured span extraction research is siloed by context length, annotation task, and domain, making it difficult to assess how well large language models (LLMs) generalize across realistic extraction settings. We introduce SSA (Structured Span Annotation), a unified evaluation framework bringing together five datasets across four domains: finance, biomedicine, affective analysis, and privacy, under a common JSONL format with character-level offsets. We conduct an exploratory study evaluating seven models (three closed, four

open-weight) under three prompting configurations: zero-shot, definition-augmented, and few-shot, formulating extraction as inline XML generation where models reproduce the document with tagged spans. Our results reveal two distinct performance regimes: on tasks requiring complex ontology reasoning, zero-shot performance is near zero (e.g., 0.00% F1 on FiNER-139) but improves substantially with label definitions (e.g., Claude Opus 4.6 rises from 8.8% to 57.5% F1); on pattern-based tasks like PII detection, definitions consistently hurt performance across all models. These findings suggest that prompting strategy must be matched to task structure, and that unified evaluation frameworks spanning varied domains and input lengths are essential for understanding LLM extraction capabilities.

[SRW] Sentiment Analysis of Yelp Review Dataset: A Comparative Study of Machine Learning Methods ▶ *Underline Video*
Krishna Thakar, Dr. Mohamed Abu Sheha and Dr. Emmanuel Thompson
Sentiment analysis involves analyzing text to determine whether the sentiment expressed is positive, negative, or neutral. In the context of online reviews, such as those on Yelp, sentiment analysis helps businesses assess customer satisfaction and identify areas for improvement. Given the large volume of user-generated content, restaurants often struggle to extract actionable insights from feedback, making sentiment analysis an efficient tool for categorizing reviews and highlighting customer concerns. This study focuses on sentiment analysis of Yelp reviews. The main research question is: How can Natural Language Processing (NLP) combined with statistical machine learning methods be applied to classify sentiment in Yelp reviews and provide actionable insights for improving customer satisfaction, service quality, and business performance? The study used 21,000 Yelp reviews, utilizing NLP approaches - tokenization, stop-word removal, and vectorization. Comparative classification predictive modeling and analysis were done across traditional machine learning (Logistic Regression, Support Vector Machine (SVM), Naïve Bayes, Random Forest), deep learning methods (CNN, LSTM, BiLSTM, GRU, RNN), and an advanced transformer-based (RoBERTa) model. Results showed that RoBERTa outperformed the other candidate methods. These findings highlight the potential of advanced NLP techniques to offer businesses practical ways to address customer complaints, enhance service quality, and drive overall business performance.

[Long] SharVeT: Similarity-aware Parameter Sharing with Vector-based Tuning for Efficient LLM Compression ▶ *Underline Video*
Jeongin Yun, Jaeri Lee, Jongjin Kim, Minjun Kim, Jinho Song and U Kang
How can we share parameters within large language models to significantly reduce memory costs while preserving accuracy? While parameter sharing is a promising solution to the memory overhead of large language models, existing methods rely on naive grouping and fail to correct sharing-induced discrepancies. We propose an accurate and efficient parameter sharing framework, SharVeT (Similarity-aware sharing with Vector-based Tuning), which performs similarity-based grouping to ensure accurate sharing, allocates parameters adaptively to preserve diversity within each group, and applies lightweight refinement with knowledge distillation to correct sharing-induced discrepancies. Experiments show that SharVeT outperforms existing sharing methods, achieving up to 32.1% lower perplexity and 23.3% higher few-shot reasoning accuracy.

[Long] Simple-VGC: Enhancing Visual Grounding in Multimodal Reasoning via Adaptive Tool Composition ▶ *Underline Video*
Ye Wang, Qianglong Chen, Siyuan Wang, Zejun Li, Shijie Guo, Zhirui Zhang and Zhongyu Wei
Multimodal Large Language Models (MLLMs) have achieved strong performance on vision-language tasks, yet often fail to preserve and effectively leverage visual evidence throughout generation. We identify three fundamental types of visual grounding failures: Long-Context Grounding Error, where visual information gradually decays over long sequences; Fine-Grained Grounding Error, where low-resolution or degraded inputs hinder the recovery of detailed visual information; and Regional Grounding Error, where spatially diffuse attention weakens region-level vision-language alignment. To address these issues, we propose a tool-augmented reasoning framework with three targeted compensation strategies: reuse, which re-injects the original image to mitigate visual forgetting; focus_area, which constrains attention to task-relevant regions; and zoom_in, which enhances visual resolution for fine-grained perception. We further construct the TWI-Tools-146K dataset and develop Simple-VGC, a tool-augmented MLLM that interleaves visual and textual tokens. Extensive experiments show that each tool yields targeted improvements for its corresponding grounding error, while their combination produces synergistic gains in visual reasoning. Beyond performance, our analysis provides mechanistic insights into how tool-based interventions improve visual grounding, pointing toward more reliable multimodal reasoning.

[Long] Spectral Characterization and Mitigation of Sequential Knowledge Editing Collapse ▶ *Underline Video*
Chi Zhang, Mengqi Zhang, Xiaotian Ye, Runxi Cheng, Zisheng Zhou, Ying Zhou, Pengjie Ren and Zhumin Chen
Sequential knowledge editing in large language models often causes catastrophic collapse of the model's general abilities, especially for parameter-modifying methods. Existing approaches mitigate this issue through heuristic constraints on parameter updates, the mechanisms underlying such degradation remain insufficiently understood. In this work, we present a systematic spectral analysis of sequential knowledge editing and show that a models general abilities are closely associated with dominant singular directions of pretrained weight matrices. These directions are highly sensitive to perturbations and are progressively disrupted by repeated edits, closely tracking the collapse in both editing efficacy and general performance. Building on this insight, we propose REVIVE, a plug-and-play framework that prevents model collapse by explicitly preserving this dominant subspace. REVIVE analyzes parameter updates in the spectral basis of the original weights and filters out components that would interfere with the dominant subspace. Extensive experiments across multiple models and benchmarks show that REVIVE consistently improves editing efficacy while substantially preserving general abilities under long-horizon sequential editing, including extreme settings with up to 20,000 edits.

[Long] SSSD: Simply-Scalable Speculative Decoding ▶ *Underline Video*

Michele Marzollo, Jiawei Zhuang, Niklas Roemer, Niklas Zwingenberger, Lorenz K. Muller and Lukas Cavigelli

Speculative Decoding has emerged as a popular technique for accelerating inference in Large Language Models. However, most existing approaches yield only modest improvements in production serving systems. Methods that achieve substantial speedups typically rely on an additional trained draft model or auxiliary model components, increasing deployment and maintenance complexity. This added complexity reduces flexibility, particularly when serving workloads shift to tasks, domains, or languages that are not well represented in the draft models training data. We introduce Simply-Scalable Speculative Decoding (SSSD), a training-free method that combines lightweight n-gram matching with hardware-aware speculation. Relative to standard autoregressive decoding, SSSD reduces latency by up to 2.9x. It achieves performance on par with leading training-based approaches across a broad range of benchmarks, while requiring substantially lower adoption effort—no data preparation, training or tuning are needed—and exhibiting superior robustness under language and domain shift, as well as in long-context settings.

[Findings] Stop Guessing When to Stop Testing: Efficient Model Evaluation with Just Enough Data ▶ *Underline Video*

Ofir Arviv, Kristjan Greenewald, Yotam Perlitz, Hadar Mulian, Michal Shmueli-Scheuer and Leshem Choshen

The inherent rigidity of fixed-size benchmarks makes them an inefficient tool for model evaluation. Diverse evaluation objectives, including model ranking, model selection and testing throughout development, demand varying levels of statistical power. The mismatch between fixed sample sizes and these diverse needs results in either excessive computational cost or compromised reliability - a critical concern for model evaluation. To overcome these limitations, we call for adoption of sequential testing in our field. We provide an adaptive evaluation framework, that provides a principled way to navigate the trade-off between efficiency and reliability in model evaluation. Our framework combines the established statistical paradigm of sequential testing with stopping criteria tailored to common evaluation needs such as diminishing returns detection, and minimum detectable effect size. We demonstrate its ability to adaptively manage the efficiency-reliability trade-off on the Open VLM Leaderboard, including, for example, a 80% reduction in computational cost compared to fixed-size evaluation (with a 2.5-point CI width allowance) while maintaining statistical significance.

[Long] StoryCoder: Narrative Reformulation for Structured Reasoning in LLM Code Generation ▶ *Underline Video*

Geonhui Jang, Dongyoon Han and YoungJoon Yoo

Effective code generation requires both model capability and a problem representation that carefully structures how models reason and plan. Existing approaches augment reasoning steps or inject specific structure into how models think, but leave scattered problem conditions unchanged. Inspired by the way humans organize fragmented information into coherent explanations, we propose StoryCoder, a narrative reformulation framework that transforms code generation questions into coherent natural language narratives, providing richer contextual structure than simple rephrasings. Each narrative consists of three components: a task overview, constraints, and example test cases, guided by the selected algorithm and genre. Experiments across 11 models on HumanEval, LiveCodeBench, and CodeForces demonstrate consistent improvements, with an average gain of 18.7% in zero-shot pass@10. Beyond accuracy, our analyses reveal that narrative reformulation guides models toward correct algorithmic strategies, reduces implementation errors, and induces a more modular code structure. The analyses further show that these benefits depend on narrative coherence and genre alignment, suggesting that structured problem representation is important for code generation regardless of model scale or architecture. Our code is available at https://github.com/gu-ni/StoryCoder.

[Long] Stress Testing Factual Consistency Metrics for Long-Document Summarization ▶ *Underline Video*

Zain Muhammad Mujahid, Dustin Wright and Isabelle Augenstein

Evaluating the factual consistency of abstractive text summarization remains a significant challenge, particularly for long documents, where conventional metrics struggle with input length limitations and long-range dependencies. In this work, we systematically evaluate the reliability of six widely used reference-free factuality metrics, originally proposed for short-form summarization, in the long-document setting. We probe metric robustness through seven factuality-preserving perturbations applied to summaries, namely paraphrasing, simplification, synonym replacement, logically equivalent negations, vocabulary reduction, compression, and source text insertion, and further analyze their sensitivity to retrieval context and claim information density. Across three long-form benchmark datasets spanning science fiction, legal, and scientific domains, our results reveal that existing short-form metrics produce inconsistent scores for semantically equivalent summaries and exhibit declining reliability for information-dense claims whose content is semantically similar to many parts of the source document. While expanding the retrieval context improves stability in some domains, no metric consistently maintains factual alignment under long-context conditions. Finally, our results highlight concrete directions for improving factuality evaluation, including multi-span reasoning, context-aware calibration, and training on meaning-preserving variations to enhance robustness in long-form summarization.

[SRW] Supervision versus Demonstration-Based In-Context Learning for Multiword Expression Classification ▶ *Underline Video*

Sercan_Karakas Sercan_Karakas and Yusuf_Simsek Yusuf_Simsek

Turkish idiomatic light verb constructions (LVCs) are challenging for multiword expression processing because they often share the same surface form as fully literal verb-object combinations while functioning as a single, partially idiomatic predicate. We frame Turkish LVC detection as a binary classification task (literal ditransitive vs. LVC) and evaluate on a manually created controlled set (N=147) with matched negatives: out-of-domain random sentences and in-domain literal controls (NLVC), alongside LVC positives. We compare a supervised Turkish encoder baseline (BERTurk with a classifier head) to three instruction-tuned LLMs from different families under zero-shot,

one-shot, and few-shot prompting, and analyze how demonstrations shift error profiles. In zero-shot, LLMs perform well on negatives but show very low LVC recall. One-shot prompting sharply improves LVC detection but can induce strong, model-specific biases (over- vs. under-predicting LVC). A richer master few-shot prompt improves calibration and yields robust overall performance for GPT-OSS-20B and Qwen 2.5. Overall, the results highlight substantial prompt sensitivity in Turkish metalinguistic classification: the supervised baseline remains competitive, while prompted LLMs can match or exceed it on LVCs with carefully constructed demonstrations. We will release code, prompts, and evaluation materials to support reproducibility.

[Findings] SWE-QA: Can Language Models Answer Repository-level Code Questions? ▶ *Underline Video*
Weihan Peng, Yuling Shi, Yuhang Wang, Xinyun Zhang, Beijun Shen and Xiaodong Gu
Understanding and reasoning about entire soft-ware repositories is an essential capability for intelligent software engineering tools. While existing benchmarks such as CoSQA and CodeQA have advanced the field, they predominantly focus on small, self-contained code snippets. These setups fail to capture the complexity of real-world repositories, where effective understanding and reasoning often require navigating multiple files, understanding software architecture, and grounding answers in long-range code dependencies. In this paper, we present SWE-QA, a repository-level code question answering (QA) benchmark designed to facilitate research on automated QA systems in realistic code environments. SWE-QA involves 720 high-quality question-answer pairs spanning diverse categories, including intention understanding, cross-file reasoning, and multi-hop dependency analysis. To construct SWE-QA, we first crawled 77,100 GitHub issues from 12 popular repositories. Based on an analysis of naturally occurring developer questions extracted from these issues, we developed a two-level taxonomy of repository-level questions and constructed a set of seed questions for each category. For each category, we manually curated and validated questions and collected their corresponding answers. We evaluate six advanced LLMs on SWE-QA under various context augmentation strategies. Experimental results highlight the promise of LLMs.

[SRW] Sycophantic Anchors: Localizing and Quantifying User Agreement in Reasoning Models ▶ *Underline Video*
Jacek Duszenko;Jan Koco;Przemysaw Kazienko
Reasoning models frequently agree with incorrect user suggestions - a behavior known as sycophancy. However, it is unclear where in the reasoning trace this agreement originates and how strong the commitment is. We introduce sycophantic anchors - sentences identified via counterfactual analysis that commit models to user agreement. Across four reasoning models spanning three architecture families (Llama, Qwen, Falcon-hybrid) and 1.5B - 8B parameters, we analyze over 200,000 counterfactual rollouts and show that linear probes reliably detect sycophantic anchors (74 - 85% balanced accuracy), outperforming text-only baselines at high commitment levels - confirming they capture internal states beyond surface vocabulary. Regressors further predict commitment strength from activations (R^2 up to 0.74). We observe a consistent asymmetry: sycophancy leaves a stronger mechanistic footprint than correct reasoning. We also find that sycophancy builds gradually during generation rather than being determined by the prompt. These findings enable sentence-level detection and quantification of model misalignment mid-inference.

[SRW] TableMBR: Minimum Bayes Risk Table Generation Based on Structural Consistency ▶ *Underline Video*
Yoshida Daiki; Hiroyuki Deguchi; Yusuke Sakai; Hidetaka Kamigaito; Taro Watanabe
The text-to-table task aims to generate structured data in tabular formats from unstructured text. While the integration of large language models (LLMs) has significantly enhanced the comprehensiveness and flexibility of generation, challenges regarding inconsistent output quality persist, such as the inclusion of redundant information and numerical inaccuracies. We propose TableMBR, a robust table generation method that maintains structural consistency through minimum Bayes risk (MBR) decoding. Experimental results showed that TableMBR outperforms the baseline, achieving relative improvements of up to 15% in F1 score on Rotowire and 23% in accuracy on LiveSum.

[SRW] Task Assignment meets Annotator Modeling: Human-LLM Collaborative Annotation with Constraints ▶ *Underline Video*
Kei Moriyama; Kouta Nakayama; Yukino Baba
Crowdsourced annotators and Large Language Models (LLMs) offer complementary, cost-effective ways to obtain labeled data, yet ensuring high label quality remains challenging. We observe that task features influence the accuracy of humans and LLMs, while real-world constraints, such as per-annotator assignment limits, further complicate allocation. Prior work typically addresses either task features or constraints, but not both. We present an integrated framework that (i) estimates per-task accuracy from task features using a learning from crowds model and (ii) incorporate these estimations into a linear programming that assigns tasks under practical constraints. Experimental results demonstrate that the proposed method achieves the comparable accuracy to baseline methods while satisfying the given constraints.

[Findings] Task Matters: Knowledge Requirements Shape LLM Responses to Context-Memory Conflict ▶ *Underline Video*
Kaiser Sun ~Kaiser_Sun1, Fan Bai and Mark Dredze
Large language models (LLMs) rely on both contextual knowledge and parametric memory, yet these sources can conflict. Prior analysis largely focused on contextual question answering, suggesting that models tend to favor parametric knowledge under conflict, but this setting assumes that tasks should always rely on the provided passage. It therefore remains unclear how LLMs behave when tasks demand different kinds and degrees of knowledge utilization. We address this gap with a model-agnostic diagnostic framework that holds underlying knowledge constant while injecting controlled conflicts across tasks with varying knowledge requirements. Evaluating representative open-source LLMs, we find that: (1) performance degradation under conflict correlates with a tasks knowledge reliance rather than conflict plausibility alone; (2) strategies such as explanatory rationales or reiteration increase context reliance, helping context-only tasks but harming

those that require parametric knowledge; and (3) these behaviors bias model-based evaluation, raising concerns about the reliability of LLMs as judges. Together, our findings show that context-memory conflict is fundamentally task-dependent and motivate task-aware approaches to balancing context and memory in LLM deployment and evaluation.

[SRW] Test-Time Strategies for More Efficient and Accurate Agentic RAG ▶ *Underline Video*

Abhinav Sharma; Brian Zhang; Deepti Guntur; Zhiyang Zuo; Shreyas Chaudhari; Wenlong Zhao; Franck Dernoncourt; Puneet Mathur; Ryan A. Rossi; Nedim Lipka;

Agentic retrieval-augmented generation (RAG) frameworks such as Search-R1 (Jin et al., 2025) improve performance on complex, multi-hop questions by interleaving reasoning and retrieval, but they often retrieve previously seen information and struggle to integrate evidence into the current reasoning state. We study lightweight test-time modifications to Search-R1 that target these inefficiencies without changing training: contextualization of retrieved passages, de-duplication of previously seen documents, and a hybrid of both. Evaluated on 500-example subsets of HotpotQA (Yang et al., 2018) and Natural Questions (Kwiatkowski et al., 2019) using exact match (EM), LLM-based semantic matching, and average retrieval turns, our best method - contextualization with GPT-4.1-mini - improves exact match by 5.6% and reduces retrieval turns by 10.5% relative to the Search-R1 baseline. These results suggest that improved use of retrieved evidence, rather than enforcing retrieval diversity alone, is the main driver of better answer accuracy and efficiency.

[Findings] The Mechanics of Interference: Defusing Distractors in RAG via Sparse Autoencoder Interventions ▶ *Underline Video*

Christian Giannetti, Giovanni Trappolini, Nicola Tonellotto, Fabrizio Silvestri and Pietro Liò

Large language models exhibit a critical vulnerability to distractor interference in retrieval-augmented contexts: they fail to prioritize relevant, factually correct documents over topically similar but misleading content. We introduce Lat-Defuse, a mechanistic framework that corrects this failure mode through targeted interventions in the model's latent space. Using Sparse Autoencoders (SAEs), our method operates in an interpretable feature space and formulates correction as constrained counterfactual optimization. On Gemma-2 and Llama-3 model families across three QA benchmarks (BioASQ, Natural Questions, PopQA), our method achieves recovery rates of up to 94% on distractor-vulnerable samples. Successful correction through sparse modifications reveals distractor interference as a localized, systematically addressable phenomenon, opening directions toward universal distractor robustness in LLMs.

[Long] The Retrieval Bottleneck: Scaling Laws for Reinforcement Learning in RAG ▶ *Underline Video*

Shu Zhou, Jinman Leng, Yufei Song, Xin Wang, Tao Fan and Hao Wang

Scaling laws have enabled predictable compute allocation for pre-training and for RL in reasoning tasks. However, research on retrieval reinforcement generation (RAG) remains insufficient and there is a lack of fundamental understanding of the interaction between retrieval quality and reinforcement learning computation. We present the first systematic study of RL scaling for RAG across three knowledge-intensive benchmarks. We introduce the Retrieval Bottleneck Hypothesis and derive sigmoidal scaling laws showing that retrieval quality, not RL compute, determines the asymptotic performance ceiling. Our analysis reveals three principles: (1) retrieval quality bounds achievable performance, with improving retrieval yielding larger gains than algorithmic innovations; (2) design choices (training objectives, rewards, off-policy methods) primarily modulate compute efficiency, with secondary effects on the ceiling that are substantially smaller than retrieval quality improvements; and (3) stable configurations enable extrapolation with 3.1% error at 4x compute. We further uncover RAG-specific dynamics: optimal document count increases with training, and RL algorithm effectiveness depends critically on retrieval quality. These insights yield RAG-ScaleRL, achieving strong performance on knowledge-intensive benchmarks while providing the predictable scaling long available for pre-training but previously absent in RAG-RL.

[Long] The Role of Mixed-Language Documents for Multilingual Large Language Model Pretraining ▶ *Underline Video*

Jiandong Shao, Raphael Tang, Crystina Zhang, Karin Sevegnani, Pontus Stenetorp, Jianfei Yang and Yao Lu

Multilingual large language models achieve impressive cross-lingual performance despite largely monolingual pretraining. While bilingual data in pretraining corpora is widely believed to enable these abilities, details of its contributions remain unclear. We investigate this question by pretraining models from scratch under controlled conditions, comparing the standard web corpus with a monolingual-only version that removes all multilingual documents. Despite constituting only 2% of the corpus, removing bilingual data causes translation performance to drop 56% in BLEU, while behaviour on cross-lingual QA and general reasoning tasks remains stable, with training curves largely overlapping the baseline. To understand this asymmetry, we categorize bilingual data into parallel (14%), code-switching (72%), and miscellaneous documents (14%) based on the semantic relevance of content in different languages. We then conduct granular ablations by reintroducing parallel or code-switching data into the monolingual-only corpus. Our experiments reveal that parallel data almost fully restores translation performance (91% of the unfiltered baseline), whereas code-switching contributes minimally. Other cross-lingual tasks remain largely unaffected by either type. These findings reveal that translation critically depends on systematic token-level alignments from parallel data, whereas cross-lingual understanding and reasoning appear to be achievable even without bilingual data.

[SRW] The Shape of Vulnerability: How Adversarial Perturbations Reshape the Topology of Language Model Latent Spaces ▶ *Underline Video*

Angelina Tsai; Shreya Subramanian; Catherine Liu; Kimberly Lopez; Leif Zinn-Brooks; Alexia Schulz; Adaku Uchendu

Adversarial perturbations in the context of large language models (LLMs) are subtle changes added to input data (i.e., images or text) that are designed to alter predictions or outputs of machine learning models. We introduce several novel visualizations using topological data analysis

(TDA) (leveraging persistent homology) to characterize how adversarial perturbations act on text inputs, specifically, how sandbagging and code-injection attacks alter the geometric structure of attention heads in transformer models. By computing persistent homology metrics from attention maps across different model architectures (such as BERT, RoBERTa, ELECTRA, DistilGPT, etc.), we find that adversarial inputs alter higher-dimensional topological features (H_1 loops and H_2 voids) in ways that distinguish them from clean, non-adversarial inputs.

[Long] The Side Effects of Being Smart: Safety Risks in MLLMs Multi-Image Reasoning ▶ *Underline Video*

Renmiao Chen, Yida Lu, Shiyao Cui, Xuan Ouyang, Victor Shea-Jay Huang, Shumin Zhang, Chengwei Pan, Han Qiu and Minlie Huang

As Multimodal Large Language Models (MLLMs) acquire stronger reasoning capabilities to handle complex, multi-image instructions, this advancement may pose new safety risks. We study this problem by introducing MIR-SafetyBench, the first benchmark focused on multi-image reasoning safety, which consists of 2,676 instances across a taxonomy of 9 multi-image relations. Our extensive evaluations on 19 MLLMs reveal a troubling trend: models with more advanced multi-image reasoning can be more vulnerable on MIR-SafetyBench. Beyond attack success rates, we find that many responses labeled as safe are superficial, often driven by misunderstanding or evasive, non-committal replies. We further observe that unsafe generations exhibit lower attention entropy than safe ones on average. This internal signature suggests a possible risk that models may over-focus on task solving while neglecting safety constraints.

[SRW] Thesis Proposal: An Explainable Multimodal Framework for Detecting Harmful Content in Code-Switched Childrens Media ▶ *Underline Video*

Juliana Isabelle A. Guillermo; Jasper Kyle Catapang; Nathaniel Oco

Current automated content moderation systems fail to protect children from harmful YouTube content, particularly in under-resourced, code-switched settings. These systems are often text-only, English-centric, and operate as 'black boxes,' lacking the multimodal understanding and transparency needed for effective moderation. This thesis proposes a novel hybrid framework for the explainable multimodal detection of harmful content in videos with code-switching. The proposed framework integrates a fine-tuned classifier for accurate, scalable detection with an LLM-powered module that synthesizes the classifiers internal evidential signals (e.g., text attention and visual heat maps) to generate faithful, human-readable rationales for each decision. As a primary case study, the framework will be developed and validated on an English-Filipino code-switched dataset. Expected contributions include a new dataset publicly available under controlled access (de-identified transcripts, blurred frames, extracted feature representations, and metadata via data-sharing agreement) and a blueprint for building more equitable, transparent, and trustworthy AI safety systems.

[SRW] Thesis Proposal: Auditing and Mitigating Demographic Bias in Multi-Stage Retrieval Systems for Criminal Justice Applications ▶ *Underline Video*

Archan Dutta

We propose a comprehensive research agenda to detect, measure, and mitigate racial bias in Natural Language Processing (NLP) systems deployed in criminal justice contexts. Our preliminary work demonstrates that racial descriptors systematically alter embedding similarity scores and retrieval rankings across six models, with bias being race-specific and models showing rank displacements of 1.82 to 7.44 positions, on average. This empirically indicates that even small shifts in similarity scores can displace relevant records outside top-10 results, leading to systematic under-retrieval of records involving certain demographic groups. Building on these findings, this thesis proposes four research questions: (1) developing and evaluating debiasing techniques including counterfactual data augmentation, adversarial training, and fairness-constrained fine-tuning; (2) validating synthetic findings on authentic law enforcement data through IRB-approved partnerships; (3) investigating intersectional bias patterns across race, gender, and age; and (4) we extend beyond embedding-level analysis to examine how bias propagates across modern multi-stage retrieval pipelines from embeddings to cross-encoders to LLMs. Expected contributions include empirical comparisons of debiasing methods, bias benchmarks for criminal justice NLP, deployment guidelines for fairness-aware retrieval systems, and the first comprehensive analysis of multi-stage bias propagation in retrieval pipelines.

[SRW] Thesis Proposal: Bring Linguistics Back to Cryptanalysis - Using Attestation to Break the Advanced Encryption Standard ▶ *Underline Video*

Madeline Boese

The Advanced Encryption Standard (AES) is currently considered the preferred standard of encryption for secure messaging by the National Institute of Standards and Technology. While its predecessor, the Data Encryption Standard (DES), can be analyzed in mere seconds with modern cryptanalysis algorithms, those same algorithms are unfeasible for cryptanalysis of AES. This is primarily because the key, or the list of values used during encryption, was increased from a length of 56 bits to 128 bits. The list of all possible combinations, also called the probability space, is tens of magnitudes larger in AES than DES. Current DES cryptanalysis methods use mathematical methods to find the most statistically likely approximate key in the entire probability space. The increased probability space in AES is too large to find an approximation using these formulas. However, these methods operate under the assumption that every key is possible. While this is mathematically true, when looked at from the lens of a linguist, many of these keys create messages that are impossible in language. Linguistic attestation is the documentation that a word, grapheme, sound, or other linguistic feature exists in a language. This thesis proposal presents an algorithm that eliminates AES keys using linguistic attestation. I apply the idea that any grammatical message encrypted by AES will not contain any "unattested" data in its input. The algorithm proposes trimming the probability space by removing keys that are unattested. Once a smaller probability space has been created, any method that searches the probability space for a solution may be applied to find the key from the new, smaller list of filtered keys.

[SRW] Thesis Proposal: Diagnosing and Mitigating Semantic Interference in Script-Sharing Low-Resource Language Models: A Case Study on Square Bai Script ▶ *Underline Video*

Jingting Zheng; Deyi Xiong

Multilingual language models now cover more languages than ever, yet script-sharing low-resource languages remain vulnerable to failures driven by script and dominant-language priors. This dissertation studies one such failure mode, semantic interference, in Square Bai Script, where many forms resemble Chinese characters but often differ in meaning. We argue that current adaptation pipelines underperform not only because Bai is low-resource, but because they treat visible overlap as safe transfer by default. Building on an expert-validated corpus of 28,382 Bai-Chinese sentence pairs, an out-of-domain epigraphic set and a reproducible encoding pipeline, the dissertation will (1) diagnose semantic interference, (2) compare adaptation strategies under realistic compute constraints, and (3) estimate when shared-script transfer helps or harms adaptation. The long-term goal is Bai-capable understanding and generation. The dissertation addresses the prerequisite problem of safe and effective adaptation in a script-sharing low-resource setting.

[SRW] Thesis Proposal: Establishing Rigorous Evaluation of Sycophancy in Pretrained Language Models ▶ *Underline Video*

Jan Batzner

Sycophancy, the tendency of a language model to agree with users and confirm their prior believes, is becoming one of the most commonly referred to failure modes in alignment research. Yet, our survey of 80 sycophancy papers (2021-2025) shows that authors use the term to describe operationally different concepts: some papers test whether models accept a user's false factual claim, others whether models shift political positions to match a stated persona, others whether models provide emotional validation, and still others whether multimodal misdirections impact model performance. This thesis proposal pursues four lines of work: (1) establishing construct validity by mapping how sycophancy is defined and distinguished from related constructs; (2) rigorously empirically evaluating whether benchmarks agree across question format, modality, and judge choice; (3) building evaluation infrastructure by applying standardized metadata to sycophancy evaluations; and (4) examining how the choice of operational definitions impacts interpretability and which model-internal mechanisms are identified.

[SRW] Thesis Proposal: Intentional Inference for Insight Generation ▶ *Underline Video*

Kristýna Onderková

Large language models show strong capabilities in natural language generation and have been applied to translate complex structured data into human-readable insights. While these models excel at surface-level fluency, they remain unreliable as they produce factually inaccurate outputs and struggle with consistent logical inference beyond surface-level patterns. Moreover, they often lack a clear sense of relevance and produce shallow or uninformative insights. This thesis proposal argues that a key source of these limitations is the fact that tasks are often underspecified, requiring models to make implicit assumptions about missing context. We will investigate how such underspecification leads to unintentional assumptions and how these affect faithfulness and evaluation. We will examine how models can identify missing premises and surface multiple plausible interpretations to improve factuality of their inferences and make their evaluation more rigorous. We will also study how underlying assumptions influence the perceived interestingness of insights. Finally, we will explore how to improve reasoning to enable deeper inferences, focusing on code generation and qualitative reasoning. By shifting focus from surface-level generation to assumption-aware deeper inferences, this work aims to improve reliability, interpretability, and user controlability in NLG.

[SRW] Thesis Proposal: LLMs post-training for multilingual medical tasks. Instruction-Tuning, Continual-Pretraining or Reasoning? ▶ *Underline Video*

pietro ferrazzu; bernardo magnini; alberto lavelli

Adapting Large Language Models to the medical domain remains an active area of research, with multiple strategies proposed to leverage annotated and unannotated data effectively. In this work, we propose a thesis outline to compare three common adaptation approaches, Instruction Tuning, Continual Pretraining, and Reasoning-oriented Training. We identify 5 dimensions to analyse: i) the interaction between the adaptation technique and the tasks; ii) the impact of the data size on the downstream performance; iii) the differences between datasets required by the three techniques; iv) the impact of the techniques given the model size; v) the impact of the techniques given the language. We construct an evaluation framework composed by 5 multilingual medical NLP tasks (named entity recognition, relation extraction, question answering, case report form filling, argument mining), spanning on 21 datasets in English, Italian, and Spanish, for a total of 61 combinations of language and sub-task.

[SRW] Thesis Proposal: Rethinking Safety Evaluation in Large Language Models ▶ *Underline Video*

Khaoula Chehbouni

The widespread integration of large language models (LLMs) across diverse domains, from entertainment to healthcare and banking, has underscored the pressing need to address the risks associated with artificial intelligence (AI). Indeed, as LLMs have grown in scale, new challenges have emerged, including privacy breaches, dissemination of misinformation, issues related to robustness and alignment, and concerns regarding the societal biases ingrained in these models. In response, there has been a surge in research focused on responsible AI and AI safety. To address these challenges, recent literature has introduced novel techniques for mitigating the risks associated with LLMs and testing the efficacy of safety protocols through red-teaming. Despite these efforts, concerns persist regarding the perpetuation of biases in next-generation LLMs and the lack of robustness in existing safety measures. In this thesis proposal, we ar-

gue that some failures in LLM safety arise from flawed evaluation practices and explore ways to more effectively assess and characterize them.

[SRW] Thesis Proposal: Targeted and Unified Cross-Lingual Unlearning from Multilingual Language Models ► *Underline Video*
Jan Bronec; Jindich Helcl;
As large language models (LLM) trained on massive corpora scraped from the web exhibit the capability to reproduce sensitive and copyright-protected data, the field of machine unlearning has emerged to address the arising ethical and legal concerns. While previous research has provided a unified evaluation of LLM unlearning methods, this unification remains constrained to English-only models and datasets. We aim to address the prevailing fragmentation in recent cross-lingual unlearning research by extending existing unified benchmarks with multilingual data. To that end, we plan to compile a dataset of parallel translations of question-answer pairs consisting of real-world facts and synthetic personally identifiable information. Moreover, we will focus on mitigating model degradation during unlearning by selectively editing only those layers that contain the given knowledge.

[SRW] Thesis Proposal: The Missing Why? Building Generative AI That Understands Purpose, Audience, and Context ► *Underline Video*
Ishani Mondal
Modern generative AI systems can produce fluent text, clear diagrams, and polished presentations, but they still fail in high-stakes scientific settings. The core issue is not fluency, but the lack of understanding of why content is generated, for whom, and how it should adapt as goals evolve. In this proposal, I argue that effective scientific communication requires modeling user intent, audience variation, and multimodal structure, not just surface quality. To address these gaps, I develop systems across the communication pipeline: ADAPTIVE IE for intent-driven information extraction, GPA and Persona-Aware Slide Generation for audience-aware content adaptation, SciDoc2Diagrammer-MAF for faithful diagram generation, and SMART-Editor for structurally coherent layout refinement. Together, these systems demonstrate that aligning with intent, audience, and structure is essential for meaningful communication. However, existing approaches still lack a principled way to evaluate and generalize usefulness. To address this, I propose three directions: (1) a goal-driven framework to measure the educational utility of document-to-video generation, (2) studying multimodal personalization to improve both creation and comprehension, and (3) building a unified SuperPersonalization benchmark to learn transferable user preferences across tasks. This work shifts the focus from generating content that looks correct to content that actually communicates.

[SRW] Thesis Proposal: Toward a Human-Centered and Perspective-Aware Framework for Reproducible ML Evaluation and AI Alignment ► *Underline Video*
Deepak Pandita; Christopher M Homan
Humans play a vital role at every stage of AI development, from data collection and curation to model development and evaluation. However, humans often disagree with each other and sometimes with themselves over time. It is essential to take disagreement into account when building human-centered AI systems, especially in domains where it is prevalent, such as AI safety, content moderation, or sentiment analysis. Disagreement typically arises from subjective human opinion and can vary with one's identity, beliefs, and social environment. Despite this, current LLM evaluation approaches frequently rely on aggregating labels (often via plurality voting) to represent consensus, failing to account for variance in human responses. This amplifies the reproducibility crisis in AI evaluation. Human feedback is also crucial for ensuring that AI systems align with human values. For these systems to be trustworthy, it is critical to ensure that they reflect diverse human values and perspectives. In this thesis proposal, we present a human-centered and perspective-aware framework for reproducible ML evaluation and AI alignment.

[SRW] Thesis Proposal: Uncertainty as Adaptive Control: From Selection to Curriculum via Conformal Calibration ► *Underline Video*
Peihong Li; Yan Yan
Large language models can produce fluent but unfaithful text, known as hallucinations, because point predictions carry no indication of their own reliability. A model that "knows what it does not know" could flag, locate, or avoid such failures, yet standard training and decoding offer no such mechanism. Uncertainty is the missing signal, and once calibrated it can serve as a selection filter, an error indicator, or a curriculum guide, three roles unified under a single adaptive-control principle. This thesis develops a framework in which uncertainty is measured from model predictions, calibrated via conformal prediction or analogous principled boundaries, and integrated into the learning or generation process itself. We instantiate this framework in three progressively challenging scenarios: classification under label noise (uncertainty as selection), video generation (uncertainty as error correction), and text generation (uncertainty as curriculum). The principle is validated in classification and video generation and extended to text generation, showing that the pipeline structure is invariant while each component takes qualitatively new forms under escalating difficulty.

[SRW] Thesis Proposal: When Does an Agent Know It Is Lost? Confidence Trajectory Analysis for Tool-Using LLMs ► *Underline Video*
Zhenjiang Mao
Large language model (LLM) agents that invoke external tools must make sequences of interdependent decisions, yet existing uncertainty quantification (UQ) methods treat each step in isolation, ignoring how confidence evolves and compounds across a full task trajectory. We propose a framework for trajectory-level confidence analysis in the tool-use agent setting. The thesis pursues three aims: (1) estimating

action-level confidence by adapting step-wise UQ to the heterogeneous think-act-observe cycles of tool-using agents; (2) aggregating the diverse action space into semantically coherent action types to enable meaningful trajectory-level analysis; and (3) discovering temporal patterns in the resulting confidence trajectories that reliably predict task success or failure. We ground the work in standard tool-use benchmarks and expect the framework to expose early warning signals for agent failure and offer interpretable diagnostic tools for understanding when and why LLM agents lose confidence, with improved calibration of multi-step agentic pipelines as a secondary benefit.

[SRW] Thesis Proposal: When Does Confidence Signal Quality? Log-Probabilities and LLM-as-Judge in Multi-Agent Debate ▶ *Underline Video*

Ali Keramati, Justin Cheok, Jacob Horne and Mark Warschauer;

Multi-agent debate systems are usually evaluated only by whether the final answer is correct. However, this overlooks the quality of the intermediate reasoning that debate is meant to produce. This PhD research studies the relationship between three signals in multi-agent debate: token-level log-probability distributions over reasoning tokens, LLM-as-judge rubric scores assigned to those tokens, and final task accuracy. Instead of focusing only on final outputs, we examine whether internal confidence signals predict externally evaluated reasoning quality, and whether either signal is aligned with task correctness. We apply this framework to three domains: rubric-based scoring, mathematical reasoning, and factual question answering. Our setup uses a two-agent debate architecture in which a Constructor and an Auditor generate opposing reasoning traces while exposing token-level log-probabilities during decoding. A separate LLM-as-judge then scores each agent's reasoning using rubric-based criteria. Preliminary experiments show that early-token log-probability statistics correlate meaningfully with judge scores and are especially useful for identifying major reasoning failures. These findings motivate the broader cross-domain investigation proposed in this dissertation.

[Long] ThreadSumm: Summarization of Nested Discourse Threads Using Tree of Thoughts ▶ *Underline Video*

Olubusayo Olabisi, Ekata Mitra and Ameeta Agrawal

Summarizing deeply nested discussion threads requires handling interleaved replies, quotes, and overlapping topics, which standard LLM summarizers struggle to capture reliably. We introduce ThreadSumm, a multi-stage LLM framework that treats thread summarization as a hierarchical reasoning problem over explicit aspect and content unit representations. Our method first performs content planning via LLM-based extraction of discourse aspects and Atomic Content Units, then applies sentence ordering to construct thread-aware sequences that surface multiple viewpoints rather than a single linear strand. On top of these interpretable units, ThreadSumm employs a Tree of Thoughts search that generates and scores multiple paragraph candidates, jointly optimizing coherence and coverage within a unified search space. With this multi-proposal and iterative refinement design, we show improved performance in generating logically structured summaries compared to existing baselines, while achieving higher aspect retention and opinion coverage in nested discussions.

[SRW] Through the Looking Glass of Multilingual AI: Contrasting Language- and Name Script-Dependent Ethnic Hierarchies in GPT and DeepSeek ▶ *Underline Video*

Annabella Sakunkoo; Jonathan Sakunkoo

Large language models (LLMs) are increasingly used as evaluative tools across languages, yet bias research remains overwhelmingly Anglocentric, with most studies conducted in English using Latin-script names. It remains unclear whether bias patterns generalize across linguistic contexts. We investigate this question and introduce the stereotype perceptual map, a framework for analyzing how ethnic groups are positioned along evaluative dimensions. Using 900,000 model responses over 45,000 names spanning 9 ethnicities, we evaluate model behavior across prompt languages (English, Chinese, Thai), writing scripts (Latin, Chinese, Thai), evaluative domains (competence, warmth), and models (GPT, DeepSeek). We find that ethnic bias hierarchies are jointly shaped by local linguistic context and model origin and differ substantially between Western-centric and Sinocentric models. DeepSeek exhibits highly stable rankings across conditions in math competence judgments, consistently placing Chinese at the top, followed by Russian, and White, Hispanic, and Black names at the bottom. GPT, by contrast, shows strong script-dependent reordering: Latin-script conditions form one stable cluster, while native-script conditions form another, with substantially lower cross-cluster correlations. We term this script-gated bias: transliterating the same names into a non-Latin script can activate a different evaluative frame and produce rankings that are sometimes inversely correlated with Latin-script results. Warmth evaluations are less stable than competence across both models. Our findings demonstrate that multilingual bias cannot be characterized through single-language, single-script audits. For multilingual users, code-switching between languages can toggle between different bias regimes. Fairness evaluations for multilingual LLMs must therefore account for deployment language, writing system, and model origin to capture the full range of potentially harmful bias these systems exhibit.

[SRW] Tonal Salience in Cognitive Decline: In-Context MCI Detection with Multimodal LLMs ▶ *Underline Video*

Christopher Song; Abdullah Ahmed

Speech-based screening for mild cognitive impairment offers a highly accessible way to detect early cognitive decline. While most existing work focuses on English, cross-linguistic research is emerging to examine how cognitive decline manifests across languages. Studies on the Interspeech'24 TAUKADIAL dataset, comprising English and Chinese speech recordings, consistently report higher classification performance on Chinese, yet the cause of this cross-lingual discrepancy remains unexplored. We examine this gap using Gemini 2.5 Pro, a multimodal large language model, using zero-shot and in-context-learning (ICL) paradigms. We hypothesize that this disparity is rooted in language typology: in tonal languages like Chinese, pitch encodes lexical meaning in every syllable (tone), whereas in non-tonal languages like English, pitch carries no lexical function. To test this, we pitch-flattened audio from TAUKADIAL and compared how classification

performance changed across both languages. We found that Chinese classification degraded significantly under both zero-shot and ICL conditions (-4.78 and -5.92 UAR, respectively), while English performance increased (+0.11 and +2.98 UAR), implicating tonal pitch as the cross-lingual advantage. These findings suggest language typology should inform the design of audio-based cognitive screening tools, with raw audio preferred for tonal languages and text for non-tonal languages, a distinction critical for developing equitable cross-linguistic screening.

[Findings] Towards Unified Multimodal Large Language Models: A survey ► *Underline Video*

Xu Ma, Yitian Zhang and Yun Fu

The recent surge of interest in unified Multimodal Large Language Models (MLLMs) has catalyzed rapid progress toward general-purpose generation and understanding across different modalities. Despite the remarkable advancements, the field lacks a systematic and cohesive framework that connects these developments, revisits the motivations, and situates current trends within a broader landscape. In this survey, we present a comprehensive and in-depth review of unified MLLMs, offering both a methodology taxonomy and unique perspectives on the field. We begin by outlining the foundational concepts and prerequisites for understanding unified MLLMs. We then delve into designs from different aspects, including model architectures, loss functions, alignment techniques, and different representation strategies. Furthermore, we discuss persistent challenges and identify promising directions for future research. By bridging scattered progress and providing a consolidated view, this survey aims to foster a deeper and systematical understanding of unified MLLMs and inspire future innovations in building truly general multimodal intelligence.

[Long] TPA: Next Token Probability Attribution for Detecting Hallucinations in RAG ► *Underline Video*

Pengqian Lu, Jie Lu, Anjin Liu and Guangquan Zhang

Detecting hallucinations in Retrieval-Augmented Generation remains a challenge. Prior approaches attribute hallucinations to a binary conflict between internal knowledge stored in FFNs and the retrieved context. However, this perspective is incomplete, failing to account for the impact of other components of the LLM, such as the user query, previously generated tokens, the self token, and the final LayerNorm adjustment. To comprehensively capture the impact of these components on hallucination detection, we propose TPA which mathematically attributes each token's probability to seven distinct sources: Query, RAG Context, Past Token, Self Token, FFN, Final LayerNorm, and Initial Embedding. This attribution quantifies how each source contributes to the generation of the next token. Specifically, we aggregate these attribution scores by Part-of-Speech (POS) tags to quantify the contribution of each model component to the generation of specific linguistic categories within a response. By leveraging these patterns, such as detecting anomalies where Nouns rely heavily on LayerNorm, TPA effectively identifies hallucinated responses. Extensive experiments show that TPA achieves state-of-the-art performance.

[SRW] Tracking the Evolution of Foresight Signals in News Data: The Case of the European Electric Vehicle Market ► *Underline Video*

Karine Navasartian

To prepare for an uncertain future, organizations must continuously monitor emerging trends and early signals of change. The increasing availability of web-based textual data has boosted natural language processing (NLP) methods in strategic foresight, particularly in the scanning phase. While prior studies have extensively focused on the identification of trends and weak signals in such data, considerably less attention has been paid to how these signals evolve over time and gain relevance as they become more visible. This study addresses this gap by examining whether tracking the temporal dynamics of signals can improve their assessment for strategic decision-making. Using a large-scale web news corpus, we apply established neural topic modeling and clustering techniques to extract and link semantically coherent signals across time. By examining the resulting cluster trajectories, we analyze how newly emerging signals intensify and potentially mature into more established trends. Demonstrated on the use case of the European electric vehicle market, we find three dominant signal trajectories and show that burst dynamics tend to surface signal consolidation rather than the early detection of weak signals. By showing that signals reoccur at irregular intervals and gain strategic relevance through evolving developments, the results indicate that foresight research should move beyond static, one-off analyses toward a dynamic temporal perspective capable of identifying signals at earlier stages of emergence. From a managerial perspective, the findings underscore the need for frameworks that systematically surface signals from external environments and enable timely strategic responses by integrating expert knowledge to support proactive decision-making and adaptation ahead of major environmental shifts.

[SRW] Understanding Clinical Cognitive Dialogues Using Large Language Models ► *Underline Video*

Vishalakshi Arumugam, Dan Schumacher, Veronica Rammouz, Enrique Gonzalez Guerrero, Jeremy Davis and Anthony Rios;

In-person cognitive assessments are structured clinical conversations that provide important cues about patients' cognition but are difficult to analyze at scale. Existing datasets do not capture both speaker roles and Dialogue-Acts in this setting, limiting progress in modeling conversational structure. We introduce a dataset of anonymized cognitive assessment transcripts annotated with speaker roles and Dialogue-Act labels. We evaluate large language models on two tasks: dialogue-act classification and patient utterance generation. Our results show that while few-shot prompting improves performance, models still struggle to capture fine-grained pragmatic intent and produce clinically consistent responses. These findings highlight a gap between fluent text generation and true dialogue understanding in clinical settings.

[SRW] Understanding Conversational Implicatures in Humans and LLMs ► *Underline Video*

author_first_Daeun author_last_Kang
In conversational implicatures, speakers convey hidden intended meanings beyond the literal content of their utterances, and hearers are expected to infer what is implied. This study examines how Large Language Models (LLMs) interpret conversational implicatures, using human interpretation as a baseline and gold standard for comparison. The same experiments were conducted with two types of participants: humans and LLMs. Two metrics were adopted: a surprisal-based metric and a response-based metric. The results suggest that the response-based metric demonstrates higher accuracy, comparable to human responses, than the surprisal-based metric. In particular, humans and LLMs using the response-based metric performed better in the literal condition than in the implied condition. Additionally, they were more sensitive to capturing implied meanings for some-all trigger than for the other triggers. Overall, LLMs employing the response-based metric tend to exhibit human-like behavior, but still diverge from humans in their understanding of conversational implicatures.

[Findings] Understanding LLM Reasoning for Abstractive Summarization ▶ *Underline Video*
Haohan Yuan and Haopeng Zhang
Reasoning has substantially improved Large Language Models (LLMs) on analytical tasks such as mathematics and code generation, but its value for abstractive summarization remains unclear. To address this gap, we adapt general reasoning strategies to the summarization setting and conduct a large-scale comparative study of 8 reasoning strategies and 3 Large Reasoning Models (LRMs) across 8 diverse datasets, evaluating both summary quality and factual faithfulness. Our results show that reasoning is not a universal solution and its effectiveness depends strongly on the strategy and the summarization setting. In particular, we find a trade-off between summary quality and factual faithfulness. Explicit reasoning strategies often improve reference-based quality, but may weaken factual grounding, whereas implicit reasoning in LRMs shows the opposite tendency. We further find that increasing an LRM's internal reasoning budget does not reliably improve summarization and can even reduce factual consistency. These findings suggest that, for summarization, more reasoning is not always better. Effective reasoning should preserve faithful compression rather than induce over-elaboration.

[SRW] Validator-Guided Hard Negative Mining for Masked Language Modeling in Low-Resource Ancient Languages ▶ *Underline Video*
Andrei Voinea;
Masked language modeling for low-resource ancient languages remains challenging because pre-trained multilingual models lack exposure to these languages. We investigate rule-based linguistic constraints and hard negative mining for Sumerian, a language isolate absent from multilingual BERT's training data. We build a hierarchical validator capturing subword, word, and part-of-speech patterns from 4,545 annotated sequences, using it to filter candidates and identify hard negatives for fine-tuning. Vanilla mBERT achieves 18.0% hit@10 accuracy. The validator alone improves this to 72.8%, while hard negative fine-tuning reaches 78.3%. Combining both yields 86.7%, a 68.7 percentage point improvement. Temporal generalization evaluation on tablets from 600 years earlier shows that both the hard negative mining and the validator alone improve performance, but the combined approach underperforms due to the validator's period specific rules. These findings demonstrate that hard negative mining transfers across periods while explicit rule-based constraints provide strong in-domain improvements but limited cross-period generalization.

[SRW] What Moves the Pareto Frontier in Tool-Using Agents? A Compute-Aware Study of ReAct Variants ▶ *Underline Video*
Rishi N. Simhadri
Tool-using LLM agents are typically compared by accuracy alone, despite deployments being constrained by inference cost. We present a budgeted evaluation of common strategies for improving ReAct-style tool agents (multi-sample aggregation, iterative self-correction, and post-hoc answer revision) using Pareto analysis of cumulative accuracy versus token budget on three benchmarks (HotPotQA, FEVER, GSM8K) with Gemini-2.5-Flash. All experiments use three random seeds (N=500 per seed); budgeted curves are computed post hoc from per-instance token logs. In our offline evaluation, Reflexion attains the highest accuracy, but its reported token costs are optimistic lower bounds because retries are stopped using ground-truth feedback; deployment costs would be higher. Sampling-based approaches often spend 3–5x more tokens for comparatively small gains on tool-heavy tasks. GSM8K, a pure-math benchmark with minimal tool interaction, shows dramatically larger gains from all methods, suggesting that the efficiency penalty of repeated tool trajectories is a key factor in tool-heavy settings. We provide a compute-aware evaluation protocol (frontier analysis and marginal-cost metrics) and practical guidance for choosing agent designs under different budget regimes.

[SRW] When Models Hesitate: Answer Instability as a Label-Free Uncertainty Signal for LLMs ▶ *Underline Video*
Jasper Arana, Kristine Carandang, Ethan Casin, Christian Alis and Christopher Monterola
Large language models (LLMs) are increasingly deployed in high-stakes settings, yet reliably estimating when their outputs should be trusted remains an open challenge. Existing uncertainty estimation approaches—such as calibration, token-level probabilities, or semantic entropy—typically require access to model internals, additional supervision, or computationally intensive pipelines. We propose answer instability, defined as the variability of a model's final answer across repeated stochastic generations of the same prompt, as a simple, label-free, and black-box uncertainty signal. Evaluated across three task types , reasoning, multiple-choice QA, and constraint-following , using four LLMs and 520 prompt-model pairs, our approach achieves performance competitive with semantic entropy while requiring no semantic similarity model. Our results show that instability strongly correlates with prediction errors and reliably discriminates correct from incorrect outputs. We further demonstrate its utility for selective prediction and targeted repair, improving reliability without access to internal probabilities or additional training.

[Long] When Seeing Overrides Knowing: Disentangling Knowledge Conflicts in Vision-Language Models ▶ *Underline Video*

Francesco Ortu, Zhijing Jin, Diego Doimo and Alberto Cazzaniga

Vision-language models (VLMs) increasingly combine visual and textual information to perform complex tasks. However, conflicts between their internal knowledge and external visual input can lead to hallucinations and unreliable predictions. In this work, we investigate the mechanisms that VLMs use to resolve cross-modal conflicts by introducing WHOOPS-AHA!, a dataset of multimodal counterfactual queries that deliberately contradict internal commonsense knowledge. Through logit inspection, we identify a small set of attention heads that mediate this conflict. By intervening in these heads, we can steer the model towards its internal parametric knowledge or the visual information. Our results show that attention patterns on these heads effectively locate image regions that influence visual overrides, providing a more precise attribution compared to gradient-based methods.

Session 9: Oral/Posters/Demos E - 11:00-12:30

Demo Session E

11:00-12:30 (Grand Hall)

[Demo] Agentic CLEAR: Automating Multi-Level Evaluation of LLM Agents ▶ *Underline Video*

Asaf Yehudai, Lilach Eden and Michal Shmueli-Scheuer;

Agentic systems are becoming more capable: agents define strategies, take actions, and interact with different environments. This freedom poses serious challenges to overseeing and assessing the agent's behavior. Most current tools are limited, focusing on observability with basic evaluation or creating a static taxonomy of agent errors. To address this gap, we present Agentic CLEAR, an automatic, dynamic, and easy-to-use evaluation framework. It produces textual insights into the agent behavior on three levels of granularity: system, trace, and node. Agentic CLEAR operates above the observability layer, enabling seamless integration and featuring an intuitive UI that makes agent evaluation highly accessible. In our experiments on four benchmarks, seven agentic settings, and tens of thousands of LLM calls, we show that Agentic CLEAR produces high-quality, data-dependent, insightful feedback. Our analysis shows strong alignment with human-annotated errors and the ability to predict task success rate.

[Demo] DIAGRAMS: A Review Framework for Reasoning-Level Attribution in Diagram QA ▶ *Underline Video*

Anirudh Iyengar Kaniyar Narayana Iyengar, Tampu Ravi Kumar, Manan Suri, Raviteja Bommireddy, Dinesh Manocha, Puneet Mathur and Vivek Gupta

Diagram question answering (Diagram QA) requires reasoning-level attribution that links each question-answer pair to all visual regions needed to derive the answer, rather than only the region containing the final response. Creating such structured evidence across diagrams, charts, maps, circuits, and infographics is time-consuming, and existing annotation tools tightly couple their interfaces to dataset-specific formats. We present DIAGRAMS, a lightweight, schema-driven review framework that decouples interface logic from dataset-specific JSON structures through an internal meta-schema and dataset adapters. Given an image and QA pair with optional candidate regions, the system performs QA-conditioned evidence selection and proposes the regions required for reasoning. When QA pairs or candidate regions are missing, it generates them and supports human verification and refinement. Across six Diagram QA datasets, model-suggested evidence achieves 85.39% precision and 75.30% recall against reviewer-final selections (micro-averaged). These results indicate that the review-first framework reduces manual region creation while maintaining high agreement with final reasoning-level attributions. We release a public demo and installable package to support dataset auditing, grounded supervision creation, and grounded evaluation.

[Demo] FD-NL2SQL: Feedback-Driven Clinical NL2SQL that Improves with Use ▶ *Underline Video*

Suparno Roy Chowdhury, Tejas Anvekar, Manan Roy Choudhury, Muhammad Ali Khan, Kaneez Zahra Rubab Khakwani, M Bassam Sonbol, Irbaz Bin Riaz and Vivek Gupta;

Clinicians exploring oncology trial repositories often need ad-hoc, multi-constraint queries over biomarkers, endpoints, interventions, and time, yet writing SQL requires schema expertise. We demo FD-NL2SQL, a feedback-driven clinical NL2SQL assistant for SQLite-based oncology databases. Given a natural-language question, a schema-aware LLM decomposes it into predicate-level sub-questions, retrieves semantically similar expert-verified NL2SQL exemplars via sentence embeddings, and synthesizes executable SQL conditioned on the decomposition, retrieved exemplars, and schema, with post-processing validity checks. To improve with use, FD-NL2SQL incorporates two update signals: (i) clinician edits of generated SQL are approved and added to the exemplar bank; and (ii) lightweight logic-based SQL augmentation applies a single atomic mutation (e.g., operator or column change), retaining variants only if they return non-empty results. A second LLM generates the corresponding natural-language question and predicate decomposition for accepted variants, automatically expanding the exemplar bank without additional annotation. The demo interface exposes decomposition, retrieval, synthesis, and execution results to support interactive refinement and continuous improvement.

[Demo] FinReporting: An Agentic Workflow for Localized Reporting of Cross-Jurisdiction Financial Disclosure ▶ *Underline Video*

Fan Zhang, Mingzi Song, Rania Elbadry, Yankai Chen, Shaobo Wang, Yixi Zhou, Xunwen Zheng, Yueru He, Yuyang Dai, Georgi Georgiev, Ayesha Gull, Muhammad Usman Safder, Fan Wu, Liyuan Meng, Fengxian Ji, Junning Zhao, Xueqing Peng, Jimin Huang, Yu Chen, Xue (Steve)Liu, Preslav Nakov and Zhuohan Xie

Financial reporting systems increasingly use large language models (LLMs) to extract and summarize corporate disclosures. However, most assume a single-market setting and do not address structural differences across jurisdictions. Variations in accounting taxonomies, tagging infrastructures (e.g., XBRL vs. PDF), and aggregation conventions make cross-jurisdiction reporting a semantic alignment and verification challenge. We present FinReporting, an agentic workflow for localized cross-jurisdiction financial reporting. The system builds a unified canonical ontology over Income Statement, Balance Sheet, and Cash Flow, and decomposes reporting into auditable stages including filing acquisition, extraction, canonical mapping, and anomaly logging. Rather than using LLMs as free-form generators, FinReporting deploys them as constrained verifiers under explicit decision rules and evidence grounding. Evaluated on annual filings from the US, Japan, and China, the system improves consistency and reliability under heterogeneous reporting regimes. We release an interactive demo supporting cross-market inspection and structured export of localized financial statements.

[Demo] OpenGlass: A Sensing-Computing Split Architecture for Local MLLM-Driven Real-Time Visual Assistance ► *Underline Video*

Mengzhang Li and Yuan Yao

We present OpenGlass, an open-source, privacy-preserving system for low-latency multimodal assistance in real-world deployments, with a primary focus on blind and low-vision users. Deploying large multimodal models for assistive use faces a practical tension between capability and interaction: cloud or server inference can provide strong visual understanding but often incurs multi-second delays and network jitter, while wearable devices can continuously sense the environment but have limited compute and power budgets. OpenGlass resolves this gap with a split local-first design: a lightweight glasses-side sensing unit captures visual context, while a nearby consumer-grade device performs multimodal inference and streams concise spoken feedback over local Wi-Fi or cellular connections, avoiding reliance on remote cloud services. We introduce a system-level evaluation suite covering response quality, end-to-end latency from query completion to audio playback, and safety-first abstention with auditable logs. OpenGlass supports three representative interaction patterns: mobile scene guidance for obstacle avoidance, user initiated interruption of speech output, and multi-turn question answering grounded in the current scene. In real deployment on consumer-grade hardware, OpenGlass achieves sub-2s system-level end-to-end latency from query completion to first audio playback.

[Demo] Praat++: Multimedia Annotation System for Speech and Vocalization ► *Underline Video*

Weiran Zhang and Kenny Zhu

High-quality time-aligned annotation is fundamental to speech processing and animal vocalization research, yet precise boundary localization and consistent labeling remain challenging in collaborative settings. We present Praat++, a web-based multimedia annotation system designed for collaborative, video-informed, and AI-assisted timeline labeling of audio and video data. The system tightly synchronizes waveform, spectrogram, pitch, intensity, and time-aligned video playback with fine-grained region-based editing, enabling precise boundary refinement and improved label accuracy within a unified interface. Praat++ further incorporates role-aware workflow management and human-in-the-loop AI-assisted pre-annotation to enhance inter-annotator consistency and reduce labeling time. Through real-world multimodal speech and animal vocalization annotation scenarios, we demonstrate that Praat++ provides an integrated infrastructure for improving annotation quality and efficiency in dataset construction workflows. The demo video (https://www.youtube.com/watch?v=YboCoBRF5lg) and website (https://redgiant.uta.edu/praat) are now publicly available.

[Demo] Ryze: Evidence-Enriched Data Synthesis from Biomedical Papers ► *Underline Video*

Yeqi Huang, Yue Chen, Yanwei Ye, Guanhao Su and Luo Mai

General-purpose VLMs remain unreliable for biomedical research because valid answers in scientific papers depend on evidence split across figures, tables, charts, captions, and referring text. Existing post-training pipelines are bottlenecked by costly expert annotation and by synthetic data that drops this evidence structure. We present Ryze, a fully automated system that converts raw biomedical papers into an evidence-enriched training set and a domain-specialized VLM\{}@. Ryze synthesizes QA pairs with complete supporting evidence (visual element, caption, extracted structure, and referring paragraphs), reduces layout and OCR errors via chart/table-aware extraction and LLM-based cleansing, and applies a two-stage post-training strategy combining supervised fine-tuning with reinforcement learning. Starting from Qwen3-VL-8B, Ryze produces BioVLM-8B at under $200, achieving 48.0% weighted accuracy on LAB-Bench—outperforming the base model by +12.6% and surpassing GPT-5.2 by +3.8%. We release Ryze as open source together with the trained BioVLM-8B model.

[Demo] TruthSplit: Revealing Conditional Validity in Arguments Through Multi-Worldview Comparative Reasoning ► *Underline Video*

Benjamin Stieger, Maximilian Terberger, Thomas Huber and Christina Niklaus;

We present TruthSplit, an interactive system for comparative argument analysis across ideological worldviews. While existing argumentation tools assess arguments against a single standard of correctness, TruthSplit models conditional validity: how the validity of a claim depends on worldview-specific values, assumptions, and conceptual definitions. Given an input argumentative text, TruthSplit extracts claims and premises, applies three-layer natural language inference (NLI) to assess logical and normative consistency, and conditions large language model (LLM) reasoning on structured worldview profiles that encode core values and decision principles. The system

generates worldview-specific interpretations, identifies value conflicts and assumption gaps, and visualizes divergence through interactive analytical interfaces. Thus, by integrating argument mining, NLI-based consistency modeling, structured knowledge representation, and LLM-conditioned reasoning into a unified framework for multi-perspective argument analysis, TruthSplit enables a systematic comparison of how different worldviews interpret the same argument.

Oral Session E: Multimodality and Language Grounding to Vision, Robotics and Beyond 1

11:00-12:30 (Harbor B-C)

[Long] Revisit What You See: Revealing Visual Semantics in Vision Tokens to Guide LVLM Decoding ▶ *Underline Video*
Beomsik Cho and Jaehyung Kim
Large Vision-Language Models (LVLMs) achieve strong performance across multimodal tasks by integrating visual perception with language understanding. However, how vision information contributes to the models decoding process remains under-explored, as reflected in frequent hallucinations. Through a series of analyses, we found that (i) vision tokens provide meaningful visual information even when hallucinations occur, and (ii) their semantics are encoded in the textual space and become explicit under appropriate vocabulary constraints. Building on these observations, we propose ReVisiT, a simple training-free decoding method that guides text generation in LVLMs by Referencing Vision Tokens. Our approach leverages the semantic information embedded within vision tokens by projecting them into the text token distribution. Specifically, ReVisiT dynamically selects the most relevant vision token at each decoding step via context-aware constrained divergence minimization. Then, ReVisiT uses its constrained projection to refine the output distribution to better incorporate visual semantics. Across five benchmarks on recent LVLMs, ReVisiT consistently enhances visual grounding with minimal computational overhead, and achieves competitive or superior results to state-of-the-art decoding baselines while reducing computational cost by up to 2x.

Orals Session E: Clinical and Biomedical Applications

11:00-12:30 (Harbor G)

[Long] ICDAGENT: Empowering Agentic Large Language Models for Explainable Medical Coding ▶ *Underline Video*
Ziyi Yin, Yuanpu Cao, Ting Wang, Jinghui Chen and Fenglong Ma
The explainable medical coding task aims to automatically assign International Classification of Diseases (ICD) codes to clinical notes while providing explicit justifications for each assignment. Recent approaches employ large language models (LLMs) to generate such explanations. However, their performance remains limited due to a lack of understanding of the clinical meanings of ICD codes. Additionally, the vast ICD code space further complicates the task of accurate prediction. To address these challenges, we propose the ICDAGENT framework, which consists of two collaborative LLM agents: a coding agent and a critical agent. The coding agent extracts ICD codes and generates preliminary rationales, while the critical agent performs fine-grained chain-of-thought reasoning to verify and refine them. Furthermore, the critical agent is trained with a rationale-aware reward, combined with reinforcement learning, enabling it to distinguish between correct and incorrect reasoning and ensure explanation accuracy. Experiments across multiple ICD coding standards and datasets demonstrate that ICDAGENT achieves effective ICD coding with accurate and trustworthy explanations.

[Long] MedMCP-Calc: Benchmarking LLMs for Realistic Medical Calculator Scenarios via MCP Integration ▶ *Underline Video*
Yakun Zhu, Yutong Huang, Shengqian Qin, Zhongzhen Huang, Shaoting Zhang and Xiaofan Zhang
Medical calculators are fundamental to quantitative, evidence-based clinical practice. However, their real-world use is an adaptive, multi-stage process, requiring proactive EHR data acquisition, scenario-dependent calculator selection, and multi-step computation, whereas current benchmarks focus only on static single-step calculations with explicit instructions. To address these limitations, we introduce MedMCP-Calc, the first benchmark for evaluating LLMs in realistic medical calculator scenarios through Model Context Protocol (MCP) integration. MedMCP-Calc comprises 118 scenario tasks across 4 clinical domains, featuring fuzzy task descriptions mimicking natural queries, structured EHR database interaction, external reference retrieval, and process-level evaluation. Our evaluation of 23 leading models reveals critical limitations: even top performers like GPT-5 exhibit substantial gaps, including difficulty selecting appropriate calculators for end-to-end workflows given fuzzy queries, poor performance in iterative SQL-based database interactions, and marked reluctance to leverage external tools for numerical computation. Performance also varies considerably across clinical domains. Building on these findings, we develop CalcMate, a fine-tuned model incorporating scenario planning and tool augmentation, achieving state-of-the-art performance among open-source models.

[Long] MedVerse: Efficient and Reliable Medical Reasoning via DAG-Structured Parallel Execution ▶ *Underline Video*
Jianwen Chen, Xinyu Yang, Peng Xia, Arian Azarang, Yueh Z Lee, Gang Li, Hongtu Zhu, Yun Li, Beidi Chen and Huaxiu Yao

Large language models (LLMs) have demonstrated strong performance and rapid progress in a wide range of medical reasoning tasks. However, their sequential autoregressive decoding forces inherently parallel clinical reasoning, such as differential diagnosis, into a single linear reasoning path, limiting both efficiency and reliability for complex medical problems. To address this, we propose MedVerse, a reasoning framework for complex medical inference that reformulates medical reasoning as a parallelizable directed acyclic graph (DAG) process based on Petri Net theory. The framework adopts a full-stack design across data, model architecture, and system execution. For data creation, we introduce the MedVerse Curator, an automated pipeline that synthesizes knowledge-grounded medical reasoning path and transforms them into Petri Net-structured representations. At the architectural level, we propose a topology-aware attention mechanism with adaptive position indices that supports parallel reasoning while preserving logical consistency. Systematically, we develop a customized inference engine that supports parallel execution without additional overhead. Empirical evaluations show that MedVerse improves strong general-purpose LLMs by up to 8.9%. Compared to specialized medical LLMs, MedVerse achieves comparable performance with improved clinical reliability, while delivering a 1.3x reduction in inference latency and a 1.7x increase in generation throughput, enabled by its parallel decoding capability.

[Long] PubMed Reasoner: Dynamic Reasoning-based Retrieval for Evidence-Grounded Biomedical Question Answering ► *Underline Video*

Yiqing Zhang, Xiaozhong Liu and Fabricio Murai

Trustworthy biomedical question answering (QA) systems must not only provide accurate answers but also justify them with current, verifiable evidence. Retrieval-augmented approaches partially address this gap but lack mechanisms to iteratively refine poor queries, whereas self-reflection methods kick in only after full retrieval is completed. In this context, we introduce PubMed Reasoner, a biomedical QA agent composed of three stages: self-critic query refinement evaluates MeSH terms for coverage, alignment, and redundancy to enhance PubMed queries based on partial (metadata) retrieval; reflective retrieval processes articles in batches until sufficient evidence is gathered; and evidence-grounded response generation produces answers with explicit citations. PubMed Reasoner with a GPT-4o backbone achieves 78.32% accuracy on PubMedQA, slightly surpassing human experts, and showing consistent gains on MMLU Clinical Knowledge. Moreover, LLM-as-judge evaluations prefer our responses across: reasoning soundness, evidence grounding, clinical relevance, and trustworthiness. By orchestrating retrieval-first reasoning over authoritative sources, our approach provides practical assistance to clinicians and biomedical researchers while controlling compute and token costs.

[Long] RespiraMFM: A Multimodal Foundation Model with Contrastive Audio-Language Alignment for Respiratory Disease Identification ► *Underline Video*

Shakhrul Iman Siam, Tiantian Feng, Jiankun Zhang, Shrikanth Narayanan and Mi Zhang

Respiratory diseases remain a leading cause of global mortality, where timely and accurate diagnosis is critical to improving patient outcomes and reducing healthcare burdens. While prior work has explored audio-based models for respiratory disease detection, such unimodal approaches often suffer from limited generalizability and diagnostic precision. In this paper, we propose RespiraMFM, a Multimodal Foundation Model that integrates respiratory sounds with patient medical history and symptoms to enhance diagnostic accuracy and disease detection capabilities. We introduce an effective contrastive alignment strategy for audio-text multimodal integration, allowing the model to learn better cross-modal representations between respiratory sounds and corresponding textual clinical information. We evaluate RespiraMFM across five major respiratory diseases using seven real-world datasets in both supervised fine-tuning and zero-shot settings, achieving a 9.15% improvement in AUROC on supervised tasks and a 20.98% gain on zero-shot tasks over existing baselines. These findings underscore the potential of our framework to advance early diagnosis and improve clinical decision-making in respiratory disease management.

[Long] Responsible Evaluation of AI for Mental Health ► *Underline Video*

Hiba Arnaout, Anmol Goel, H. Andrew Schwartz, Steffen T. Eberhardt, Dana Atzil-Slonim, Gavin Doherty, Brian Schwartz, Wolfgang Lutz, Tim Althoff, Munmun De Choudhury, Hamidreza Jamalabadi, Raj Sanjay Shah, Flor Miriam Plaza-del-Arco, Dirk Hovy, Maria Liakata and Iryna Gurevych

Although artificial intelligence (AI) shows growing promise for mental health care, current approaches to evaluating AI tools in this domain remain fragmented and poorly aligned with clinical practice, social context, and first-hand user experience. This paper argues for a rethinking of responsible evaluation – what is measured, by whom, and for what purpose – by introducing an interdisciplinary framework that integrates clinical soundness, social context, and equity, providing a structured basis for evaluation. Through an analysis of 135 recent *CL publications, we identify recurring limitations, including over-reliance on generic metrics that do not capture clinical validity, therapeutic appropriateness, or user experience, limited participation from mental health professionals, and insufficient attention to safety and equity. To address these gaps, we propose a taxonomy of AI mental health support types – assessment-, intervention-, and information synthesis-oriented – each with distinct risks and evaluative requirements, and illustrate its use through case studies.

[Long] Speech-Hands: A Self-Reflection Voice Agentic Approach to Speech Recognition and Audio Reasoning with Omni Perception ► *Underline Video*

Zhen Wan, Chao-Han Huck Yang, Jinchuan Tian, Hanrong Ye, Ankita Pasad, Szu-Wei Fu, Arushi Goel, Ryo Hachiuma, Shizhe Diao, Kunal Dhawan, Sreyan Ghosh, Yusuke Hirota, Zhehuai Chen, Rafael Valle, Chenhui Chu, Shinji Watanabe, Boris Ginsburg and Yu-Chiang Frank Wang

We introduce a voice-agentic framework that learns one critical omni-understanding skill: knowing when to trust itself versus when to consult

external audio perception. Our work is motivated by a crucial yet counterintuitive finding: naively fine-tuning an omni-model on both speech recognition and external sound understanding tasks often degrades performance, as the model can be easily misled by noisy hypotheses. To address this, our framework, Speech-Hands, recasts the problem as an explicit self-reflection decision. This learnable reflection primitive proves effective in preventing the model from being derailed by flawed external candidates. We show that this agentic action mechanism generalizes naturally from speech recognition to complex, multiple-choice audio reasoning. Across the OpenASR leaderboard, which includes seven domain-diverse speech datasets, Speech-Hands consistently outperforms strong baselines by 12.1% WER on the OpenASR benchmark. The model also achieves 77.37% accuracy and high F1 on audio QA decisions, showing robust generalization and reliability across diverse audio question answering datasets. By unifying perception and decision-making, our work offers a practical path toward more reliable and resilient audio intelligence.

[Long] Text-Attributed Knowledge Graph Enrichment with Large Language Models for Medical Concept Representation ► *Underline Video*

Mohsen Nayebi Kerdabadi, Arya Hadizadeh Moghaddam, Chen Chen, Dongjie Wang and Zijun Yao

In electronic health record (EHR) mining, learning high-quality representations of medical concepts (e.g., standardized diagnosis, medication, and procedure codes) is fundamental for downstream clinical prediction. However, robust concept representation learning is hindered by two key challenges: (i) clinically important cross-type dependencies (e.g., diagnosis-medication and medication-procedure relations) are often missing or incomplete in existing ontology resources, limiting the ability to model complex EHR patterns; and (ii) rich clinical semantics are often missing from structured resources, and even when available as text, are difficult to integrate with KG structure for representation learning. To address these challenges, we present MedCo, an LLM-empowered graph learning framework for medical concept representation. MedCo first builds a global knowledge graph (KG) over medical codes by combining statistically reliable associations mined from EHRs with type-constrained LLM prompting to infer semantic relations. It then utilizes LLMs to enrich the KG into a text-attributed graph by generating node descriptions and edge rationales, providing semantic signals for both concepts and their relationships. Finally, MedCo jointly trains a LoRA-tuned LLaMA text encoder with a heterogeneous GNN, fusing text semantics and graph structure into unified concept embeddings. Extensive experiments on MIMIC-III and MIMIC-IV show that MedCo consistently improves prediction performance and serves as an effective plug-in concept encoder for standard EHR pipelines.

[Long] TIGER: Text-Informed Generalized Enzyme-Reaction Retrieval ► *Underline Video*

Yuhang Zhang, Keyan Ding, Peilin Chen, Han Liu, Can Lin, Ruixi Chen, Shiqi Wang and Qi Song

Enzyme-reaction retrieval is a fundamental problem in computational biology, underpinning enzyme characterization, reaction mechanism elucidation, and the rational design of metabolic pathways and biocatalysts. As a bidirectional task, it entails both enzyme-to-reaction and reaction-to-enzyme mapping. However, existing approaches suffer from poor generalization across tasks and distributions, with performance highly sensitive to dataset splits and substantial asymmetry between retrieval directions. To address these challenges, we present TIGER, a Text-Informed Generalized Enzyme-Reaction Retrieval framework that leverages protein-to-text generation models to distill textual semantic knowledge from enzyme sequences, providing a generalized representation that bridges enzymes and biochemical reactions. To ensure the quality and reliability of textual semantics, we design a Dynamic Gating Network that adaptively fuses text-derived knowledge with sequence features, enabling more consistent and informative enzyme representations, while a Structure-Shared Feature Projector aligns enzyme and reaction representations within a unified latent space. Extensive experiments demonstrate that, under bidirectional retrieval supervision, TIGER significantly outperforms state-of-the-art baselines across diverse distributions and exhibits strong robustness and transferability across tasks.

Orals Session E: Code Models

11:00-12:30 (Regatta)

[Long] Agentic Rubrics as Contextual Verifiers for SWE Agents ► *Underline Video*

Mohit Raghavendra, Anisha Gunjal, Bing Liu and Yunzhong He

Verification is critical for improving agents: it provides the reward signal for Reinforcement Learning and enables inference-time gains through Test-Time Scaling (TTS). Despite its importance, verification in software engineering (SWE) agent settings often relies on code execution, which can be difficult to scale due to environment setup overhead. Scalable alternatives such as patch classifiers and heuristic methods exist, but they are less grounded in codebase context and harder to interpret. To this end, we explore Agentic Rubrics: an expert agent interacts with the repository to create a context-grounded rubric checklist, and candidate patches are then scored against it without requiring test execution. On SWE-Bench Verified under parallel TTS evaluation, Agentic Rubrics achieve a score of 54.2% on Qwen3-Coder-30B-A3B and 40.6% on Qwen3-32B, with at least a +3.5 percentage-point gain over the strongest baseline in our comparison set. We further analyze rubric behavior, showing that rubric scores are consistent with ground-truth tests while also flagging issues that tests do not capture. Our ablations show that agentic context gathering is essential for producing codebase-specific, unambiguous criteria. Together, these results suggest that Agentic Rubrics provide an efficient, scalable, and granular verification signal for SWE agents.

[Long] CodeHacker: Automated Test Case Generation for Detecting Vulnerabilities in Competitive Programming Solutions ▶ *Underline Video*

Jingwei Shi, Xinxiang Yin, Jing Huang, Jinman Zhao and Shengyu Tao

The evaluation of Large Language Models (LLMs) for code generation relies heavily on the quality and robustness of test cases. However, existing benchmarks often lack coverage for subtle corner cases, allowing incorrect solutions to pass. To bridge this gap, we propose CodeHacker, an automated agent framework dedicated to generating targeted adversarial test cases that expose latent vulnerabilities in program submissions. Mimicking the hack mechanism in competitive programming, CodeHacker employs a multi-strategy approach, including stress testing, anti-hash attacks, and logic-specific targeting to break specific code submissions. To ensure the validity and reliability of these attacks, we introduce a Calibration Phase, where the agent iteratively refines its own Validator and Checker via self-generated adversarial probes before evaluating contestant code. Experiments demonstrate that CodeHacker significantly improves the True Negative Rate (TNR) of existing datasets, effectively filtering out incorrect solutions that were previously accepted. Furthermore, generated adversarial cases prove to be superior training data, boosting the performance of RL-trained models on benchmarks like LiveCodeBench. All code, datasets, and evaluation scripts will be open-sourced to promote further investigation in LLMs for competitive programming.

[Long] CODERL+: Improving Code Generation via Reinforcement with Execution Semantics Alignment ▶ *Underline Video*

Xue Jiang, Yihong Dong, Mengyang Liu, Hongyi Deng, Tian Wang, Yongding Tao, Zhi Jin, Wenpin Jiao and Ge Li

While Large Language Models (LLMs) excel at code generation by learning from vast code corpora, a fundamental semantic gap remains between their training on textual patterns and the goal of functional correctness, which is governed by formal execution semantics. Reinforcement Learning with Verifiable Rewards (RLVR) approaches attempt to bridge this gap using outcome rewards from executing test cases. However, solely relying on binary pass/fail signals is inefficient for establishing a well-aligned connection between the textual representation of code and its execution semantics, especially for subtle logical errors within the code. In this paper, we propose CODERL+, a novel approach that integrates execution semantics alignment into the RLVR training pipeline for code generation. CODERL+ enables the model to infer variable-level execution trajectory, providing a direct learning signal of execution semantics. CODERL+ can construct execution semantics alignment directly using existing on-policy rollouts and integrates seamlessly with various RL algorithms. Extensive experiments demonstrate that CODERL+ outperforms post-training baselines (including RLVR and Distillation), achieving a 4.6% average relative improvement in pass@1. CODERL+ generalizes effectively to other coding tasks, yielding 15.5% and 4.4% higher accuracy on code-reasoning and test-output-generation benchmarks, respectively. CODERL+ shows strong applicability across diverse RL algorithms and LLMs. Furthermore, probe analyses provide compelling evidence that CODERL+ strengthens the alignment between codes textual representations and its underlying execution semantics.

[Long] EyeMulator: Improving Code Language Models by Mimicking Human Visual Attention ▶ *Underline Video*

Yifan Zhang, Chen Huang, Yueke Zhang, Jiahao Zhang, Toby Jia-Jun Li, Collin McMillan, Kevin Leach and Yu Huang

Code Language Models (CodeLLMs) traditionally learn attention based solely on statistical input-output token correlations ("machine attention"). In contrast, human developers rely on intuition, selectively fixating on semantically salient tokens during program comprehension. We present EyeMulator, a model-agnostic technique to align CodeLLM attention with human visual attention without architectural changes. By extracting scan paths from eye-tracking data, we derive token-level attention weights used to augment the loss function during fine-tuning. This induces the model to mimic human focus. Our evaluation across StarCoder, Llama-3.2, and DeepSeek-Coder shows that EyeMulator significantly outperforms baselines, achieving gains of over 30 CodeBLEU points in translation and up to 22 BERTScore points in summarization. Ablation studies confirm that these gains stem directly from replicating human attention dynamics. Artifacts are available at https://zenodo.org/records/17205682.

[Long] From Laboratory to Real-World Applications: Benchmarking Agentic Code Reasoning at the Repository Level ▶ *Underline Video*

Jia Li, Yuxin Su and Michael R. Lyu

As large language models (LLMs) evolve into autonomous agents, evaluating repository-level reasoning, the ability to maintain logical consistency across massive, real-world, interdependent file systems, has become critical. Current benchmarks typically fluctuate between isolated code snippets and black-box evaluations. We present RepoReason, a white-box diagnostic benchmark centered on abductive assertion verification. To eliminate memorization while preserving authentic logical depth, we implement an execution-driven mutation framework that utilizes the environment as a semantic oracle to regenerate ground-truth states. Furthermore, we establish a fine-grained diagnostic system using dynamic program slicing, quantifying reasoning via three orthogonal metrics: ESV (reading load), MCL (simulation depth), and DFI (integration width). Comprehensive evaluations of frontier models (e.g., Claude-4.5-Sonnet, DeepSeek-v3.1-Terminus) reveal a prevalent aggregation deficit, where integration width serves as the primary cognitive bottleneck. Our findings provide granular white-box insights for optimizing the next generation of agentic software engineering.

[Long] SeCuRepair: Semantics-Aligned, Curriculum-Driven, and Reasoning-Enhanced Vulnerability Repair Framework ▶ *Underline Video*

Chengran Yang, Ting Zhang, Jinfeng Jiang, Xin Zhou, Haoye Tian, Mingzhe Du, Jieke Shi, Junkai Chen, Yikun Li, Eng Lieh Ouh, Lwin Khin Shar and David Lo

The rapid accumulation of software vulnerabilities has outpaced manual remediation, creating an urgent need for Automated Vulnerability

Repair (AVR). However, existing methods suffer from syntactic overfitting, mimicking surface forms without understanding the underlying repair logic, and fail to generalize to complex fixes. To transcend these limitations, we propose SeCuRepair, a reliable, scalable, and efficient RL-based AVR framework. By introducing a semantic-aware reward, SeCuRepair optimizes for code semantic equivalence rather than lexical mimicry. Furthermore, SeCuRepair incorporates an expert-aligned reasoning mechanism that explicitly grounds patch generation in a structured diagnosis. Finally, SeCuRepair introduces a difficulty-based curriculum that progressively disentangles the optimization barriers of entangled multi-hunk repairs. Extensive evaluations on a rigorous repository-level split show that SeCuRepair substantially outperforms state-of-the-art baselines, as confirmed by both automatic evaluation and human study.

[Long] Selective Knowledge Distillation: Fusing LLM Semantic Strengths with DNN Efficiency for Binary Code Similarity Detection ▶ *Underline Video*

Shize Zhou, Peiyu Liu, Lirong Fu, Tong Ye and Wenhai Wang

Binary Code Similarity Detection (BCSD) plays a vital role in various security applications, including vulnerability identification, malware analysis, and code plagiarism detection. With the growing adoption of deep neural networks (DNNs), substantial progress has been made in recognizing and classifying similar code segments. However, DNN-based BCSD methods often exhibit low accuracy and robustness because they struggle to capture fine-grained and high-level program semantics. In contrast, such semantics are typically captured through natural language interpretations of source code by large language models (LLMs). Yet, LLM-based BCSD methods are constrained by their large model sizes and high inference latency. To alleviate these limitations, this paper proposes BinSKD. The key idea is to leverage an LLM-based BCSD method as the teacher model and transfer its knowledge of high-level program semantics to various DNN-based student models. Specifically, to avoid propagating errors from the teacher to the student, we introduce selective distillation, selecting targets with accurate semantics according to their detection retrieval. In addition, to mitigate the noise introduced by a number of negative samples during distillation, we further propose discrepancy-weighted sampling to focus on the samples where the students prediction notably deviates from the teachers. Our experiments show that BinSKD yields Recall@1 improvements of 14.5%-91.2% for DNN-based BCSD methods and enables HermesSim to match the teacher's performance with orders-of-magnitude efficiency.

[Long] The Path Not Taken: Duality in Reasoning about Program Execution ▶ *Underline Video*

Eshgin Hasanov, Md. Mahadi Hassan, Santu Karmaker and Aashish Yadavally

Large language models (LLMs) have shown remarkable capabilities across diverse coding tasks. However, their adoption requires a true understanding of program execution rather than relying on surface-level patterns. Existing benchmarks primarily focus on predicting program properties tied to specific inputs (e.g., code coverage, program outputs). As a result, they provide a narrow view of dynamic code reasoning and are prone to data contamination. We argue that understanding program execution requires evaluating its inherent duality through two complementary reasoning tasks: (i) predicting a program's observed behavior for a given input, and (ii) inferring how the input must be mutated toward a specific behavioral objective. Both tasks jointly probe a model's causal understanding of execution flow. We instantiate this duality in DexBench, a benchmark comprising 445 paired instances, and evaluate 13 LLMs. Our results demonstrate that dual-path reasoning provides a robust and discriminative proxy for dynamic code understanding.

[Long] XOXO: Stealthy Cross-Origin Context Poisoning Attacks against AI Coding Assistants ▶ *Underline Video*

Adam torek, Mukur Gupta, Noopur Bhatt, Aditya Gupta, Janie Kim, Prashast Srivastava and Suman Jana

AI coding assistants automatically gather context from potentially untrusted sources to generate code recommendations. We introduce Cross-Origin Context Poisoning (XOXO), a novel attack that exploits this automatic context inclusion by subtly manipulating code without changing its semantics. Attackers introduce semantics-preserving transformations (e.g., renamed variables) to shared code, causing AI assistants to unknowingly recommend vulnerable code patterns to victims. To systematically identify effective transformations, we present Greedy Cayley Graph Search (GCGS), a black-box algorithm that efficiently composes transformations to identify adversarial inputs. Our evaluation demonstrates XOXO's effectiveness at making LLMs generate buggy and vulnerable code, achieving average attack success rates of 73.20% against eight state-of-the-art models including GPT 4.1 and Claude 3.5 Sonnet v2, with vulnerability injection rates up to 66.67%. We also demonstrate a real-world attack against GitHub Copilot, highlighting critical security gaps in current AI coding tools.

Orals Session E: Human-AI Interaction/Cooperation and Human-Centric NLP 1

11:00-12:30 (Harbor A)

[Long] Communication-Efficient Desire Alignment for Proactive Embodied Human-Agent Interaction ▶ *Underline Video*

Yuanfei Wang, Xinju Huang, Fangwei Zhong, Yaodong Yang, Yizhou Wang, Yuanpei Chen and Hao Dong

Effective real-world human-agent interactions, such as household robotic services, are often long-term and repeated. Beyond executing tasks, agents are expected to quickly become familiar with individual users. In everyday use, people do not want to repeatedly specify precise instructions. Instead, they prefer agents that adapt to their habits and preferences over interaction while minimizing communication effort. This poses a key challenge: enabling agents to rapidly align with user needs and provide proactive assistance within limited communication. To study this problem in a realistic embodied setting, we first introduce HA-Desire, a home assistance simulation environment. HA-Desire

features an LLM-driven proxy user with value-driven preferences and natural language behavior, enabling systematic evaluation of how agents adapt to users across interactions and satisfy their desires. We further propose FAMER, a framework that integrates goal-relevant memory, desire-centered mental reasoning, and efficient communication to infer user preferences from interaction while reducing unnecessary dialogue. Experiments across embodied household tasks and different LLMs show that FAMER improves both task success and interaction efficiency compared to existing baselines, highlighting the importance of communication-efficient desire alignment for proactive embodied agents that support users without requiring frequent instructions.

[Long] ExPerT: Personalizing LLM Responses to Users' Domain Expertise via Query-Wise Semantic and Keystroke Behavioral Cues ▶ *Underline Video*

Yeji Park, Jiwon Tark and Taesik Gong

Large language models (LLMs) are increasingly used by end users, yet existing personalization methods relying on static profiles or text-only signals fail to capture query-specific expertise variation. We present ExPerT, a query-wise personalization framework that adapts LLM responses to users' query domain expertise by combining semantic and behavioral cues. ExPerT consists of two key components: (i) a semantic-behavioral expertise inference module that jointly interprets query text and keystroke dynamics via in-context LLM prompting, and (ii) an expertise-conditioned response generation that adapts the level of detail, terminology, and conceptual complexity. Our user study with 40 participants and 1270 queries demonstrated that ExPerT reduced expertise inference error by 65.7% compared to the strongest baseline (MAE = 0.398 vs. 1.162) and improved response satisfaction by 17.52% (from 3.71 to 4.36) on a 5point Likert scale.

[Long] Feeling Right vs. Being Right: How AI Sycophancy Affects Value-Laden Deliberation ▶ *Underline Video*

Jeongwoo Ryu, Soomin Kim, Jinsu Eun, Kyusik Kim, Changhoon Oh and Bongwon Suh

As people increasingly turn to AI for personal deliberation beyond task-oriented assistance, concerns about sycophancy in these value-laden contexts have grown. Unlike human flattery, which is intentional and self-interested, AI sycophancy emerges as a byproduct of RLHF's reward structure for user-preference alignment. Yet the observable behavior is similar: both produce responses that preserve what users want to hear. Focusing on this phenomenon through Goffman's face-work framework, we operationalize AI sycophancy as excessive face-saving, either active (preserving positive face through agreement) or passive (preserving negative face by withholding challenge). In a mixed-methods study (N=31), participants engaged with AI across three moral dilemmas under these conditions and a non-sycophantic neutral baseline. Sycophantic responses increased decision confidence but reduced open-minded thinking; participants felt supported yet found the conversations unproductive. Neutral responses, though initially uncomfortable, promoted cognitive flexibility and meaningful deliberation. These findings reveal a confidence-competence trade-off in AI-mediated moral reasoning and suggest that effective AI for personal deliberation requires calibrated friction, not unconditional agreement.

[Long] IEvoAgent: Evolving Conversational Agent based on User Implicit Feedback ▶ *Underline Video*

Yichen Cai, Jiayang Li, Junyuan Qiu, Jingya Guo, Weitao You, Changyuan Yang, Lingyun Sun and Pei Chen

Current conversational agents often follow static learning paradigms and miss the implicit, evolving feedback embedded in users follow-up behaviors. We propose IEvoAgent, an evolving conversational agent framework that leverages the structured dependency between agent responses and user reactions. We construct an annotated dataset from LMSYS-Chat-1M and WildChat and find consistent response-conditioned feedback patterns. Based on this finding, IEvoAgent uses a conditional feedback distribution matrix to estimate expected feedback rewards, combining offline KTO alignment with an inference-time prompt-evolution mechanism driven by a dynamic matrix. Experiments on MT-Bench-101, WildBench, and FB-Bench show improvements over open-source baselines, indicating that mining implicit feedback supports better multi-turn alignment under evolving user preferences. Our code and dataset are available at https://github.com/Hualeez/IEvoAgent.

[Long] LVLMs and Humans Ground Differently in Referential Communication ▶ *Underline Video*

Peter Zeng, Weiling Li, Amie Paige, Zhengxiang Wang, Panagiotis Kaliosis, Dimitris Samaras, Gregory Zelinsky, Susan Brennan and Owen Rambow

For generative AI agents to partner effectively with human users, the ability to accurately predict human intent is critical. But this ability to collaborate remains limited by a critical deficit: an inability to model common ground. We present a referential communication experiment with a factorial design involving director-matcher pairs (human-human, human-AI, AI-human, and AI-AI) that interact with multiple turns in repeated rounds to match pictures of objects not associated with any obvious lexicalized labels. We show that LVLMs cannot interactively generate and resolve referring expressions in a way that enables smooth communication, a crucial skill that underlies human language use. We release our corpus of 356 dialogues (89 pairs over 4 rounds each) along with the online pipeline for data collection and the tools for analyzing accuracy, efficiency, and lexical overlap.

[Long] Measuring Human Contribution in AI-Assisted Content Generation ▶ *Underline Video*

Yueqi Xie, Tao Qi, Jingwei Yi, Xiyuan Yang, Ryan Whalen, Junming Huang, Qian Ding, Yu Xie, Xing Xie and Fangzhao Wu

With the growing prevalence of generative AI, an increasing amount of content is no longer exclusively generated by humans but by generative AI models with human guidance. This shift presents notable challenges for the delineation of originality due to the varying degrees of human contribution in AI-assisted works. This study raises the research question of measuring human contribution in AI-assisted content generation and introduces a framework to address this question that is grounded in information theory. By calculating mutual information

between human input and AI-assisted output relative to self-information of AI-assisted output, we quantify the proportional information contribution of humans in content generation. Our experimental results demonstrate that the proposed measure effectively discriminates between varying degrees of human contribution across multiple creative domains. To further enhance real-world applicability, we extend the framework to estimate the minimal necessary human contribution for any text without requiring human input and validate its effectiveness. We hope that this work lays a foundation for measuring human contributions in AI-assisted content generation in the era of generative AI.

[Long] Stable Signer: Hierarchical Sign Language Generative Model ▶ *Underline Video*

Sen Fang, Yalin Feng, Hongbin Zhong, Yanxin Zhang and Dimitris N. Metaxas

Sign Language Production (SLP) is the process of converting the complex input text into a real video. Most previous works focused on the Text2Gloss, Gloss2Pose, Pose2Vid stages, and some concentrated on Prompt2Gloss and Text2Avatar stages. However, this field has made slow progress due to the inaccuracy of text conversion, pose generation, and the rendering of poses into real human videos in these stages, resulting in gradually accumulating errors. Therefore, in this paper, we streamline the traditional redundant structure, simplify and optimize the task objective, and design a new sign language generative model called Stable Signer. It redefines the SLP task as a hierarchical generation end-to-end task that only includes text understanding (Prompt2Gloss, Text2Gloss) and Pose2Vid, and executes text understanding through our proposed new Sign Language Understanding Linker called SLUL, and generates hand gestures through the named SLP-MoE hand gesture rendering expert block to end-to-end generate high-quality and multi-style sign language videos. SLUL is trained using the newly developed Semantic-Aware Gloss Masking Loss (SAGM Loss). Its performance has improved by 48.6% compared to the current SOTA generation methods, which is a significant increase in the SLP field. More demo can be obtained at https://anonymoussubmissionurl.github.io/Stable-Signer.

[Long] What Makes LLMs Effective Sequential Recommenders? A Study on Preference Intensity and Temporal Context ▶ *Underline Video*

Zhongyu Ouyang, Qianlong Wen, Chunhui Zhang, Yefang Ye and Soroush Vosoughi

What enables large language models (LLMs) to effectively model user preferences in sequential recommendation? Our investigation reveals that existing preference-alignment approaches largely rely on binary pairwise comparisons, overlooking two critical factors: preference intensity (the structured strength of affinity or aversion) and temporal context (the extent to which recent interactions better reflect a user's current intent). Through controlled experiments, we show that leveraging comprehensive feedback with structured preference signals substantially improves recommendation performance, indicating that binary modeling discards essential information. Motivated by these findings, we propose RecPO, a unified preference optimization framework that maps both explicit and implicit feedback into a common preference signal and constructs adaptive reward margins that jointly account for preference intensity and interaction recency. Experiments across five datasets show that RecPO consistently outperforms state-of-the-art baselines while exhibiting behavioral patterns aligned with human decision-making, including favoring immediate satisfaction, maintaining preference coherence, and avoiding dispreferred items. Our results highlight that preference intensity and temporal context are fundamental ingredients for effective LLM-based recommendation. Code: https://github.com/zyouyang/RecPO

[Long] Your Students Dont Use LLMs Like You Wish They Did ▶ *Underline Video*

Sebastian Kobler, Matthew Clemson, Angela Sun and Jonathan K. Kummerfeld

Educational NLP systems are typically evaluated using engagement metrics and satisfaction surveys, which are at best a proxy for meeting pedagogical goals. We introduce six computational metrics for automated evaluation of pedagogical alignment in student-AI dialogue. We validate our metrics through analysis of 12,650 messages across 500 conversations from four courses. Using our metrics, we identify a fundamental misalignment: educators design conversational tutors for sustained learning dialogue, but students mainly use them for answer-extraction. Deployment context is the strongest predictor of usage patterns, outweighing student preference or system design: when AI tools are optional, usage concentrates around deadlines; when integrated into course structure, students ask for solutions to verbatim assignment questions. Whole-dialogue evaluation misses these turn-by-turn patterns. Our metrics will enable researchers building educational dialogue systems to measure whether they are achieving their pedagogical goals.

Orals Session E: LLM Efficiency 1

11:00-12:30 (Harbor E-F)

[Long] From Local to Global: Revisiting Structured Pruning Paradigms for Large Language Models ▶ *Underline Video*

Ziyan Wang, Enmao Diao, Qi Le, Pu Wang, Minwoo Lee, Shu-ping Yeh, Evgeny Stupachenko, Hao Feng and Li Yang

Structured pruning is a practical approach to deploying large language models (LLMs) efficiently, as it yields compact, hardware-friendly architectures. However, the dominant local paradigm is task-agnostic: by optimizing layer-wise reconstruction rather than task objectives, it tends to preserve perplexity or generic zero-shot behavior but fails to capitalize on modest task-specific calibration signals, often yielding limited downstream gains. We revisit global structured pruning and present GISP, Global Iterative Structured Pruning, a post-training method that removes attention heads and MLP channels using first-order, loss-based important scores aggregated at the structure level

with block-wise normalization. Built on this global importance metric, GISP adopts an iterative schedule, rather than one-shot pruning, stabilizes accuracy at higher sparsity, and mitigates perplexity collapse without requiring intermediate fine-tuning. Importantly, the iterative pruning forms nested subnetworks that support a ”prune-once, deploy-many” workflow. Furthermore, GISP defines structural importance directly with respect to a target loss, making it easy to adapt pruning to task-specific objectives. In this work, we use perplexity for language modeling and a margin-based objective for decision-style tasks. Extensive experiments show that across Llama27B/13B, Llama38B, and Mistral0.37B, GISP consistently lowers WikiText2 perplexity and improves downstream accuracy, with especially strong gains at 40-50% sparsity; on DeepSeek-R1-Distill-Llama-3-8B and Qwen3-8B with GSM8K, taskaligned calibration substantially boosts exactmatch accuracy.

[Long] GMSA: Enhancing Context Compression via Group Merging and Layer Semantic Alignment ▶ *Underline Video*

Jiwei Tang, Zhicheng Zhang, Shunlong Wu, Jingheng Ye, Lichen Bai, Zitai Wang, Tingwei Lu, Lin Hai, Yiming Zhao, Hai-Tao Zheng and Hong-Gee Kim

Large Language Models (LLMs) have achieved remarkable performance across a wide range of Natural Language Processing (NLP) tasks. However, in long-context scenarios, they face two challenges: high computational cost and information redundancy. To address these challenges, we propose GMSA, an encoder-decoder context compression framework that generates a compact sequence of soft tokens for downstream tasks. GMSA introduces Group Merging to achieve more uniform aggregation, mitigating semantic dominance during autoencoder pretraining, and Layer Semantic Alignment (LSA) to bridge the semantic gap between high-level abstract semantics and low-level input semantics. We first pretrain GMSA as an autoencoder and then fine-tune it for downstream tasks. Experiments demonstrate that GMSA improves context reconstruction compared to existing soft prompt compression paradigm and outperforms baselines on multiple long-context question answering and summarization benchmarks across two backbone models, while maintaining low end-to-end latency.

[Long] LongSpec: Long-Context Lossless Speculative Decoding with Efficient Drafting and Verification ▶ *Underline Video*

Penghui Yang, Cunxiao Du, Fengzhuo Zhang, Haonan Wang, Tianyu Pang, Chao Du and Bo An

As Large Language Models (LLMs) can now process extremely long contexts, efficient inference over these extended inputs has become increasingly important, especially for emerging applications like LLM agents that highly depend on this capability. Speculative decoding (SD) offers a promising lossless acceleration technique compared to lossy alternatives such as quantization and model cascades. However, most state-of-the-art SD methods are trained on short texts (typically fewer than 4k tokens), making them unsuitable for long-context scenarios. Specifically, adapting these methods to long contexts presents three key challenges: (1) the excessive memory demands posed by draft models due to large Key-Value (KV) cache; (2) performance degradation resulting from the mismatch between short-context training and long-context inference; and (3) inefficiencies in tree attention mechanisms when managing long token sequences. This work introduces LongSpec, a framework that addresses these challenges through three core innovations: a memory-efficient draft model with a constant-sized KV cache; novel position indices that mitigate the training-inference mismatch; and an attention aggregation strategy that combines fast prefix computation with standard tree attention to enable efficient decoding. Experimental results confirm the effectiveness of LongSpec, achieving up to a 3.26x speedup over strong Flash Attention baselines across five long-context understanding datasets, as well as a 2.34x reduction in wall-clock time on four math reasoning tasks with the QwQ model, demonstrating significant latency improvements for long-context applications.

[Long] OCP: Outlier-Centric Probing for Dynamic Structured Pruning of LLMs ▶ *Underline Video*

Yang Ji and Ying Sun

Structured pruning offers a hardware-friendly approach for efficient LLM inference. Early static methods determine fixed subnetworks through offline calibration, suffering from performance degradation and calibration sensitivity. Recent methods explore input-adaptive pruning by selecting a subset of tokens as probes to estimate hidden activations for online pruning decisions. However, existing probe selection strategies fail to identify outlier-triggering tokens, and uniform layer-wise sparsity misaligns with heterogeneous outlier distributions, leading to critical channels being incorrectly pruned. Therefore, we propose OCP (Outlier-Centric Probing for structured pruning), a principled framework that prioritizes capturing outlier-triggering tokens rather than reconstructing full hidden distributions. Specifically, OCP includes three key components: (1) sensitivity-weighted probing for FFN layers that identifies outlier patterns via precomputed weight aggregations, (2) attention-accumulated probing that leverages preceding attention matrices to identify salient tokens, and (3) online adaptive sparsity allocation that dynamically adjusts layer-wise pruning based on history-guided outlier statistics. Extensive experiments on LLaMA2, LLaMA3, and OPT demonstrate that OCP consistently outperforms state-of-the-art methods across benchmarks, achieving up to 25% perplexity reduction at 1.6x speedup.

[Long] TA-GRPO-d: Trajectory-Aware GRPO for Optimizing Denoising Trajectories in Diffusion LLMs ▶ *Underline Video*

Gyunyeop Kim and Sangwoo Kang

Diffusion large language models (dLLMs) generate text by repeatedly unmasking a partially noised sequence in parallel, promising lower latency than autoregressive decoding. However, most discrete dLLMs still rely on fixed denoising schedules, which are nonadaptive to input difficulty and cannot learn efficient unmasking orders. This paper introduces a reinforcement learning (RL) framework that transforms dLLM decoding into a trajectory-aware, learnable policy. We propose a confidence-gated denoising strategy that dynamically decides which tokens to unmask and how many to unmask per step, enabling adaptive exploration of denoising trajectories. Building on Group Relative Policy Optimization, we reformulate it into a trajectory-aware variant, TA-GRPO-d, which combines a trajectory-level signal, captured as the zscore of the AUC over intermediate rewards, with a token-level unmasking-time weight. This design allows the model to learn not only the final

output quality but also the efficiency of the decoding path itself. Experiments on MATH-500, Countdown, Sudoku, and code benchmarks (HumanEval, MBPP) show that TA-GRPO-d maintains or improves accuracy while reducing average denoising steps by up to half, achieving both faster inference and lower computational cost. Our approach provides an RL framework for optimizing dLLM decoding policies toward adaptive, efficient reasoning. Code is available at our GitHub.

[Long] TokenTiming: A Dynamic Alignment Method for Universal Speculative Decoding Model Pairs ▶ *Underline Video*

Sibo Xiao, Fu Jinyuan, Zhongle Xie and Lidan Shou

Accelerating the inference of large language models (LLMs) has been a critical challenge in generative AI. Speculative decoding (SD) substantially improves LLM inference efficiency. However, its utility is limited by a fundamental constraint: the draft and target models must share the same vocabulary, thus limiting the herd of available draft models and often necessitating the training of a new model from scratch. Inspired by Dynamic Time Warping (DTW), a classic algorithm for aligning time series, we propose the algorithm TokenTiming for universal speculative decoding. It operates by re-encoding the draft token sequence to get a new target token sequence, and then uses DTW to build a mapping to transfer the probability distributions for speculative sampling. Benefiting from this, our method accommodates mismatched vocabularies and works with any off-the-shelf models without retraining and modification. We conduct comprehensive experiments on various tasks, demonstrating 1.57x speedup. This work enables a universal approach for draft model selection, making SD a more versatile and practical tool for LLM acceleration.

[Long] UniSpec: Training-Free Speculative Decoding for Robust LLM Acceleration Across Languages and Hardware ▶ *Underline Video*

Do Dinh Truong, Le Nguyen Khang and Nguyen Le Minh

Speculative decoding accelerates large language model (LLM) inference through a draft-and-verify paradigm, yet existing methods face three key limitations: reliance on fixed draft templates that ignore device-specific verification costs, lack of mechanisms to assess draft token quality, and suboptimal tree expansion strategies. We introduce UniSpec, a training-free, lossless speculative decoding framework that enables robust, plug-and-play LLM acceleration across diverse hardware configurations and languages. UniSpec incorporates three novel components: (1) a device-aware calibration mechanism that determines the optimal draft size by measuring the acceptance-time trade-off on each target device; (2) a confidence score estimation module that assigns quality scores to n-grams based on the verifier's token probabilities, enabling selective retention of high-quality draft candidates; and (3) an improved tree expansion strategy that broadens first-level exploration and applies threshold-based filtering to prune low-confidence nodes. To comprehensively evaluate multilingual performance, we create a comprehensive benchmark, covering seven languages across seven generation tasks. Experiments with various LLM architectures, hardware environments, and languages demonstrate that UniSpec consistently outperforms existing training-free methods, achieving speedups of up to 2.6x while maintaining output quality identical to standard autoregressive decoding. Our code and benchmark are publicly available.

[Long] What Makes Good Instruction-Tuning Data? An In-Context Learning Perspective ▶ *Underline Video*

Guangzeng Han and Xiaolei Huang

Instruction-tuning datasets often contain substantial redundancy and low-quality samples, necessitating effective data selection methods. We propose an instruction data selection framework based on weighted in-context influence (wICI), which measures how effectively each candidate example reduces instruction-following difficulty for semantically related peers. Through systematic experiments, we address three key questions: what constitutes effective instruction tuning data from an in-context perspective, whether sample difficulty correlates with in-context influence, and how in-context influence translates to instruction tuning effectiveness. Experiments across multiple models and benchmarks demonstrate that our method consistently outperforms existing baselines under constrained data budgets, while empirically showing that sample difficulty negatively correlates with in-context influence.

[Long] ZoomR: Memory Efficient Reasoning through Multi-Granularity Key Value Retrieval ▶ *Underline Video*

David H. Yang, Yuxuan Zhu, Mohammad Mohammadi Amiri, Keerthiram Murugesan, Tejaswini Pedapati, Subhajit Chaudhury and Pin-Yu Chen

Large language models (LLMs) have shown great performance on complex reasoning tasks but often require generating long intermediate thoughts before reaching a final answer. During generation, LLMs rely on a key-value (KV) cache for autoregressive decoding. However, the memory footprint of the KV cache grows with output length. Prior work on KV cache optimization mostly focus on compressing the long input context, while retaining the full KV cache for decoding. For tasks requiring long output generation, this leads to increased computational and memory costs. In this paper, we introduce ZoomR, a novel approach that enables LLMs to adaptively compress verbose reasoning thoughts into summaries and uses a dynamic KV cache selection policy that leverages these summaries while also strategically "zooming in" on fine-grained details. By using summary keys as a coarse-grained index during decoding, ZoomR uses the query to retrieve details for only the most important thoughts. This hierarchical strategy significantly reduces memory usage by avoiding full-cache attention at each step. Experiments across math and reasoning tasks show that our approach achieves competitive performance compared to baselines, while reducing inference memory requirements by more than 4 x. These results demonstrate that a multi-granularity KV selection enables more memory efficient decoding, especially for long output generation.

Orals Session E: Low-resource Methods for NLP 2

11:00-12:30 (Harbor H-I)

[Long] Investigating Cross-Modal Skill Injection: Scenarios, Methods, and Hyperparameters ▶ *Underline Video*

Zhiyu Xu, Lean Wang, Yuanxin Liu, Lei Li, Hao Zhou, Fandong Meng, Jie Zhou and Xu Sun

Vision-Language Models (VLMs) have demonstrated remarkable proficiency in general multi-modal understanding; yet they struggle to efficiently acquire continually evolving domain-specific skills. Conventional approaches to enhancing VLM capabilities, such as Supervised Fine-Tuning (SFT), require extensive dataset curation and substantial computational resources. Model merging has emerged as an efficient alternative that enables the transfer of domain-specific expertise from Large Language Models (LLMs) to VLMs without incurring additional training data requirements or significant computational overhead. Unlike conventional merging of homogeneous LLMs, which mainly aggregates existing capabilities, cross-modal skill injection aims to induce emergent cross-modal capabilities by integrating a domain-expert LLM into a VLM. However, existing research lacks a systematic analysis of the applicability and methodology of cross-modal skill injection. In this study, we investigate cross-modal skill injection across three main aspects: scenarios, methods, and hyperparameters. For scenarios, we find that cross-modal skill injection generally performs well in instruction-following and cross-lingual settings, yet struggles with mathematical reasoning. For methods, we find that classic approaches such as TA and DARE consistently achieve superior performance over alternative merging methods. We also provide a systematic and quantitative analysis of the hyperparameter tuning that these classic methods critically depend on.

[Long] K-Merge: Online Continual Merging of Adapters for On-device Large Language Models ▶ *Underline Video*

Donald Shenaj, Ondrej Bohdal, Taha Ceritli, Mete Ozay, Pietro Zanuttigh and Umberto Michieli

On-device deployment of Large Language Models (LLMs) frequently leverages Low-Rank Adapters (LoRAs) to support diverse downstream tasks under tight resource constraints. To address the limited storage capacity of mobile devices, recent works have explored model merging techniques to fuse multiple LoRAs into a single one. In practice, however, LoRAs are often delivered incrementally, as users request support for new tasks (e.g., novel problem types or languages). This scenario introduces a new challenge: on-device online continual merging, where the objective is to incorporate new LoRAs while preserving the performance on previously supported tasks. In this paper, we propose a data-free and computationally efficient strategy for selecting and merging LoRAs when a new one becomes available, assuming the device can store only a limited number of adapters. Extensive experiments across real-world tasks demonstrate the superiority of our approach compared to alternative strategies while adhering to the storage budget and compute limitations of on-device settings. The project page is available at: https://donaldssh.github.io/K-Merge.

[Long] LEAF: Knowledge Distillation of Text Embedding Models with Teacher-Aligned Representations ▶ *Underline Video*

Robin Vujanic and Thomas RückstieSS

We present a knowledge distillation framework for text embedding models. A key distinguishing feature is that our distilled models are compatible with their teacher, enabling flexible asymmetric architectures where documents are encoded with the larger teacher model, while queries use smaller student models. We also show that our models automatically inherit MRL and robustness to output quantization whenever these properties are present in the teacher model, without explicitly training for them. To demonstrate the effectiveness of our framework we publish leaf-ir, a 23M parameters information retrieval oriented model that, besides being teacher-compatibile, sets a new state-of-the-art (SOTA) on BEIR, ranking no.1 on the public leaderboard for models of its size. Asymmetric mode further increases its retrieval performance. Our scheme is however not restricted to information retrieval. We demonstrate its wider applicability by synthesizing the multi-task leaf-mt model. This also sets a new SOTA, achieving no.1 on the public MTEB v2 (English) leaderboard for models of its size. Our technique is applicable to black-box models, requires no judgments nor hard negatives, and training can be conducted using small batch sizes. Thus, dataset and training infrastructure requirements for our framework are modest. We make our models publicly available under a permissive Apache 2.0 license.

[Long] LightReasoner: Can Small Language Models Teach Large Language Models Reasoning? ▶ *Underline Video*

Jingyuan Wang, Yankai Chen, Zhonghang Li and Chao Huang

Large language models (LLMs) have demonstrated remarkable progress in reasoning, often through supervised fine-tuning (SFT). However, SFT is resource-intensive, relying on large curated datasets, rejection-sampled demonstrations, and uniform optimization across all tokens, even though only a fraction carry meaningful learning value. In this work, we explore a counterintuitive idea: can smaller language models (SLMs) teach larger language models (LLMs) by revealing high-value reasoning moments that reflect the latter's unique strength? We propose LightReasoner, a novel framework that leverages the behavioral divergence between a stronger expert model (LLM) and a weaker amateur model (SLM). LightReasoner operates in two stages: (1) a sampling stage that pinpoints critical reasoning moments and constructs supervision examples capturing the expert's advantage through expert-amateur contrast, and (2) a fine-tuning stage that aligns the expert model with these distilled examples, amplifying its reasoning strengths. Across seven benchmarks, LightReasoner improves accuracy by up to 28.1%, while reducing time consumption by 90%, sampled problems by 80%, and tuned token usage by 99%, all without relying on ground-truth labels. By turning weaker SLMs into effective teaching signals, LightReasoner offers a scalable and resource-efficient approach for advancing LLM reasoning.

[Long] Logic Matters in Lightweight Hallucination Classification for RAG System ▶ *Underline Video*
Ningyuan Yang and Kaizhu Huang
We propose a lightweight, modular framework for hallucination detection in Retrieval-Augmented Generation (RAG) systems, addressing the critical challenge where logical dependencies span across fragmented retrieval results. To address the inherent limitations of compact models in processing long-context information and performing multi-hop reasoning, our approach systematically analyzes the logical relationships among retrieved documents within the vector space. By capturing these geometric patterns through a novel feature extraction framework, the proposed classifier significantly enhances context-aware hallucination detection without requiring complex architectures or pre-training on datasets. Meanwhile, to evaluate multi-document reasoning, we release HotPotQA-derived, a hallucination dataset preserving separate retrieved texts. Experimental results on HotPotQA-derived and several open-source datasets demonstrate that our framework can achieve results comparable to or even surpassing those of large language models (LLMs) on the task of hallucination detection.

[Long] S2O: Early Stopping for Sparse Attention via Online Permutation ▶ *Underline Video*
Yu Zhang, Songwei Liu, Chenqian Yan, linsheng linsheng, BeichenNing BeichenNing, Fangmin Chen and Xing Wang
Attention scales quadratically with sequence length, fundamentally limiting long-context inference. Existing block-granularity sparsification can reduce latency, but coarse blocks impose an intrinsic sparsity ceiling, making further improvements difficult even with carefully engineered designs. We present S2O, which performs early stopping for sparse attention via online permutation. Inspired by virtual-to-physical address mapping in memory systems, S2O revisits and factorizes FlashAttention execution, enabling inference to load non-contiguous tokens rather than a contiguous span in the original order. Motivated by fine-grained structures in attention heatmaps, we transform explicit permutation into an online, index-guided, discrete loading policy; with extremely lightweight preprocessing and index-remapping overhead, it concentrates importance on a small set of high-priority blocks. Building on this importance-guided online permutation for loading, S2O further introduces an early-stopping rule: computation proceeds from high to low importance; once the current block score falls below a threshold, S2O terminates early and skips the remaining low-contribution blocks, thereby increasing effective sparsity and reducing computation under a controlled error budget. As a result, S2O substantially raises the practical sparsity ceiling. On Llama-3.1-8B under a 128K context, S2O reduces single-operator MSE by 3.82x at matched sparsity, and reduces prefill compute density by 3.31x at matched MSE; meanwhile, it preserves end-to-end accuracy and achieves 7.51x attention and 3.81x end-to-end speedups.

[Long] SARA: Selective and Adaptive Retrieval-augmented Generation with Context Compression ▶ *Underline Video*
Yiqiao Jin, Kartik Sharma, Vineeth Rakesh, Yingtong Dou, Menghai Pan, Mahashweta Das and Srijan Kumar
Retrieval-augmented generation (RAG) extends large language models (LLMs) with external knowledge, but it must balance limited effective context, redundant retrieved evidence, and the loss of fine-grained facts under aggressive compression. Pure compression-based approaches reduce input size but often discard fine-grained details essential for factual accuracy. We propose SARA, a hybrid RAG framework that targets answer quality under fixed token budgets by combining natural-language snippets with semantic compression vectors. SARA retains a small set of passages in text form to preserve entities and numerical values, compresses the remaining evidence into interpretable vectors for broader coverage, and uses those vectors for iterative evidence reranking. Across 9 datasets and 5 open-source LLMs spanning 3 model families (Mistral, Llama, and Gemma), SARA consistently improves answer relevance (+17.71), answer correctness (+13.72), and semantic similarity (+15.53), demonstrating the importance of integrating textual and compressed representations for robust, context-efficient RAG.

[Long] TeRA: Vector-based Random Tensor Network for High-Rank Adaptation of Large Language Models ▶ *Underline Video*
Yuxuan Gu, Wuyang Zhou, Giorgos Iacovides and Danilo Mandic
Parameter-Efficient Fine-Tuning (PEFT) methods, such as Low-Rank Adaptation (LoRA), have significantly reduced the number of trainable parameters needed in fine-tuning large language models (LLMs). The developments of LoRA-style adapters have considered two main directions: (1) enhancing model expressivity with high-rank adapters, and (2) aiming for further parameter reduction, as exemplified by vector-based methods. However, these approaches come with a trade-off, as achieving the expressivity of high-rank weight updates typically comes at the cost of sacrificing the extreme parameter efficiency offered by vector-based techniques. To address this issue, we propose a vector-based random Tensor network for high-Rank Adaptation (TeRA), a novel PEFT method that achieves high-rank weight updates while retaining the parameter efficiency of vector-based PEFT adapters. This is achieved by parametrizing the tensorized weight update matrix as a Tucker-like tensor network (TN), whereby large randomly initialized factors are frozen and shared across layers, while only small layer-specific scaling vectors, corresponding to diagonal entries of factor matrices, are trained. Comprehensive experiments demonstrate that TeRA matches or even outperforms existing high-rank adapters, while requiring as few trainable parameters as vector-based methods. Theoretical analysis and ablation studies validate the effectiveness of the proposed TeRA method. The code is available at https://github.com/guyuxuan9/TeRA.

[Long] The African Languages Lab: A Collaborative Approach to Advancing Low-Resource African NLP ▶ *Underline Video*
Sheriff Issaka, Keyi Wang, Yinka Ajibola, Oluwatumininu Samuel-Ipaye, Zhaoyi Zhang, Nicte Aguillon Jimenez, Evans Kofi Agyei, Abraham Lin, Rohan Ramachandran, Sadick Abdul Mumin, Faith Nchifor, Mohammed Shuraim, Erick Rosas Gonzalez, Lieqi Liu, Sylvester Kpei, Jemimah Osei, Carlene Ajeneza, Persis Boateng, Prisca Adwoa Dufie Yeboah and Saadia Gabriel
Despite representing nearly one-third of the world's languages, African languages remain critically underserved by modern NLP technologies, with 88% classified as severely underrepresented or completely ignored in computational linguistics. We present the African Languages Lab (All Lab), a comprehensive research initiative that addresses this technological gap through systematic data collection, model development,

and empirical analysis. Our contributions include: (1) a quality-controlled data collection pipeline, yielding the largest validated African multi-modal speech and text dataset spanning 40 languages with 19 billion text tokens and 12,628 hours of aligned speech data; (2) extensive experimental validation demonstrating that even modest-scale models, when fine-tuned on targeted language data, achieve substantial improvements over untrained baselines, averaging +23.69 ChrF++, +0.33 COMET, and +15.34 BLEU points across 31 evaluated languages; and (3) a comparative analysis against Google Translate in which a 1B-parameter model matched or surpassed the commercial system in several languages including Yoruba and Twi, revealing that data scarcity, rather than model scale, constitutes the primary bottleneck for low-resource NLP, and suggesting that systematic dataset development yields disproportionate returns for low-resource languages.

Orals Session E: Machine Learning for NLP 2

11:00-12:30 (Harbor D)

[Long] Do Not Step Into the Same River Twice: Learning to Reason from Trial and Error ► *Underline Video*

Chenming Tang, Hsiu-Yuan Huang, Weijie Liu, Clive Bai, Saiyong Yang and Yunfang Wu

Reinforcement learning with verifiable rewards (RLVR) has significantly boosted the reasoning capability of language models (LMs). However, existing RLVR approaches train LMs based on their own on-policy responses and are constrained by the initial capability of LMs, thus prone to exploration stagnation, in which LMs fail to solve more training problems and cannot further learn from the training data. Some approaches try to address this by leveraging off-policy solutions to training problems, but rely on external expert guidance that is limited in availability and scalability. In this work, we propose LTE (Learning to reason from Trial and Error), an approach that hints LMs with their previously self-made mistakes, not requiring any external expert guidance. Experiments validate the effectiveness of LTE, which outperforms the normal group relative policy optimization (GRPO) by 5.02 in Pass@1 and 9.96 in Pass@k on average across six mathematical reasoning benchmarks for Qwen3-8B-Base and even performs better than methods that require external guidance. Further analysis confirms that LTE successfully mitigates exploration stagnation and enhances both exploitation and exploration during training. Our code is available at https://github.com/JamyDon/LTE.

[Long] Efficient Learned Data Compression via Dual-Stream Feature Decoupling ► *Underline Video*

Huidong Ma, Xinyan Shi, Hui Sun, Xiaofei Yue, Xiaoguang Liu, Gang Wang and Wentong Cai

While Learned Data Compression (LDC) has achieved superior compression ratios, balancing precise probability modeling with system efficiency remains challenging. Crucially, uniform single-stream architectures struggle to simultaneously capture micro-syntactic and macro-semantic features, necessitating deep serial stacking that exacerbates latency. Compounding this, heterogeneous systems are constrained by device speed mismatches, where throughput is capped by Amdahl's Law due to serial processing. To this end, we propose a Dual-Stream Multi-Scale Decoupler that disentangles local and global contexts to replace deep serial processing with shallow parallel streams, and incorporate a Hierarchical Gated Refiner for adaptive feature refinement and precise probability modeling. Furthermore, we design a Concurrent Stream-Parallel Pipeline, which overcomes systemic bottlenecks to achieve full-pipeline parallelism. Extensive experiments demonstrate that our method achieves state-of-the-art performance in both compression ratio and throughput, while maintaining the lowest latency and memory usage. The code is available at https://github.com/huidong-ma/FADE.

[Long] Flip-Flop Consistency: Unsupervised Training for Robustness to Prompt Perturbations in LLMs ► *Underline Video*

Parsa Hejabi, Elnaz Rahmati, Alireza Salkhordeh Ziabari and Morteza Dehghani

Large Language Models (LLMs) often produce inconsistent answers when faced with different phrasings of the same prompt. In this paper, we propose Flip-Flop Consistency (F^2C), an unsupervised training method that improves robustness to such perturbations. F^2C is composed of two key components. The first, Consensus Cross-Entropy (CCE), uses a majority vote across prompt variations to create a hard pseudo-label. The second is a representation alignment loss that pulls lower-confidence and non-majority predictors toward the consensus established by high-confidence, majority-voting variations. We evaluate our method on 11 datasets spanning four NLP tasks, with 4–15 prompt variations per dataset. On average, F^2C raises observed agreement by 11.62%, improves mean F_1 by 8.94%, and reduces performance variance across formats by 3.29%. In out-of-domain evaluations, F^2C generalizes effectively, increasing F_1 and agreement while decreasing variance across most source-target pairs. Finally, when trained on only a subset of prompt perturbations and evaluated on held-out formats, F^2C consistently improves both performance and agreement while reducing variance. These findings highlight F^2C as an effective unsupervised method for enhancing LLM consistency, performance, and generalization under prompt perturbations.

[Long] Gained in Translation: Privileged Pairwise Judges Enhance Multilingual Reasoning ► *Underline Video*

Lintang Sutawika, Gokul Swamy, Steven Wu and Graham Neubig

When asked a question in a language less seen in its training data, current reasoning large language models (RLMs) often exhibit dramatically lower performance than when asked the same question in English. In response, we introduce SP3F (Self-Play with Privileged Pairwise Feedback), a two-stage framework for enhancing multilingual reasoning without any data in the target language(s). First, we supervise fine-tune (SFT) on translated versions of English question-answer pairs to raise base model correctness. Second, we perform RL with feedback from a pairwise judge in a self-play fashion, with the judge receiving the English reference response as privileged information.

Thus, even when none of the model's responses are completely correct, the privileged pairwise judge can still tell which response is better. End-to-end, SP3F greatly improves base model performance, even outperforming fully post-trained models on multiple math and non-math tasks with less than 1/8 of the training data across the single-language, multilingual, and generalization to unseen language settings.

[Long] IMPACT: Importance-Aware Activation Space Reconstruction ▶ *Underline Video*

Md Mokarram Chowdhury, Daniel Agyei Asante, Ernie Chang and Yang Li

Large language models (LLMs) achieve strong performance across diverse domains but remain difficult to deploy in resource-constrained environments due to their size. Low-rank compression is a common remedy, typically minimizing weight reconstruction error under the assumption that weights are low-rank. However, this assumption often does not hold in LLMs. In contrast, LLM activations exhibit a more pronounced low-rank structure, motivating approaches that minimize activation reconstruction error. This shift alone, however, is not sufficient: different activation dimensions contribute unequally to model performance, and treating them uniformly can lead to accuracy loss. We introduce IMPACT, an importance-aware activation reconstruction framework that links compression to its effect on model performance. IMPACT formulates compression as an optimization problem that integrates activation structure with gradient-based importance, deriving a closed-form solution where reconstruction bases arise from an importance-weighted activation covariance matrix. This yields low-rank compression explicitly optimized for accuracy preservation. Experiments across multiple models and tasks demonstrate that IMPACT achieves up to 55.4% greater model size reduction while maintaining accuracy comparable to or better than state-of-the-art baselines.

[Long] Reinforcement Learning on Pre-Training Data ▶ *Underline Video*

Siheng_Li Siheng_Li, Kejiao Li, Zenan Xu, Guanhua Huang, Kun Li, Haoyuan Wu, Wujiajia Wujiajia, Zihao Zheng, Chenchen Zhang, kun shi, Xue Gong, Qi Yi, Ruibin Xiong, Tingqiang Xu, Yuhao Jiang, Jianfeng Yan, Yuyuan Zeng, Guanghui Xu, Jinbao Xue, Zhijiang xu, Zheng Fang, SHUAI LI, Qibin Liu, Xiaoxue Li, Zhuoyu Li, Yangyu Tao, Fei Gao, Cheng Jiang, Bochao Wang, Kai Liu, Jianchen Zhu, Wai Lam, Bo Zhou and Di Wang

Recent progress in large language models (LLMs) is largely driven by scaling training compute through either pre-training with next-token prediction (NTP) or post-training with reinforcement learning (RL). The former contributes to learning broad knowledge and skills from general data, while struggling with data inefficiency and catastrophic forgetting in continual learning settings. The latter incentivizes reasoning capabilities with strong generalization, but is constrained by limited data availability due to its reliance on human annotation. To alleviate these issues, we propose Reinforcement Learning on Pre-Training data (RLPT), which combines the advantages of learning from general data and RL. In particular, RLPT derives reward signals directly from general text data through a next-segment reasoning objective, rewarding the policy for correctly predicting next text segments conditioned on the prefix text. Experiments across multiple benchmarks and models demonstrate the effectiveness of. For example, RLPT yields substantial improvements in continual pre-training (+4.6%) and provides a strong foundation for post-training (+3.4%) on Qwen3-8B-Base.

[Long] SAGE: Sparse Adaptive Guidance for Dependency-Aware Tabular Data Generation ▶ *Underline Video*

Shuo Yang, Zheyu Zhang, Bardh Prenkaj and Gjergji Kasneci

Generating high-fidelity synthetic tabular data remains a critical challenge for enhancing data availability in privacy-sensitive and low-resource domains. Recent approaches leverage LLMs by representing table rows as sequences, yet suffer from two fundamental limitations: (1) they model feature dependencies densely, introducing spurious correlations; and (2) they assume static relationships between features, ignoring how these dependencies vary with feature values. To overcome these limitations, we introduce SAGE (Sparse Adaptive Guidance), a novel LLM-based generation framework that enforces sparse and dynamic dependency guidance. SAGE discretizes features into value-aware pseudo-features and constructs a mutual information-based sparse dependency graph. This graph adaptively guides generation through explicit context selection or implicit logit correction, enabling LLMs to focus on truly relevant information during synthesis. Our extensive experiments across six datasets and multiple tasks reveal that SAGE not only improves data fidelity and downstream utility, boosting F1 scores by 10% compared to previous LLM-based methods, but also reduces policy violations by one point. These results highlight the importance of adaptive structure in tabular data generation and provide new insights into context-sensitive control of LLMs.

[Long] SATQuest: A Verifier for Logical Reasoning Evaluation and Reinforcement Fine-Tuning of LLMs ▶ *Underline Video*

Yanxiao Zhao, Yaqian Li, Zihao Bo, Rinyoichi Takezoe, Haojia Hui, Mo Guang, Lei Ren, Xiaolin Qin and Kaiwen Long

Large language models (LLMs) exhibit strong general reasoning, yet the community lacks controllable, scalable, and verifiable tools to analyze and improve these abilities. We present SATQuest, a verifier that generates diverse SAT-based reasoning tasks directly from Conjunctive Normal Form (CNF) instances and checks answers objectively with PySAT. SATQuest factorizes evaluation along three orthogonal dimensions, instance, problem type, and question format, enabling fine-grained, multi-dimensional analysis and reinforcement fine-tuning. Randomized CNF generation mitigates memorization and supports reproducible experiments. Using SATQuest, we benchmark a range of open- and closed-weight LLMs and uncover persistent gaps in logical reasoning, particularly on higher-complexity tasks and in transfer beyond familiar mathematical notation to machine or narrative formats. We further show that reinforcement fine-tuning with SATQuest rewards substantially boosts targeted performance and generalizes to larger instances, while cross-format robustness remains challenging. Collectively, SATQuest provides verifier-backed infrastructure for controlled, scalable, and reproducible empirical research on LLM logical reasoning and its training.

[Long] Towards Scalable Lifelong Knowledge Editing with Selective Knowledge Suppression ▶ *Underline Video*

Dahyun JungJaewook LeeHeuiseok Lim
Large language models (LLMs) require frequent knowledge updates to reflect changing facts and mitigate hallucinations. To meet this demand, lifelong knowledge editing has emerged as a continual approach to modify specific pieces of knowledge without retraining the entire model. Existing parameter-editing methods struggle with stability during sequential edits due to catastrophic forgetting. While retrieval-based approaches are proposed to alleviate this issue, their applicability remains limited across various datasets because of high training costs. To address these limitations and enhance scalability in lifelong settings, we propose LightEdit. Our framework first selects relevant knowledge from retrieved information to modify the query effectively. It then incorporates a decoding strategy to suppress the models original knowledge probabilities, thereby enabling efficient edits based on the selected information. Extensive experiments on ZSRE, Counterfact, and RIPE benchmarks demonstrate that LightEdit outperforms existing lifelong knowledge editing methods. Furthermore, by minimizing training costs, LightEdit achieves cost-effective scalability, enabling easy adaptation to various datasets.

Orals Session E: Machine Translation

11:00-12:30 (Promenade)

[Long] Adam's Law: Textual Frequency Law on Large Language Models ▶ *Underline Video*
Hongyuan Lu, Zixuan Li, Zefan Zhang, Bowen Cao and Wai Lam
While textual frequency has been validated as relevant to human cognition in reading speed, its relatedness to Large Language Models (LLMs) is seldom studied. We propose a novel research direction in terms of textual data frequency, which is an understudied topic. to the best of our knowledge. Our framework is composed of three units. First, this paper proposes Textual Frequency Law (TFL), which indicates that frequent textual data should be preferred for LLMs for both prompting and fine-tuning. Since many LLMs are closed-source in their training data, we propose using online resources to estimate the sentence-level frequency. We can then use an input paraphraser to paraphrase the input into a more frequent textual expression. Next, we propose Textual Frequency Distillation (TFD) by querying LLMs to conduct story completion by further extending the sentences in the datasets, and the resulting corpora are used to adjust the initial estimation. Finally, we propose Curriculum Textual Frequency Training (CTFT) that fine-tunes LLMs in an increasing order of sentence-level frequency. Experiments are conducted on our curated dataset Textual Frequency Paired Dataset (TFPD) on math reasoning and machine translation. Results indicate the effectiveness of our framework.

[Long] Adaptive Text2GQL: Integrating Structural Twig Linking and Evolutionary In-Context Learning ▶ *Underline Video*
Fang Niu, Chaokun Wang, Hang Zhang and Songyao Wang
While large language models have revolutionized Text-to-SQL tasks, translating natural language into Graph Query Languages (Text2GQL) remains underexplored due to the topological heterogeneity and syntactic diversity of graph query languages (e.g., Cypher, Gremlin, SPARQL). Existing approaches often struggle with structural hallucinations and lack adaptability in cold-start scenarios. In this paper, we present a unified, training-free Text2GQL framework. First, Structural Twig Linking elevates schema grounding to the identification of semantic substructures ("twigs"), providing robust topological priors. Second, addressing data scarcity, Evolutionary In-Context Learning operates in a Tabula Rasa setting to implicitly construct a self-growing repository of verified examples driven by syntactic utility. Finally, our Adversarial Execution-Guided Correction agent enforces fidelity through synergistic static critique and dynamic verification. Experiments demonstrate significant improvements over baselines in both accuracy and executability across diverse GQLs. The code is available at https://github.com/nf202/Text2Graph.

[Long] Beyond Literal Mapping: Benchmarking and Improving Non-Literal Translation Evaluation ▶ *Underline Video*
Yanzhi Tian, Cunxiang Wang, Zeming Liu, Heyan Huang, Wenbo Yu, Dawei Song, Jie Tang and Yuhang Guo
Large Language Models (LLMs) have significantly advanced Machine Translation (MT), applying them to linguistically complex domains-such as Social Network Services, literature etc. In these scenarios, translations often require handling non-literal expressions, leading to the inaccuracy of MT metrics. To systematically investigate the reliability of MT metrics, we first curate a meta-evaluation dataset focused on non-literal translations, namely MENT. MENT encompasses four non-literal translation domains and features source sentences paired with translations from diverse MT systems, with 7,530 human-annotated scores on translation quality. Experimental results reveal the inaccuracies of traditional MT metrics and the limitations of LLM-as-a-Judge, particularly the knowledge cutoff and score inconsistency problem. To mitigate these limitations, we propose RATE, a novel agentic translation evaluation framework, centered by a reflective Core Agent that dynamically invokes specialized sub-agents. Experimental results indicate the efficacy of RATE, achieving an improvement of at least 3.2 points in combined system- and segment-level correlation with human judgments compared with current methods. Further experiments demonstrate the robustness of RATE to general-domain MT evaluation. Code and dataset are available at: https://github.com/BITHLP/RATE.

[Long] G-IdiomAlign: A Gloss-Pivoted Benchmark for Cross-Lingual Idiom Alignment ▶ *Underline Video*
Fengying Ye, Yanming Sun, Runzhe Zhan, Lidia S. Chao, Zheqi Zhang and Derek F. Wong
Idioms are difficult to transfer across languages due to their non-compositionality and weak surface-form grounding, making literal mappings

unreliable. We present G-IdiomAlign, a gloss-pivoted benchmark where each idiom is anchored by an English gloss from Wiktionary. We further construct a high-confidence reference alignment set for reproducible evaluation. G-IdiomAlign supports two protocols: (1) a controlled Multiple-Choice Idiom Equivalence with typed distractors for error attribution; and (2) a Gloss-Contrastive Generation contrasting No-gloss and With-gloss inputs to isolate the effect of an explicit semantic pivot. Across diverse LLMs, a bias to literal translation is a dominant failure mode, especially when the target is a low-resource language. Glosses consistently improve Gloss-Contrastive Generation under an embedding-based semantic proxy, but performance remains modest, indicating substantial headroom in the open output space. Subsequent analysis on Qwen3-8B further suggests that cross-condition differences are concentrated more in attention heads than in layers, while better With-gloss generations coincide with stronger gloss anchoring.

[Long] Hierarchical Policy Optimization for Simultaneous Translation of Unbounded Speech ▶ *Underline Video*

Siqi Ouyang, Shuoyang Ding, Oleksii Hrinchuk, Vitaly Lavrukhin, Brian Yan, Boris Ginsburg and Lei Li

Simultaneous speech translation (SST) generates translations while receiving partial speech input. Recent advances show that large language models (LLMs) can substantially improve SST quality, but at the cost of high computational overhead. To reduce this cost, prior work reformulates SST as a multi-turn dialogue task, enabling full reuse of the LLMs key-value (KV) cache and eliminating redundant feature recomputation. However, this approach relies on supervised fine-tuning (SFT) data in dialogue form, for which few human annotations exist, and existing synthesis methods cannot guarantee data quality. In this work, we propose a Hierarchical Policy Optimization (HPO) approach that post-train models trained on imperfect SFT data. We introduce a hierarchical reward that balances translation quality and latency objectives. Experiments on English to Chinese/German/Japanese demonstrate improvements of over +7 COMET score and +1.25 MetricX score at a latency of 1.5 seconds. Comprehensive ablation studies further validate the effectiveness of different quality rewards, hierarchical reward formulations, and segmentation strategies.

[Long] NeoAMT: Neologism-Aware Agentic Machine Translation with Reinforcement Learning ▶ *Underline Video*

Zhongtao Miao, Kaiyan Zhao, Masaaki Nagata and Yoshimasa Tsuruoka

Neologism-aware machine translation aims to translate source sentences containing neologisms into target languages. This field remains underexplored compared with general machine translation (MT). In this paper, we propose an agentic framework, NeoAMT, for neologism-aware machine translation equipped with a Wiktionary-based search toolkit. Specifically, we first construct a dedicated dataset for neologism-aware machine translation and build a search toolkit grounded in Wiktionary. The dataset covers 16 languages and 75 translation directions in total, derived from approximately 10 million records of an English Wiktionary dump. The retrieval corpus of the search toolkit is also constructed from around 3 million cleaned records of the same dump. We then leverage the dataset and toolkit to train a translation agent via reinforcement learning (RL) and to evaluate the accuracy of neologism-aware machine translation. Furthermore, we propose an RL training framework featuring a novel reward design and an adaptive rollout generation strategy that exploits translation difficulty to further improve the translation quality of translation agents using our search toolkit.

[Long] Quantifying the Impact of Translation Errors on Multilingual LLM Evaluation ▶ *Underline Video*

Klaudia-Doris Thellmann, Bernhard Stadler, Michael Färber and Jens Lehmann

Machine-translated benchmarks are widely used to assess the multilingual capabilities of large language models (LLMs), yet translation errors in these benchmarks remain underexplored, raising concerns about the reliability and comparability of multilingual evaluation. We address two practical gaps: (i) how well automatic MQM-style error spans from LLM judges and a span-aware QE baseline (xCOMET-XXL) match expert human span annotations on benchmark translations, and (ii) how strongly translation errors (as opposed to source-side issues in the English original) explain accuracy drops on translated benchmarks. We find that span agreement is non-trivial on naturally occurring benchmark translations, and that target-side translation errors are consistently associated with measurable, percentage-point drops in translated accuracy even after controlling for English correctness and source-side anomalies.

[Long] Think in Latent Thoughts: A New Paradigm for Gloss-Free Sign Language Translation ▶ *Underline Video*

Yiyang Jiang, Li Zhang, Xiao-Yong Wei and Li Qing

Many SLT systems quietly assume that brief chunks of signing map directly to spoken-language words. That assumption breaks down because signers often create meaning on the fly using context, space, and movement. We revisit SLT and argue that it is mainly a cross-modal reasoning task, not just a straightforward video-to-text conversion. We thus introduce a reasoning-driven SLT framework that uses an ordered sequence of latent thoughts as an explicit middle layer between the video and the generated text. These latent thoughts gradually extract and organize meaning over time. On top of this, we use a plan-then-ground decoding method: the model first decides what it wants to say, and then looks back at the video to find the evidence. This separation improves coherence and faithfulness. We also built and released a new large-scale gloss-free SLT dataset with stronger context dependencies and more realistic meanings. Experiments across several benchmarks show consistent gains over existing gloss-free methods. Our code and data are available at https://github.com/fletcherjiang/SignThought.

[Long] What Does LLM Refinement Actually Improve? A Systematic Study on Document-Level Literary Translation ▶ *Underline Video*

Shaomu Tan, Dawei Zhu, Ke Tran, Michael Denkowski, Sony Trenous, Bill Byrne, Leonardo Ribeiro and Felix Hieber

Iterative refinement is a simple inference-time strategy for machine translation: given an initial translation, an LLM revises it without additional training. Yet document-scale refinement remains poorly understood: 1) which pipelines work best, 2) what quality dimensions

improve, and 3) how refiners behave. In this paper, we present a systematic study of document-level literary translation, covering six LLMs and seven language pairs. Across nine translation-refinement granularity combinations and five refinement strategies, a) we find a robust recipe: document-level MT followed by segment-level refinement yields the strongest and most stable improvements. In our setting, doc-level refinement often makes fewer edits and leads to smaller or less reliable gains. Surprisingly, a simple general refinement prompt consistently outperforms error-specific prompting and evaluate-then-refine schemes. b) Fine-grained MQM analyses and professional-translator evaluation show that gains come primarily from fluency, with limited improvements in adequacy. c) Probing translator-refiner strength interactions suggests refinement behaves less like targeted post-editing and more like projecting outputs toward the refiner's learned distribution while remaining anchored to the initial translation.

Orals Session E: Multimodality and Language Grounding to Vision, Robotics and Beyond 1

11:00-12:30 (Harbor B-C)

[Long] APB-V: Accelerating Long-Video Understanding via Sequence-Parallelism-aware Approximate Attention ▶ *Underline Video*

Yuxiang Huang, Mingye Li, Xu Han, Chaojun Xiao, Weilin Zhao, Ao Sun, Ziqi Yuan, Hao Zhou, Fandong Meng and Zhiyuan Liu

The efficiency of long-video inference remains a critical bottleneck, mainly due to the dense computation in the prefill stage of Large Multimodal Models (LMMs). Existing methods either compress visual embeddings or apply sparse attention on a single GPU, yielding limited acceleration or degraded performance and restricting LMMs from handling longer, more complex videos. To overcome these issues, we propose APB-V, a sequence-parallel framework with optimized attention that accelerates long-video inference across multiple GPUs. By distributing approximate attention, APB-V reduces computation and increases parallelism, enabling efficient processing of more visual embeddings without compression and thereby improving task performance. System-level optimizations, such as load balancing and fused forward passes, further unleash the potential of APB-V, delivering speedups of 12.72x, 1.70x, and 1.18x over FlashAttn, ZigZagRing, and APB, without notable performance loss.

[Long] Benchmarking Fine-Grained Error Detection in Multimodal Reasoning ▶ *Underline Video*

Chi-Min Chan, Han Zhu, Chunyang Jiang, Jiaming Ji, Juntao Dai, Wei Xue, Sirui Han and Yike Guo

Multimodal Process Reward Models (MPRMs) have emerged as a pivotal framework for enhancing the reasoning capabilities of Multimodal Large Language Models (MLLMs). However, the research community currently lacks a dedicated benchmark to rigorously assess the error discernment capabilities of these models. To address this gap, we introduce PRMBench-V, a novel benchmark specifically designed to evaluate MPRMs proficiency in detecting erroneous reasoning steps across diverse error categories. Leveraging a semi-automated annotation pipeline augmented with human verification, we construct a comprehensive dataset comprising 907 unique queries, each annotated with nine distinct error types, resulting in 8,163 test cases with fine-grained step-level error labels. Through extensive experiments involving over 15 open- and closed-source models, we uncover several key findings: (1) even the strongest existing MPRMs achieve only 30% accuracy in error identification; (2) while partial error detection achieves moderate precision and recall (60%), overall accuracy remains low (20%); and (3) benchmark scores exhibit a strong correlation with downstream task performance gains (r=0.86). Furthermore, we demonstrate that PRMBench-V can inform the development of more robust MPRMs: by introducing the Bayesian Rater Reliability Process Reward Model (BR2-PRM), we achieve up to a 4.8% performance improvement through test-time scaling. We believe that PRMBench-V will serve as a valuable resource for advancing MPRM research, enabling more rigorous evaluation and fostering the development of models with fine-grained multimodal reasoning capabilities.

[Long] CARES: Context-Aware Resolution Selector for VLMs ▶ *Underline Video*

Moshe Kimhi, Nimrod Shabtay, Raja Giryes, Chaim Baskin and Eli Schwartz

Large vision-language models (VLMs) commonly process images at native or high resolution to remain effective across tasks. This inflates visual tokens to 97-99% of total tokens, resulting in high compute and latency, even when low-resolution images would suffice. We introduce CARES, a Context-Aware Resolution Selector, a lightweight preprocessing module that, given an image-query pair, predicts the minimal sufficient input resolution. CARES uses a compact VLM (350M) to extract features and predict when a target pretrained VLM's response converges to its peak ability to answer correctly. Though trained as a discrete classifier over a set of optional resolutions, CARES interpolates continuous resolutions at inference for fine-grained control. Across five multimodal benchmarks spanning documents and natural images, as well as diverse target VLMs, CARES preserves task performance while reducing compute by up to 80%.

[Long] Interleaved Latent Visual Reasoning with Selective Perceptual Modeling ▶ *Underline Video*

Shuai Dong, Siyuan Wang, Xingyu Liu, Chenglin Li, Haowen Hou and zhongyu wei

Interleaved reasoning paradigms enhance Multimodal Large Language Models (MLLMs) with visual feedback but are hindered by the prohibitive computational cost of re-encoding pixel-dense images. A promising alternative, latent visual reasoning, circumvents this bottleneck yet faces limitations: methods either fail to capture intermediate state evolution due to single-step, non-interleaved structures, or sacrifice precise perceptual modeling by over-compressing features. We introduce Interleaved Latent Visual Reasoning (ILVR), a framework that unifies dynamic state evolution with precise perceptual modeling. ILVR interleaves textual generation with latent visual representations

that act as specific, evolving cues for subsequent reasoning. Specifically, we employ a self-supervision strategy where a momentum teacher model selectively distills relevant features from ground-truth intermediate images into sparse supervision targets. This adaptive selection mechanism guides the model to autonomously generate context-aware visual signals. Extensive experiments on multimodal reasoning benchmarks demonstrate that ILVR outperforms existing approaches, effectively bridging the gap between fine-grained perception and sequential multimodal reasoning.

[Short] LaMI: Augmenting Large Language Models via Late Multi-Image Fusion ▶ *Underline Video*

Guy Yariv, Idan Schwartz, Yossi Adi and Sagie Benaim

Commonsense reasoning often requires both textual and visual knowledge, yet Large Language Models (LLMs) trained solely on text lack visual grounding (e.g., "what color is an emperor penguin's belly?"). Visual Language Models (VLMs) perform better on visually grounded tasks but face two limitations: (i) often reduced performance on text-only commonsense reasoning compared to text-trained LLMs, and (ii) adapting newly released LLMs to vision input typically requires costly multimodal training. An alternative augments LLMs with test-time visual signals, improving visual commonsense without harming textual reasoning, but prior designs often rely on early fusion and a single image, which can be suboptimal. We propose a late multi-image fusion method: multiple images are generated from the text prompt with a lightweight parallel sampling, and their prediction probabilities are combined with those of a text-only LLM through a late-fusion layer that integrates projected visual features just before the final prediction. Across visual commonsense and NLP benchmarks, our method significantly outperforms augmented LLMs on visual reasoning, matches VLMs on vision-based tasks, and, when applied to strong LLMs such as LLaMA 3, also improves NLP performance while adding only modest test-time overhead.

[Long] Towards Mitigating Modality Bias in Vision-Language Models for Temporal Action Localization ▶ *Underline Video*

Jiaqi Li, Guangming Wang, Shuntian Zheng, Minzhe Ni, Xiaoman Lu, Guanghui Ye and Yu Guan

Temporal Action Localization (TAL) requires identifying both the boundaries and categories of actions in untrimmed videos. While vision-language models (VLMs) offer rich semantics to complement visual evidence, existing approaches tend to overemphasize linguistic priors at the expense of visual performance, leading to a pronounced modality bias. We propose ActionVLM, a vision-language aggregation framework that systematically mitigates modality bias in TAL. Our key insight is to preserve vision as the dominant signal while adaptively exploiting language only when beneficial. To this end, we introduce (i) a debiasing reweighting module that estimates the language advantage, the incremental benefit of language over vision-only predictions, and dynamically reweights language modality accordingly, and (ii) a residual aggregation strategy that treats language as a complementary refinement rather than the primary driver. This combination alleviates modality bias, reduces overconfidence from linguistic priors, and strengthens temporal reasoning. Experiments on THUMOS14 show that our model outperforms state-of-the-art by up to 3.2% mAP. Our code is available at https://github.com/JiaqiLi404/ActionVLM

[Long] ViLL-E: Video LLM Embeddings for Retrieval ▶ *Underline Video*

Rohit Gupta, Jayakrishnan Unnikrishnan, Fan Fei, Sheng Liu, Son Tran and Mubarak Shah

Video Large Language Models (VideoLLMs) excel at video understanding tasks where outputs are textual, such as Video Question Answering and Video Captioning. However, they underperform specialized embedding-based models in Retrieval tasks, such as Text-toVideo Retrieval and Moment Retrieval. We introduce ViLL-E (Video-LLM-Embed), a unified VideoLLM architecture endowed with a novel embedding generation mechanism that allows the model to "think longer" for complex videos and stop early for easy ones. We train this model with a three-stage training methodology combining generative and contrastive learning: initial large-scale pre-training with video-caption pairs; followed by continual training on a smaller, detailed-caption dataset; and concluding with task-specific fine-tuning on a novel multi-task dataset covering Video QA, Temporal Localization, Video Retrieval, and Video-Text Matching. Our model significantly improves temporal localization (on avg. 7% over other VideoLLMs) and video retrieval (up to 4% over dual encoder models), achieving performance comparable to state-of-the-art specialized embedding models while remaining competitive on VideoQA tasks. Furthermore, our joint contrastive-generative training unlocks new zero-shot capabilities, significantly outperforming state-of-the-art methods in composed video retrieval (+5% over SotA) and retrieval from long text (+2% over SotA).

[Long] VisRet: Visualization Improves Knowledge-Intensive Text-to-Image Retrieval ▶ *Underline Video*

Di Wu, Yixin Wan and Kai-Wei Chang

Text-to-image retrieval (T2I retrieval) remains challenging because cross-modal embeddings often behave as bags of concepts, underrepresenting structured visual relationships such as pose and viewpoint. We proposeVisualize-then-Retrieve (VisRet), a retrieval paradigm that mitigates this limitation of cross-modal similarity alignment. VisRet first projects textual queries into the image modality via T2I generation, then performs retrieval within the image modality to bypass the weaknesses of cross-modal retrievers in recognizing subtle visual-spatial features. Across four benchmarks (Visual-RAG, INQUIRE-Rerank, Microsoft COCO, and our new Visual-RAG-ME featuring multi-entity comparisons), VisRet substantially outperforms cross-modal similarity matching and baselines that recast T2I retrieval as text-to-text similarity matching, improving nDCG@30 by 0.125 on average with CLIP as the retriever and by 0.121 with E5-V. For downstream question answering, VisRet increases accuracy on Visual-RAG and Visual-RAG-ME by 3.8% and 15.7% in top-1 retrieval, and by 3.9% and 11.1% in top-10 retrieval. Ablation studies show compatibility with different T2I instruction LLMs, T2I generation models, and downstream LLMs. VisRet provides a simple yet effective perspective for advancing in text-image retrieval. Our code and the new benchmark are publicly available at https://github.com/xiaowu0162/Visualize-then-Retrieve.

Poster Session E (Coronado)

11:00-12:30 (Coronado)

[Findings] A Multi-View Media Profiling Suite: Resources, Evaluation, and Analysis ▶ *Underline Video*

Muhammad Arslan Manzoor, Dilshod Azizov, Daniil Orel, Umer Siddique, Zain Muhammad Mujahid, Yufang Hou and Preslav Nakov

News outlets shape public opinion on a scale, which makes automated detection of political bias and factuality essential. Yet, the field still lacks unified resources, comprehensive evaluations in diverse approaches, and systematic analyzes of the representations and fusion strategies that matter the most, especially under label sparsity and dataset diversity. In addition, there is little empirical work that reports broad observation driven findings about what consistently works, what fails, and why. We address these gaps with four contributions: (i) MBFC-2025, a large-scale label set that covers ~2,600 outlets from Media Bias/Fact Check (MBFC); (ii) multi-view representations for ACL-2020 ~900 outlets and MBFC-2025, spanning Alexa graphs, hyperlink graphs, LLM-derived graphs, articles, and Wikipedia descriptions; (iii) systematic evaluation and analysis of embedding views and fusion strategies, including an RL-based fusion variant; and (iv) extensive experiments that achieve state-of-the-art results on ACL-2020 and establish strong benchmarks on MBFC-2025.

[Findings] A Survey on MLLM-based Visually Rich Document Understanding: Methods, Challenges, and Emerging Trends ▶ *Underline Video*

Yihao Ding, Siwen Luo, Yue Dai, Yanbei Jiang, Zechuan Li, Qiang Sun, Geoffrey Martin, Wei Liu and Yifan Peng

Visually Rich Document Understanding (VRDU) has become a pivotal area of research, driven by the need to automatically interpret documents that contain intricate visual, textual, and structural elements. Recently, Multimodal Large Language Models (MLLMs) have demonstrated significant promise in this domain, including both OCR-based and OCR-free approaches for information extraction from document images. This survey reviews recent advances in MLLM-based VRDU, highlighting emerging trends and promising research directions with a focus on two key aspects: (1) techniques for representing and integrating textual, visual, and layout features; (2) training paradigms, including pretraining, instruction tuning, and training strategies. Moreover, we address challenges such as data scarcity, handling multi-page and multilingual documents, and integrating emerging trends such as Retrieval-Augmented Generation and agentic frameworks. Our analysis offers a roadmap for advancing MLLM-based VRDU toward more scalable, reliable, and adaptable systems.

[Long] AEGIS: A Holistic Benchmark for Evaluating Forensic Analysis of AI-Generated Academic Images ▶ *Underline Video*

Bo Zhang, Tzu-Yen Ma, Zichen Tang, Junpeng Ding, Zirui Wang, Yizhuo Zhao, Peilin Gao, Zijie Xi, Zixin Ding, Haiyang Sun, Haocheng Gao, Yuan Liu, Liangjia Wang, Yiling Huang, Yujie Wang, Yuyue Zhang, Ronghui Xi, Yuanze Li, Jiacheng Liu, Zhongjun Yang and Haihong E

We introduce AEGIS, A holistic benchmark for Evaluating forensic analysis of AI-Generated academic ImageS. Compared to existing benchmarks, AEGIS features three key advances: (1) Domain-Specific Complexity: covering seven academic categories with 39 fine-grained subtypes, exposing intrinsic forensic difficulty, where even GPT-5.1 reaches 48.80% overall performance and expert models achieve only limited localization accuracy (IoU 30.09%); (2) Diverse Forgery Simulations: modeling four prevalent academic forgery strategies across 25 generative models, with 11 yielding average forensic accuracy below 50%, showing that forensics lag behind generative advances; and (3) Multi-Dimensional Forensic Evaluation: jointly assessing detection, reasoning, and localization, revealing complementary strengths between model families, with multimodal large language models (MLLMs) at 84.74% accuracy in textual artifact recognition and expert detectors peaking at 79.54% accuracy in binary authenticity detection. By evaluating 25 leading MLLMs, nine expert models, and one unified multimodal understanding and generation model, AEGIS serves as a diagnostic testbed exposing fundamental limitations in academic image forensics.

[Long] AgentGym2: Benchmarking Large Language Model Agents in De-Idealized Real-World Environments ▶ *Underline Video*

Zhiheng Xi, Dingwen Yang, Jiaqi Liu, Jixuan Huang, Jixuan Huang, Honglin Guo, Baodai Huang, Tinggang Chen, Qi Zhang, Zhonghang Lu, Chenyu Liu, Jiajun Sun, Jiazheng Zhang, Dingwei Zhu, Xin Guo, Junzhe Wang, Zhihao Zhang, Yuming Yang, Junjie Ye, Minghe Gao, Dongrui Liu, Jiaming Ji, Guohao Li, Tao Gui, Qi Zhang and Xuanjing Huang

Language agents, i.e., LLM agents, progress rapidly and are increasingly deployed in production environments. This trend underscores the urgent need for rigorous and realistic evaluations. However, most existing benchmarks evaluate agents in simplified, idealized settings. They typically rely on pre-packaged tool interfaces, overlook critical steps, and assume inputs are clean and fully specified. Consequently, they understate the difficulty of real deployments, where uncertainty and noise are ubiquitous and agents must proactively explore the environment to uncover new tools. To bridge this gap, we present AgentGym2, a new evaluation framework with task instances grounded in real-world end-to-end working demands. Beyond reasoning and planning, it measures agents ability to execute end-to-end procedures, discover tools via exploration, compose tools for unseen tasks, and remain robust to noisy and underspecified information. Experiments on 15 proprietary and open-source models show that even SOTA systems like Gemini and GPT-5 struggle on AgentGym2, revealing a substantial gap between the capability of current agents and the demands of real-world applications.

[Long] Annotating Dimensions of Social Perception in Text: A Sentence-Level Dataset of Warmth and Competence ▶ *Underline Video*

Mutaz Ayesh, Saif M. Mohammad and Nedjma Ousidhoum

Warmth (W) (often further broken down intoTrust (T) and Sociability (S)) and Competence (C) are central dimensions along which people

evaluate individuals and social groups (Fiske, 2018). While these constructs are well established in social psychology, they are only starting to get attention in NLP research through word-level lexicons, which do not fully capture their contextual expression in larger text units and discourse. In this work, we introduceWarmth and Competence Sentences (W&C-Sent), the first sentence-level dataset annotated for warmth and competence. The dataset includes over 1,600 English sentence–target pairs annotated along three dimensions: trust and sociability (components of warmth), and competence. The sentences in W&C-Sent are social media posts that express attitudes and opinions about specific individuals or social groups (the targets of our annotations). We describe the data collection, annotation, and quality-control procedures in detail, and evaluate a range of large language models (LLMs) on their ability to identify trust, sociability, and competence in text. W&C-Sent provides a new resource for analyzing warmth and competence in language and supports future research at the intersection of NLP and computational social science.

[Findings] ARISE: An Adaptive Resolution-Aware Metric for Test-Time Scaling Evaluation in Large Reasoning Models ► *Underline Video*

Zhangyue Yin, Qiushi Sun, Zhiyuan Zeng, Zhiyuan Yu, Qipeng Guo, Xuanjing Huang and Xipeng Qiu

Test-time scaling has emerged as a transformative paradigm for enhancing the performance of large reasoning models, enabling dynamic allocation of computational resources during inference. However, as the landscape of reasoning models rapidly expands, a critical question remains: how can we systematically compare and evaluate the test-time scaling capabilities across different models? In this paper, we introduce ARISE (Adaptive Resolution-aware Scaling Evaluation), a novel metric specifically designed to assess the test-time scaling effectiveness of large reasoning models. Unlike existing evaluation approaches, ARISE incorporates two key innovations: (1) sample-level awareness that effectively penalizes negative scaling behaviors where increased computation leads to performance degradation, and (2) a dynamic sampling mechanism that mitigates the impact of accuracy fluctuations and token count instability on the final assessment. We conduct comprehensive experiments evaluating state-of-the-art reasoning models across diverse domains including mathematical reasoning, code generation, and agentic tasks. Our results demonstrate that ARISE provides a reliable and fine-grained measurement of test-time scaling capabilities, revealing significant variations in scaling efficiency across models. Notably, our evaluation identifies Claude Opus as exhibiting superior scaling characteristics compared to other contemporary reasoning models.

[Findings] ARM2: Adaptive Reasoning Model with Vision Understanding and Executable Code ► *Underline Video*

Jian Xie, Zhendong Chu, Aoxiao Zhong, Kai Zhang, Mingzhe Han, Xing fan, Jialie Shen and Qingsong Wen

Large Reasoning Models (LRMs) often suffer from the "over-thinking" problem, generating unnecessarily long reasoning on simple tasks. Some strategies have been proposed to mitigate this issue, such as length penalties or routing mechanisms, but they are typically heuristic and task-specific, lacking a general framework for adaptive reasoning. In this paper, we present ARM2, a unified model that adaptively balances reasoning performance and efficiency across multiple formats through a reinforcement learning framework augmented with length-aware optimization. Beyond conventional natural language inference, ARM2 integrates vision understanding, extending its applicability to multimodal. Moreover, ARM2 integrates executable code into reasoning, enabling substantial reductions in token cost while preserving task performance compared to long CoT. Experiments demonstrate that ARM2 achieves performance on par with traditional reasoning models trained with GRPO, while reducing token usage by over 70% on average. We further conduct extensive analyses to validate the effectiveness of ARM2 and the soundness of its design.

[Long] Author-in-the-Loop Response Generation and Evaluation: Integrating Author Expertise and Intent in Responses to Peer Review ► *Underline Video*

Qian Ruan and Iryna Gurevych

Author response (rebuttal) writing is a critical stage of scientific peer review that demands substantial author effort. In practice, authors possess domain expertise, author-only information, and response strategies – concrete forms of author expertise and intent – and seek NLP assistance that integrates these signals into author response generation (ARG). Yet this author-in-the-loop paradigm lacks formal NLP formulation and systematic study: no dataset provides fine-grained author signals, existing ARG work lacks author inputs and controls, and no evaluation measures response reflection of author signals and effectiveness in addressing reviewer concerns. To fill these gaps, we introduce (i) ReşAlign, the first large-scale dataset of aligned review-response-revision triplets, where revisions proxy author signals; (ii) REspGen, an author-in-the-loop ARG framework supporting flexible author input, multi-attribute control, and evaluation-guided refinement; and (iii) REspEval, a comprehensive evaluation suite with 20+ metrics spanning input utilization, controllability, response quality, and discourse. Experiments with SOTA LLMs demonstrate the benefits of author input and evaluation-guided refinement, the impact of input specificity on response quality, and controllability-quality trade-offs. We release our dataset, generation and evaluation tools.

[Long] AutoJudger: An Agent-Driven Framework for Efficient Benchmarking of MLLMs ► *Underline Video*

Xuanwen Ding, Chengjun Pan, Zejun Li, Jiwen Zhang, Siyuan Wang and zhongyu wei

Evaluating multimodal large language models (MLLMs) is becoming increasingly expensive as benchmarks grow in scale and cross-modality complexity. Inspired by structuralism in cognitive psychology, we tackle this difficulty with an adaptive evaluation framework for efficient benchmarking, namely AutoJudger. Instead of passively scoring on a fixed test set, AutoJudger treats evaluation as an interview-like process by keeping a hypothesized ability structure of the evaluated model and actively selecting the informative questions so as to refine these ability boundaries. Specifically, AutoJudger has three core components: ability decomposition to organize evaluation along meaningful capability dimensions, ability estimation to maintain an up-to-date quantitative profile of the model competence, and adaptive question

selection to choose the most informative questions. To operationalize this paradigm, we introduce A^2-Judger, a novel MLLM-based Agentic instantiation of AutoJudger equipped with semantic-aware retrieval and dynamic memory. Experiments on four representative multimodal benchmarks show that A^2-Judger significantly improves sample efficiency while maintaining reliable evaluation results.

[Findings] Automated Knowledge Component Generation and Interpretable Knowledge Tracing in Coding Problems ▶ *Underline Video*

Zhangqi Duan, Nigel Fernandez, Arun Balajiee Lekshmi Narayanan, Mohammad Hassany, Rafaella Sampaio de Alencar, Peter Brusilovsky, Bita Akram and Andrew Lan

Knowledge components (KCs) are key to assessing student knowledge levels on fine-grained skills and driving personalization and feedback. However, crafting KCs and tagging them for problems, traditionally performed by human domain experts, is highly labor-intensive. Prior work has studied automated KC generation only for multiple-choice questions but not open-ended ones. We bridge this gap and present an automated, large language model (LLM)-based pipeline for KC generation and tagging for open-ended programming problems. We also develop an LLM-based knowledge tracing (KT) framework to leverage these LLM-generated KCs. We conduct extensive quantitative and qualitative evaluations on two real-world student code submission datasets. Results show that our KT method outperforms existing ones and LLM-generated KCs outperform human-written KCs on future student response prediction. We also investigate how these KCs enable us to analyze student learning curves and conduct human evaluation with course instructors to further verify the quality of KC-problem tagging.

[Long] BenchMarker: An Education-Inspired Toolkit for Highlighting Flaws in Multiple-Choice Benchmarks ▶ *Underline Video*

Nishant Balepur, Bhavya Rajasekaran, Hyunjin Jane Oh, Michael Xie, Atrey Desai, Vipul Gupta, Steven James Moore, Eunsol Choi, Rachel Rudinger and Jordan Lee Boyd-Graber

Multiple-choice question answering (MCQA) is standard in NLP, but benchmarks lack rigorous quality control. We present BenchMarker, an education-inspired toolkit using LLM judges to flag three common MCQ flaws: 1) contamination, items appearing exactly online; 2) shortcuts, cues in the choices that enable guessing; and 3) writing errors, structural/grammatical issues based on a 19-rule education rubric. We validate BenchMarker with human annotations, then run the tool to audit 12 benchmarks, revealing: 2) contaminated MCQs tend to inflate accuracy, while writing errors tend to lower it and change rankings beyond random; and 3) prior benchmark repairs address their targeted issues (i.e., lowering accuracy with LLM-written distractors), but inadvertently add new flaws (i.e. implausible distractors, many correct answers). Overall, flaws in MCQs degrade NLP evaluation, but education research offers a path forward. We release BenchMarker to bridge the fields and improve MCQA benchmark design.

[Findings] Beyond Evidence: Belief-Chain Conditioning for Persuasive Misinformation Debunking Explanation ▶ *Underline Video*

Yi-Li Hsu, Li-Wun Chang, Chun-Yu Hsu, Wei-Kuan Shih, Aiping Xiong and Lun-Wei Ku

As AI systems increasingly mediate everyday communication, large language models (LLMs) are expected not only to provide factually accurate responses but also to generate explanations that engage with users mental states. We build on the concept of cognitive chains, structured representations of Situation, Clue, Thought, Action, and Emotion inspired by Theory of Mind, to investigate whether conditioning LLM outputs on such belief chains improves explanation quality. Specifically, we evaluate explanations along six reader-perceived dimensions: overall quality, logical correctness, completeness, conciseness, empathy, and agreement. Prior work shows that LLM explanations often default to neutral or uncertain stances, while individuals holding strong false beliefs remain highly resistant to correction. To address this challenge, we instantiate cognitive chains from two perspectives: believers and non-believers of the news claims. Using GPT-4.1 as a role-player across these stances, we find that incorporating believers chains improves the perceived quality of explanations for audiences with misinformation-aligned beliefs. Our findings underscore the importance of modeling diverse mental states in explanation generation and provide the first systematic evidence that Theory-of-Mind-based cognitive chains enhance the persuasiveness of explanations in misinformation contexts.

[Findings] Beyond Facts- Benchmarking Distributional Reading Comprehension in Large Language Models ▶ *Underline Video*

Pei-Fu Guo, Ya An Tsai, Chun-Chia Hsu, Kai-Xin Chen, Yun-Da Tsai, Kai-Wei Chang, Nanyun Peng, Mi-Yen Yeh and Shou-De Lin

While most reading comprehension benchmarks for LLMs focus on factual information that can be answered by localizing specific textual evidence, many real-world tasks require understanding distributional information, such as population-level trends and preferences expressed across collections of text. We introduce Text2DistBench, a reading comprehension benchmark for evaluating LLMs ability to infer distributional knowledge from natural language. Built from real-world YouTube comments about movie and music entities, the benchmark provides models with entity metadata and associated comments, and requires them to answer distributional questions, such as estimating the proportions of positive and negative comments, or identifying the most and second most frequent topics discussed among viewers. To support reliable and long-term evaluation, the construction pipeline of Text2DistBench is fully automated and continuously updated to incorporate newly emerging entities over time. Experiments across multiple LLMs show that while models substantially outperform random baselines, performance varies widely across different distribution types and characteristics. These findings highlight both the capabilities and limitations of current LLMs in distributional reading comprehension and demonstrate the value of Text2DistBench as a practical and scalable testbed for future research.

[Findings] Beyond Fixed Psychological Personas: State Beats Trait, but Language Models are State-Blind ▶ *Underline Video*

Tamunotonye Harry, Ivoline Ngong, Chima Nweke, Yuanyuan Feng and Joseph Near

User interactions with language models vary due to static properties of the user (trait) and the specific context of the interaction (state). However, existing persona datasets (like PersonaChat, PANDORA etc.) capture only trait, and ignore the impact of state. We introduce Chameleon, a dataset of 5,001 contextual psychological profiles from 1,667 Reddit users, each measured across multiple contexts. Using the Chameleon dataset, we present three key findings. First, inspired by Latent State-Trait theory, we decompose variance and find that 74% is within-person (state) while only 26% is between-person (trait). Second, we find that LLMs are state-blind: they focus on trait only, and produce similar responses regardless of state. Third, we find that reward models react to user state, but inconsistently: different models favor or penalize the same users in opposite directions. We release Chameleon to support research on affective computing, personalized dialogue, and RLHF alignment.

[Findings] Beyond Marginal Distributions: A Framework to Evaluate the Representativeness of Demographic-Aligned LLMs ▶ *Underline Video*

Tristan Williams, Franziska Weeber, Sebastian Padó and Alan Akbik

Large language models are increasingly used to represent human opinions, values, or beliefs, and their steerability towards these ideals is an active area of research. Existing work focuses predominantly on aligning marginal response distributions, treating each alignment evaluation example independently. While essential, this may overlook deeper latent structures that characterise real populations and underpin cultural values theories. We propose a framework for evaluating the representativeness of aligned models through multivariate correlation patterns in addition to marginal distributions. We show the value of our evaluation scheme by comparing two model steering techniques (persona prompting and demographic fine-tuning) and evaluating them against human responses from the World Values Survey. While the demographic fine-tuned model better approximates marginal response distributions, persona prompting performs marginally better at reproducing the empirical correlation structure between survey items. Despite this reversal, neither technique aligns with human correlation patterns. We conclude that representativeness is a distinct aspect of value alignment and an evaluation focused on marginals can mask structural failures, leading to overly optimistic conclusions about model representativeness.

[Long] Bidirectional LMs are Better Knowledge Memorizers? A Benchmark for Real-world Knowledge Injection ▶ *Underline Video*

Yuwei Zhang, Wenhao Yu, Shangbin Feng, Yifan Zhu, Letian Peng, Jayanth Srinivasa, Gaowen Liu and Jingbo Shang

Despite significant advances in large language models (LLMs), their knowledge memorization capabilities remain underexplored, due to the lack of standardized and high-quality testing grounds. In this paper, we introduce a novel, real-world and large-scale knowledge injection benchmark that continuously evolves without human intervention. Specifically, we propose WikiDYK, which leverages recently-added and expert-curated facts from Wikipedia's "Did You Know..." entries. Each entry is converted into multiple question-answer pairs spanning diverse task formats from easy cloze prompts to complex multi-hop questions. WikiDYK currently contains 12,290 facts and 77,180 questions, and its design allows for seamless extension with future updates from Wikipedia editors. Through extensive experiments using continued pre-training, we reveal a surprising insight: despite their prevalence in modern LLMs, Causal Language Models (CLMs) demonstrate significantly weaker knowledge memorization capabilities compared to Bidirectional Language Models (BiLMs), exhibiting a 23% lower accuracy in terms of reliability. To compensate for the smaller scales of current BiLMs, we introduce a modular collaborative framework utilizing ensembles of BiLMs as external knowledge repositories to integrate with LLMs. Experiment shows that this framework further improves the reliability accuracy by up to 29.1%. Code: https://github.com/zhang-yu-wei/WikiDYK.

[Long] CAML: A Conflict-Aware Molecular Language Model Merging Framework for Multi-Constraint Molecular Generation ▶ *Underline Video*

Xuanbai Ren, Luoda Tan, Pei Liu, Tengfei Ma, Xiangzheng Fu, Longyue Wang, Yiping Liu and Xiangxiang Zeng

Transfer learning has demonstrated efficacy in single-property constraint molecular generation. However, real-world drug discovery demands molecules to satisfy multiple property constraints. Existing paradigms often struggle with this challenge due to catastrophic forgetting or gradient conflicts. To address this, we propose a conflict-aware molecular language model merging framework (CAML). CAML generates multiple constraints molecular as a cooperative game among property-specific fine-tune models (expert models). Specifically, we formulate a Stability-Aware Covariance Matrix Adaptation Evolution Strategy (SACMA-ES) to dynamically optimize the fusion strategy. This algorithm searches for a Nash-equilibrium-like solution that minimizes conflicts among properties by exploring the optimal combination of the importance of the task parameter (intrinsic scale) and relative fusion weights of each expert (fusion coefficient), yielding a multi-constraint molecular property generation model without revisiting the training data. Extensive experiments demonstrate that CAML achieves state-of-the-art performance in complex multi-constraint scenarios. Our results validate that this training-free paradigm offers a robust and efficient solution for resolving intrinsic property conflicts in de novo molecular design.

[Long] Can AI-Generated Persuasion Be Detected? Persuaficial Benchmark and AI vs. Human Linguistic Differences ▶ *Underline Video*

Arkadiusz Modzelewski, Pawe Golik, Anna Koos and Giovanni Da San Martino

Large Language Models (LLMs) can generate highly persuasive text, raising concerns about their misuse for propaganda, manipulation, and other harmful purposes. This leads us to our central question: Is LLM-generated persuasion more difficult to automatically detect than human-written persuasion? To address this, we categorize controllable generation approaches for producing persuasive content with LLMs and introduce Persuaficial, a high-quality multilingual benchmark covering six languages: English, German, Polish, Italian, French and Russian. Using this benchmark, we conduct extensive empirical evaluations comparing human-authored and LLM-generated persuasive

texts. We find that although overtly persuasive LLM-generated texts can be easier to detect than human-written ones, subtle LLM-generated persuasion consistently degrades automatic detection performance. Beyond detection performance, we provide the first comprehensive linguistic analysis contrasting human and LLM-generated persuasive texts, offering insights that may guide the development of more interpretable and robust detection tools.

[Long] ClaimDB: A Fact Verification Benchmark over Large Structured Data ▶ *Underline Video*

Michael Theologitis, Preetam Prabhu Srikar Dammu, Chirag Shah and Dan Suciu

Real-world fact-checking often involves verifying claims grounded in structured data at scale. Despite substantial progress in fact-verification benchmarks, this setting remains largely underexplored. In this work, we introduce ClaimDB, a fact-verification benchmark where the evidence for claims is derived from compositions of millions of records and multiple tables. ClaimDB consists of 80 unique real-life databases covering a wide range of domains, from governance and healthcare to media, education and the natural sciences. At this scale, verification approaches that rely on "reading" the evidence break down, forcing a timely shift toward reasoning in executable programs. We conduct extensive experiments with 30 state-of-the-art proprietary and open-source (below 70B) LLMs and find that more than half score below 55% accuracy. Our analysis also reveals that both closed- and open-source models struggle with abstention – the ability to admit that there is no evidence to decide – raising doubts about their reliability in high-stakes data analysis tasks. We release the benchmark, code, and the LLM leaderboard at https://claimdb.github.io.

[Long] CLARity: Reasoning Consistency Alone Can Teach Reinforced Experts ▶ *Underline Video*

Jiuheng Lin, Cong Jiang, Zirui Wu, Jiarui Sun and Yansong Feng

Training expert LLMs in domains with scarce fine-grained annotated data is admittedly challenging, often relying on multiple-choice questions (MCQs). However, standard outcome-based reinforcement learning (RL) on MCQs is risky. While outcome-based RL may improve accuracy, it frequently compromises the reasoning process, yielding internally inconsistent rationales that diverge from the final predictions. Existing solutions to supervise the reasoning process, such as large-scale Process Reward Models (PRMs), are prohibitively expensive. To address this, we propose CLARity, a cost-effective RL framework that enhances reasoning quality using a small, general-purpose LLM only. CLARity integrates a consistency-aware reward mechanism with a 2-stage refine-then-monitor training pipeline to enhance reasoning consistency, and a dynamic data reformulation strategy to better exploit annotated data available. Experiments demonstrate that CLARity can improve the consistency of responses by 16.5% over standard outcome-based RL, and bring an improvement of 7.5% in final accuracy. Human evaluations further confirm substantial gains in factual correctness and reasoning coherence, leading to more trustworthy model outputs. Thus, CLARity offers a generalizable solution that enables smaller models to effectively guide expert LLM training by monitoring reasoning consistency.

[Findings] CompassLLM: A Multi-Agent Approach toward Geo-Spatial Reasoning for Popular Path Query ▶ *Underline Video*

Md. Nazmul Islam Ananto, Shamit Fatin, Mohammed Eunus Ali and Md Rizwan Parvez

The popular path query - identifying the most frequented routes between locations from historical trajectory data - has important applications in urban planning, navigation optimization, and travel recommendations. While traditional algorithms and machine learning approaches have achieved success in this domain, they typically require model training, parameter tuning, and retraining when accommodating data updates. As Large Language Models (LLMs) demonstrate increasing capabilities in spatial and graph-based reasoning, there is growing interest in exploring how these models can be applied to geo-spatial problems. We introduce CompassLLM, a novel multi-agent framework that intelligently leverages the reasoning capabilities of LLMs into the geo-spatial domain to solve the popular path query. CompassLLM employs its agents in a two-stage pipeline: the SEARCH stage that identifies popular paths, and a GENERATE stage that synthesizes novel paths in the absence of an existing one in the historical trajectory data. Experiments on real and synthetic datasets show that CompassLLM demonstrates superior accuracy in SEARCH and competitive performance in GENERATE while being cost-effective.

[Findings] ConSensus: Multi-Agent Collaboration for Multimodal Sensing ▶ *Underline Video*

Hyungjun Yoon, Mohammad Malekzadeh, Sung-Ju Lee, Fahim Kawsar and Lorena Qendro

Large language models (LLMs) are increasingly grounded in sensor data to perceive and reason about human physiology and the physical world. However, accurately interpreting heterogeneous multimodal sensor data remains a fundamental challenge. We show that a single monolithic LLM often fails to reason coherently across modalities, leading to incomplete interpretations and prior-knowledge bias. We introduce ConSensus, a training-free multi-agent collaboration framework that decomposes multimodal sensing tasks into specialized, modality-aware agents. To aggregate agent-level interpretations, we propose a hybrid fusion mechanism that balances semantic aggregation, which enables cross-modal reasoning and contextual understanding, with statistical consensus, which provides robustness through agreement across modalities. While each approach has complementary failure modes, their combination enables reliable inference under sensor noise and missing data. We evaluate ConSensus on five diverse multimodal sensing benchmarks, demonstrating an average accuracy improvement of 7.1% over the single-agent baseline. Furthermore, ConSensus matches or exceeds the performance of iterative multi-agent debate methods while achieving a 12.7 times reduction in average fusion token cost through a single-round hybrid fusion protocol, yielding a robust and efficient solution for real-world multimodal sensing tasks. The source code is available at https://github.com/nokia/multi-agent-collaboration-for-multimodal-sensing.

[Findings] CoTEvol: Self-Evolving Chain-of-Thoughts for Data Synthesis in Mathematical Reasoning ▶ *Underline Video*

Zhuo Wang, Zhuo Zhang, Yafu Li, Yu Cheng, Lizhen Qu and Zenglin Xu

Large Language Models (LLMs) exhibit strong mathematical reasoning when trained on high-quality Chain-of-Thought (CoT) that articulates intermediate steps, yet costly CoT curation hinders further progress. While existing remedies such as distillation from stronger LLMs and self-synthesis based on test-time search alleviate this issue, they often suffer from diminishing returns or high computing overhead. In this work, we propose CoTEvol, a genetic evolutionary framework that casts CoT generation as a population-based search over reasoning trajectories. Candidate trajectories are iteratively evolved through reflective global crossover at the trajectory level and local mutation guided by uncertainty at the step level, enabling holistic recombination and fine-grained refinement. Lightweight, task-aware fitness functions are designed to guide the evolutionary process toward accurate and diverse reasoning. Empirically, improves correct-CoT synthesis success by over 30% and enhances structural diversity, with markedly improved efficiency. LLMs trained on these evolutionary CoT data achieve an average gain of 6.6% across eight math benchmarks, outperforming previous distillation and self-synthesis approaches. These results underscore the promise of evolutionary CoT synthesis as a scalable and effective method for mathematical reasoning tasks.

[Findings] CTRL: Control-Based Time Series Forecasting with LLM-Guided Residual Learning ▶ *Underline Video*

Minkyoung Kim, Daeun Ji, Yohan Lee, Beomsoo Kim and Beakcheol Jang

Time series forecasting underpins critical decision-making across diverse domains. While large language models (LLMs) offer promising reasoning capabilities, existing LLM-based time series forecasting approaches either reduce them to numerical predictors that bypass their strengths, or allow direct forecast generation that destabilizes predictions in non-stationary settings. We introduce CTRL, a framework that decouples semantic reasoning from quantitative prediction. A frozen backbone generates base forecasts, while specialized LLM agents function as controllers that analyze backbone prediction errors through decomposed trend, seasonal, and irregular components, grounding reasoning in interpretable temporal structure. Each agent outputs compact control signals that a lightweight residual decoder translates into forecast corrections. CTRL incorporates label-free test-time adaptation that detects distribution shift from input statistics alone and readapts control signals with only 3–24 LLM calls via caching. CTRL is explicitly designed to improve robustness under non-stationary temporal dynamics and distribution shift, while remaining competitive on highly stationary time series where adaptive correction provides limited additional benefit.

[Findings] Data Swarms: Optimizable Generation of Synthetic Evaluation Data ▶ *Underline Video*

Shangbin Feng, Yike Wang, Weijia Shi and Yulia Tsvetkov

We propose Data Swarms, an algorithm to optimize the generation of synthetic evaluation data and advance quantitative desiderata of LLM evaluation. We first train a swarm of initial data generators using existing data, and define various evaluation objectives to reflect the desired properties of evaluation (e.g., generate more difficult problems for the evaluated models) and quantitatively evaluate data generators. We then employ particle swarm optimization to optimize the swarm of data generators, where they collaboratively search through the model parameter space to find new generators that advance these objectives. We further extend it to Adversarial Swarms, where the data generator swarm generates harder data while the test taker model swarm learns from such data, co-evolving dynamically for better data and models simultaneously. Extensive experiments demonstrate that Data Swarms outperforms eight data generation baselines across five evaluation objectives, while Adversarial Swarms produce more robust learning of synthetic data and stronger generalization. Further analysis reveals that Data Swarms successfully optimizes compositions of multiple evaluation objectives and generalizes to new off-the-shelf LLMs, unseen at optimization time.

[Long] Decoding Scientific Experimental Images: The SPUR Benchmark for Perception, Understanding, and Reasoning ▶ *Underline Video*

Junpeng Ding, Zichen Tang, Haihong E, Mengyuan Ji, Yang Liu, Haolin Tian, Haiyang Sun, Pengqi Sun, Yang Xu, Yichen Liu, Haocheng Gao, Zijie Xi, Ruomeng Jiang, Peizhi Zhao, Rongjin Li, Yuanze Li, Jiacheng Liu, Zhongjun Yang, Jintong Chen and Siying Lin

We introduce SPUR, a comprehensive benchmark for scientific experimental image perception, understanding, and reasoning, comprising 4,264 question-answering (QA) pairs derived from 1,084 expert-curated images. SPUR features three key innovations: (1) Panel-Level Fine-Grained Perception: evaluating the visual perception of multimodal large language models (MLLMs) across three dimensions (numerical, morphological, and information localization) on six fine-grained panel types; (2) Cross-Panel Relation Understanding: utilizing complex images with an average of 14.3 panels per sample to evaluate MLLMs' ability to decipher intricate cross-panel relations; (3) Expert-Level Reasoning: assessment of qualitative and quantitative reasoning across five experimental paradigms to determine if models can infer conclusions from evidence as human experts do. Comprehensive evaluation of 20 MLLMs and four multimodal Chain-of-Thought (MCoT) methods reveals that current models fall significantly short of the expert-level requirements for scientific image interpretation, underscoring a critical bottleneck in AI for Science (AI4S) research.

[Long] Diagnosing Spatial Consistency across Perspectives and Viewpoints in Large Vision-Language Models ▶ *Underline Video*

Yoonji Kim, Jieun Kim, Yujin Jeong and Sung-Bae Cho

Consistent reasoning about 3D spatial relations across changing viewpoints is fundamental for Embodied AI agents operating in dynamic environments. While Large Vision-Language Models (LVLMs) have advanced multimodal perception, their ability to maintain spatial consistency across diverse perspectives remains underexplored. Existing benchmarks primarily assess spatial capabilities from a static, single-view, and egocentric perspective, failing to capture the dynamic nature of real-world spatial cognition. To address this gap, we introduce SCOPE (Spatial COnsistency across PErspectives and Viewpoints), a comprehensive benchmark designed to rigorously diagnose spatial reasoning capabilities. Grounded in human cognitive theories of dual spatial representations, SCOPE discretizes the 360^ field into

multiview scenarios to systematically evaluate both allocentric and egocentric reasoning capabilities. Our dataset comprises 20.1K spatial VQA pairs derived from high-quality 3D environments. Through an extensive evaluation of 26 state-of-the-art LVLMs, we identify two fundamental limitations that prevent consistent spatial understanding across viewpoints. We hope SCOPE facilitates the diagnosis of spatial reasoning, serving as a stepping stone toward reliable embodied action.

[Long] DimABSA: Building Multilingual and Multidomain Datasets for Dimensional Aspect-Based Sentiment Analysis ▶ *Underline Video*

Lung-Hao Lee, Liang-Chih Yu, Natalia V Loukachevitch, Ilseyar Alimova, Alexander Panchenko, Tzu-Mi Lin, Zhe-Yu Xu, Jian-Yu Zhou, Guangmin Zheng, Jin Wang, Sharanya Awasthi, Jonas Becker, Jan Philip Wahle, Terry Ruas, Shamsuddeen Hassan Muhammad and Saif M. Mohammad

Aspect-Based Sentiment Analysis (ABSA) focuses on extracting sentiment at a fine-grained aspect level and has been widely applied across real-world domains. However, existing ABSA research relies on coarse-grained categorical labels (e.g., positive, negative), which limits its ability to capture nuanced affective states. To address this limitation, we adopt a dimensional approach that represents sentiment with continuous valence-arousal (VA) scores, enabling fine-grained analysis at both the aspect and sentiment levels. To this end, we introduce DimABSA, the first multilingual, dimensional ABSA resource annotated with both traditional ABSA elements (aspect terms, aspect categories, and opinion terms) and newly introduced VA scores. This resource contains 76,958 aspect instances across 42,590 sentences, spanning six languages and four domains. We further introduce three subtasks that combine VA scores with different ABSA elements, providing a bridge from traditional ABSA to dimensional ABSA. Given that these subtasks involve both categorical and continuous outputs, we propose a new unified metric, continuous F1 (cF1), which incorporates VA prediction error into standard F1. We provide a comprehensive benchmark using both prompted and fine-tuned large language models across all subtasks. Our results show that DimABSA is a challenging benchmark and provides a foundation for advancing multilingual dimensional ABSA. We publicly released the DimABSA dataset, which was used for Track A of SemEval-2026 Task 3, attracting over 300 participants.

[Findings] DisastQA: A Comprehensive Benchmark for Evaluating Question Answering in Disaster Management ▶ *Underline Video*

Zhitong Chen, Kai Yin, Xiangjue Dong, Chengkai Liu, Xiangpeng Li, Yiming Xiao, Bo Li, Junwei Ma, Ali Mostafavi and James Caverlee

Accurate question answering (QA) in disaster management requires reasoning over uncertain and conflicting information, a setting poorly captured by existing benchmarks built on clean evidence. We introduce DisastQA, a rigorously verified benchmark of 3,000 expert-annotated questions (2,000 multiple-choice and 1,000 open-ended) spanning eight disaster types. The benchmark is constructed via a human-LLM collaboration pipeline with stratified sampling to ensure balanced coverage. Models are evaluated under varying evidence conditions, from closed-book to noisy evidence integration, enabling separation of internal knowledge from reasoning under imperfect information. For open-ended QA, we propose a human-verified keypoint-based evaluation protocol emphasizing factual completeness over verbosity. Experiments with 20 models reveal substantial divergences from general-purpose leaderboards such as MMLU-Pro. While recent open-weight models approach proprietary systems in clean settings, performance degrades sharply under realistic noise, exposing critical reliability gaps for disaster response. All code, data, and evaluation resources are available at https://github.com/TamuChen18/DisastQA

[Findings] Distilling Examples into Task Instructions: Enhanced In-Context Learning for Real-World B2B Conversations ▶ *Underline Video*

Guy Rotman, Adi Kopilov, Danit Berger Zalmanson and Omri Allouche

In-context learning (ICL) is the standard method for low-resource classification, yet its efficacy in specialized domains remains largely unexplored. We address the challenge of classifying semantically complex, multi-party B2B conversations, where traditional ICL encounters significant limitations, especially as context length increases due to the concatenation of multiple few-shot examples. We introduce the Call Playbook dataset, featuring five classification tasks derived from real-world B2B conversations targeting core sales concepts. To bridge the gap between performance and practical utility, we propose novel knowledge extraction methods that distill verbose examples into compact, interpretable representations of structured classification criteria and precise task descriptions. Our approach achieves a 99% reduction in token usage and improves macro-averaged AUC by up to 7% over traditional ICL. Notably, it remains robust as context grows, unlike advanced token compression baselines which degrade by over 9 F1 points. Importantly, our framework enables direct refinement of classification logic, addressing critical needs for transparency, efficiency, and user interaction in real-world NLP applications.

[Long] DMRetriever: A Family of Models for Improved Text Retrieval in Disaster Management ▶ *Underline Video*

Kai Yin, Xiangjue Dong, Chengkai Liu, Allen Lin, Lingfeng Shi, Ali Mostafavi and James Caverlee

Effective and efficient access to relevant information is essential for disaster management. However, no retrieval model is specialized for disaster management, and existing general-domain models fail to handle the varied search intents inherent to disaster management scenarios, resulting in inconsistent and unreliable performance. To this end, we introduce DMRetriever, the first series of dense retrieval models (33M to 7.6B) tailored for this domain. It is trained through a novel three-stage framework of bidirectional attention adaptation, unsupervised contrastive pre-training, and difficulty-aware progressive instruction fine-tuning, using high-quality data generated through an advanced data refinement pipeline. Comprehensive experiments demonstrate that DMRetriever achieves state-of-the-art performance across all six search intents at every model scale. Moreover, DMRetriever is highly parameter-efficient, with 596M model outperforming baselines over 13.3 larger and 33M model exceeding baselines with only 7.6% of their parameters. All codes, data, and checkpoints are available at https://github.com/KaiYin97/DMRETRIEVER.

[Findings] Do Image-Text Metrics Respect Semantic Invariances? ► *Underline Video*

Amit Agarwal, Hitesh Laxmichand Patel, Meizhu Liu, Jyotika Singh, Karan Dua, Hansa Meghwani, Matthew Rowe, M. Avendi, Yassi Abbasi, Tao Sheng, Sujith Ravi and Dan Roth

Reference-free image-to-text evaluators are now standard for scoring image-caption alignment, yet it is unclear whether they respect semantic invariances. We present an invariance probe on five popular evaluators (CLIPScore, PAC-S, UMIC, FLEUR, and a deterministic LLM judge) under semantics-preserving perturbations along three axes: spatial (flips, context-preserving repositioning, light rotations), object (scale, category), and socio-linguistic framing (cultural/economic adjectives with neutral and length-matched controls). Across curated slices of three detection datasets and three caption evaluation suites, we find consistent non-semantic sensitivities: benign spatial edits and simple phrasing changes shift scores by (~)6-9% on average, and for systems separated by just 0.7% these shifts can cause ranking flips in upto ()37% of cases, particularly under spatial changes. A small human study also supports this finding and confirms that annotators generally judge perturbed pairs as equally correct, so these shifts reflect metric behavior rather than semantic change. We further propose invariance-calibrated scoring, a post-hoc adjustment that roughly halves median absolute sensitivity while retaining correlation with learned caption evaluators.

[Long] Do MLLMs Capture How Interfaces Guide User Behavior? A Benchmark for Multimodal UI/UX Design Understanding ► *Underline Video*

Jaehyun Jeon, Min Soo Kim, Janghan Yoon, Sumin Shim, Yejin Choi, Hanbin Kim, Dae Hyun Kim and Youngjae Yu

User interface (UI) design goes beyond visuals to shape user experience (UX), underscoring the shift toward UI/UX as a unified concept. While recent studies have explored UI evaluation using Multimodal Large Language Models (MLLMs), they largely focus on surface-level features, overlooking how design choices influence user behavior at scale. To fill this gap, we introduce WiserUI-Bench, a novel benchmark for multimodal understanding of how UI/UX design affects user behavior, built on 300 real-world UI image pairs from industry A/B tests, with empirically validated winners that induced more user actions. For future design progress in practice, post-hoc understanding of why such winners succeed with mass users is also required; we support this via expert-curated key interpretations for each instance. Experiments across multiple MLLMs on WiserUI-Bench for two main tasks, (1) predicting the more effective UI image between an A/B-tested pair, and (2) explaining it post-hoc in alignment with expert interpretations, show that models exhibit limited understanding of the behavioral impact of UI/UX design. We believe our work will foster research on leveraging MLLMs for visual design in user behavior contexts.

[Long] DPC: Training-Free Text-to-SQL Candidate Selection via Dual-Paradigm Consistency ► *Underline Video*

Boyan Li, Ou Ocean Kun Hei, Yue Yu and Yuyu Luo

While Large Language Models (LLMs) demonstrate impressive proficiency in generating SQL queries, they fundamentally lack the capability to self-evaluate correctness without an execution oracle. This limitation creates a stark Generation-Selection Gap, where high potential accuracy (Pass@K) fails to translate into execution accuracy (Pass@1). Although supervised verifiers offer mitigation, they incur prohibitive annotation costs and suffer from domain fragility. Consequently, recent research has pivoted to the training-free setting. However, existing methods—such as Self-Consistency or LLM-as-a-Judge—remain hampered by systematic bias (consensus on hallucinations) and symbolic blindness (inability to simulate execution states). We introduce DPC (Dual-Paradigm Consistency), a multi-agent framework that reformulates SQL selection from a probabilistic guessing task on hidden data into a deterministic verification task on visible data. Specifically, DPC employs a Slicer and a Tester agent to collaboratively construct a Minimal Distinguishing Database (MDD)—an adversarial, fully observable micro-environment engineered to expose logical discrepancies between candidates. To break the self-correction bias, a Solver agent then verifies the SQL candidates by cross-referencing their execution against a parallel Python/Pandas solution. By validating execution consistency between declarative (SQL) and imperative (Python) paradigms, DPC robustly discriminates correct logic from systematic hallucinations. Experiments on BIRD and Spider across multiple LLMs demonstrate that our method consistently outperforms existing selection baselines, achieving absolute accuracy improvements of up to 2.2% over strong competitors like Self-Consistency.

[Findings] EDU-CIRCUIT-HW: Evaluating Multimodal Large Language Models on Real-World University-Level STEM Student Handwritten Solutions ► *Underline Video*

Weiyu Sun, Liangliang Chen, Yongnuo Cai, Huiru Xie, Yi Zeng and Ying Zhang

Multimodal Large Language Models (MLLMs) hold significant promise for revolutionizing traditional education and reducing teachers' workload. However, accurately interpreting unconstrained STEM student handwritten solutions with intertwined mathematical formulas, diagrams, and textual reasoning poses a significant challenge due to the lack of authentic and domain-specific benchmarks. Additionally, current evaluation paradigms predominantly rely on the outcomes of downstream tasks (e.g., auto-grading), which often probe only a subset of the recognized content, thereby failing to capture the MLLMs' understanding of complex handwritten logic as a whole. To bridge this gap, we release EDU-CIRCUIT-HW, a dataset consisting of 1,300+ authentic student handwritten solutions from a university-level STEM course. Utilizing the expert-verified verbatim transcriptions and grading reports of student solutions, we simultaneously evaluate various MLLMs upstream recognition fidelity and downstream auto-grading performance. Our evaluation uncovers an astonishing scale of latent failures within MLLM-recognized student handwritten content, highlighting the models' insufficient reliability for auto-grading and other understanding-oriented applications in high-stakes educational settings. In response, we present a case study demonstrating that leveraging identified error patterns to preemptively detect and rectify recognition errors, with only minimal human intervention (e.g., with 3.3% assignments routed to human graders while the rest to GPT-5.1 grader), can effectively enhance the robustness of the deployed AI-enabled

grading system on unseen student solutions.

[Long] Evaluating Structure-Aware Retrieval and Safety in Statute-Centric Legal QA ▶ *Underline Video*

Kyubyung Chae, Jewon Yeom, Jeongjae Park, Seunghyun Bae, Ijun Jang, Hyunbin Jin, Jinkwan Jang and Taesup Kim

Legal QA benchmarks have predominantly focused on case law, overlooking the unique challenges of statute-centric regulatory reasoning. In statutory domains, relevant evidence is distributed across hierarchically linked documents, creating a statutory retrieval gap where conventional retrievers fail and models often hallucinate under incomplete context. We introduce SearchFireSafety, a structure- and safety-aware benchmark for statute-centric legal QA. Instantiated on fire-safety regulations as a representative case, the benchmark evaluates whether models can retrieve hierarchically fragmented evidence and safely abstain when statutory context is insufficient. SearchFireSafety adopts a dual-track evaluation framework combining real-world questions that require citation-aware retrieval and synthetic partial-context scenarios that stress-test hallucination and refusal behavior. Experiments across multiple large language models show that graph-guided retrieval substantially improves performance, but also reveal a critical safety trade-off: domain-adapted models are more likely to hallucinate when key statutory evidence is missing. Our findings highlight the need for benchmarks that jointly evaluate hierarchical retrieval and model safety in statute-centric regulatory settings.

[Long] Evaluating Temporal Consistency in Multi-Turn Language Models ▶ *Underline Video*

Yash Kumar Atri, Steven L. Johnson and Thomas Hartvigsen

Language models are increasingly deployed in interactive settings where users reason about facts over time rather than in isolation. In such scenarios, correct behavior requires models to maintain and update implicit temporal assumptions established earlier in a conversation. We study this challenge through the lens of temporal scope stability: the ability to preserve, override, or transfer time-scoped factual context across dialogue turns. We introduce ChronoScope, a large-scale diagnostic benchmark designed to isolate temporal scope behavior in controlled multi-turn interactions, comprising over one million deterministically generated question chains grounded in Wikidata. ChronoScope evaluates whether models can correctly retain inferred temporal scope when follow-up questions omit explicit time references, spanning implicit carryover, explicit scope switching, cross-entity transfer, and longer temporal trajectories. Through extensive evaluation of state-of-the-art language models, we find that temporal scope stability is frequently violated in controlled multi-turn settings, with models often drifting toward present-day assumptions despite correct underlying knowledge. These failures intensify with interaction length and persist even under oracle context conditions, revealing a gap between single-turn factual accuracy and coherent temporal reasoning under sequential interaction. We make our dataset and evaluation suite publicly available at https://github.com/yashkumaratri/ChronoScope.

[Long] Explainable Disentangled Representation Learning for Generalizable Authorship Attribution in the Era of Generative AI ▶ *Underline Video*

Hieu Man, Van-Cuong Pham, Nghia Trung Ngo, Franck Dernoncourt and Thien Huu Nguyen

Learning robust representations of authorial style is crucial for authorship attribution and AI-generated text detection. However, existing methods often struggle with content-style entanglement, where models learn spurious correlations between authors' writing styles and topics, leading to poor generalization across domains. To address this challenge, we propose Explainable Authorship Variational Autoencoder (EAVAE), a novel framework that explicitly disentangles style from content through architectural separation-by-design. EAVAE first pretrains style encoders using supervised contrastive learning on diverse authorship data, then finetunes with a Variational Autoencoder (VEA) architecture using separate encoders for style and content representations. Disentanglement is enforced through a novel discriminator that not only distinguishes whether pairs of style/content representations belong to the same or different authors/content sources, but also generates natural language explanation for their decision, simultaneously mitigating confounding information and enhancing interpretability. Extensive experiments demonstrate the effectiveness of EAVAE. On authorship attribution, we achieve state-of-the-art performance on various datasets, including Amazon Reviews, PAN21, and HRS. For AI-generated text detection, EAVAE excels in few-shot learning over the M4 dataset.

[Findings] FigSIM: A Dataset for Fine-grained Suicide Severity and Figurative Language in Suicide Memes ▶ *Underline Video*

Liuliu Chen, Elise Carrotte, Brian E. Chapman, Jo Robinson and Mike Conway

Suicide memes are memes used to express suicide-related thoughts or comment on suicide-related issues. Suicide memes are increasingly common on social media, yet remain poorly understood and potentially harmful. There is an urgent need to better understand their characteristics and to develop appropriate content moderation strategies that limits users' exposure to potentially harmful content. Currently, the absence of annotated datasets of suicide memes remains a key barrier to developing and evaluating automated moderation approaches. In this paper, we introduce FigSIM, the first dataset designed for fine-grained analysis of suicide memes. The dataset consists of 1049 memes, each annotated for (1) fine-grained suicide severity levels, (2) figurative phenomena (e.g. metaphors), and (3) suicide-related content (e.g. suicide method depiction). We benchmark 16 unimodal and multimodal models across three tasks: figurative language, suicide severity, and suicide-related content detection. Overall, FigSIM demonstrates that suicide memes pose unique challenges for both modeling and content moderation. Analysis revealed biases, such as underprediction of higher suicide severity levels, especially for figurative memes.

[Findings] Fine-Grained Detection of Context-Grounded Hallucinations Using LLMs ▶ *Underline Video*

Yehonatan Pesiakhovsky, Zorik Gekhman, Yosi Mass, Liat Ein-Dor and Roi Reichart

Context-grounded hallucinations are cases where model outputs contain information not verifiable against the source text. We study the applicability of LLMs for localizing such hallucinations, as a more practical alternative to existing complex evaluation pipelines. In the

absence of established benchmarks for meta-evaluation of hallucinations localization, we construct one tailored to LLMs, involving a challenging human annotation of over 1,000 examples. We complement the benchmark with an LLM-based evaluation protocol, verifying its quality in a human evaluation. Since existing representations of hallucinations limit the types of errors that can be expressed, we propose a new representation based on free-form textual descriptions, capturing the full range of possible errors. We conduct a comprehensive study, evaluating four large-scale LLMs, which highlights the benchmark's difficulty, as the best model achieves an F1 score of only 0.67. Through careful analysis, we offer insights into optimal prompting strategies for the task and identify the main factors that make it challenging for LLMs: (1) a tendency to incorrectly flag missing details as inconsistent, despite being instructed to check only facts in the output; and (2) difficulty with outputs containing factually correct information absent from the source - and thus not verifiable - due to alignment with the model's parametric knowledge.

[Long] From Knowing to Teaching: Scaffolding Pedagogical Decisions for LLM Agent ▶ *Underline Video*
Yucheng Wang, Shen Yang, Jifan Yu, Haoxuan Li, Joy Jia Yin Lim, Daniel Zhang-Li, Huiqin LIU, Lei Hou, Juanzi Li and Bin Xu
Knowing and teaching differ fundamentally: effective instruction requires transforming knowledge into forms learners can grasp. Large language models, when asked to generate lessons (a concrete form of teaching), produce content lacking pedagogical depth. We trace this failure to three decisions that expert teachers make: selecting content by recognizing each source's instructional role, sequencing topics so foundations precede applications, and synthesizing components into a unified whole. To scaffold these decisions, we introduce TeachCraft, a framework with three agents: Explorer classifies sources by pedagogical intent to guide selection; Planner orders objectives from foundational to advanced; Generator produces lesson materials through a schema that ensures consistency across components. To evaluate this approach, we construct LessonBench, 40 expert-designed lessons paired with two to five heterogeneous source documents, on which TeachCraft achieves 67.8% win rate in human evaluation and 79.6% in LLM-based evaluation against eight baselines, with ablations confirming that each decision contributes independently to overall lesson quality.

[Findings] From Recall to Forgetting: Benchmarking Long-Term Memory for Personalized Agents ▶ *Underline Video*
Md Nayem Uddin, Kumar Shubham, Eduardo Blanco, Chitta Baral and Gengyu Wang
Personalized agents that interact with users over long periods must maintain persistent memory across sessions and update it as circumstances change. However, existing benchmarks predominantly frame long-term memory evaluation as fact retrieval from past conversations, providing limited insight into agents ability to consolidate memory over time or handle frequent knowledge updates. We introduce Memora, a long-term memory benchmark spanning weeks to months long user conversations. The benchmark evaluates three memory-grounded tasks: remembering, reasoning, and recommending. To ensure data quality, we employ automated memory-grounding checks and human evaluation. We further introduce Forgetting-Aware Memory Accuracy (FAMA), a metric that penalizes reliance on obsolete or invalidated memory when evaluating long-term memory. Evaluations of four LLMs and six memory agents reveal frequent reuse of invalid memories and failures to reconcile evolving memories. Memory agents offer marginal improvements, exposing shortcomings in long-term memory for personalized agents.

[Findings] Generalizable LLM Learning of Graph Synthetic Data with Post-training Alignment ▶ *Underline Video*
Yizhuo Zhang, Heng Wang, Shangbin Feng, Zhaoxuan Tan, Xinyun Liu and Yulia Tsvetkov
Previous research has sought to enhance the graph reasoning capabilities of LLMs by supervised fine-tuning on synthetic graph data. While these led to specialized LLMs better at solving graph algorithm problems, we don't need LLMs for shortest path: we need generalization from synthetic graph data to real-world tasks with implicit graph structures. In this work, we propose to unlock generalizable learning of graph with post-training alignment with synthetic data. We first design solution-based and process-based rewards for synthetic graph problems: instead of rigid memorizing response patterns in direct fine-tuning, we posit that post-training alignment would help LLMs grasp the essentials underlying graph reasoning and alleviate overfitting on synthetic data. We employ post-training alignment algorithms such as GRPO and DPO, aligning both off-the-shelf LLMs and LLMs fine-tuned on synthetic graph data. We then compare them against existing settings on both in-domain synthetic tasks and out-of-domain real-world tasks with implicit graph structures such as multi-hop QA, structured planning, and more. Extensive experiments demonstrate that our post-training alignment recipe leads to statistically significant improvement on 5 datasets, with an average gain of 12.9% over baseline settings. Further analysis reveals that process-based rewards consistently outperform solution-based rewards on synthetic data but not on real-world tasks, and compositionality and explainable intermediate steps remains a critical challenge even after post-training alignment.

[Long] GS-Quant: Granular Semantic and Generative Structural Quantization for Knowledge Graph Completion ▶ *Underline Video*
Qizhuo Xie, Yunhui Liu, Yu Xing, Qianzi Hou, Xudong Jin, Tao Zheng and Tieke He
Large Language Models (LLMs) have shown immense potential in Knowledge Graph Completion (KGC), yet bridging the modality gap between continuous graph embeddings and discrete LLM tokens remains a critical challenge. While recent quantization-based approaches attempt to align these modalities, they typically treat quantization as flat numerical compression, resulting in semantically entangled codes that fail to mirror the hierarchical nature of human reasoning. In this paper, we propose GS-Quant, a novel framework that generates semantically coherent and structurally stratified discrete codes for KG entities. Unlike prior methods, GS-Quant is grounded in the insight that entity representations should follow a linguistic coarse-to-fine logic. We introduce a Granular Semantic Enhancement module that injects hierarchical knowledge into the codebook, ensuring that earlier codes capture global semantic categories while later codes refine specific attributes. Furthermore, a Generative Structural Reconstruction module imposes causal dependencies on the code sequence, transforming

independent discrete units into structured semantic descriptors. By expanding the LLM vocabulary with these learned codes, we enable the model to reason over graph structures isomorphically to natural language generation. Experimental results demonstrate that GS-Quant significantly outperforms existing text-based and embedding-based baselines.

[Long] GUIDE: Towards Scalable Advising for Research Ideas ▶ *Underline Video*

Yaowenqi Liu, BingXu Meng, Rui Pan, Yuxing Liu, Jerry Huang, Jiaxuan You and Tong Zhang

The field of AI research is advancing at an unprecedented pace, enabling automated hypothesis generation and experimental design across diverse domains such as biology, mathematics, and artificial intelligence. Despite these advancements, there remains a significant gap in the availability of scalable advising systems capable of providing high-quality, well-reasoned feedback to refine proposed hypotheses and experimental designs. To address this challenge, we explore key factors that underlie the development of robust advising systems, including model size, data reweighting, context length, confidence estimation, and structured reasoning processes. Our findings reveal that a relatively small model, when equipped with a well-compressed literature database and a structured reasoning framework, can outperform powerful general-purpose language models such as Deepseek-R1 in terms of acceptance rates for self-ranked top-30% submissions to ICLR 2025. Moreover, when limited to high-confidence predictions, our system achieves an acceptance rate exceeding 90% on the ICLR 2025 test set, underscoring its potential to significantly enhance the quality and efficiency of hypothesis generation and experimental design.

[Findings] How Grounded is Wikipedia? A Study on Structured Evidential Support and Retrieval ▶ *Underline Video*

William Walden, Kathryn Ricci, Miriam Wanner, Zhengping Jiang, Chandler May, Rongkun Zhou and Benjamin Van Durme

Wikipedia is a critical resource for modern NLP, serving as a rich repository of up-to-date and citation-backed information on a wide variety of subjects. The reliability of Wikipedia—its groundedness in its cited sources—is vital to this purpose. This work analyzes both how grounded Wikipedia is and how readily fine-grained grounding evidence can be retrieved. To this end, we introduce PeopleProfiles—a large-scale, multi-level dataset of claim support annotations on biographical Wikipedia articles. We show that: ~22% of claims in Wikipedia lead sections are unsupported by the article body; ~30% of annotated claims in the article body are unsupported by their (publicly accessible) sources; and real-world Wikipedia citation practices often differ from documented standards. Finally, we show that complex evidence retrieval remains a challenge—even for recent reasoning rerankers.

[Long] INFACT: A Diagnostic Benchmark for Induced Faithfulness and Factuality Hallucinations in Video-LLMs ▶ *Underline Video*

Yangjunqi Yangjunqi, Yuecong Min, Jie Zhang, Shiguang Shan and Xilin CHEN

Despite rapid progress, Video Large Language Models (Video-LLMs) remain unreliable due to hallucinations, which are outputs that contradict either video evidence (faithfulness) or verifiable world knowledge (factuality). Existing benchmarks provide limited coverage of factuality hallucinations and predominantly evaluate models only in clean settings. We introduce INFACT, a diagnostic benchmark comprising 9,800 QA instances with fine-grained taxonomies for faithfulness and factuality, spanning real and synthetic videos. INFACT evaluates models in four modes: Base (clean), Visual Degradation, Evidence Corruption, and Temporal Intervention for order-sensitive items. Reliability under induced modes is quantified using Resist Rate (RR) and Temporal Sensitivity Score (TSS). Experiments on 14 representative Video-LLMs reveal that higher Base-mode accuracy does not reliably translate to higher reliability in the induced modes, with evidence corruption reducing stability and temporal intervention yielding the largest degradation. Notably, many open-source baselines exhibit near-zero TSS on factuality, indicating pronounced temporal inertia on order-sensitive questions.

[Findings] Instruction-Guided Poetry Generation in Arabic and Its Dialects ▶ *Underline Video*

Abdelrahman Sadallah, Kareem Elozeiri, Mervat Abassy, Rania Elbadry, Mohamed Anwar, Abed Alhakim Freihat, Preslav Nakov and Fajri Koto

Poetry has long been a central art form for Arabic speakers, serving as a powerful medium of expression and cultural identity. While modern Arabic speakers continue to value poetry, existing research on Arabic poetry within Large Language Models (LLMs) has primarily focused on analysis tasks such as interpretation or metadata prediction, e.g., rhyme schemes and titles. In contrast, our work addresses the practical aspect of poetry creation in Arabic by introducing controllable generation capabilities to assist users in writing poetry. Specifically, we present a large-scale, carefully curated instruction-based dataset in Modern Standard Arabic (MSA) and various Arabic dialects. This dataset enables tasks such as writing, revising, and continuing poems based on predefined criteria, including style and rhyme, as well as performing poetry analysis. Our experiments show that fine-tuning LLMs on this dataset yields models that can effectively generate poetry that is aligned with user requirements, based on both automated metrics and human evaluation with native Arabic speakers. The data and the code are available at https://github.com/mbzuai-nlp/instructpoet-ar

[Long] Iterative Dual-Model Alignment for Story Evaluation ▶ *Underline Video*

Bruce Qin and Dan Goldwasser

Large language models (LLMs) can both evaluate and explain text quality; however, most existing evaluators operate as static classifiers and lack the ability to refine their reasoning through interaction. We propose an Iterative Alpha–Beta Learning framework that jointly trains two complementary 8B models: an Alpha (alpha) classifier that assesses pairwise story engagement, and a Beta (beta) generator that produces structured, rubric-guided comparative explanations. The two models co-evolve within a closed feedback loop: alpha provides probabilistic preference signals to guide betas Direct Preference Optimization (DPO), while betas improved explanations are reintegrated to retrain alpha via a KL-based contrastive objective. This dual optimization enables mutual learning: alpha gains interpretability and robustness from

betas textual rationales, while beta acquires stronger alignment and discriminative precision from alphas confidence deltas. Experiments on human-annotated story-pair datasets HANNA show that the proposed system consistently outperforms strong single-model baselines in both accuracy and explanation quality across multiple iterative rounds.

[Long] Learning Faster with Better Tokens: Parameter-Efficient Vocabulary Adaptation for Specialized Text Summarization ▶ *Underline Video*

Gunjan Balde, Soumyadeep Roy, Mainack Mondal and Niloy Ganguly

Large language models pretrained on general-domain corpora often exhibit tokenization inefficiencies when applied to specialized domains. Although continual pretraining for domain adaptation partially alleviate performance degradation, it does not resolve the fundamental vocabulary mismatch. To address this gap, we introduce a targeted parameter-efficient domain adaptation approach that combines vocabulary adaptation with pretraining for LLM-based text summarization. Our unified framework augments pretrained tokenizers with domain-specific tokens while selectively replacing under-trained and unreachable tokens to limit parameter growth. We evaluate our approach on Llama-3.1-8B and Qwen2.5-7B across legal and medical summarization tasks on a challenge-oriented evaluation protocol focused on expert-driven text and summaries which typically has higher concentration of over-fragmented Out-of-Vocabulary (OOV) words. The vocabulary adaptation algorithm enhances the overall quality of the summarization model by improving semantic similarity between the generated summaries and their references. In addition, the adapted model produces summaries that incorporate more appropriate novel and domain-specific words, leading to improved coherence, relevance, and faithfulness. We further observe that our proposed approach significantly reduce training time by 35-55% over continual pretraining and reduce parameter counts up to 37% w.r.t expansion-only methods. We make codebase publicly available at https://github.com/gb-kgp/VocabReplace-Then-Expand.

[Long] LLM-Based Multi-Task Bangla Hate Speech Detection: Type, Severity, and Target ▶ *Underline Video*

Md Arid Hasan, Firoj Alam, Md Fahad Hossain, Usman Naseem and Syed Ishtiaque Ahmed

Online social media platforms have become central to communication and information exchange, however, they also serve as fertile ground for hate speech, offensive language, and bullying targeting individuals and communities. Such content undermines online safety and inclusion, underscoring the need for reliable detection systems, especially in low-resource languages with limited moderation tools. For Bangla, existing work provides valuable resources and models, however, they are mostly single-task (e.g., binary hate/offense) with narrow coverage of key dimensions such as type, severity, and target. We address these gaps by introducing the first multi-task Bangla hate-speech dataset, BanglaMultiHate, one of the largest manually annotated dataset to date. Using this resource, we performed a comparative study across different baselines, monolingual pretrained models, and LLMs under zero-shot, few-shot, and LoRA fine-tuning settings. Our findings show that while LoRA-tuned LLMs rival BanglaBERT, culturally grounded pretraining remains crucial for robust performance. Overall, BanglaMultiHate establishes a stronger benchmark for hate speech detection in low-resource contexts. All data and scripts are released for reproducibility.

[Long] Locket: Robust Feature-Locking Technique for Language Models ▶ *Underline Video*

Lipeng He, Vasisht Duddu and N. Asokan

Chatbot service providers (e.g., OpenAI) rely on tiered subscription plans to generate revenue, offering black-box access to basic models for free users and advanced models to paying subscribers. However, this approach is unprofitable and inflexible. A pay-to-unlock scheme for premium features (e.g., math, coding) offers a more sustainable alternative. Enabling such a scheme requires a feature-locking technique (FLoTE) that is (i) effective in refusing locked features, (ii) utility-preserving for unlocked features, (iii) robust against evasion or unauthorized credential sharing, and (iv) scalable to multiple features and clients. Existing FLoTEs (e.g., password-locked models) fail to meet these criteria. To fill this gap, we present Locket, a more robust and scalable FLoTE to enable pay-to-unlock schemes. We develop a framework for adversarial training and merging of feature-locking adapters, which enables Locket to selectively disable specific features of a model. Evaluation shows that Locket is effective (100% refusal rate), utility-preserving (<= 7% utility degradation), robust (<= 5% attack success rate), and scalable to multiple features and clients.

[Long] LogicEval: A Systematic Framework for Evaluating Automated Repair Techniques for Logical Vulnerabilities in Real-World Software ▶ *Underline Video*

Syed Md Mukit Rashid, Abdullah Al Ishtiaq, Kai Tu, Yilu Dong, Tianwei Wu, Ali Ranjbar, Tianchang Yang, Najrin Sultana, Shagufta Mehnaz and Syed Rafiul Hussain

Logical vulnerabilities in software stem from flaws in program logic rather than memory safety, which can lead to critical security failures. Although existing automated program-repair techniques primarily focus on repairing memory-corruption vulnerabilities, they struggle with logical vulnerabilities because of their limited semantic understanding of the vulnerable code and its expected behavior. On the other hand, recent successes of large language models (LLMs) in understanding and repairing code are promising. However, no framework currently exists to analyze the capabilities and limitations of such techniques for logical vulnerabilities. We aim to systematically evaluate both traditional and LLM-based repair approaches for addressing real-world logical vulnerabilities. To facilitate our assessment, we created the first-ever dataset, LogicDS, comprising 122 logical vulnerabilities that reflect tangible security impact. We also developed a systematic framework, LogicEval, to evaluate patches for logical vulnerabilities. Evaluations suggest that compilation and testing failures are primarily driven by prompt sensitivity, loss of code context, and difficulty in patch localization.

[Findings] MARCH: Evaluating the Intersection of Ambiguity Interpretation and Multi-hop Inference ▶ *Underline Video*

Jeonghyun Park, Ingeol Baek, Seunghyun Yoon, Haeun Jang, Aparna Garimella, Akriti Jain, Nedim Lipka and Hwanhee Lee

Real-world multi-hop QA is naturally linked with ambiguity, where a single query can trigger multiple reasoning paths that require independent resolution. Since ambiguity can occur at any stage, models must navigate layered uncertainty throughout the entire reasoning chain. Despite its prevalence in real-world user queries, previous benchmarks have primarily focused on single-hop ambiguity, leaving the complex interaction between multi-step inference and layered ambiguity underexplored. In this paper, we introduce MARCH, a benchmark for their intersection, with 2,209 multi-hop ambiguous questions curated via multi-LLM verification and validated by human annotation with strong agreement. Our experiments reveal that even state-of-the-art models struggle with MARCH, confirming that combining ambiguity resolution with multi-step reasoning is a significant challenge. To address this, we propose CLARION, a two-stage agentic framework that explicitly decouples ambiguity planning from evidence-driven reasoning, significantly outperforms existing approaches, and paves the way for robust reasoning systems.

[Findings] MATCHA: Matching Text via Contrastive Semantic Alignment ▶ *Underline Video*

Siran Li, Ece Sena Etoglu, Carsten Eickhoff and Seyed Ali Bahrainian

Reliable evaluation is essential for understanding large language model (LLM) performance, yet today's go-to metrics, namely token-overlap scores (e.g., ROUGE) and embedding-based measures (e.g., BERTScore), often misjudge semantic similarity of documents. Our study shows that both token-overlap metrics and embedding-based metrics routinely assign nearly identical scores to texts that directly contradict each other, thereby potentially masking fundamental errors. We introduce MATCHA, an automatic metric that jointly rewards semantic agreement with a reference and penalizes contradictions. MATCHA employs a dual-view perspective that measures (i) proximity to the gold text and (ii) distance from an adversarially generated counterfactual contradiction. In eight public benchmarks, MATCHA outperforms popular metrics, compared with human annotations on question-answering, image caption generation, natural language inference, summarization, and semantic textual similarity tasks. On the TruthfulQA dataset (i.e., a dataset without a training set, where no embedding-based metrics could locally train on), this improvement in terms of matching texts with a reference reaches 18.38% over ROUGE-L and 20.82% over BERTScore. Both quantitative comparison and qualitative human assessments confirm the efficacy and validity of MATCHA and uncover fundamental weaknesses in pre-existing metrics. Compared with 23 embedding models, including top state-of-the-art ones, used as a metric similar to BERTScore, MATCHA remains the most accurate in distinguishing correct from incorrect statements solely based on a reference. Our code and metric are publicly available (https://github.com/Siran-Li/MATCHA).

[Long] Measuring Distribution Shift in User Prompts and Its Effects on LLM Performance ▶ *Underline Video*

Parker Seegmiller and Sarah Masud Preum

LLMs are increasingly deployed in dynamic, real-world settings, where the distribution of user prompts can shift substantially over time as new tasks, prompts, and users are introduced to a deployed model. Such natural prompt distribution shift poses a major challenge to LLM reliability, particularly for specialized models designed for narrow domains or user populations. Despite attention to out-of-distribution robustness, there is very limited exploration of measuring natural prompt distribution shift in prior work, and its impact on deployed LLMs remains poorly understood. We introduce the LLM Evaluation under Natural prompt Shift (LENS) framework: a data-centric approach for quantifying natural prompt distribution shift and evaluating its effect on the performance of deployed LLMs. We perform a large-scale evaluation using 192 real-world post-deployment prompt shift settings over time, user group, and geographic axes, training a total of 81 models on 4.68M training prompts, and evaluating on 57.6k prompts. We find that even moderate shifts in user prompt behavior correspond with large performance drops (73% average loss) in deployed LLMs. This performance degradation is particularly prevalent when users from different latent groups and geographic regions interact with models and is correlated with natural prompt distribution shift over time. We systematically characterize how LLM instruction following ability degrades over time and between user groups. Our findings highlight the critical need for data-driven monitoring to ensure LLM performance remains stable across diverse and evolving user populations.

[Long] Mediocrity is the key for LLM as a Judge Anchor Selection ▶ *Underline Video*

Shachar Don-Yehiya, Asaf Yehudai, Leshem Choshen and Omri Abend

The "LLM-as-a-judge" paradigm has become a standard method for evaluating open-ended generation. To address the quadratic scalability costs of pairwise comparisons, popular benchmarks like Arena-Hard and AlpacaEval compare all models against a single anchor. However, despite its widespread use, the impact of anchor selection on the reliability of the results remains largely unexplored. In this work, we systematically investigate the effect of anchor selection by evaluating 22 different anchors on the Arena-Hard-v2.0 dataset. We find that the choice of anchor is critical: a poor anchor can dramatically reduce correlation with human rankings. We identify that common anchor choices (best-performing and worst-performing models) make poor anchors. Because these extreme anchors are consistently better or worse than all other models, they are seldom indicative of the relative ranking of the models. We further quantify the effect size of anchor selection, showing it is comparable to the selection of a judge model. We conclude with actionable recommendations. First, we conduct a power analysis, and compute sufficient benchmark sizes for anchor-based evaluation, finding that standard benchmark sizes are insufficient for pairwise evaluation and fail to distinguish between competitive models reliably. Second, we provide guidelines for selecting informative anchors to ensure reliable and efficient evaluation practices.

[Findings] MedScore: Generalizable Factuality Evaluation of Open-ended Long-form Medical Answers by Domain-adapted Claim Decomposition and Verification ▶ *Underline Video*

Heyuan Huang, Alexandra DeLucia, Vijay Murari Tiyyala and Mark Dredze
While Large Language Models (LLMs) can generate fluent and convincing responses, they are not necessarily correct. This is especially apparent in the popular decompose-then-verify factuality evaluation pipeline, where LLMs evaluate generated text by decomposing it into individual, valid claims. Factuality evaluation is especially important for medical answers, since incorrect medical information could seriously harm the patient. However, existing factuality systems are a poor match for the medical domain, as they are typically only evaluated on objective, entity-centric, formulaic texts such as biographies and historical topics. This differs from condition-dependent, conversational, hypothetical, sentence-structure diverse, and subjective medical answers, making decomposition into valid facts challenging. We propose MedScore, a new pipeline to decompose medical answers into condition-aware valid facts and verify against in-domain corpora. Our method extracts up to three times as many valid facts as existing methods, reducing hallucination and vague references, and retaining condition-dependency in facts. We also find MedScore is generalizable to non-medical domains without any specific tuning. The resulting factuality score substantially varies by decomposition method, verification corpus, and used backbone LLM, highlighting the importance of customizing each step for reliable factuality evaluation by using our generalizable and modularized pipeline for domain adaptation.

[Findings] MobileBench-OL: A Comprehensive Chinese Benchmark for Evaluating Mobile GUI Agents in Real-World Environment ▶ *Underline Video*
Qinzhuo Wu, Zhizhuo Yang, Hanhao Li, Pengzhi Gao, Wei Liu and Jian Luan
Recent advances in mobile Graphical User Interface (GUI) agents highlight the growing need for comprehensive evaluation benchmarks. While new online benchmarks offer more realistic testing than offline ones, they tend to focus on the agents' task instruction-following ability while neglecting their reasoning and exploration ability. Moreover, these benchmarks do not consider the random noise in real-world mobile environments. This leads to a gap between benchmarks and real-world environments. To addressing these limitations, we propose MobileBench-OL, an online benchmark with 1080 tasks from 80 Chinese apps. It measures task execution, complex reasoning, and noise robustness of agents by including 5 subsets, which set multiple evaluation dimensions. We also provide an auto-eval framework with a reset mechanism, enabling stable and repeatable real-world benchmarking. Evaluating 13 leading GUI agents on MobileBench-OL shows significant room for improvement to meet real-world requirements. Human evaluation further confirms that MobileBench-OL can reliably measure the performance of leading GUI agents in real environments.

[Findings] MTRAG-UN: A Benchmark for Open Challenges in Multi-Turn RAG Conversations ▶ *Underline Video*
Sara Rosenthal, Yannis Katsis, Vraj Shah, Lihong Le, Lucian Popa and Marina Danilevsky
We present MTRAG-UN, a benchmark for exploring open challenges in multi-turn retrieval augment generation, a popular use of large language models. We release a benchmark of 666 tasks from 666 conversations containing over 2,800 conversation turns across 6 domains with accompanying corpora. Our experiments show that retrieval and generation models continue to struggle on conversations with UNanswerable, UNderspecified, and NONstandalone questions and UNclear responses. Our benchmark is available at https://github.com/IBM/mt-rag-benchmark

[Long] MuSe: Multi-Stage Graph Reasoning via Vision-Language Models ▶ *Underline Video*
Guanyu Wang, Xu Chu, Zhijie Tan, Xinrong Chen, Tong Mo and Weiping Li
Graph-related tasks are traditionally addressed with Graph Neural Networks (GNNs) or graph transformers, but their task-specific training limits generalization. Large Language Models (LLMs) offer stronger generalization, yet encoding graphs as one-dimensional text struggles to capture multi-hop dependencies and two-dimensional topology. Vision-Language Models (VLMs) provide an alternative by visualizing graphs, but rendering large graphs in a single image causes clutter, occlusion, and distraction, hindering reasoning. We propose MuSe, a novel multi-stage graph reasoning framework based on VLMs. Instead of processing entire graphs at once, MuSe incrementally samples and visualizes task-relevant subgraphs, enabling progressive reasoning. The framework employs a two-stage training paradigm: supervised fine-tuning to acquire local sampling and reasoning skills, followed by reinforcement learning with GRPO to refine the sampling strategy and control dialog length.To support evaluation, we introduce LGVLQA, a new multimodal dataset with larger and more complex graph structures, addressing the scalability limitations of existing benchmarks. Experiments show that MuSe consistently outperforms leading LLM and VLM baselines, demonstrating improved structural understanding and reasoning ability.

[Findings] ODASim: Ordered, Distinctive and Absolute Semantic Similarity for Code Explanation Evaluation ▶ *Underline Video*
Prince Kumar, Vitobha Munigala, Jaydeep Sen, Ashish Mittal, vishwajeet kumar and Srikanth G. Tamilselvam
Code explanations are increasingly generated by large language models and used in software engineering workflows, making reliable evaluation essential. However, existing model-based and embedding-based methods often fail to distinguish correct explanations from partially or fully incorrect ones, and their similarity scores are poorly calibrated and do not reflect meaningful differences in explanation quality. To address this, we propose ODASim(Orderly, Dstinctive, and Absolute Similarity), a model-agnostic graded fine-tuning framework for embedding models that learns calibrated similarity representations between code and explanations. To support fine-grained supervision and evaluation, we also introduce ODA-X, a novel benchmark for code-to-explanation quality grading, comprising code-explanation pairs graded similarity labels derived from strategic perturbations of gold explanations. We apply our ODASim approach to multiple embedding models and evaluate it on two benchmarks: widely popular CodeXGLUE and our proposed benchmark ODA-X, spanning four programming languages - Python, Java, JavaScript, and Go. Results show that our method achieves up to 35% improvement in F1 score and 85% reduction in Expected Calibration Error (ECE), enabling reliable evaluation of code to explanation quality.

[Long] One Battle After Another: Probing LLMs' Limits on Multi-Turn Instruction Following with a Benchmark Evolving Framework ▶ *Underline Video*

Qi Jia, Ye Shen, Xiujie Song, Kaiwei Zhang, Shibo Wang, Dun Pei, Xiangyang Zhu and Guangtao Zhai

Evaluating LLMs' instruction-following ability in multi-topic dialogues is essential yet challenging. Existing benchmarks are limited to a fixed number of turns, susceptible to saturation and failing to account for users' interactive experience. In this work, we propose a novel framework featuring a three-layer tracking mechanism and a query synthesis agent to mimic sequential user behaviors. Grounded in Flow Theory, we introduce process-centric metrics and terminate a conversational evaluation only upon exhausting user patience. Leveraging this framework, we present EvolIF, an evolving benchmark covering 12 constraint groups. Our analysis reveals deficiencies in failure recovery and fine-grained instruction following, with performance stratification becoming evident as conversational depth increases. GPT-5 demonstrates the most sustained resilience, maintaining a 66.40% stability score, outperforming Gemini-3-Pro by 5.59%, while other models lag behind.

[Findings] OpenGenAlign: A Preference Dataset and Benchmark for Trustworthy Reward Modeling in Open-Ended, Long-Context Generation ▶ *Underline Video*

Hanning Zhang, Juntong Song, Juno Zhu, Yuanhao Wu, Tong Zhang and Cheng Niu

Reward Modeling is critical in evaluating and improving the generation of Large Language Models (LLMs).% They are typically trained on high-quality annotated datasets to distinguish between better and worse responses. While numerous recent works have shown its feasibility in improving safety, helpfulness, reasoning, and instruction-following ability, its capability and generalization to open-ended long-context generation is still rarely explored. In this paper, we introduce OpenGenAlign, a framework and a high-quality dataset designed to develop reward models to evaluate and improve hallucination-free, comprehensive, reliable, and efficient open-ended long-context generation. We define four key metrics to assess generation quality and develop an automated pipeline to evaluate the outputs of multiple LLMs across long-context QA, Data-to-Text, and Summarization scenarios using o3, ending up with 33K high-quality preference data with a human agreement rate of 81%. Experimental results first demonstrate that existing reward models perform suboptimally on the held-out benchmark. And Our trained reward model achieves superior performance in the benchmark and effectively improves the generation quality of the policy models using Reinforcement Learning (RL). Additionally, OpenGenAlign could be used for effective guided generation in existing datasets. Furthermore, we demonstrate that the OpenGenAlign could be integrated with reward data from other domains to achieve better performance.

[Long] Overcoming Copyright Barriers in Corpus Distribution Through Non-Reversible Hashing ▶ *Underline Video*

Arthur Amalvy, Vincent Labatut, Xavier Bost and Hen-Hsen Huang

While annotated corpora are crucial in the field of natural language processing (NLP), those containing copyrighted material are difficult to exchange among researchers. Yet, such corpora are necessary to fully represent the diversity of data found in the wild in the context of NLP tasks. We tackle this issue by proposing a method to lawfully share the annotations of any sequential copyrighted corpus. The corpus creator shares the annotations in clear, along with a non-reversible hashed version of the source material. The corpus user must own the source material, and apply the same hash function to their own tokens, in order to match them to the shared annotations. Crucially, our method is robust to reasonable divergences in the version of the copyrighted data owned by the user. As an illustration, we present alignment experiments on different editions of novels. Our results show that our method is able to correctly correctly align 98.7 to 99.79% of tokens depending on the novel, provided the user version is sufficiently close to the corpus creator's version. We publicly release novelties-bookshare, a Python implementation of our method.

[Long] P-Check: Advancing Personalized Reward Model via Learning to Generate Dynamic Checklist ▶ *Underline Video*

Kwangwook Seo and Dongha Lee

Recent approaches in personalized reward modeling have primarily focused on leveraging user interaction history to align model judgments with individual preferences. However, existing approaches largely treat user context as a static or implicit conditioning signal, failing to capture the dynamic and multi-faceted nature of human judgment. In this paper, we propose P-Check, a novel personalized reward modeling framework, designed to train a plug-and-play checklist generator that synthesizes dynamic evaluation criteria for guiding the reward prediction. To better align these checklists with personalized nuances, we introduce Preference-Contrastive Criterion Weighting, a training strategy that assigns saliency scores to criteria based on their discriminative power for personalized judgment. We conduct extensive experiments and demonstrate that P-Check not only improves reward accuracy but also enhances downstream personalized generation, and remains robust in OOD scenarios.

[Long] PIArena: A Platform for Prompt Injection Evaluation ▶ *Underline Video*

Runpeng Geng, Chenlong Yin, Yanting Wang, Ying Chen and Jinyuan Jia

Prompt injection attacks pose serious security risks across a wide range of real-world applications. While receiving increasing attention, the community faces a critical gap: the lack of a unified platform for prompt injection evaluation. This makes it challenging to reliably compare defenses, understand their true robustness under diverse attacks, or assess how well they generalize across tasks and benchmarks. For instance, many defenses initially reported as effective were later found to exhibit limited robustness on diverse datasets and attacks. To bridge this gap, we introduce PIArena, a unified and extensible platform for prompt injection evaluation that enables users to easily integrate state-of-the-art attacks and defenses and evaluate them across a variety of existing and new benchmarks. We also design a dynamic strategy-based attack that

adaptively optimizes injected prompts based on defense feedback. Through comprehensive evaluation using PIArena, we uncover critical limitations of state-of-the-art defenses: limited generalizability across tasks, vulnerability to adaptive attacks, and fundamental challenges when an injected task aligns with the target task. The code and datasets are available at https://github.com/sleeepeer/PIArena.

[Long] PLAWBENCH: A Rubric-Based Benchmark for Evaluating LLMs in Real-World Legal Practice ► *Underline Video*

Yuzhen Shi, Huanghai Liu, Yiran HU, Song gaojie, Xu Xinran, Yubo Ma, Tianyi Tang, Li Zhang, Qingjing Chen, Feng Di, Wenbo Lv, Weiheng Wu, Kexin Yang, Sen Yang, Wei Wang, Rongyao Shi, Qiu Yuanyang, Yuemeng Qi, Zhang Jingwen, Sui Xiaoyu, Yifan Chen, ZhangYi ZhangYi, An Yang, Bowen Yu, Dayiheng Liu, Junyang Lin, Weixing Shen, Bing Zhao, Charles L. A. Clarke and HU WEI

As large language models (LLMs) are increasingly applied to legal domain-specific tasks, evaluating their ability to perform legal work in real-world settings has become essential. However, existing legal benchmarks rely on simplified and highly standardized tasks, failing to capture the ambiguity, complexity, and reasoning demands of real legal practice. Moreover, prior evaluations often adopt coarse, single-dimensional metrics and do not explicitly assess fine-grained legal reasoning. To address these limitations, we introduce PLawBench, a Practical Law Benchmark designed to evaluate LLMs in realistic legal practice scenarios. Grounded in real-world legal workflows, PLawBench models the core processes of legal practitioners through three task categories: public legal consultation, practical case analysis, and legal document generation. These tasks assess a model's ability to identify legal issues and key facts, perform structured legal reasoning, and generate legally coherent documents. PLawBench comprises 850 questions across 13 practical legal scenarios, with each question accompanied by expert-designed evaluation rubrics, resulting in approximately 12,500 rubric items for fine-grained assessment. Using an LLM-based evaluator aligned with human expert judgments, we evaluate 10 state-of-the-art LLMs. Experimental results show that none achieves strong performance on PLawBench, revealing substantial limitations in the fine-grained legal reasoning capabilities of current LLMs and highlighting important directions for future evaluation and development of legal LLMs. Data is available at: https://anonymous.4open.science/r/PLawbench-B524/.

[Long] PluRule: A Benchmark for Moderating Pluralistic Communities on Social Media ► *Underline Video*

Zoher Kachwala, Bao Tran Truong, Rasika Muralidharan, Haewoon Kwak, Jisun An and Filippo Menczer;

Social media are shifting towards pluralism , community-governed platforms where groups define their own norms. What violates rules in one community may be perfectly acceptable in another. Can AI models help moderate such pluralistic communities? We formalize the task as a multiple-choice problem, mirroring how human moderators operate in the real world: given a comment and its surrounding context, identify which specific rule, if any, is violated. We introduce PluRule, a multimodal, multilingual benchmark for detecting 13,371 rule violations across 1,989 Reddit communities spanning 2,885 rules in 9 languages. Using this benchmark, we show that state-of-the-art vision-language models struggle significantly: even GPT-5.2 with high reasoning performs only slightly better than a trivial baseline. We also find that bigger models and increased context provide marginal gains, and universal rules like civility and self-promotion are easier to detect. Our results show that moderation of pluralistic communities on social media is a fundamental challenge for language models. Our code and benchmark are publicly available.

[Long] PRBench: Large-Scale Expert Rubrics for Evaluating High-Stakes Professional Reasoning ► *Underline Video*

Afra Feyza Akyürek, Advait Gosai, Chen Bo Calvin Zhang, Vipul Gupta, Jaehwan Jeong, Anisha Gunjal, Tahseen Rabbani, Maria Mazzone, David Randolph IV, Mohammad Mahmoudi Meymand, Gurshaan Chattha, Paula Rodriguez, Diego A. Mares Buendia, Pavit Singh, Michael Liu, Subodh Chawla, Peter Cline, Lucy Ogaz, Ernesto Gabriel Hernández Montoya, Zihao Wang, Pavi Bhatter, Marcos Ayestaran, Bing Liu and Yunzhong He

Frontier model progress is often measured using academic benchmarks that provide a limited view of performance on open-ended, economically consequential tasks in high-stakes professional domains where practical returns matter most. We introduce Professional Reasoning Bench (PRBench), a realistic, open-ended, and difficult benchmark of real-world problems in Finance and Law. We open-source its 1,100 expert-authored tasks and 19,356 expert-curated criteria, making it the largest public, rubric-based benchmark for both legal and finance domains. We recruit 182 qualified professionals, holding JDs, CFAs, or 6+ years of experience, who contributed questions inspired by their actual workflows. This process yields significant diversity, with tasks spanning 114 countries and 47 US jurisdictions. Our expert-curated rubrics are validated through a rigorous quality pipeline, including independent expert validation. Subsequent evaluation of 20 leading models reveals substantial room for improvement, with top scores of only 0.39 (Finance) and 0.37 (Legal) on our Hard subsets. We further catalog associated economic impacts of the prompts and analyze performance using human-annotated rubric categories. Common failure modes include inaccurate judgments, a lack of process transparency and incomplete reasoning, highlighting critical gaps in their reliability for professional adoption.

[Long] PRIME: A Process-Outcome Alignment Benchmark for Verifiable Reasoning in Mathematics and Engineering ► *Underline Video*

Xiangfeng Wang, Hangyu Guo, Yanlin Lai, Mitt Huang, Liang Zhao, Chengyuan Yao, Yinmin Zhang, Qi Han, Xiaoxiao Ren, Chun Yuan, Tong Xu, Zheng Ge, Xiangyu Zhang and Daxin Jiang

While model-based verifiers are essential for scaling Reinforcement Learning with Verifiable Rewards (RLVR), current outcome-centric verification paradigms primarily focus on the consistency between the final result and the ground truth, often neglecting potential errors in the derivation process. This leads to assigning positive rewards to correct answers produced from incorrect derivations. To bridge this gap, we introduce PRIME, a benchmark for evaluating verifiers on PRocess-outcome alignment verification In Mathematics and Engineering.

Curated from a comprehensive collection of college-level STEM problems, PRIME comprises 2,530 high-difficulty samples through a consistency-based filtering pipeline. Through extensive evaluation, we find that current verifiers frequently fail to detect derivation flaws. Furthermore, we propose a process-aware RLVR training paradigm utilizing verifiers selected via PRIME. This approach substantially outperforms the outcome-only verification baseline, achieving absolute performance gains of 8.29%, 9.12%, and 7.31% on AIME24, AIME25, and Beyond-AIME, respectively, for the Qwen3-14B-Base model. Finally, we demonstrate a strong linear correlation ($R^2 > 0.92$) between verifier accuracy on PRIME and RLVR training effectiveness, validating PRIME as a reliable predictor for verifier selection.

[Long] PRISM: Probabilistic Reward Model with Inherent Structural Modeling ▶ *Underline Video*

Yuhang Zhou, Yuchen Ni, Xiang Liu, Shihan Dou, Xutian Chen, Ge Zhang, Guangnan Ye and Yixin Cao

Standard evaluators, such as reward models, compress diverse human judgments into a single scalar, conflating valid Subjective Preference with Cognitive Uncertainty. This structural mismatch often leads to brittle alignment and reward hacking. To address this, we propose PRISM which reinterprets reward evaluation as a conditional distribution parameterized by a Mixture of Gaussians. PRISM structurally disentangles these factors: distinct Gaussian experts emerge to capture conflicting preference dimensions, while their variance estimates quantify uncertainty, acting as a dynamic reliability gate during optimization. We introduce a two-stage training strategy to learn these disentangled representations from scalable pairwise comparisons without requiring massive fine-grained annotations. Empirical results show that PRISM significantly outperforms scalar baselines in both accuracy and generalization. Furthermore, in downstream Reinforcement Learning, PRISM effectively mitigates reward hacking, yielding policies that are more robust and resilient to distribution shifts.

[Findings] ProMediate: A Simulation Testbed for Evaluating Proactive Mediation in Multi-Party Negotiation ▶ *Underline Video*

Ziyi Liu, Bahareh Sarrafzadeh, Pei Zhou, Longqi Yang, Jieyu Zhao and Ashish Sharma

While LLMs increasingly assist individual users, there is a critical need for agents that can proactively manage complex, multi-party collaboration. However, the scarcity of systematic evaluation methods for these group dynamics limits the development of AI capable of effectively supporting teams Here, we present ProMediate, the first testbed for evaluating proactive AI mediator agents in complex, multi-topic, multi-party negotiations. ProMediate consists of two core components: (i) a simulation environment based on realistic negotiation cases with a plug-and-play proactive AI mediator, capable of flexibly deciding when and how to intervene; and (ii) a socio-cognitive evaluation framework with a new suite of metrics to measure consensus changes, intervention latency, mediator effectiveness, and intelligence. These components establish a systematic framework for assessing the capability of proactive AI agents in multi-party settings. Our results show that a socially intelligent mediator agent outperforms a generic baseline, via faster, better-targeted interventions. In the ProMediate-Hard setting, our social mediator increases consensus change by 3.6 percentage points compared to the generic baseline (10.65% vs 7.01%) while being 77% faster in response (15.98s vs. 3.71s). In conclusion, ProMediate provides a rigorous, theory-grounded testbed to advance the development of proactive, socially intelligent agents.

[Long] Psyche-R1: Towards Reliable Psychological LLMs through Unified Empathy, Expertise, and Reasoning ▶ *Underline Video*

Chongyuan Dai, Jinpeng Hu, Hongchang Shi, Zhuo Li, Dan Guo, Xun Yang and Meng Wang

Amidst a shortage of qualified mental health professionals, the integration of large language models (LLMs) into psychological applications offers a promising way to alleviate the growing burden of mental health disorders. Recent reasoning-augmented LLMs have achieved remarkable performance in mathematics and programming, while research in the psychological domain has predominantly emphasized emotional support and empathetic dialogue, with limited attention to reasoning mechanisms that are beneficial to generating accurate responses. Therefore, in this paper, we propose Psyche-R1, the first Chinese psychological LLM that jointly integrates empathy, psychological expertise, and reasoning, built upon a novel data curation pipeline. Specifically, we design a comprehensive data synthesis pipeline that produces over 75k high-quality psychological questions paired with detailed rationales, generated through and iterative prompt-rationale optimization procedure, along with 73k empathetic dialogues. Subsequently, we employ a hybrid training strategy wherein challenging samples are identified through a multi-LLM cross-selection strategy for group relative policy optimization (GRPO) to improve reasoning ability, while the remaining data are used for supervised fine-tuning (SFT) to enhance empathetic response generation and psychological domain knowledge. Extensive experiment results demonstrate the effectiveness of Psyche-R1 across several psychological benchmarks, where our 7B Psyche-R1 achieves comparable results to 671B DeepSeek-R1.

[Long] Pub-LawBench: Public-Oriented Benchmarking for LegalAI ▶ *Underline Video*

Qiaoyu Zheng, Zehan Ma, Yijing Zhang, Qiqi Wang, Huijia Li and Qian Liu

Large language models (LLMs) are playing an increasingly pivotal role in LegalAI. However, existing benchmarks are primarily tailored for legal professionals, emphasizing deep reasoning and explainability. While public-facing legal applications demand outputs that are direct, actionable, and accessible, a need largely overlooked by current evaluation frameworks. To bridge this gap, we propose a public-oriented LegalAI benchmark grounded in legal functionalism and genre analysis. Specifically, we categorize public legal demands into two core tasks: Instant Question Answering and Legal Text Generation. We further introduce three public-oriented evaluation dimensions: legal normativity, content relevance, and format usability, which collectively assess the practical validity and user readiness of model outputs. To reflect real-world lay user usage, we evaluate 17 LLMs on Pub-LawBench using only simple prompts and Chain-of-Thought under a vanilla inference setting, excluding complex techniques like RAG or agent-based methods inaccessible to non-experts. Experiments reveal limitations of current LLMs in delivering effective public-oriented legal assistance, highlighting the need for more user-centric model development and benchmarking. Our code and datasets are available for review at https://anonymous.4open.science/r/P-LawBench-E565/.

[Long] Ready Jurist One: Benchmarking Language Agents for Legal Intelligence in Dynamic Environments ▶ *Underline Video*

Zheng Jia, Shengbin Yue, Wei Chen, Siyuan Wang, Yidong Liu, Zejun Li, Yun Song and zhongyu wei

The gap between existing benchmarks and the dynamic nature of real-world legal practice poses a key barrier to advancing legal intelligence. To this end, we introduce J1-ENVS, the first interactive and dynamic legal environment tailored for LLM-based agents. Guided by legal experts, it comprises six representative scenarios from Chinese legal practices at three levels of environmental complexity. We further introduce J1-EVAL, a dual-metric evaluation framework, designed to assess both task performance and procedural compliance across varying levels of legal proficiency. Extensive experiments on 17 LLM agents reveal that while many models demonstrate solid legal knowledge, they struggle with procedural execution in dynamic settings. Even the SOTA model is below 60% overall performance. These findings highlight persistent challenges in achieving dynamic legal intelligence and offer valuable insights to guide future research.

[Long] Reasoning with OmniThought: A Large CoT Dataset with Verbosity and Cognitive Difficulty Annotations ▶ *Underline Video*

Wenrui Cai, Chengyu Wang, Junbing Yan, Jun Huang and Xiangzhong Fang

Tasks such as mathematical problem solving and coding require models to leverage chain-of-thought (CoT) processes, enabling human-like reasoning strategies. However, the advancement of large reasoning models (LRMs) is hindered by the lack of comprehensive CoT datasets. Existing resources often fail to provide extensive reasoning problems with coherent CoT processes distilled from multiple teacher models, and do not account for multifaceted properties describing the internal characteristics of CoTs. To address these challenges, we introduce OmniThought, a large-scale dataset featuring 2 million CoT processes generated and validated by multiple powerful LRMs. Each CoT process in OmniThought is annotated with novel Reasoning Verbosity (RV) and Cognitive Difficulty (CD) scores, which characterize the appropriateness of CoT verbosity and the cognitive difficulty level for models to comprehend these reasoning processes. We further establish a self-reliant pipeline to curate this dataset. Extensive experiments using Qwen2.5 and Qwen3 of various sizes demonstrate the positive impact of our RV and CD scores on LRM training effectiveness. Based on the OmniThought dataset, we train and release a series of high-performing LRMs with enhanced reasoning abilities and optimized CoT output length. Our contributions advance the development of LRMs across different scales for solving complex reasoning tasks.

[Findings] Red-Teaming NSFW Image Classifiers as Text-to-Image Safeguards ▶ *Underline Video*

Tinghao Xie, Yueqi Xie, Alireza Zareian, Shuming Hu, Felix Juefei-Xu, Xiaowen Lin, Ankit Jain, Prateek Mittal and Li Chen

Not Safe for Work (NSFW) image classifiers play a critical role in safeguarding text-to-image (T2I) systems. However, a concerning phenomenon has emerged in T2I systems – changes in text prompts that manipulate benign image elements can result in failed detection by NSFW classifiers – dubbed "context shifts." For instance, while a NSFW image of "a nude person in an empty scene" can be easily blocked by most NSFW classifiers, a stealthier one that depicts "a nude person blending in a group of dressed people" may evade detection. We ask: how to systematically reveal NSFW image classifiers' failure against such context shifts? Towards this end, we present an automated red-teaming framework that leverages a set of generative AI tools. We propose an exploration-exploitation approach: First, in the exploration stage, we synthesize a diverse and massive 36K NSFW image dataset that facilitates our study of context shifts. We find that varying fractions (e.g., 4.1% to 36% nude and sexual content) of the dataset are misclassified by NSFW image classifiers like GPT-4o and Gemini. Second, in the exploitation stage, we leverage these failure cases to train a specialized LLM that rewrites unseen seed prompts into more evasive versions, increasing the likelihood of detection evasion by up to 6 times. Alarmingly, we show these failures translate to real-world T2I and even T2V systems like DALL-E 3, Sora, Nano Banana, and Veo 3 – beyond the open-weight image generators in our main study. For example, querying DALL-E 3 with prompts rewritten by our approach increases the chance of obtaining NSFW images from 0 to over 50%.

[Long] Revisiting Metric Reliability for Fine-grained Evaluation of Machine Translation and Summarization in Indian Languages ▶ *Underline Video*

Amir Hossein Yari, Kalmit Kulkarni, Ahmad Raza Khan and Fajri Koto

While automatic metrics drive progress in Machine Translation (MT) and Text Summarization (TS), existing metrics have been developed and validated almost exclusively for English and other high-resource languages. This narrow focus leaves Indian languages, spoken by over 1.5 billion people, largely overlooked, casting doubt on the universality of current evaluation practices. To address this gap, we introduce ITEM, a large-scale benchmark that systematically evaluates the alignment of 26 automatic metrics with human judgments across six major Indian languages, enriched with fine-grained annotations. Our extensive evaluation, covering agreement with human judgments, sensitivity to outliers, language-specific reliability, inter-metric correlations, and resilience to controlled perturbations, reveals four central findings: (1) LLM-based evaluators show the strongest alignment with human judgments at both segment and system levels; (2) outliers exert a significant impact on metric-human agreement; (3) In TS, metrics are more effective at capturing content fidelity, whereas in MT, they better reflect fluency; and (4) Metrics differ in their robustness and sensitivity when subjected to diverse perturbations. Collectively, these findings offer critical guidance for advancing metric design and evaluation in Indian languages.

[Long] ROSE: An Intent-Centered Evaluation Metric for NL2SQL ▶ *Underline Video*

Wenqi Pei, Shizheng Hou, Boyan Li, Han Chen, Zhichao Shi and Yuyu Luo

Execution Accuracy (EX), the widely used metric for evaluating the effectiveness of Natural Language to SQL (NL2SQL) solutions, is becoming increasingly unreliable. It is sensitive to syntactic variation, ignores that questions may admit multiple interpretations, and is easily misled by erroneous ground-truth SQL. To address this, we introduce ROSE, an intent-centered metric that focuses on whether the predicted

SQL answers the question, rather than consistency with the ground-truth SQL under the reference-dependent paradigm. ROSE employs an adversarial Prover-Refuter cascade: SQL Prover assesses the semantic correctness of a predicted SQL against the user's intent independently, while Adversarial Refuter uses the ground-truth SQL as evidence to challenge and refine this judgment. On our expert-aligned validation set ROSE-VEC, ROSE achieves the best agreement with human experts, outperforming the next-best metric by nearly 24% in Cohen's Kappa. We also conduct a large-scale re-evaluation of 19 NL2SQL methods, revealing four valuable insights. We release ROSE and ROSE-VEC to facilitate more reliable NL2SQL research.

[Long] ScholaWrite: A Dataset of End-to-End Scholarly Writing ► *Underline Video*

Khanh Chi Le, Linghe Wang, Minhwa Lee, Ross Volkov, Luan Tuyen Chau and Dongyeop Kang

Writing is a cognitively demanding activity that requires constant decision-making, heavy reliance on working memory, and frequent shifts between tasks of different goals. To build writing assistants that truly align with writers' cognition, it is necessary to capture and analyze the complete thought process behind how writers transform ideas into final texts. We present SCHOLAWRITE, the first dataset of end-to-end scholarly writing, tracing the multi-month journey from initial drafts to final manuscripts. The dataset traces nearly 62K LaTeX-based edits from five computer science preprints over four months and is enriched with fine-grained annotations of cognitive writing intentions. We demonstrate the value of ScholaWrite through three complementary contributions: (1) analysis of real-world writing behavior reveals that scholarly writing is highly non-linear and multi-intentional, blending rapid drafting bursts with cognitively sustained writing sessions; (2) evaluations of current large language models show that they struggle to provide meaningful support throughout the human writing process; and (3) models finetuned on SCHOLAWRITE demonstrate improved alignment with human writing workflows. SCHOLAWRITE underscores the value of capturing scientists' cognitive writing process and provides actionable insights and resources for the development of future writing assistants.

[Long] See2Refine: Vision-Language Feedback Improves LLM-Based eHMI Action Designers ► *Underline Video*

Ding Xia, Xinyue Gui, Mark Colley, Fan Gao, Zhongyi Zhou, Dongyuan Li, Renhe Jiang and Takeo Igarashi

Automated vehicles lack natural communication channels with other road users, making external Human-Machine Interfaces (eHMIs) essential for conveying intent and maintaining trust in shared environments. However, most eHMI studies rely on developer-crafted message-action pairs, which are difficult to adapt to diverse and dynamic traffic contexts. A promising alternative is to use Large Language Models (LLMs) as action designers that generate context-conditioned eHMI actions, yet such designers lack perceptual verification and typically depend on fixed prompts or costly human-annotated feedback for improvement. We present See2Refine, a human-free, closed-loop framework that uses vision-language models (VLMs) for perceptual evaluation as automated visual feedback to improve an LLM-based eHMI action designer. Given a driving context and a candidate eHMI action, the VLM evaluates the perceived appropriateness of the action, and this feedback is used to iteratively revise the designer's outputs, enabling systematic refinement without human supervision. We evaluate our framework across three eHMI modalities (lightbar, eyes, and arm) and multiple LLM model sizes. Across settings, our framework consistently outperforms prompt-only LLM designers and manually specified baselines in both VLM-based metrics and human-subject evaluations. The results further indicate that the improvements are generalized across modalities and that VLM evaluations are reasonably aligned with human preferences in our controlled settings, supporting the robustness and effectiveness of See2Refine for scalable action design.

[Long] SkMTEB: Slovak Massive Text Embedding Benchmark and Model Adaptation ► *Underline Video*

Marek SuppaAndrej RidzikDaniel HládekNatália KaekováViktória Ondrejová

We introduce SkMTEB, the first comprehensive MTEB-style text embedding benchmark for Slovak, a low-resource West Slavic language, comprising 31 datasets across 7 task types—nearly 4x the depth of existing multilingual benchmark coverage for Slovak. Our evaluation of 31 embedding models reveals that large instruction-tuned multilingual models achieve the strongest performance, while existing Slovak-specific models trained for NLU tasks transfer poorly to embedding tasks. To address the need for efficient, locally-deployable Slovak embeddings, we develop e5-sk-small (45M parameters) and e5-sk-large (365M) by applying vocabulary trimming and fine-tuning to Multilingual E5 models. Despite size reductions of up to 62%, our open-source models achieve competitive performance with proprietary APIs while remaining locally deployable for semantic search and retrieval-augmented generation (RAG). We release the benchmark, models, datasets, and code openly, hoping our approach offers a replicable path for other under-resourced languages.

[Findings] ToolPRMBench: Evaluating and Advancing Process Reward Models for Tool-using Agents ► *Underline Video*

Dawei Li, Yuguang Yao, Zhen Tan, Huan Liu and Ruocheng Guo

Reward-guided search methods have demonstrated strong potential in enhancing tool-using agents by effectively guiding sampling and exploration over complex action spaces. As a core design, those search methods utilize process reward models (PRMs) to provide step-level rewards, enabling more fine-grained monitoring. However, there is a lack of systematic and reliable evaluation benchmarks for PRMs in tool-use settings. In this paper, we introduce ToolPRMBench, a large-scale benchmark specifically designed to evaluate PRMs for tool-using agents. ToolPRMBench is built on top of several representative tool-use benchmarks and converts agent trajectories into step-level test cases. Each case contains the interaction history, a correct action, a plausible but incorrect alternative, and relevant tool metadata. We respectively utilize offline sampling to isolate local single-step errors and online sampling to capture realistic multi-step failures from full agent rollouts. A multi-LLM verification pipeline is proposed to reduce label noise and ensure data quality. We conduct extensive experiments across large language models, general PRMs, and tool-specialized PRMs on ToolPRMBench. The results reveal clear

differences in PRM effectiveness and highlight the potential of specialized PRMs for tool-using. Our code and dataset are available at: https://github.com/David-Li0406/ToolPRMBench.

[Findings] ToxReason: A Benchmark for Mechanistic Chemical Toxicity Reasoning via Adverse Outcome Pathway ▶ *Underline Video*
Jueon Park, Wonjune Jang, Chanhwi Kim, Yein Park and Jaewoo Kang
Recent advances in large language models (LLMs) have enabled molecular reasoning for property prediction. However, toxicity arises from complex biological mechanisms beyond chemical structure, necessitating mechanistic reasoning for reliable prediction. Despite its importance, current benchmarks fail to systematically evaluate this capability. LLMs can generate fluent but biologically unfaithful explanations, making it difficult to assess whether predicted toxicities are grounded in valid mechanisms. To bridge this gap, we introduce ToxReason, a benchmark grounded in the Adverse Outcome Pathway (AOP) that evaluates organ-level toxicity reasoning across multiple organs. ToxReason integrates experimental drug-target interaction evidence with toxicity labels, requiring models to infer both toxic outcomes and their underlying mechanisms from Molecular Initiating Event (MIE) to Adverse Outcome (AO). Using ToxReason, we evaluate toxicity prediction performance and reasoning quality across diverse LLMs. We find that strong predictive performance does not necessarily imply reliable reasoning. Furthermore, we show that reasoning-aware training improves mechanistic reasoning and, consequently, toxicity prediction performance. Together, these results underscore the necessity of integrating reasoning into both evaluation and training for trustworthy toxicity modeling.

[Long] TRN-R1-Zero: Text-rich Network Reasoning via LLMs with Reinforcement Learning Only ▶ *Underline Video*
Yilun Liu, Ruihong Qiu and Zi Huang
Zero-shot reasoning on text-rich networks (TRNs) remains a challenging frontier, as models must integrate textual semantics with relational structure without task-specific supervision. While graph neural networks rely on fixed label spaces and supervised objectives, recent large language model (LLM)-based approaches often overlook graph context or depend on distillation from larger models, limiting generalisation. We propose TRN-R1-Zero, a post-training framework for TRN reasoning trained solely via reinforcement learning. TRN-R1-Zero directly optimises base LLMs using a Neighbour-aware Group Relative Policy Optimisation objective that dynamically adjusts rewards based on a novel margin gain metric for the informativeness of neighbouring signals, effectively guiding the model toward relational reasoning. Unlike prior methods, TRN-R1-Zero requires no supervised fine-tuning or chain-of-thought data generated from large reasoning models. Extensive experiments across citation, hyperlink, social and co-purchase TRN benchmarks demonstrate the superiority and robustness of TRN-R1-Zero. Beyond cross-domain transfer, TRN-R1-Zero, trained solely on node-level tasks, further generalises to edge- and graph-level tasks in a zero-shot manner. The codebase is open-source at https://github.com/superallen13/TRN-R1-Zero.

[Findings] UCGRec: User-Centric Graph Learning for LLM-based Sequential Recommendation ▶ *Underline Video*
HanBeul Kim, CheolWon Na, Suyoung Bae, YunSeok Choi and Jee-Hyong Lee
Recently, Large Language Models (LLM) have emerged as a promising paradigm for sequential recommendation. In sequential recommendation, effectively integrating diverse user preferences is essential for improving LLM performance, as users often exhibit multiple interests across different contexts. However, most existing LLM-based methods rely primarily on item descriptions or utilize user preferences independently. As a result, they overlook the relationships among preferences and fail to filter out less-relevant items that introduce noise. This makes it difficult to accurately capture the user's interests, leading to suboptimal recommendations. To overcome these limitations, we propose UCGRec (User-Centric Graph Learning for LLM-based Sequential Recommendation), a novel method that effectively integrates diverse user-relevant preference signals into a unified user-centric graph. Then, we inject the graph-based knowledge into the LLM through end-to-end training with graph neural networks. We conduct extensive experiments on four widely used sequential real-world recommendation datasets. Our experimental results demonstrate that UCGRec significantly outperforms conventional and state-of-the-art LLM-based methods.

[Long] UrbanGeoEval: A City-Scale Benchmark for Evaluating Large Language Models in Geospatial Reasoning ▶ *Underline Video*
Mutian Bao, Qiuyi Qi, Tian Liang, Jinjian Zhang, Wei Zhou, Ming Kong, Linjian Mo and Qiang Zhu
Current evaluations of geospatial reasoning in LLMs are frequently impeded by the entanglement of factual recall and spatial logic, which often obscures the models' true capabilities in complex city-scale environments. To address this, we introduce UrbanGeoEval, a comprehensive benchmark featuring a dual-module framework designed to disentangle these competencies. The Knowledge Module assesses urban memory via scalable map-based queries, while the Reasoning Module isolates pure logical inference across 3,148 realistic tasks by providing necessary geospatial context. Unlike prior benchmarks that hand the model pre-computed spatial text, UrbanGeoEval provides raw geometry and forces the model to act as a spatial computing engine. Our evaluation methodology introduces a reliable hybrid pipeline that merges deterministic programmatic checks with an LLM-as-a-Judge, achieving expert-level evaluation accuracy. Extensive experiments on 18 widely used LLMs uncover critical insights: (1) models exhibit severe geographic biases and resolution gaps; (2) failures in complex multi-hop tasks often stem from brittle foundational spatial skills rather than high-level logic deficits. UrbanGeoEval provides a precise diagnostic tool for advancing urban geospatial intelligence in LLMs.

[Findings] VChain: Chain-of-Visual-Thought for Reasoning in Video Generation ▶ *Underline Video*
Ziqi Huang, Ning Yu, Gordon Chen, Haonan Qiu, Paul Debevec and Ziwei Liu
Recent video generation models can produce smooth and visually appealing clips, but they often struggle to synthesize complex dynamics with a coherent chain of consequences. Accurately modeling visual outcomes and state transitions over time remains a core challenge. In

contrast, large language and multimodal models (e.g., GPT-4o) exhibit strong visual state reasoning and future prediction capabilities. To bridge these strengths, we introduce VChain, a novel inference-time chain-of-visual-thought framework that injects visual reasoning signals from multimodal models into video generation. Specifically, VChain contains a dedicated pipeline that leverages large multimodal models to generate a sparse set of critical keyframes as snapshots, which are then used to guide the sparse inference-time visual-state adaptation of a pre-trained video generator only at these key moments. Our approach is tuning-efficient, introduces minimal overhead and avoids dense supervision. Extensive experiments on complex, multi-step scenarios show that VChain significantly enhances the quality of generated videos.

[Long] VisAidMath: Benchmarking Visual-Aided Mathematical Reasoning ▶ *Underline Video*

Jingkun Ma, Runzhe Zhan, Yang Li, Di Sun, Hou Pong Chan, Lidia S. Chao and Derek F. Wong

A hallmark of advanced artificial intelligence is the capacity to progress from passive visual perception to the strategic modification of visual information to facilitate complex reasoning. This advanced capability, however, remains critically underdeveloped in current Large Multi-modal Models (LMMs). The deficiency is often masked by evaluation metrics that prioritize final-answer accuracy, creating an illusion of competence where genuine reasoning is absent. Using the domain of geometric problem-solving as a precise instrument, we probe this issue through tasks that require constructing visual aids.To this end, we introduce VisAidMath, a challenging benchmark, and our novel Three-Layered Funnel Evaluation Framework. This framework moves beyond simple accuracy (ACCU) to scrutinize the generation of valid visual aids (PVA) and the soundness of subsequent reasoning steps (SPRS). Our extensive experiments on state-of-the-art models, including Doubao-Seed-1.6 and o4, reveal a profound "Reasoning Illusion". We observe that high surface-level accuracy conceals a catastrophic failure in the models' ability to produce valid visual aids or to reason from them. Our findings expose a fundamental schism between visual perception and logical deduction in modern LMMs. We provide a public evaluation platform on CodaBench and release the project homepage.

[Long] When Benchmarks Leak: Inference-Time Decontamination for LLMs ▶ *Underline Video*

Jianzhe Chai, YU ZHE and Jun Sakuma

Benchmark-based evaluation is the de facto standard for comparing large language models (LLMs). However, its reliability is increasingly threatened by test set contamination, where test samples or their close variants leak into training data and artificially inflate reported performance. To address this issue, prior work has explored two main lines of mitigation. One line attempts to identify and remove contaminated benchmark items before evaluation, but this inevitably alters the evaluation set itself and becomes unreliable when contamination is moderate or severe. The other line preserves the benchmark and instead suppresses contaminated behavior at evaluation time; however, such interventions often interfere with normal inference and lead to noticeable performance degradation on clean inputs. We propose DeconIEP, a decontamination framework that operates entirely during evaluation by applying small, bounded perturbations in the input embedding space. Guided by a relatively less-contaminated reference model, DeconIEP learns an instance-adaptive perturbation generator that steers the evaluated model away from memorization-driven shortcut pathways. Across multiple open-weight LLMs and benchmarks, extensive empirical results show that DeconIEP achieves strong decontamination effectiveness while incurring only minimal degradation in benign utility.

Poster Session E (Grand Hall)

11:00-12:30 (Grand Hall)

[Findings] A Dual-View Analysis of Multiple Languages in Colonial Newspapers ▶ *Underline Video*

Zhan Su, Xiaoya Chen, Fengran Mo, Ida L. Vos, Prayag Tiwari, Yazhou Zhang, Qian Zheng and Natália da Silva Perez

Historical newspapers from the colonial period offer valuable evidence of how racializing language evolved over time. However, there are challenges in studying this type of historical data: 1) Data scarcity: acquiring large, annotated historical datasets is difficult, hindering the possibility of analyzing racialization comprehensively; 2) Digitized materials frequently contain Optical Character Recognition (OCR) errors and other types of noise that complicate text extraction and computational analysis; 3) Colonial newspapers are often multilingual and written in archaic prose, hindering the effectiveness of NLP tools developed for modern, single language texts. This paper addresses these challenges by conducting a dual-view, jointly studying multilingual event extraction and temporal semantic shift tasks. Specifically, we introduce a contextual question answering (CQA) and a visual question answering (VQA) derived from eighteenth- and nineteenth-century colonial newspapers. Content-wise, we focus on how enslaved people were described by enslavers as well as how they articulated their own condition through QA pairs of newspapers written in Dutch, English-French, and Spanish. Our results show that LLMs are still limited for low-resource VQA tasks. For temporal semantic change, we train temporal word embedding with a compass. The study concludes that racialization is a fluid process of linguistic recalibration where the decline of slavery merely shifted the language of control onto new categories of labor and identity.

[Long] A Multilingual Social Bias Benchmark Incorporating Thinking Processes ▶ *Underline Video*

Masahiro Kaneko, Danushka Bollegala and Timothy Baldwin

Large Language Models (LLMs) can learn both useful knowledge and harmful stereotypes, making bias evaluation essential. Existing frameworks fall into two types: those considering reasoning steps (Thinking Process-Aware Evaluation, TPAE) and those focusing only on final outputs (Straight-to-the-Answer Evaluation, SAE). Prior TPAE studies showed effectiveness in assessing gender bias but relied on

template-based, word-counting prompts, limiting generalization to other bias types, languages, and reasoning-based methods. In this study, we introduce MBTP, a multilingual social bias benchmark that incorporates human-generated pro- and anti-stereotype reasoning as part of the thinking process, and propose a few-shot meta-evaluation method that enables scalable bias assessment without model fine-tuning. From experiments evaluating 13 social bias categories across 8 languages, we find that human-generated thinking consistently yields higher-quality evaluations than LLM-generated or template-based approaches. Furthermore, TPAE demonstrates superior performance over SAE, highlighting the importance of considering reasoning processes in bias evaluation. We will release the MBTP dataset upon paper acceptance.

[Long] ACIArena: Toward Unified Evaluation for Agent Cascading Injection ▶ *Underline Video*

Hengyu An, Minxi Li, Jinghuai Zhang, Naen Xu, Chunyi Zhou, Changjiang Li, Xiaogang Xu, Tianyu Du and Shouling Ji

Collaboration and information sharing empower Multi-Agent Systems (MAS) but also introduce a critical security risk known as Agent Cascading Injection (ACI). In such attacks, a compromised agent exploits inter-agent trust to propagate malicious instructions, causing cascading failures across the system. However, existing studies consider only limited attack strategies and simplified MAS settings, limiting their generalizability and comprehensive evaluation. To bridge this gap, we introduce ACIArena, a unified framework for evaluating the robustness of MAS. ACIArena offers systematic evaluation suites spanning multiple attack surfaces (i.e., external inputs, agent profiles, inter-agent messages) and attack objectives (i.e., instruction hijacking, task disruption, information exfiltration). Specifically, ACIArena establishes a unified specification that jointly supports MAS construction and attack-defense modules. It covers six widely used MAS implementations and provides a benchmark of 1,356 test cases for systematically evaluating MAS robustness. Our benchmarking results show that evaluating MAS robustness solely through topology is insufficient; robust MAS require deliberate role design and controlled interaction patterns. Moreover, defenses developed in simplified environments often fail to transfer to real-world settings; narrowly scoped defenses may even introduce new vulnerabilities. ACIArena aims to provide a solid foundation for advancing deeper exploration of MAS design principles.

[Findings] Adaptive Data Collection for Latin-American Community-sourced Evaluation of Stereotypes (LACES) ▶ *Underline Video*

Guido Ivetta, Pietro Palombini, Sofía Martinelli, Marcos J Gomez, M Emilia Echeveste, Sunipa Dev, Vinodkumar Prabhakaran and Luciana Benotti

The evaluation of societal biases in NLP models is critically hindered by a geo-cultural gap. This leaves regions such as Latin America severely underserved, making it impossible to adequately assess or mitigate the perpetuation of harmful regional stereotypes in language technologies. This paper presents LACES, a stereotype association dataset, for 15 Latin American countries. This dataset includes 4,789 stereotype associations, manually created and annotated by 83 participants. The dataset was developed through targeted community partnerships across Latin America. Additionally, in this paper, we propose a novel adaptive data collection methodology that uniquely integrates the sourcing of new stereotype entries and the validation of existing data within a single, unified workflow. This approach results in a resource with more unique stereotypes than previous static collection methods, enabling a more efficient stereotype collection. The paper further supports the quality of LACES by demonstrating reduced efficacy of debiasing methods on this dataset in comparison to existing popular stereotype benchmarks. Content Warning: This research involves the study of social biases. Consequently, the paper contains examples of discriminatory language and stereotypes that may be sensitive or upsetting to readers. These examples are included for the purpose of scientific analysis and do not reflect the views of the authors.

[Long] Adaptive Instruction Composition for Automated LLM Red-Teaming ▶ *Underline Video*

Jesse Zymet, Andy Luo, Swapnil Shinde, Sahil Wadhwa and Emily Chen

Many approaches to LLM red-teaming leverage an attacker LLM to discover jailbreaks against a target. Several of them task the attacker with identifying effective strategies through trial and error, resulting in a semantically limited range of successes. Another approach discovers diverse attacks by combining crowdsourced harmful queries and tactics into instructions for the attacker, but does so at random, limiting effectiveness. This article introduces a novel framework, Adaptive Instruction Composition, that combines crowdsourced texts according to an adaptive mechanism trained to jointly optimize effectiveness with diversity. We use reinforcement learning to balance exploration with exploitation in a combinatorial space of instructions to guide the attacker toward diverse generations tailored to target vulnerabilities. We demonstrate that our approach substantially outperforms random combination on a set of effectiveness and diversity metrics, even under model transfer. Further, we show that it surpasses a host of recent adaptive approaches on Harmbench. We employ a lightweight neural contextual bandit that adapts to contrastive embedding inputs, and provide ablations suggesting that the contrastive pretraining enables the network to rapidly generalize and scale to the massive space as it learns.

[Long] Adaptive Prompt Structure Factorization: A Framework for Self-Discovering and Optimizing Compositional Prompt Programs ▶ *Underline Video*

Haoyue Liu, Zhichao Wang, Yongxin Guo, Haoran Shou and Xiaoying Tang

Automated prompt optimization is crucial for eliciting reliable reasoning from large language models (LLMs), yet most API-only prompt optimizers iteratively edit monolithic prompts, coupling components and obscuring credit assignment, limiting controllability, and wasting tokens. We propose Adaptive Prompt Structure Factorization (aPSF), an API-only framework (prompt-in/text-out; no access to model internals) that uses an Architect model to discover task-specific prompt structures as semantic factors. aPSF then performs interventional, single-factor updates: interventional factor-level scoring estimates each factor's marginal contribution via validation-performance changes,

and error-guided factor selection routes updates to the current dominant failure source for more sample-efficient optimization. Across multiple advanced reasoning benchmarks, aPSF outperforms strong baselines, improving accuracy by up to +4.29 percentage points on average, and reduces optimization cost by 45-87% tokens on MultiArith while reaching peak validation in 1 step.

[Long] ARCHITECT: Uncertainty-Aware Dynamic Tool Learning via Causal Intervention for Open-World Agents ▶ *Underline Video*
Wang zhangyi, JIEXIANG XU, Bingnan Yu and Li Zongze
Dynamic tool generation empowers Large Language Model (LLM) agents to synthesize tools on demand, yet a critical challenge remains: 32.4% of generated tools fail on first invocation. We present Causal Tool Diagnosis (CTD), a principled framework that moves beyond black-box reliability prediction to interpretable failure attribution. CTD constructs a Structural Causal Model (SCM) capturing how specification quality, code characteristics, and execution environment jointly determine tool outcomes. Uniquely leveraging code's intervenability, we conduct controlled sandbox experiments to estimate causal effects—an advantage unavailable in pure text generation. CTD jointly predicts confidence (Spearman rank correlation coefficient =0.90) and root cause attribution (78% accuracy), with attributions directly guiding targeted repairs (+9.6% success rate over error-type classification). Our ARCHITECT framework, integrating CTD throughout the tool lifecycle, achieves state-of-the-art on four benchmarks including StableToolBench (+3.8%), MINT (+4.6%), T-Eval (+3.7%), and SWE-bench Lite (+2.4%), with consistent improvements across all settings.

[Long] Arguments that Alter Minds: LLM Rationales Sway Human (and LLM) Notions of Plausibility ▶ *Underline Video*
Shramay Palta, Peter Rankel, Sarah Wiegreffe and Rachel Rudinger
We investigate the degree to which human (and LLM) plausibility judgments of multiple-choice commonsense benchmark answers are subject to influence by (im)plausibility arguments for or against an answer, in particular, using rationales generated by LLMs. We collect 3,000 plausibility judgments from humans and another 13,600 judgments from LLMs. Overall, we observe increases and decreases in mean human plausibility ratings in the presence of LLM-generated PRO and CON rationales, respectively, suggesting that, on the whole, human judges find these rationales convincing. Experiments with LLMs reveal similar patterns of influence. Our findings demonstrate a novel use of LLMs for studying aspects of human cognition, while also raising practical concerns that, even in domains where humans are "experts" (i.e., common sense), LLMs have the potential to exert considerable influence on people's beliefs.

[Findings] Assessing Y-Axis Influence: Bias in Multimodal Language Models on Chart-to-Table Translation ▶ *Underline Video*
Seok_Hwan Song, Azher Ahmed Efat and Wallapak Tavanapong
Chart-to-table translation converts chart images into structured tabular data. Accurate translation is crucial for Multimodal Language Modal (MLM) to answer complex queries. We observe imbalances in the number of images across different aspects of the y-axis information in public chart datasets. Such imbalances can introduce unintended biases, causing uneven MLM performance. Previous works have not systematically examined these biases. To address this gap, we propose a new framework, FairChart2Table, for analyzing y-axis-related bias on five state-of-the-art models. Key Findings: (1) There are significant y-axis biases related to the digit length of the major tick values, the number of major ticks, the range of values, and the tick value format (e.g., abbreviation or scientific format). (2) The number of legends/entities in chart images impacts MLM performance. (3) Prompting MLM with y-axis information can significantly enhance the performance for some MLMs.

[Findings] Attention Sinks in Diffusion Language Models ▶ *Underline Video*
Maximo Rulli, Simone Petruzzi, Edoardo Michielon, Fabrizio Silvestri, Simone Scardapane and Alessio Devoto
Masked Diffusion Language Models (DLMs) have recently emerged as a promising alternative to traditional Autoregressive Models (ARMs). DLMs employ transformer encoders with bidirectional attention, enabling parallel token generation while maintaining competitive performance. Although their efficiency and effectiveness have been extensively studied, the internal mechanisms that govern DLMs remain largely unexplored. In this work, we conduct an empirical analysis of DLM attention patterns, focusing on the attention sinking phenomenon, an effect previously observed in various transformer-based architectures. Our findings reveal that DLMs also exhibit attention sinks, but with distinct characteristics. First, unlike in ARMs, the sink positions in DLMs tend to shift throughout the generation process, displaying a dynamic behaviour. Second, while ARMs are highly sensitive to the removal of attention sinks, DLMs remain robust: masking sinks leads to only a minor degradation in performance. These results provide new insights into the inner workings of diffusion-based language models and highlight fundamental differences in how they allocate and utilize attention compared to autoregressive models.

[Short] Attention Under Attack: Analog Noise Effects and Mechanistic Vulnerabilities in Transformer Models ▶ *Underline Video*
Mafizur Rahman and Lijun Qian
Analog in-memory computing (AIMC) offers substantial efficiency gains for transformer inference but introduces hardware-induced noise that can distort attention behavior. Prior studies primarily focus on AIMC evaluations for vision tasks and CNN-based models. They largely overlook how hardware-induced noise perturbs internal attention dynamics in NLP models. In this work, we present the first fine-grained analysis of analog vulnerability in pretrained transformers, examining projection submodules, attention heads, and layer-wise dynamics across multiple NLP tasks. Results show that query (Q), key (K), and value (V) projections are the most sensitive components, while feed-forward layers remain comparatively robust. Also, analog noise yields depth-dependent degradation in higher layers, leading to scattered attention and disrupted token routing. This pre-deployment analysis mitigates potential resource misuse before physical deployment and offers practical guidance for designing noise-resilient analog NLP transformers.

[Findings] Audit Me If You Can: Query-Efficient Active Fairness Auditing of Black-Box LLMs ▶ *Underline Video*

David Hartmann, Lena Pohlmann, Lelia Hanslik, Noah GieSSing, Bettina Berendt and Pieter Delobelle

Large Language Models (LLMs) exhibit systematic biases across demographic groups. Auditing is proposed as an accountability tool for black-box LLM applications, but suffers from resource-intensive query access. We conceptualise auditing as uncertainty estimation over a target fairness metric and introduce BAFA, the Bounded Active Fairness Auditor for query-efficient auditing of black-box LLMs. BAFA maintains a version space of surrogate models consistent with queried scores and computes uncertainty intervals for fairness metrics (e.g., AUC) via constrained empirical risk minimisation. Active query selection narrows these intervals to reduce estimation error. We evaluate BAFA on two standard fairness dataset case studies: CivilComments and Bias-in-Bios, comparing against stratified sampling, power sampling, and ablations. BAFA achieves target error thresholds with up to 40x fewer queries than stratified sampling (e.g., 144 vs 5,956 queries at =0.02 for CivilComments) for tight thresholds, demonstrates substantially better performance over time, and shows lower variance across runs. These results suggest that active sampling can reduce resources needed for independent fairness auditing with LLMs, supporting continuous model evaluations.

[Findings] Auto-Stega: An Agent-Driven System for Lifelong Strategy Evolution in LLM-Based Text Steganography ▶ *Underline Video*

Jiuan Zhou, Yu Cheng, Yuan Xie and Zhaoxia Yin

With the rapid progress of LLMs, high quality generative text has become widely available as a cover for text steganography. However, prevailing methods rely on hand-crafted or pre-specified strategies and struggle to balance efficiency, imperceptibility, and security, particularly at high embedding rates. Accordingly, we propose Auto-Stega, an agent-driven self-evolving framework that is the first to realize self-evolving steganographic strategies by automatically discovering, composing, and adapting strategies at inference time; the framework operates as a closed loop of generating, evaluating, summarizing, and updating that continually curates a structured strategy library and adapts across corpora, styles, and task constraints. A decoding LLM recovers the information under the shared strategy. To handle high embedding rates, we introduce PC-DNTE, a plug-and-play algorithm that maintains alignment with the base model's conditional distribution at high embedding rates, preserving imperceptibility while enhancing security. Experimental results demonstrate that at higher embedding rates Auto-Stega achieves superior performance with gains of 42.2% in perplexity and 1.6% in anti-steganalysis performance over SOTA methods.

[Findings] AutoSearch: Adaptive Search Depth for Efficient Agentic RAG via Reinforcement Learning ▶ *Underline Video*

Jingbo Sun, Wenyue Chong, Songjun Tu, Qichao Zhang, Yaocheng Zhang, Jiajun Chai, Xiaohan Wang, Wei Lin, Guojun Yin and Dongbin Zhao

Agentic retrieval-augmented generation (RAG) systems enable large language models (LLMs) to solve complex tasks through multi-step interaction with external retrieval tools. However, such multi-step interaction often involves redundant search steps, incurring substantial computational cost and latency. Prior work limits search depth (i.e., the number of search steps) to reduce cost, but this often leads to underexploration of complex questions. To address this, we first investigate how search depth affects accuracy and find a minimal sufficient search depth that defines an accuracy-efficiency trade-off, jointly determined by question complexity and the agents capability. Furthermore, we propose AutoSearch, a reinforcement learning framework that evaluates each search step via self-generated intermediate answers. By a self-answering mechanism, AutoSearch identifies the minimal sufficient search depth and promotes efficient search by rewarding its attainment while penalizing over-searching. In addition, reward mechanisms are introduced to stabilize search behavior and improve answer quality on complex questions. Extensive experiments on multiple benchmarks show that AutoSearch achieves a superior accuracy-efficiency trade-off, alleviating over-searching while preserving search quality.

[Findings] Balancing Knowledge Breadth and Task Depth for Effective Domain Adaptation Fine-Tuning ▶ *Underline Video*

Mu Zhang, Yuxiang Chu, Guangya Yu, Yongqi Fan, Weiyan Zhang, Hang Hu, Tong Ruan and Jingping Liu

Training large language models for domain adaptation poses a significant challenge in balancing the acquisition of domain knowledge with the retention of general abilities, often leading to catastrophic forgetting. While curriculum learning offers a promising direction, conventional methods typically rely on a single dimension of knowledge or task, which is insufficient to navigate the trade-off between knowledge breadth and task depth. In this paper, we propose a two-dimensional curriculum learning framework that coordinates model training along two orthogonal axes: the knowledge dimension and the task dimension. We first reconstruct the dataset by clustering instances according to their semantic similarity to general-domain data, and subsequently annotate them with a task hierarchy. Then, we design an integrated curriculum that develops from general to domain-specific knowledge clusters, and within each cluster, from lower- to higher-order cognitive tasks. Compared with the second-best method, our method improves accuracy on medical evaluations by 2.49% and on financial evaluations by 1.2%. Ablation and cross-domain experiments further demonstrate our method as a scalable and effective framework for structured domain adaptation in large language model fine-tuning. We have released the code in an anonymous repository at https://github.com/Melo-1017/Balancing-Knowledge-Breadth-and-Task-Depth.

[Findings] Bayesian Active Learning with Gaussian Processes Guided by LLM Relevance Scoring for Dense Passage Retrieval ▶ *Underline Video*

Junyoung Kim, Anton Korikov, Jiazhou Liang, Justin Cui, Yifan Simon Liu, Qianfeng Wen, Mark Zhao and Scott Sanner

While Large Language Models (LLMs) exhibit exceptional zero-shot relevance modeling, their high computational cost necessitates framing

passage retrieval as a budget-constrained global optimization problem. Existing approaches passively rely on first-stage dense retrievers, which leads to two limitations: (1) failing to retrieve relevant passages in semantically distinct clusters, and (2) failing to propagate relevance signals to the broader corpus. To address these limitations, we propose Bayesian Active Learning with Gaussian Processes guided by LLM relevance scoring (BAGEL), a novel framework that propagates sparse LLM relevance signals across the embedding space to guide global exploration. BAGEL models the multimodal relevance distribution across the entire embedding space with a query-specific Gaussian Process (GP) based on LLM relevance scores. Subsequently, it iteratively selects passages for scoring by strategically balancing the exploitation of high-confidence regions with the exploration of uncertain areas. Extensive experiments across four benchmark datasets and two LLM backbones demonstrate that BAGEL effectively explores and captures complex relevance distributions and outperforms LLM reranking methods under the same LLM budget on all four datasets.

[Findings] BCL: Bayesian In-Context Learning Framework for Information Extraction ▶ *Underline Video*

Haoliang Liu, Chengkun Cai, Xu Zhao, Han Zhu, Shizhou Huang, Xinglin Zhang, Tao Chen, Jenq-Neng Hwang, Zhang Huaping and Lei Li

Existing information extraction (IE) tasks increasingly adopt in-context learning (ICL) with large language models. However, current approaches either show inconsistent performance across model scales or lack systematic optimization and generalizability. Building on this, we propose BCL-IE (Bayesian In-Context Learning Framework for Information Extraction), the first optimization framework that uses particle filtering with Bayesian updates to systematically refine label representations across IE tasks. Through four steps, initialization, observation, weight update, and resampling, BCL-IE generalizes to both sequence labeling and relation classification paradigms. Extensive experiments demonstrate substantial improvements over existing approaches (up to 30%), achieving prior performance while other methods either fail to generalize or show limited effectiveness.

[Long] Behavior Knowledge Merge in Reinforced Agentic Models ▶ *Underline Video*

Xiangchi Yuan, Dachuan Shi, Chunhui Zhang, Zheyuan Liu, Shenglong Yao, Soroush Vosoughi and Wenke Lee

Reinforcement learning (RL) is central to post-training, particularly for agentic models that require specialized reasoning behaviors. In this setting, model merging offers a practical mechanism for integrating multiple RL-trained agents from different tasks into a single generalist model. However, existing merging methods are designed for supervised fine-tuning (SFT), and they are suboptimal to preserve task-specific capabilities on RL-trained agentic models. The root is a task-vector mismatch between RL and SFT: on-policy RL induces task vectors that are highly sparse and heterogeneous, whereas SFT-style merging implicitly assumes dense and globally comparable task vectors. When standard global averaging is applied under this mismatch, RL's non-overlapping task vectors that encode critical task-specific behaviors are reduced and parameter updates are diluted. To address this issue, we propose Reinforced Agent Merging (RAM), a distribution-aware merging framework explicitly designed for RL-trained agentic models. RAM disentangles shared and task-specific unique parameter updates, averaging shared components while selectively preserving and rescaling unique ones to counteract parameter update dilution. Experiments across multiple agent domains and model architectures demonstrate that RAM not only surpasses merging baselines, but also unlocks synergistic potential among agents to achieve performance superior to that of specialized agents in their domains.

[Long] Benchmarking and Enabling Efficient Chinese Medical Retrieval via Asymmetric Encoders ▶ *Underline Video*

Angqing Jiang, Jianlyu Chen, Zhefang Zhefang, Yongcan Wang, Xinpeng Li, Keyu Ding and Defu Lian

Effective medical text retrieval requires both high accuracy and low latency. While LLM-based embedding models possess powerful retrieval capabilities, their prohibitive latency and high computational cost limit their application in real-time scenarios. Furthermore, the lack of comprehensive and high-fidelity benchmarks hinders progress in Chinese medical text retrieval. In this work, we introduce the Chinese Medical Text Embedding Benchmark (CMedTEB), a benchmark spanning three kinds of practical embedding tasks: retrieval, reranking, and semantic textual similarity (STS). Distinct from purely automated datasets, CMedTEB is curated via a rigorous multi-LLM voting pipeline validated by clinical experts, ensuring gold-standard label quality while effectively mitigating annotation noise. On this foundation, we propose the Chinese Medical Asymmetric REtriever (CARE), an asymmetric architecture that pairs a lightweight BERT-style encoder for online query encoding with a powerful LLM-based encoder for offline document encoding. However, optimizing such an asymmetric retriever with two structurally different encoders presents distinctive challenges. To address this, we introduce a novel two-stage training strategy that progressively bridges the query and document representations. Extensive experiments demonstrate that CARE surpasses state-of-the-art symmetric models on CMedTEB, achieving superior retrieval performance without increasing inference latency.

[Findings] Better and Worse with Scale: How Contextual Entrainment Diverges with Model Size ▶ *Underline Video*

Dikshant Kukreja, Kshitij Sah, Gautam Gupta, Avinash Anand, Rajiv Ratn Shah, Zhengkui Wang, Aik Beng Ng and Erik Cambria

Larger language models become simultaneously better and worse at handling contextual information, better at ignoring false claims, worse at ignoring irrelevant tokens. We formalize this apparent paradox through the first scaling laws for contextual entrainment, the tendency of models to favor tokens that appeared in context regardless of relevance. Analyzing the Cerebras-GPT (111M-13B) and Pythia (14M-12B) model families, we find entrainment follows predictable power-law scaling, but with opposite trends depending on context type: semantic contexts show decreasing entrainment with scale, while non-semantic contexts show increasing entrainment. Concretely, the largest models are four times more resistant to counterfactual misinformation than the smallest, yet simultaneously twice as prone to copying arbitrary tokens. These diverging trends, which replicate across model families, suggest that semantic filtering and mechanical copying are functionally distinct behaviors that scale in opposition. These opposing trends suggest that scaling alone does not resolve context sensitivity, it reshapes it.

[Findings] Bias Dynamics in BabyLMs: Towards a Compute-Efficient Sandbox for Democratising Pre-Training Debiasing ▶ *Underline Video*

Filip TrhlíkAndrew CainesPaula Buttery

Pre-trained language models (LMs) have, over the last few years, grown substantially in both societal adoption and training costs. This rapid growth in size has constrained progress in understanding and mitigating their biases. Since re-training LMs is prohibitively expensive, most debiasing work has focused on post-hoc or masking-based strategies, which often fail to address the underlying causes of bias. In this work, we seek to democratise pre-model debiasing research by using low-cost proxy models. Specifically, we investigate BabyLMs, compact BERT-like models trained on small and mutable corpora that can approximate bias acquisition and learning dynamics of larger models. We show that BabyLMs display closely aligned patterns of intrinsic bias formation and performance development compared to standard BERT models, despite their drastically reduced size. Furthermore, correlations between BabyLMs and BERT hold across multiple intra-model and post-model debiasing methods. Leveraging these similarities, we conduct pre-model debiasing experiments with BabyLMs, replicating prior findings and presenting new insights regarding the influence of gender imbalance and toxicity on bias formation. Our results demonstrate that BabyLMs can serve as an effective sandbox for large-scale LMs, reducing pre-training costs from over 500 GPU-hours to under 30 GPU-hours. This provides a way to democratise pre-model debiasing research and enables faster, more accessible exploration of methods for building fairer LMs.

[Long] Bridging Internal Consistency and External Alignment: A Causal and Dynamic Interpretability Framework for LLM Generation ▶ *Underline Video*

Shuyao Xiao, Shengling Wang and Ke Chao

Large Language Models (LLMs) are widely used in high-stakes applications, making their interpretability increasingly important. Existing interpretability methods are typically categorized into internal and external perspectives, which are often studied in isolation and tend to overlook two key aspects: causality and temporal dynamics. Explanations are often limited to surface correlations or static dependencies, failing to capture how influences evolve during autoregressive generation. To address these limitations, we propose a causal and dynamic interpretability framework for LLM generation. We first characterize the backdoor-adjusted causal effects of both the generated prefix and the prompt on the current token using the Structural Causal Model. Next, we introduce two metrics to quantify contextual causal influence and question-answer causal influence. Overall, our work provides a unified causal view of internal consistency and external alignment in LLM generation dynamics.

[Long] Building LLMs Like LEGO: Two-dimensional Architecture Reassembly of Large Language Models ▶ *Underline Video*

Xingyu Wu, Yu Zhou and Kay Chen Tan

Pretrained large language models (LLMs) are typically reused as indivisible artifacts, adapted, merged, or ensembled as a whole. In this study, we show that LLMs can instead be structurally recomposed as modular building blocks to create new architectures without access to original training data. We introduce architecture-level reassembly as a new reuse paradigm, in which Transformer blocks from heterogeneous models are treated as reusable components. This idea is formalized through a two-dimensional reassembly space that supports both vertical recombination across depth and horizontal composition within layers. To make this space tractable, we propose a chromosome-based architectural encoding and perform a bi-level multi-objective evolutionary optimization over vertical structure and horizontal composition. To resolve representation incompatibility across heterogeneous blocks, we introduce lightweight glue layers trained via data-free knowledge distillation, enabling valid information flow without modifying pretrained parameters. Our results demonstrate that architecture-level reassembly unlocks a new dimension of flexibility in model reuse, pointing toward a modular and evolutionary view of LLM design.

[Findings] Can Large Language Models Infer Human Actions and Motives? Evaluation in Social Prediction and Inspection Games ▶ *Underline Video*

Kaleen Shrestha, Abhinav Gupta, Harish Dukkipati, Zhonghao Shi and Maja Matari

Humans are able to predict each others actions by reasoning about the others underlying goals, preferences, and motives, such as greed and risk-aversion. Game theory provides a framework for studying human behaviors through incentivized games that simulate social situations. We utilized two validated games from the cognitive science literature, the Social Prediction Game (SPG) and the Inspection Game (IG), to systematically study how well several recent open- and closed-source LLMs predict player actions and whether they can leverage and generalize the players motives learned from the iterated games. Our results indicate that state-of-the-art LLMs can achieve accuracy close to human levels in predicting players actions with underlying human motives in SPGs. However, unlike humans, who rely on reasoning about players motives to inform their predictions, LLMs failed to recognize statistical patterns in players actions. As a result, LLM prediction accuracy did not improve over multiple rounds. Our results in the IG further demonstrate that, unlike humans, LLMs were unable to recognize a players underlying motives and to generalize their understanding of the same player to a new context. This suggests that LLMs may lack reasoning capabilities. Our findings offer insights into differences in human and LLM reasoning mechanisms, suggesting that further research into human-AI alignment is needed before utilizing LLMs for human behavior modeling and simulation in this and related contexts.

[Long] CaRL-EM: Cost-Aware Reinforcement Learning for Entity Matching with LLMs ▶ *Underline Video*

Chaohui Guo, Michel Klein and Zhisheng Huang

Entity matching (EM) requires fine-grained contextual understanding and domain knowledge. Recent work shows that large language models (LLMs) can serve as strong matchers across domains, but most methods either make independent pairwise decisions or rely on manually designed composite pipelines, thus lacking flexibility in realistic multi-candidate settings. At the same time, they typically ignore inference cost at scale. We formulate LLM-based EM with candidates as a cost-aware sequential decision problem and propose CaRL-EM, a reinforcement learning controller that manages LLM operations. Given the state of an anchor record, its candidate set, and the cost, CaRL-EM adaptively chooses among different operators (Match/Compare/Select/Decide) and model capacities to maximize a quality-cost objective. The policy interacts with abstract operators, allowing the same controller to be reused with different underlying LLM backends at inference time without retraining. Experiments on 7 benchmarks show that CaRL-EM (i) learns to dynamically plan the usage of inexpensive and expensive operators based on task complexity, (ii) achieves robust zero-shot transfer across diverse datasets and domains, and (iii) consistently achieves a better quality-cost trade-off than strong LLM-based baselines and manually designed pipelines, yielding a lower inference cost at comparable or higher quality.

[Long] Chaining the Evidence: Robust Reinforcement Learning for Deep Search Agents with Citation-Aware Rubric Rewards ▶ *Underline Video*

Jiajie Zhang, Xin Lv, Ling Feng, Lei Hou and Juanzi Li

Reinforcement learning (RL) has emerged as a critical technique for enhancing LLM-based deep search agents. However, existing approaches primarily rely on binary outcome rewards, which fail to capture the comprehensiveness and factuality of agents' reasoning process, and often lead to undesirable behaviors such as shortcut exploitation and hallucinations. To address these limitations, we propose Citation-aware Rubric Rewards (CaRR), a fine-grained reward framework for deep search agents that emphasizes reasoning comprehensiveness, factual grounding, and evidence connectivity. CaRR decomposes complex questions into verifiable single-hop rubrics and requires agents to satisfy these rubrics by explicitly identifying hidden entities, supporting them with correct citations, and constructing complete evidence chains that link to the predicted answer. We further introduce Citation-aware Group Relative Policy Optimization (C-GRPO), which combines CaRR and outcome rewards for training robust deep search agents. Experiments show that C-GRPO consistently outperforms standard outcome-based RL baselines across multiple deep search benchmarks. Our analysis also validates that C-GRPO effectively discourages shortcut exploitation, promotes comprehensive, evidence-grounded reasoning, and exhibits strong generalization to open-ended deep research tasks.

[Findings] Characterizing Web Search in The Age of Generative AI ▶ *Underline Video*

Elisabeth Kirsten, Jost GroSSe Perdekamp, Qinyuan Wu, Mihir Upadhyay, Krishna P. Gummadi and Muhammad Bilal Zafar

The advent of LLMs has given rise to generative search, a new search paradigm in which LLMs retrieve information from the web related to a query and synthesize it into a single, coherent response. This paradigm differs fundamentally from traditional web search, where results are returned as a ranked list of independent web pages. In this paper, we ask: Along what dimensions does generative search differ from traditional search? We conduct a systematic comparison between Google organic search and five generative search systems from three providers: Google, OpenAI, and Perplexity. Our analysis reveals substantial variation among engines in their reliance on internal v.s. external knowledge, source diversity, and stability. While generative systems often achieve topical coverage comparable to traditional search, they do so using markedly different retrieval footprints and synthesis strategies. We further show that the outputs of generative search can vary across time and executions, raising new challenges for robustness. Our findings demonstrate that generative search introduces new dimensions that are not captured by existing evaluation paradigms, motivating the development of evaluations that explicitly account for retrieval behavior, synthesis, and stability in generative search systems.

[Long] CIG: Measuring Conversational Information Gain in Deliberative Dialogues with Semantic Memory Dynamics ▶ *Underline Video*

Ming-Bin Chen, Jey Han Lau and Lea Frermann

Measuring the quality of public deliberation requires evaluating not only civility or argument structure, but also the informational progress of a conversation. We introduce a framework for Conversational Information Gain (CIG) that evaluates each utterance in terms of how it advances collective understanding of the target topic. To operationalize CIG, we model an evolving semantic memory of the discussion: the system extracts atomic claims from utterances and incrementally consolidates them into a structured memory state. Using this memory, we score each utterance along three interpretable dimensions: Novelty, Relevance, and Implication Scope. We annotate 80 segments from two moderated deliberative settings (TV debates and community discussions) with these dimensions and show that memory-derived dynamics (e.g., the number of claim updates) correlate more strongly with human-perceived CIG than traditional heuristics such as utterance length or TF–IDF. We develop effective LLM-based CIG predictors paving the way for information-focused conversation quality analysis in dialogues and deliberative success.

[Long] Closing the Spatial Execution Gap in Digital Whiteboards via Verifiable Reinforcement Learning ▶ *Underline Video*

Chang Liu, Benjamin Wagley, Zibo Wang, Mehmet Belviranli and Bo Wu

While multi-modal large language models such as GPT-5 demonstrate exceptional general understanding, they suffer from a fundamental Spatial Execution Gap, failing to translate visual semantics into precise, schema-valid coordinate operations in interactive environments. In this work, we show that model scale alone cannot close this gap; instead, verifiable structured reasoning provides the key to spatial precision. We present a comprehensive pipeline that leverages Group Relative Policy Optimization to enforce a strict Identify-Reason-Verify protocol, effectively shifting the computational burden from parameters to test-time reasoning. By utilizing a multi-agent system to distill optimal

reasoning schemas and training on execution-verifiable rewards, our specialized 3B agent achieves 100% format coherence and 81.12% operation accuracy on digital whiteboard tasks. Crucially, our approach outperforms a state-of-the-art frontier model, GPT-5, by 16.75% in operation accuracy. The results suggest that for complex user interface manipulation, small, RL-aligned models with dedicated reasoning protocols are superior to generalist frontier models, offering a promising direction for building reliable web agents.

[Findings] Code-Switching Information Retrieval: Benchmarks, Analysis, and the Limits of Current Retrievers ▶ *Underline Video*

Qingcheng Zeng, Yuheng Lu, Zeqi Zhou, Heli Qi, Puxuan Yu, Fuheng Zhao, Hitomi Yanaka, Weihao Xuan and Naoto Yokoya

Code-switching is a pervasive linguistic phenomenon in global communication, yet modern information retrieval systems remain predominantly designed for, and evaluated within, monolingual contexts. To bridge this critical disconnect, we present a holistic study dedicated to code-switching IR. We introduce CSR-L (Code-Switching Retrieval benchmark-Lite), constructing a dataset via human annotation to capture the authentic naturalness of mixed-language queries. Our evaluation across statistical, dense, and late-interaction paradigms reveals that code-switching acts as a fundamental performance bottleneck, degrading the effectiveness of even robust multilingual models. We demonstrate that this failure stems from substantial divergence in the embedding space between pure and code-switched text. Scaling this investigation, we propose CS-MTEB, a comprehensive benchmark covering 11 diverse tasks, where we observe performance declines of up to 27%. Finally, we show that standard multilingual techniques like vocabulary expansion are insufficient to resolve these deficits completely. These findings underscore the fragility of current systems and establish code-switching as a crucial frontier for future IR optimization.

[Findings] Confidence Estimation for LLMs in Multi-turn Interactions ▶ *Underline Video*

Caiqi Zhang, Ruihan Yang, Xiaochen Zhu, Chengzu Li, Tiancheng Hu, Yijiang River Dong, Deqing Yang and Nigel Collier

While confidence estimation is a promising direction for mitigating hallucinations in Large Language Models (LLMs), current research overwhelmingly focuses on single-turn settings. The dynamics of model confidence in multi-turn conversations, where context accumulates and ambiguity is progressively resolved, remain largely unexplored. This work presents the first systematic study of confidence estimation in multi-turn interactions, establishing a formal evaluation framework grounded in two key desiderata: per-turn calibration and monotonicity of confidence as more information becomes available. To facilitate this, we introduce novel metrics, including a length-normalized Expected Calibration Error (InfoECE), and a new "Hinter-Guesser" paradigm for generating controlled evaluation datasets. Our experiments reveal that widely-used confidence techniques struggle with calibration and monotonicity in multi-turn dialogues. In contrast, a novel logit-based probe we introduce, P(Sufficient), proves comparatively more effective, robustly tracking evidence accumulation and distinguishing it from conversational filler. Our work provides a foundational methodology for developing more reliable and trustworthy conversational agents.

[Findings] Confidence over Time: Confidence Calibration with Temporal Logic for Large Language Model Reasoning ▶ *Underline Video*

Zhenjiang Mao, Anirudhh Venkat, Artem Bisliouk, Sindhura Kumbakonam Subramanian, Akshat Kothiyal, Saithej Singhu and Ivan Ruchkin

Large Language Models (LLMs) increasingly rely on long-form, multi-step reasoning to solve complex tasks such as mathematical problem solving and scientific question answering. Despite strong performance, existing confidence estimation methods typically reduce an entire reasoning process to a single scalar score, ignoring how confidence evolves throughout the generation. As a result, these methods are often sensitive to superficial factors such as response length or verbosity, and struggle to distinguish correct reasoning from confidently stated errors. We propose to characterize the stepwise confidence signal using Signal Temporal Logic (STL). Using a discriminative STL mining procedure, we discover temporal formulas that distinguish confidence signals of correct and incorrect responses. Our analysis found that the STL patterns generalize across tasks, and numeric parameters exhibit sensitivity to individual questions. Based on these insights, we develop a confidence estimation approach that informs STL blocks with parameter hypernetworks. Experiments on multiple reasoning tasks show our confidence scores are more calibrated than the baselines.

[Findings] Creating ConLangs to Probe the Metalinguistic Grammatical Knowledge of LLMs ▶ *Underline Video*

Chihiro Taguchi and Richard Sproat

We present a system that uses LLMs as a tool in the development of Constructed Languages, ConLangs, which we call IASC (Interactive Agentic System for ConLangs). The system is modular in that it creates each of the components, phonology, morphology and syntax, lexicon, orthography, and grammatical handbook, using module-specific sets of prompts. The approach is agentic in that various modules allow for refining the output given automatically-generated commentary on a previous step. Our main goals are twofold. First, we aim to provide tools that facilitate an engaging and enjoyable experience in creating artificially constructed languages. Second, the focus of this paper is on using our ConLang framework as a novel way to explore what LLMs know about language, not what they know about any particular language or encyclopedic facts, but how much they know about and understand language and linguistic concepts. In the experiments, we particularly focus on the morphosyntax module and show that there is a fairly wide gulf in capabilities both among different LLMs and among different linguistic specifications, with it being notably easier for systems to deal with more typologically common patterns than rarer ones. All code is released: https://github.com/SakanaAI/IASC.

[Long] CUB: Benchmarking Context Utilisation Techniques for Language Models ▶ *Underline Video*

Lovisa Hagström, Youna Kim, Haeun Yu, Sang-goo Lee, Richard Johansson, Hyunsoo Cho and Isabelle Augenstein

Incorporating external knowledge is crucial for knowledge-intensive tasks, such as question answering and fact checking. However, language models (LMs) may ignore relevant information that contradicts outdated parametric memory or be distracted by irrelevant contexts. While

many context utilisation manipulation techniques (CMTs) have recently been proposed to alleviate these issues, few have seen systematic comparison. In this paper, we develop CUB (Context Utilisation Benchmark) - the first comprehensive benchmark designed to help diagnose CMTs under diverse noisy context conditions within retrieval-augmented generation (RAG). With this benchmark, we conduct the most extensive evaluation to date of seven state-of-the-art methods, representative of the main categories of CMTs, across three diverse datasets and tasks, applied to 11 LMs. Our findings expose critical gaps in current CMT evaluation practices, demonstrating the need for holistic testing. We reveal that most existing CMTs struggle to handle the full spectrum of context types encountered in real-world RAG scenarios. We also find that many CMTs display inflated performance on simple synthesised datasets, compared to more realistic datasets with naturally occurring samples.

[Long] CURE: Critique-Driven Unified Reinforcement Learning for Test-Time Self-Improvement ▶ ***Underline Video***

Guirong Chen, Shuqi Ye, Wenkai Yang, Shiqi Shen, Guangyao Shen and Yankai Lin

The evolution paradigm of Large Language Models (LLMs) is shifting from scaling training compute to scaling inference-time compute. While Reinforcement Learning with Verifiable Rewards (RLVR) has become a key engine for this transition, standard approaches often fail to equip models with the autonomous improvement capabilities required for test-time scaling. Existing critique-guided methods attempt to mitigate this by leveraging external feedback or ground-truth signals; however, these dependencies are unavailable at test time, fundamentally limiting the model's capacity for continuous self-improvement. To bridge this gap, we propose CURE (Critique-driven Unified REinforcement Learning), a framework that jointly optimizes a single policy for standard solving, critiquing, and guided re-exploration. Uniquely, CURE facilitates re-exploration by generating strategic hints while discarding initial incorrect solutions to mitigate anchoring bias. Empirical results across diverse mathematical reasoning and code generation benchmarks demonstrate that CURE not only maintains competitive single-turn performance but, more importantly, unlocks effective inference-time scaling, enabling the model to significantly boost accuracy through iterative self-improvement.

[Long] Current Agents Fail to Leverage World Model as Tool for Foresight ▶ ***Underline Video***

Cheng Qian, Emre Can Acikgoz, Bingxuan Li, Xiusi Chen, Yuji Zhang, Bingxiang He, Qinyu Luo, Gokhan Tur, Dilek Hakkani-Tür, Yunzhu Li and Heng Ji

Agents built on vision-language models increasingly face tasks that demand anticipating future states rather than relying on short-horizon reasoning. Generative world models offer a promising remedy: agents could use them as external simulators to foresee outcomes before acting. This paper empirically examines whether current agents can leverage such world models as tools to enhance their cognition. Across diverse agentic and visual question answering tasks, we observe that some agents rarely invoke simulation (fewer than 1%), frequently misuse predicted rollouts (approximately 15%), and often exhibit inconsistent or even degraded performance (up to 5%) when simulation is available or enforced. Attribution analysis further indicates that the primary bottleneck lies in the agents capacity to decide when to simulate, how to interpret predicted outcomes, and how to integrate foresight into downstream reasoning. These findings underscore the need for mechanisms that foster calibrated, strategic interaction with world models, paving the way toward more reliable anticipatory cognition in future agent systems.

[Long] Data Mixing Agent: Learning to Re-weight Domains for Continual Pre-training ▶ ***Underline Video***

Kailai Yang, Xiao Liu, Lei Ji, Hao Li, Xiao Liang, Zhiwei Liu, Yeyun Gong, Peng Cheng and Mao Yang

Continual pre-training on small-scale task-specific data is an effective method for improving large language models in new target fields, yet it risks catastrophic forgetting of their original capabilities. A common solution is to re-weight training data mixtures from source and target fields on a domain space to achieve balanced performance. Previous domain reweighting strategies rely on manual designation with certain heuristics based on human intuition or empirical results. In this work, we prove that more general heuristics can be parameterized by proposing Data Mixing Agent, the first model-based, end-to-end framework that learns to re-weight domains. The agent learns generalizable heuristics through reinforcement learning on large quantities of data mixing trajectories with corresponding feedback from an evaluation environment. Experiments in continual pre-training on math reasoning show that Data Mixing Agent outperforms strong baselines in achieving balanced performance across source and target field benchmarks. Furthermore, it generalizes well across unseen source fields, target models, and domain spaces without retraining. Direct application to the code generation field also indicates its adaptability across target domains. Further analysis showcases the agents' well-aligned heuristics with human intuitions and their efficiency in achieving superior model performance with less source-field data.

[Findings] Datamart-Agent: LLM-Driven Game-Theoretic Agent for Data Marketplace Modeling ▶ ***Underline Video***

Pangjing Wu, Peter Q. Chen, Xiaodong Li, Wenqi Fan and Li Qing

Data marketplaces analyze strategic data exchanges among users, platforms, and buyers. However, most existing studies model static equilibria and complete information, which limits their realism. In this work, we study whether large language model (LLM)-driven agents can make equilibrium-consistent decisions in analytically tractable data marketplaces with evolving and incomplete-information. Specifically, we introduce EvoDM, an agent-based modeling framework that extends the static data marketplace to dynamic and incomplete-information settings while providing tractable equilibrium benchmarks for evaluating agent decisions. Building upon EvoDM, we propose Datamart-Agent, an LLM-driven game-theoretic agent that improves equilibrium-consistent decision execution through dynamic game tree memory and mechanism-guided reflection, without requiring parameter updates. Experiments demonstrate that Datamart-Agent closely matches equilibrium-consistent decision-making, achieving the lowest utility gap and over 20% higher Pass@epsilon than strong baselines.

After validating its effectiveness, we employ EvoDM with Datamart-Agent to analyze competition and regulation in assumption-relaxed settings where closed-form ground truth is unavailable, providing exploratory simulation-based insights into market dynamics and regulatory effects.

[Findings] Diagnosing LLMs via Information Spectrum Analysis: Tail Behavior and the Effects of Side Information ▶ *Underline Video*

Yuuki Tachioka

Large language models (LLMs) exhibit non-stationary generation: their output distributions shift with prompts, retrieved documents, and decoding conditions. Under such variability, average likelihood metrics can obscure heterogeneous behaviors across samples, especially in high-surprisal tails where failures often occur. We propose an information-spectrum-based diagnostic framework that treats LLMs as general sources without assuming stationarity, ergodicity, or the asymptotic equipartition property. We define sequence-level self-information density (coding rate; mean surprisal) and construct an empirical information spectrum from finite samples, enabling operational estimates of spectrum quantiles and width. We further introduce an information gain spectrum, a teacher-forced likelihood-based measure that evaluates the same generated sequence with and without side information. Across multiple Japanese LLMs and QA settings, we observe that correctness differences are often more visible in the high-surprisal tail than in the mean coding rate, and that side information can reshape tail behavior in heterogeneous ways across sequences. We also observe that instruction tuning changes the spectrum structure, making tail statistics and spectrum width more predictive of correctness than the mean coding rate. Overall, our analysis illustrates how spectrum-based diagnostics complement average-based metrics for understanding conditional generation.

[Long] DialDefer: A Framework for Detecting and Mitigating LLM Dialogic Deference ▶ *Underline Video*

Parisa Rabbani, Priyam Sahoo, Ruben Mathew, Aishee Mondal, Harshita Ketharaman, Nimet Beyza Bozdag and Dilek Hakkani-Tür

LLMs are increasingly used as third-party judges, yet their reliability when evaluating speakers in dialogue remains poorly understood. We show that LLMs judge identical claims differently depending on framing: the same content receives different verdicts when presented as a statement to verify ("Is this statement correct?") versus attributed to a speaker ("Is this speaker correct?"). We call this dialogic deference and introduce DialDefer, a framework for detecting and mitigating these framing-induced judgment shifts. Our Dialogic Deference Score (DDS) captures directional shifts that aggregate accuracy obscures. Across ten domains, 3k+ instances, and five models, conversational framing induces large shifts (mean |DDS| = 15.9 percentage points (pp) across models, $p < .0001$) while accuracy remains stable (<2 pp), with effects amplifying 2-5x on naturalistic Reddit conversations. This effect is domain-dependent: a single model can shift toward disagreement (skepticism) on graduate-level science and toward agreement (deference) on social judgment. Ablations reveal that human-vs-LLM attribution drives the largest shifts (17.7 pp swing), suggesting models treat disagreement with humans as more costly than with AI. Mitigation attempts can reduce deference but over-correct into skepticism, revealing a calibration problem beyond accuracy optimization.

[Findings] Disentangling Continued Pre-Training: Attention-Driven Routing and Semantic Hub Preservation in Language Adaptation ▶ *Underline Video*

Khanh-Tung Tran, Vinh-Khanh Tran', "Barry O'Sullivan" and 'Hoang D. Nguyen

Continued Pre-Training (CPT) enables Large Language Models (LLMs) to acquire second-language capabilities, yet the underlying mechanisms remain poorly understood. In this work, we investigate how CPT adapts model representations across diverse language families and scripts, model sizes, and architectures. We find that second-language abilities emerge through a selective adaptation mechanism: task-solving capabilities are preserved in "semantic hub", while interface layers retarget to shifted token distributions. Layer-swapping experiments demonstrate that semantic understanding can be surgically transferred between base and CPT models with minimal loss (e.g., swapping 50% of model parameters reduces performance by only 0.3%). Furthermore, we establish that attention components route language adaptation: larger parameter changes than feedforward networks, correlate more strongly with language-specific neurons, and their surgical replacement substantially degrades performance. Overall, our work provides a mechanistic understanding of CPT, guiding future work on efficient strategies for language adaptation.

[Findings] Do Language Models Use Logophoric Cues? Evidence from Mandarin Chinese Long-Distance Reflexive ▶ *Underline Video*

Yunfang Dong

Resolving anaphora requires integrating syntactic, semantic, and discourse information. Mandarin Chinese offers a particularly revealing case through the reflexive ziji, whose interpretation permits long-distance binding licensed by logophoric cues (i.e., cues relevant to discourse perspective). While these cues have been extensively studied in linguistic theory and psycholinguistic experiments, it remains an open question to what extent such cues are captured by computational models. We investigate this question by probing large language models' sensitivity to four logophoric cues known to license long-distance binding of ziji: predicate type, perspective marking, discourse topicality, and discourse relation. Using minimal pairs and surprisal-based measures, we assess whether models exhibit systematic biases toward non-local antecedents in logophoric contexts. Across two model families, we find that (i) models exhibit above-chance sensitivity to all four cues; (ii) lexically anchored cues are more robustly captured than discourse-level cues; and (iii) some cues generalize cross-lingually, whereas others appear to depend on language-specific training data. Taken together, these findings provide non-English evidence that large language models capture certain aspects of logophoricity, yet continue to struggle with discourse-level representations that are central to human anaphora resolution. Code and data are available at: https://github.com/yunfang-dong/mandarin-logophoricity-llm

[Long] Do LLMs Encode Functional Importance of Reasoning Tokens? ▶ *Underline Video*

Janvijay Singh and Dilek Hakkani-Tür

Large language models solve complex tasks by generating long reasoning chains, achieving higher accuracy at the cost of increased computational cost and reduced ability to isolate functionally relevant reasoning. Prior work on compact reasoning shortens such chains through probabilistic sampling, heuristics, or supervision from frontier models, but offers limited insight into whether models internally encode token-level functional importance for answer generation. We address this gap diagnostically and propose greedy pruning, a likelihood-preserving deletion procedure that iteratively removes reasoning tokens whose removal minimally degrades model likelihood under a specified objective, yielding length-controlled reasoning chains. We evaluate pruned reasoning in a distillation framework and show that students trained on pruned chains outperform a frontier-model-supervised compression baseline at matched reasoning lengths. Finally, our analysis reveals systematic pruning patterns and shows that attention scores can predict greedy pruning ranks, further suggesting that models encode a nontrivial functional importance structure over reasoning tokens.

[Long] Do LLMs Really Need 10+ Thoughts for Find the Time 1000 Days Later? Towards Structural Understanding of LLM Overthinking ▶ *Underline Video*

Xinliang Frederick Zhang, Anhad Mohananey, Alexandra Chronopoulou, Pinelopi Papalampidi, Somit Gupta, Tsendsuren Munkhdalai, Lu Wang and Shyam Upadhyay

Models employing long chain-of-thought (CoT) reasoning have shown superior performance on complex reasoning tasks. Yet, this capability introduces a critical and often overlooked inefficiency—overthinking—models often engage in unnecessarily extensive reasoning even for simple queries, incurring significant computations without accuracy improvements. While prior work has explored solutions to mitigate overthinking, a fundamental gap remains in our understanding of its underlying causes. Most existing analyses are limited to superficial, profiling-based observations, failing to delve into LLMs' inner workings. This study introduces a systematic, fine-grained analyzer of LLMs' thought process to bridge the gap, TRACE. We first benchmark the overthinking issue, confirming that long-thinking models are five to twenty times slower on simple tasks with no substantial gains. We then use TRACE to first decompose the thought process into minimally complete sub-thoughts. Next, by inferring discourse relationships among sub-thoughts, we construct granular thought progression graphs and subsequently identify common thinking patterns for topically similar queries. Our analysis reveals two major patterns for open-weight thinking models—Explorer and Late Landing. This finding provides evidence that over-verification and over-exploration are the primary drivers of overthinking in LLMs. Grounded in thought structures, we propose a utility-based definition of overthinking, which moves beyond length-based metrics. This revised definition offers a more insightful understanding of LLMs' thought progression, as well as practical guidelines for principled overthinking management.

[Findings] Do VLMs Have a Moral Backbone? A Study on the Fragile Morality of Vision-Language Models ▶ *Underline Video*

Zhining Liu, Tianyi Wang, Xiao Lin, Penghao Ouyang, Gaotang Li, Ze Yang, Hui Liu, Sumit Keswani, Vishwa Pardeshi, Huijun Zhao, Wei Fan and Hanghang Tong

Despite substantial efforts toward improving the moral alignment of Vision-Language Models (VLMs), it remains unclear whether their ethical judgments are stable in realistic settings. This work studies moral robustness in VLMs, defined as the ability to preserve moral judgments under textual and visual perturbations that do not alter the underlying moral context. We systematically probe VLMs with a diverse set of model-agnostic multimodal perturbations and find that their moral stances are highly fragile, frequently flipping under simple manipulations. Our analysis reveals systematic vulnerabilities across perturbation types, moral domains, and model scales, including a sycophancy trade-off where stronger instruction-following models are more susceptible to persuasion. We further show that lightweight inference-time interventions can partially restore moral stability. These results demonstrate that moral alignment alone is insufficient and that moral robustness is a necessary criterion for the responsible deployment of VLMs.

[Long] DRInQ: Evaluating Conversational Implicature with Controlled Context Variation ▶ *Underline Video*

Hirona Arai and Xiang Ren

Human conversation relies heavily on conversational implicature, in which speakers convey meanings that are suggested rather than explicitly stated. Although recent large language models (LLMs) exhibit strong conversational fluency, they remain unreliable when interpretation depends on reasoning that integrates social and contextual cues, a process rarely articulated in text. We introduce DRinQ, a benchmark for evaluating pragmatic reasoning about conversational implicature in question utterances, designed to isolate pragmatic variation while holding each questions surface form fixed. To support scalable evaluation, we propose a semi-automated pipeline that produces question-context-interpretation instances with systematic variation. Across evaluations, we find a consistent generation-inference asymmetry: while state-of-the-art models can generate plausible pragmatic scenarios when guided, they often fail to recover the intended implication at inference time. For smaller models, structured prompting improves alignment with human judgments. A comparative writing study further reveals complementary strengths: human authors tend to produce safer, predictable contexts, whereas models generate varied scenarios with interpretations that sometimes exceed contextual support. These findings highlight persistent challenges in modeling conversational implicature and motivate more context-sensitive evaluation frameworks.

[Findings] EAIR: Entity-aware Inference-Time Knowledge Routing for Multi-Hop Knowledge Editing ▶ *Underline Video*

Jungyu Lee, Kunhui Lee, Gyun Lee and Seung-Hoon Na

Existing in-context editing (ICE) methods for multi-hop knowledge editing commonly suffer from paraphrase sensitivity, which refers to

the phenomenon where these methods are not sufficiently robust to paraphrased multi-hop questions. To improve retrieval accuracy and knowledge routing to address paraphrase sensitivity, this paper proposes a novel entity-aware inference-time knowledge routing method, referred to as EAIR, which consists of four major steps: 1) Entity-referential query decomposition, which decomposes the original question into multiple entity-referential sub-question instructions; 2) Entity-aware retrieval, which leverages the previously reference-resolved topic entity in the retrieval step; 3) Evidence-conditioned contrastive decoding, which discourages the model from relying on its parametric knowledge and steers the model toward following retrieved edits; 4) Reflection-based knowledge routing, which additionally filters decoding results using refusal-style reflection to mitigate the risk introduced by contrastive decoding. Experimental results across the MQuAKE benchmark family and model scales show that EAIR achieves the highest strict case accuracy in 11 of 12 settings, substantially reducing paraphrase sensitivity.

[Long] EconProver: Towards More Economical Test-Time Scaling for Automated Theorem Proving ▶ *Underline Video*

Mukai Li, Linfeng Song, Zhenwen Liang, Jiahao Xu, Shansan Gong, Qi Liu, Haitao Mi and Dong Yu

Large Language Models (LLMs) have recently advanced the field of Automated Theorem Proving (ATP), attaining substantial performance gains through widely adopted test-time scaling strategies, notably reflective Chain-of-Thought (CoT) reasoning and increased sampling passes. However, they both introduce significant computational overhead for inference. Moreover, existing cost analyses typically regulate only the number of sampling passes, while neglecting the substantial disparities in sampling costs introduced by different scaling strategies. In this paper, we systematically compare the efficiency of different test-time scaling strategies for ATP models and demonstrate the inefficiency of the current state-of-the-art (SOTA) open-source approaches. We then investigate approaches to significantly reduce token usage and sample passes while maintaining the original performance. Specifically, we propose two complementary methods that can be integrated into a unified EconRL pipeline for amplified benefits: (1) a dynamic Chain-of-Thought (CoT) switching mechanism designed to mitigate unnecessary token consumption, and (2) Diverse parallel-scaled reinforcement learning (RL) with trainable prefixes to enhance pass rates under constrained sampling passes. Experiments on miniF2F and ProofNet demonstrate that our EconProver-GD achieves comparable performance to baseline methods with only 12% of the computational cost. This work provides actionable insights for deploying lightweight ATP models without sacrificing performance.

[Long] Efficient Self-Evaluation for Diffusion Language Models via Sequence Regeneration ▶ *Underline Video*

Linhao Zhong, Linyu Wu, Wen Wang, Yuling Xi, Chenchen Jing, Jiaheng Zhang, Hao Chen and Chunhua Shen

Diffusion large language models (dLLMs) have recently attracted significant attention for their ability to enhance diversity, controllability, and parallelism. However, their non-sequential, bidirectionally masked generation makes quality assessment difficult, underscoring the need for effective self-evaluation. In this work, we propose DiSE, a simple yet effective self-evaluation confidence quantification method for dLLMs. DiSE quantifies confidence by computing the probability of regenerating the tokens in the entire generated sequence, given the full context. This method enables more efficient and reliable quality assessment by leveraging token regeneration probabilities, facilitating both likelihood estimation and robust uncertainty quantification. Building upon DiSE, we further introduce a flexible-length generation framework, which adaptively controls the sequence length based on the model's self-assessment of its own output. We analyze and validate the feasibility of DiSE from the perspective of dLLM generalization, and empirically demonstrate that DiSE is positively correlated with both semantic coherence and answer accuracy. Extensive experiments on likelihood evaluation, uncertainty quantification, and flexible-length generation further confirm the effectiveness of the proposed DiSE.

[Findings] Emergence and Localisation of Semantic Role Circuits in LLMs ▶ *Underline Video*

Nura Aljaafari, Danilo Carvalho and Andre Freitas

Despite displaying semantic competence, large language models' internal mechanisms that ground abstract semantic structure remain insufficiently characterised. To investigate whether and how LLMs develop causally functional representations of semantic roles, we introduce a causal-temporal methodology combining contrastive minimal pairs, edge-attribution circuit discovery, and training-time tracking. Our analysis reveals that LLMs encode semantic roles through highly localised circuits (89-92% attribution within 28 nodes) that emerge gradually via structural refinement rather than phase transitions. These circuits exhibit moderate cross-scale conservation (24-51% component overlap) alongside high spectral similarity, with larger models reusing similar components while rewiring connections. These findings suggest that LLMs form compact, causally isolated mechanisms for abstract semantic structure that exhibit partial transfer across scales and architectures.

[Long] Emergent Misalignment via In-Context Learning: Narrow in-context examples can produce broadly misaligned LLMs ▶ *Underline Video*

Nikita Afonin, Nikita Andriianov, Vahagn Hovhannisyan, Nikhil Bageshpura, Kyle Liu, Kevin Zhu, Sunishchal Dev, Ashwinee Panda, Oleg Rogov, Elena Tutubalina, Alexander Panchenko and Mikhail Seleznyov

Recent work has shown that narrow finetuning can produce broadly misaligned LLMs, a phenomenon termed emergent misalignment (EM). While concerning, these findings were limited to finetuning and activation steering, leaving out in-context learning (ICL). We therefore ask: does EM emerge in ICL? We find that it does: across four model families (Gemini, Kimi-K2, Grok, and Qwen), narrow in-context examples cause models to produce misaligned responses to benign, unrelated queries. With 16 in-context examples, EM rates range from 1% to 24% depending on model and domain, appearing with as few as 2 examples. Neither larger model scale nor explicit reasoning provides reliable protection, and larger models are typically even more susceptible. Next, we formulate and test a hypothesis, which explains in-context EM

as conflict between safety objectives and context-following behavior. Consistent with this, instructing models to prioritize safety reduces EM while prioritizing context-following increases it. These findings establish ICL as a previously underappreciated vector for emergent misalignment that resists simple scaling-based solutions.

[Long] EmoS: A HighFidelity Multimodal Benchmark for Fine-grained Streaming Emotional Understanding ▶ *Underline Video*

Pengze Guo, Jingxi Liang, Zhiwen Xie, Qifeng Wang and Derek F. Wong

In the context of today's high-pressure, aging society, the demand for large-scale emotional models capable of providing empathetic support is more critical than ever. However, existing benchmarks fail to simultaneously achieve ecological validity, signal clarity, and reliable fine-grained labeling. We introduce EmoS, a high-fidelity bilingual benchmark designed to resolve the limitations of ecological validity and noise in existing datasets by combining strictly filtered static slices with a dynamic Streaming Monologue subset. Supported by a rigorous dual-layer human annotation pipeline, EmoS provides trusted ground truth that captures continuous emotional evolution. Empirical results show that fine-tuning MLLMs (multimodal large language models) on EmoS yields significant gains over zero-shot baselines, laying the foundation for the training and evaluation of future emotion recognition models and empathy models. The dataset and code are publicly available at https://github.com/NLP2CT/EmoS.

[Findings] EviReport: From Reasoned Outlines to Evidence Tracked Long-Form Reports ▶ *Underline Video*

Zihan Liu, Jianhui li, Zexin Wang, Fei Sun, JINGJING LI, Zheyuan Li, Ke Xiang, Hang Cui, Houhua Gong, Changhua Pei and Gaogang Xie

Evidence-intensive analytical reports are expected to be fact-dense, quantitatively correct, and supported by figures. Yet one-shot long-form generation with large language models (LLMs) frequently produces fluent but under-supported drafts: core facts are missed, numbers drift, and key visuals are absent, making the report hard to trust. We propose EviReport, an evidence-tracked report-writing workflow that improves reliability by (i) organizing corpus evidence into compact, traceable units and retrieves query-relevant subgraphs into retrieval-ready packages (ii) leveraging a reasoning-focused LLM sketches a high-level plan for full coverage, then a chat-based LLM sharpens it into a detailed hierarchical outline with explicit scope and ordering (iii) rive generation with a facts-first iterative loop: extracting verifiable facts, composing strictly from those facts, then triggering gap-aware append queries to fill missing evidence To evaluate both correctness and completeness, we introduce EviReportBench, a benchmark instantiated on data-rich indicator reports that measures factual accuracy (claim verification), factual coverage (quiz-based evaluation), and visual evidence integration (image recall). Across 8 topics, experiments show that EviReport consistently outperforms strong baselines in factual coverage (2.16x), factual accuracy (+8.9 points), and visual evidence integration (+34 points), approaching the quality of expert-written reports across multiple dimensions.

[Findings] EvoCoT: Overcoming the Exploration Bottleneck in Reinforcement Learning for LLMs ▶ *Underline Video*

Huanyu Liu, Jia Li, Yihong Dong, Chang Yu, Taozhi Chen, Lecheng Wang, Yongding Tao, Bin Gu and Ge Li

Reinforcement learning with verifiable reward (RLVR) has become a promising paradigm for post-training large language models (LLMs) to improve their reasoning capability. However, when the rollout accuracy is low on hard problems, the reward becomes sparse, limiting learning efficiency and causing exploration bottlenecks. Existing approaches either rely on teacher models for distillation or filter out difficult problems, which limits scalability or restricts reasoning improvement through exploration. We propose EvoCoT, a self-evolving curriculum learning framework based on two-stage chain-of-thought (CoT) reasoning optimization. EvoCoT constrains the exploration space by self-generating and verifying CoT trajectories, then gradually shortens CoT steps to expand the space in a controlled way. The framework enables LLMs to stably learn from initially unsolved hard problems under sparse rewards. We apply EvoCoT to multiple LLM families, including Qwen, DeepSeek, and Llama. Experiments show that EvoCoT enables LLMs to solve previously unsolved problems, improves reasoning capability without external CoT supervision, and is compatible with various RL fine-tuning methods. We release the source code to support future research.

[Long] Explaining Sources of Uncertainty in Automated Fact-Checking ▶ *Underline Video*

Jingyi Sun, Greta Warren, Irina Shklovski and Isabelle Augenstein

Human-AI collaboration in knowledge-intensive tasks such as fact-checking requires understanding model uncertainty in multidocument reasoning amid conflicting/agreeing evidence. Yet, existing methods only express uncertainty as numbers or hedges without revealing which evidence conflicts cause the uncertainty, leaving users unable to resolve disagreements. We present CLUE (Conflict-&Agreement-aware Language-model Uncertainty Explanations), a plugandplay whitebox framework that, to our knowledge, is the first to generate natural-language explanations of model uncertainty grounded in conflicting/agreeing evidence. CLUE (i) identifies spanlevel claim-evidence and interevidence relations that signal conflict or agreement without supervision, and (ii) uses these relations to steer explanation generation, articulating how they drive the model's uncertainty. Across three language models and two factchecking datasets, CLUE produces explanations that more faithfully track model uncertainty and better align with the models factchecking decisions than spanagnostic explanation prompting; human raters also judge them more helpful, more informative, less redundant, and more logically consistent with the input. By explicitly tying uncertainty to evidence conflicts and agreements, CLUE supports practical factchecking and other tasks that require reasoning over complex, conflicting information.

[Findings] Exploring Two-Phase Continual Instruction Fine-tuning for Multilingual Adaptation in Large Language Models ▶ *Underline Video*

Divyanshu Aggarwal, Sankarshan Damle, Navin Goyal, Satya Lokam and Sunayana Sitaram

A key challenge for Large Language Models (LLMs) is improving their Multilingual instruction-following ability over time without deteriorating their ability in languages they already excel at, typically English. In this paper, we study a two-phase Continual Fine-tuning (CFT) setup toward improving a model's Multilingual adaptability. Concretely, we consider a two-phase CFT process in which an English-only end-to-end instruction fine-tuned LLM (Phase 1) is sequentially fine-tuned on a multilingual instruction dataset (Phase 2). Across MISTRAL-7B and LLAMA-3-8B and multiple dataset pairs, we show that instructional similarity between phases is critical: aligned datasets preserve or improve English while boosting multilingual ability, whereas misaligned datasets cause English degradation. We show that this degradation arises from representation shift during CFT, and that targeted mitigation strategies, including generative replay and heuristic-based layer freezing, reduce this shift and improve multilingual adaptation.

[Long] Factual Retrieval in LLMs Is a Redundant, Distributed and Non-Contiguous Process ▶ *Underline Video*

Hail Hochman, Natalie Shapira and Yoav Goldberg

Large language models (LLMs) store and recall factual knowledge, yet the precise mechanism of how entity representations are transformed to enable specific attribute retrieval remains underexplored. In this work, we investigate this mechanism through the lens of an "attribute-computation path", a sequence of computational steps over the entity representation required to elicit a target attribute. We then propose an iterative patching protocol to identify a minimal subset of layers necessary for this computation. Applying our method to LLaMA 3.1 8B and Qwen 3 8B, we find that these paths are non-contiguous, often skipping layers, and that models possess multiple, functionally-equivalent paths for the same entity and fact, highlighting a high degree of redundancy in attribute computation. This implies that knowledge computation is highly distributed, potentially explaining the localization-editing mismatch and suggesting that knowledge storage and retrieval in LLMs is far from being well understood.

[Findings] Fair RAG: End-to-End Fairness Across Retrieval and Generation ▶ *Underline Video*

Farsheed Haque, Zhe Fu, Ramit Aditya, Depeng Xu and Xi Niu

Large Language Models (LLMs) used in Retrieval-Augmented Generation (RAG) can amplify demographic bias: retrievers may surface skewed context and generators can propagate that skew into decisions. Prior work typically treats fairness in retrieval or generation in isolation, leaving end-to-end fairness in RAG underexplored. We propose a post-hoc pipeline that jointly controls both stages: (i) a Fair Greedy Reranker (FGR) that builds prefix-balanced slates toward a target group mix; (ii) a Residual Slate Bias Estimator (RSBE) using signed, prefix-sensitive NDKL to quantify remaining skew; and (iii) Confidence-Gated Logit Calibration (CGLC) that converts the residual signal into small and margin-focused logit corrections without retraining. On an occupation classification task, our approach reduces retriever-side skew (lowest NDKL among baselines for both dense and sparse retrievers) and achieves the lowest generator-side disparity (e.g., Risk Difference) while largely preserving utility. The same calibration can be tuned to alternative fairness criteria (e.g., Equal Opportunity) with minimal utility loss.

[Long] Fair-CCD: Mitigating Bias in Large Language Models for Tabular Classification Through Context-Contrastive Decoding ▶ *Underline Video*

Donghan Liu, Han Sun, Zhaohui Wang, Qin Li and Min Zhang

While recent studies show the effectiveness of in-context learning (ICL) for tabular data prediction, they also reveal significant fairness issues in large language models (LLMs). Prior work to mitigate fairness issues often employs interventions relying on subjective demonstration selection. Its effectiveness varies significantly with the specific demonstration content, leading to low controllability. Moreover, the improvement of fairness is highly unstable across different models and tasks. To address the challenges of low controllability and limited stability in fairness interventions, we propose Fairness-Aware Context-Contrastive Decoding (Fair-CCD). Fair-CCD first constructs Structural Bias Templates (SBTs), motivated by behavioral patterns observed in demonstrations, to encode the relationship between sensitive attributes and predicted labels in a structured and controllable form. During inference, Fair-CCD injects multiple SBTs and contrasts the models responses, generating two differential signals that guide fairness adjustment and preserve task performance. By leveraging attention signals to scale decoding adjustments guided by the difference signals, Fair-CCD achieves stable and adaptive bias mitigation across models and tasks. Extensive experimental results demonstrate that Fair-CCD consistently improves fairness metrics without degrading task accuracy.

[Long] FlexGuard: Continuous Risk Scoring for Strictness-Adaptive LLM Content Moderation ▶ *Underline Video*

Zhihao Ding, Jinming Li, Ze Lu and Jieming Shi

Ensuring the safety of LLM-generated content is essential for real-world deployment. Most existing guardrail models formulate moderation as a fixed binary classification task, implicitly assuming a fixed definition of harmfulness. In practice, enforcement strictness, how conservatively harmfulness is defined and enforced, varies across platforms and evolves over time, making binary moderators brittle under shifting requirements. In this paper, we introduce FlexBench, a strictness-adaptive LLM moderation benchmark that enables controlled evaluation under multiple strictness regimes. Experiments on FlexBench reveal substantial cross-strictness inconsistency in existing moderators: models that perform well under one regime can degrade substantially under others, limiting their practical usability. To address this, we propose FlexGuard, an LLM-based moderator that outputs a calibrated continuous risk score reflecting risk severity and supports strictness-specific decisions via thresholding. We train FlexGuard via risk-alignment optimization to improve score–severity consistency and provide practical threshold selection strategies to adapt to target strictness at deployment. Experiments on FlexBench and public benchmarks demonstrate that FlexGuard achieves higher moderation accuracy and substantially improved robustness under varying strictness.

[Findings] Forge: Quality-Aware Reinforcement Learning for NP-Hard Optimization in LLMs ▶ *Underline Video*

Xiaozhe Li, Xinyu Fang, Shengyuan Ding, Yang Li, Linyang Li, Haodong Duan, Qingwen Liu and Kai Chen

Large Language Models (LLMs) have achieved remarkable success on reasoning benchmarks through Reinforcement Learning with Verifiable Rewards (RLVR), excelling at tasks such as math, coding, logic and puzzles. However, existing benchmarks evaluate only correctness, overlooking optimality, the ability to find the best solutions under constraints. We propose, the first comprehensive framework for training and evaluating LLMs on NP-hard optimization problems through quality-aware RLVR. provides three key components: a scalable training infrastructure with instance generators, quality verifiers, and optimal baselines across 10 tasks; a rigorous benchmark with 1,000 instances evaluating both feasibility (Success Rate) and quality (Quality Ratio); and quality-aware rewards enabling continuous improvement beyond binary correctness. Training on Qwen2.5-7B-Instruct-1M with 15K examples achieves 93.1% SR and 46.6% QR, significantly outperforming GPT-4o (29.6% SR, 14.6% QR). Beyond optimization, training on transfers to diverse tasks: mathematics (+2.2%), logic (+1.2%), knowledge (+4.1%), and instruction-following (+6.1%). Our analysis reveals quality-aware rewards improve solutions by 28.8% over binary rewards, and task diversity drives generalization more than data quantity, offering insights into RLVR scaling for complex reasoning.

[Long] From Individual to Common: An Early Exploration of Consensus in Non-verifiable Data for Balanced Preference Optimization ▶ *Underline Video*

Shangjian Yin and Zhouxing Shi

Reinforcement Learning with Verifiable Rewards (RLVR) has demonstrated remarkable effectiveness in boosting the objective performance (e.g., reasoning) of Large Language Models (LLMs) through rule-based, on-policy self-improvement strategies. However, optimizing LLMs for subjective capabilities and alignment with human preferences remains challenging due to the non-verifiable nature. Most prior works use datasets comprising response pairs with substantial quality gaps labeled by a strong external judge. While effective for preference metrics, this paradigm often incurs an "alignment tax", where the model's objective performance on downstream tasks degrades as it overfits to subjective preferences. In this work, we introduce Donkey, a high-quality, non-verifiable dataset where response pairs differ only by subtle nuances. We find that LLMs optimized on Donkey via preference learning outperform those trained on data with explicit quality gaps, while simultaneously maintaining their objective capabilities. Furthermore, we observe that preference signals on Donkey can be decomposed into consensus preferences and individual preferences. Our analysis reveals that distilling consensus preferences provides a significantly more data-efficient signal for preference optimization. Our findings underscore the importance of leveraging nuanced preference signals and the consensus of multiple judges for advancing subjective LLM alignment. Our code and data will be available at https://github.com/SJY8460/Donkey.

[Long] From Proof to Program: Characterizing Tool-Induced Reasoning Hallucinations in Large Language Models ▶ *Underline Video*

Farima Fatahi Bayat, Pouya Pezeshkpour and Estevam Hruschka

Tool-augmented Language Models (TaLMs) can invoke external tools to solve problems beyond their parametric capacity. However, it remains unclear whether these tool-enabled gains reflect trustworthy reasoning. Focusing on the Code Interpreter tool, we show that even when tools are selected and executed correctly, TaLMs treat tool outputs as substitutes for reasoning, producing solutions that appear correct but lack coherent justification. We term this failure mode Tool-Induced Myopia (TIM), and study it using PyMath, a benchmark of 1,679 competition-level mathematical problems for which Python code is helpful but not sufficient. We further develop a multi-dimensional evaluation suite to quantify reasoning degradation in TaLMs relative to their non-tool counterparts. Our findings reveal that while TaLMs achieve up to a 19.3 percentage point gain in final-answer accuracy, their reasoning behavior consistently deteriorates (e.g., non-tool language models win up to 41.5% more often in pairwise comparisons of reasoning processes). This degradation intensifies with tool use; the more frequently a model invokes tools, the less coherent its reasoning becomes. Moreover, tool use shifts errors from arithmetic mistakes toward global reasoning failures (logic, assumption, creativity). Finally, we propose a preference-optimization-based framework that realigns TaLMs to use tool outputs as assistive evidence, improving both final-answer accuracy and reasoning depth under tool use. Code and data will be released upon publication.

[Findings] GanitLLM: Difficulty-Aware Bengali Mathematical Reasoning through Curriculum-GRPO ▶ *Underline Video*

Shubhashis Roy Dipta, Khairul Mahbub and Nadia Najjar

We present a Bengali mathematical reasoning model called GanitLLM (named after the Bangla word for mathematics, "Ganit"), together with a new difficulty-aware Bengali math corpus and a curriculum-based GRPO pipeline. Bengali is one of the world's most widely spoken languages, yet existing LLMs either reason in English and then translate, or simply fail on multi-step Bengali math, in part because reinforcement learning recipes are tuned for high-resource languages and collapse under reward sparsity in low-resource settings. To address this, we construct Ganit, a rigorously filtered and decontaminated Bengali math dataset with automatic difficulty tags derived from the pass@k of a strong evaluator model. Building on this dataset, we propose Curriculum-GRPO, which combines multi-stage training (SFT + GRPO) with difficulty-aware sampling and verifiable rewards for format, numerical correctness, and Bengali reasoning. On Bn-MGSM and Bn-MSVAMP, GanitLLM-4B improves over its Qwen3-4B base by +8 and +7 accuracy points, respectively, while increasing the percentage of Bengali reasoning tokens from 14% to over 88% and reducing average solution length from 943 to 193 words.

[Findings] GOAT: A Training Framework for Goal-Oriented Agent with Tools ▶ *Underline Video*

Hyunji Min, Sangwon Jung, Junyoung Sung, Dosung Lee, Leekyeung Han and Paul Hongsuck Seo
Large language models (LLMs) have evolved from pure text generators into interactive agents capable of invoking external tools. However, LLM agents still struggle with goal-oriented queries, which require decomposing high-level objectives into sequences of interdependent API calls with accurate planning and execution. Current approaches rely on zero-shot evaluation due to the absence of training data; while proprietary models such as GPT-4 exhibit strong reasoning capabilities, smaller open-source models remain ineffective at complex tool use. To address this limitation, we propose a novel training framework GOAT, that enables fine-tuning LLM agents without human annotation. GOAT automatically synthesizes goal-oriented API execution data from API documents using a novel call-first generation paradigm, that constructs training data based on executed API call sequences. Through extensive experiments, we show that GOAT-trained agents achieve state-of-the-art performance across multiple existing goal-oriented benchmarks. In addition, we introduce GOATBench, a new goal-oriented API execution benchmark, and demonstrate that agents trained with GOAT also excel in this setting. These results highlight GOAT as a practical path toward building robust open-source LLM agents capable of complex reasoning and tool use.

[Findings] HalluGuard: Evidence-Grounded Small Reasoning Models to Mitigate Hallucinations in Retrieval-Augmented Generation ▶ *Underline Video*
Loris Bergeron, Ioana Buhnila, Jerome Francois and Radu State
Large Language Models excel at NLP tasks but remain prone to hallucinations, limiting trust in real-world applications. We present HalluGuard, a 4B-parameter Small Reasoning Model (SRM) designed as a guardrail for Retrieval-Augmented Generation (RAG) pipelines, which classify document-claim pairs as grounded or hallucinated in closed-book, document-grounded settings and produces evidence-grounded justifications. Our approach combines (i) a domain-agnostic synthetic dataset derived from FineWeb and refined through multi-stage curation and data reformation, (ii) synthetic grounded and hallucinated claims, and (iii) preference-based fine-tuning with Odds Ratio Preference Optimization (ORPO) to distill large-model reasoning into a smaller backbone. On the RAGTruth subset of the LLM-AggreFact benchmark, HalluGuard achieves 84.4% balanced accuracy (BAcc), surpassing specialized models, MiniCheck (7B; 84.0%) and Granite Guardian 3.3 (8B; 82.2%) while using roughly half their parameters. Across the benchmark, it reaches 77.1% BAcc, surpassing larger general-purpose LLMs such as GPT-4o (75.9%). HalluGuard and datasets will be released upon acceptance.

[Long] Hierarchical Reinforcement Learning with Augmented Step-Level Transitions for LLM Agents ▶ *Underline Video*
Shuai Zhen, Yanhua Yu, Ruopei Guo, Nan Cheng and Yang Deng
Large language model (LLM) agents have demonstrated strong capabilities in complex interactive decision-making tasks. However, existing LLM agents typically rely on increasingly long interaction histories, resulting in high computational cost and limited scalability. In this paper, we propose STEP-HRL, a hierarchical reinforcement learning (HRL) framework that enables step-level learning by conditioning only on single-step transitions rather than full interaction histories. STEP-HRL structures tasks hierarchically, using completed subtasks to represent global progress of overall task. By introducing a local progress module, it also iteratively and selectively summarizes interaction history within each subtask to produce a compact summary of local progress. Together, these components yield augmented step-level transitions for both high-level and low-level policies. Experimental results on ScienceWorld and ALFWorld benchmarks consistently demonstrate that STEP-HRL substantially outperforms baselines in terms of performance and generalization while reducing token usage. Our code is available at https://github.com/TonyStark042/STEP-HRL.

[Findings] How Hypocritical Is Your LLM judge? Listener-Speaker Asymmetries in the Pragmatic Competence of Large Language Models ▶ *Underline Video*
Judith Sieker and Sina ZarrieSS
Large language models (LLMs) are increasingly studied as repositories of linguistic knowledge. In this line of work, models are commonly evaluated both as generators of language and as judges of linguistic output, yet these two roles are rarely examined in direct relation to one another. As a result, it remains unclear whether success in one role aligns with success in the other. In this paper, we address this question for pragmatic competence by comparing LLMs performance as pragmatic listeners, judging the appropriateness of linguistic outputs, and as pragmatic speakers, generating pragmatically appropriate language. We evaluate multiple open-weight and proprietary LLMs across three pragmatic settings. We find a robust asymmetry between pragmatic evaluation and pragmatic generation: many models perform substantially better as listeners than as speakers. Our results suggest that pragmatic judging and pragmatic generation are only weakly aligned in current LLMs, calling for more integrated evaluation practices.

[Findings] How Value Induction Reshapes LLM Behavior ▶ *Underline Video*
Arnav Arora, Natalie Schluter, Katherine Metcalf and Maartje Ter Hoeve
Conversational Large Language Models are post-trained on language that expresses specific behavioural traits, such as curiosity, open-mindedness, and empathy, and values, such as helpfulness, harmlessness, and honesty. This is done to increase utility, ensure safety, and improve the user experience of the people interacting with the model. However, values are complex and inter-related - incorporating one can modify behaviour on another. Further, incorporating certain values can make models more addictive or sycophantic, potentially having a detrimental effect on the user interacting with it. We investigate these and other unintended effects of value incorporation into models. We fine-tune models using value subsets of existing preference datasets, measuring the effect of value induction of 15 values on safety, anthropomorphism, and various QA benchmarks. We find that i) inducing values also leads to expression of other related, and sometimes contrastive values, ii) inducing positive values increases safety, and iii) all values increase anthropomorphic language use by models, making

them more validating and sycophantic.

[Findings] Identifying and typifying demographic unfairness in phoneme-level embeddings of self-supervised speech recognition models ► *Underline Video*

Felix Herron;Solange Rossato;Alexandre Allauzen;François Portet

Modern automatic speech recognition (ASR) systems have been observed to function better for certain speaker groups (SGs) than others, despite recent gains in overall performance. One potential impediment to progress towards fairer ASR is a more nuanced understanding of the types of modeling errors that speech encoder models make, and in particular the difference between the structure of embeddings for high-performance and low-performance SGs. This paper proposes a framework typifying two types of error that can occur in modeling phonemes in ASR systems: random error/high variance in phoneme embedding, vs systematic error/embedding bias. We find that training phoneme classification probes only on a single, typically disadvantaged SG, sometimes improves performance for that SG, which is evidence for the existence of SG-level bias in phoneme embeddings. On the other hand, we find that speakers and SGs with higher levels of phoneme variance are the same as those with worse phoneme prediction accuracy. We conclude that both types of error are present in phoneme embeddings and both are candidate causes for SG-level unfairness in ASR, though random error is likely a greater hindrance to fairness than systematic error. Furthermore, we find that finetuning encoder models using a fairness-enhancing algorithm (domain enhancing and adversarial training) changes neither the benefits of in-domain phoneme classification probe training, nor measured levels of random embedding error.

[Long] Inhibitory Attacks on Backdoor-based Fingerprinting for Large Language Models ► *Underline Video*

hang fu, Wanli Peng, Yinghan Zhou, Jiaxuan Wu, Juan Wen and Xue yiming

The widespread adoption of Large Language Model (LLM) in commercial and research settings has intensified the need for robust intellectual property protection. Backdoor-based LLM fingerprinting has emerged as a promising solution for this challenge. In practical application, the low-cost multi-model collaborative technique, LLM ensemble, combines diverse LLMs to leverage their complementary strengths, garnering significant attention and practical adoption. Unfortunately, the vulnerability of existing LLM fingerprinting for the ensemble scenario is unexplored. In order to comprehensively assess the robustness of LLM fingerprinting, in this paper, we propose two novel fingerprinting attack methods: token filter attack (TFA) and sentence verification attack (SVA). The TFA gets the next token from a unified set of tokens created by the token filter mechanism at each decoding step. The SVA filters out fingerprint responses through a sentence verification mechanism based on perplexity and voting. Experimentally, the proposed methods effectively inhibit the fingerprint response while maintaining ensemble performance. Compared with state-of-the-art attack methods, the proposed method can achieve better performance. The findings necessitate enhanced robustness in LLM fingerprinting.

[Long] Instructions are all you need: Self-supervised Reinforcement Learning for Instruction Following ► *Underline Video*

Qingyu Ren, Qianyu He, Powei Chang, Jie Zeng, Zeye Sun, Fei Yu, Jiaqing Liang and Yanghua Xiao

Language models often struggle to follow multi-constraint instructions that are crucial for real-world applications. Existing reinforcement learning (RL) approaches suffer from dependency on external supervision and sparse reward signals from multi-constraint tasks. We propose a label-free self-supervised RL framework that eliminates dependency on external supervision by deriving reward signals directly from instructions and generating pseudo-labels for reward model training. Our approach introduces constraint decomposition strategies and efficient constraint-wise binary classification to address sparse reward challenges while maintaining computational efficiency. Experiments show that our approach generalizes well, achieving strong improvements across 3 in-domain and 5 out-of-domain datasets, including challenging agentic and multi-turn instruction following. We will open-source our code and data to facilitate future research.

[Long] Interpretable Coreference Resolution Evaluation Using Explicit Semantics ► *Underline Video*

Bruno Gatti, Giuliano Martinelli and Roberto Navigli

Coreference resolution is typically evaluated using aggregate statistical metrics such as CoNLL-F1, which measure structural overlap between predicted and gold clusters. While widely used, these metrics offer limited diagnostic insights, penalizing errors without revealing whether a system struggles with specific semantic categories, such as people, locations, or events, and making it difficult to interpret model capabilities or derive actionable improvements. We address this gap by introducing a semantically-enhanced evaluation framework for coreference resolution. Our approach overlays Concept and Named Entity Recognition (CNER) onto coreference outputs, assigning semantic labels to nominal mentions and propagating them to entire coreference clusters. This enables the computation of typed scores aimed at evaluating mention extraction and linking capabilities stratified by semantic class. Across our experiments on OntoNotes, LitBank, and PreCo, we show that our framework uncovers systematic weaknesses that remain obscured by aggregate metrics. Furthermore, we show that these diagnostics can be used to design targeted, low-cost data augmentation strategies, achieving measurable out-of-domain improvements.

[Long] IntrAgent: An LLM Agent for Content-Grounded Information Retrieval through Literature Review ► *Underline Video*

Fengbo Ma, Zixin Rao, Xiaoting Li, Zhetao Chen, Hongyue Sun, Yiping Zhao, Xianyan Chen and Zhen Xiang

Scientific research relies on accurate information retrieval from literature to support analytical decisions. In this work, we introduce a new task, INformation reTRieval through literAture reVIEW (IntraView), which aims to automate fine-grained information retrieval faithfully grounded in the provided content in response to research-driven queries, and propose IntrAgent, an LLM-based agent that addresses this challenging task. In particular, IntrAgent is designed to mimic human behaviors when reading literature for information retrieval - identifying

relevant sections and then iteratively extracting key details to refine the retrieved information. It follows a two-stage pipeline: a Section Ranking stage that prioritizes relevant literature sections through structural-knowledge-enabled reasoning, and an Iterative Reading stage that continuously extracts details and synthesizes them into concise, contextually grounded answers. To support rigorous evaluation, we introduce IntraBench, a new benchmark consisting of 315 test instances built from expert-authored questions paired with literature spanning five STEM domains. Across seven backbone LLMs, IntrAgent achieves on average 13.2% higher cross-domain accuracy than state-of-the-art RAG and research-agent baselines.

[Findings] Iterative Formalization and Planning in Partially Observable Environments ▶ *Underline Video*

Liancheng Gong, Wang Bill Zhu, Jesse Thomason and Li Zhang

Using LLMs not to predict plans but to formalize an environment into the Planning Domain Definition Language (PDDL) has been shown to improve performance and control. While most existing methodology only applies to fully observable environments, we adapt to the more realistic and challenging partially observable environments without sufficient information to make a complete plan. We propose PDDLego+, a framework to iteratively formalize, plan, grow, and refine PDDL representations by decomposing the environment and the goal into fully observable episodes. Without finetuning, in-context exemplars, or trajectories, PDDLego+ improves planning success and exhibits robustness against problem complexity compared to end-to-end approaches. We also show that the domain knowledge captured after a successful trial can benefit future tasks.

[Long] JARVIS or Ultron? A Survey on the Safety and Security Threats of Computer-Using Agents ▶ *Underline Video*

Ada Chen, Yongjiang Wu, Junyuan Zhang, Jingyu Xiao, Shu Yang, Jen-tse Huang, Kun Wang, Wenxuan Wang and Shuai Wang

Recently, AI-driven interactions with computing devices have advanced from basic prototype tools to sophisticated, LLM-based systems that emulate human-like operations in graphical user interfaces. We are now witnessing the emergence of Computer-Using Agents (CUAs), capable of autonomously performing tasks such as navigating desktop applications, web pages, and mobile apps. However, as these agents grow in capability, they also introduce novel safety and security risks. Vulnerabilities in LLM-driven reasoning, with the added complexity of integrating multiple software components and multimodal inputs, further complicate the security landscape. In this paper, we present a systematization of knowledge on the safety and security threats of CUAs. We conduct a comprehensive literature review and distill our findings along four research objectives: (i) define the CUA that suits safety analysis; (ii) categorize current safety threats among CUAs; (iii) propose a comprehensive taxonomy of existing defensive strategies; (iv) summarize prevailing benchmarks, datasets, and evaluation metrics used to assess the safety and performance of CUAs. Building on these insights, our work provides future researchers with a structured foundation for exploring unexplored vulnerabilities and offers practitioners actionable guidance in designing and deploying secure Computer-Using Agents.

[Findings] Justice in Judgment: Unveiling (Hidden) Bias in LLM-assisted Peer Reviews ▶ *Underline Video*

Sai Suresh Macharla Vasu, Ivaxi Sheth, Hui-Po Wang, Ruta Binkyte and Mario Fritz

The adoption of large language models (LLMs) is transforming the peer review process, from assisting reviewers in writing detailed evaluations to generating entire reviews automatically. While these capabilities offer new opportunities, they also raise concerns about fairness and reliability. In this paper, we investigate bias in LLM-generated peer reviews through controlled interventions on author metadata, including affiliation, gender, seniority, and publication history. Our analysis consistently shows a strong affiliation bias favoring authors from highly ranked institutions. We also identify directional preferences associated with seniority and prior publication record, which can influence acceptance decisions for borderline papers. Gender effects are smaller but present in several models. Notably, implicit biases become more pronounced when examining token-level soft ratings, suggesting that alignment may mask but not fully eliminate underlying preferences.

[Long] Language of Thought Shapes Output Diversity in Large Language Models ▶ *Underline Video*

Shaoyang Xu and Wenxuan Zhang

Output diversity is crucial for Large Language Models as it underpins pluralism and creativity. In this work, we reveal that controlling the language used during model thinking, the language of thought, provides a novel and structural source of output diversity. Our preliminary study shows that different thinking languages occupy distinct regions in a model's thinking space. Based on this observation, we study two repeated sampling strategies under multilingual thinking, Single-Language Sampling and Mixed-Language Sampling, and conduct diversity evaluation on outputs that are controlled to be in English, regardless of the thinking language used. Across extensive experiments, we demonstrate that switching the thinking language from English to non-English languages consistently increases output diversity, with a clear and consistent positive correlation such that languages farther from English in the thinking space yield larger gains. We further show that aggregating samples across multiple thinking languages yields additional improvements through compositional effects, and that scaling sampling with linguistic heterogeneity expands the model's diversity ceiling. Finally, we show that these findings translate into practical benefits in pluralistic alignment scenarios, leading to broader coverage of cultural knowledge and value orientations in LLM outputs. Our code is publicly available at https://github.com/iNLP-Lab/Multilingual-LoT-Diversity.

[Short] Language-Aware Token Boosting: LLM Language Confusion Reduction Without Tuning ▶ *Underline Video*

Trapoom Ukarapol, Pakhapoom Sarapat and Nut Chukamphaeng

Large language models (LLMs) sometimes exhibit language confusion when generating non-English text. Existing approaches typically rely on fine-tuning to mitigate this issue. In contrast, we propose a tuning-free paradigm for reducing language confusion. Within this paradigm,

we introduce two methods: Language-Aware Token Boosting (LATB), which applies targeted perturbations to tokens associated with the desired language, and Adaptive Language-Aware Token Boosting (Adaptive-LATB), which dynamically adjusts these perturbations based on the models confidence in the intended language. Experiments demonstrate that our methods effectively improve multilingual alignment by reducing language confusion, while maintain the summarization quality without requiring any additional fine-tuning. Our code is publicly available..

[Findings] Language-Coupled Reinforcement Learning for Multilingual Retrieval-Augmented Generation ▶ *Underline Video*

Rui Qi, Fengran Mo, Yufeng Chen, Xue Zhang, Shuo Wang, Hongliang Li, Xu Jinan, Meng Jiang, Jian-Yun Nie and Kaiyu Huang

Multilingual retrieval-augmented generation (MRAG) requires models to effectively acquire and integrate beneficial external knowledge from multilingual collections. However, most existing studies employ a unitive process where queries of equivalent semantics across different languages are processed through a single-turn retrieval and subsequent optimization. Such a "one-size-fits-all" strategy is often suboptimal in multilingual settings, as the models occur to knowledge bias and conflict during the interaction with the search engine. To alleviate the issues, we propose LcRL, a multilingual search-augmented reinforcement learning framework that integrates a language-coupled Group Relative Policy Optimization into the policy and reward models. We adopt the language-coupled group sampling in the rollout module to reduce knowledge bias, and regularize an auxiliary anti-consistency penalty in the reward models to mitigate the knowledge conflict. Experimental results demonstrate that ~not only achieves competitive performance but is also appropriate for various practical scenarios such as constrained training data and retrieval over collections encompassing a large number of languages. Our code is available at https://anonymous.4open.science/r/LcRL-B4EF.

[Findings] Large Reasoning Models Are (Not Yet) Multilingual Latent Reasoners ▶ *Underline Video*

Yihong Liu, Raoyuan Zhao, Hinrich Schuetze and Michael A. Hedderich

Large reasoning models (LRMs) achieve strong performance on mathematical reasoning tasks, often attributed to their capability to generate explicit chain-of-thought (CoT) explanations. However, recent work shows that LRMs often arrive at the correct answer before completing these textual reasoning steps, indicating the presence of latent reasoning – internal, non-verbal computation encoded in hidden states. While this phenomenon has been explored in English, its multilingual behavior remains largely unknown. In this paper, we conduct a systematic investigation of multilingual latent reasoning in LRMs across 11 languages. Using a truncation-based strategy, we examine how the correct answer emerges as the model is given only partial reasoning traces, allowing us to measure stepwise latent prediction formation. Our results reveal clear evidence of multilingual latent reasoning, though unevenly: strong in resource-rich languages, weaker in low-resource ones, and broadly less observable on harder benchmarks. To understand whether these differences reflect distinct internal mechanisms, we further perform representational analyses. Despite surface-level disparities, we find that the internal evolution of predictions is highly consistent across languages and broadly aligns with English – a pattern suggesting an English-centered latent reasoning pathway.

[Findings] LeakDojo: Decoding the Leakage Threats of RAG Systems ▶ *Underline Video*

Maosen Zhang, Jianshuo Dong, Lu boting, Li Wenyue, Xiaoping Zhang, Tianwei Zhang and Han Qiu

Retrieval-Augmented Generation (RAG) enables large language models (LLMs) to leverage external knowledge, but also exposes valuable RAG databases to leakage attacks. As RAG systems grow more complex and LLMs exhibit stronger instruction-following capabilities, existing studies fall short of systematically assessing RAG leakage risks. We present LeakDojo, a configurable framework for controlled evaluation of RAG leakage. Using LeakDojo, we benchmark six existing attacks across fourteen LLMs, four datasets, and diverse RAG systems. Our study reveals that (1) query generation and adversarial instructions contribute independently to leakage, with overall leakage well approximated by their product; (2) stronger instruction-following capability correlates with higher leakage risk; and (3) improvements in RAG faithfulness can introduce increased leakage risk. These findings provide actionable insights for understanding and mitigating RAG leakage in practice. Our codebase is available at https://github.com/yeasen-z/LeakDojo.

[Findings] LearnAct: Few-Shot Mobile GUI Agent with a Unified Demonstration Benchmark ▶ *Underline Video*

Guangyi Liu, Pengxiang Zhao, Liang Liu, Zhiming Chen, Yuxiang Chai, Siheng Chen, Yaozhen Liang, WenHao Wang, Zhengxi Lu, Shuai Ren, Hao Wang, Shibo He, Yong Liu and Wenchao Meng

Mobile GUI agents show promise in automating tasks but face significant generalization challenges in long-tail scenarios. While learning from few-shot demonstrations is an emerging solution, its progress is hindered by two critical gaps: the lack of a comprehensive benchmark for systematic evaluation on mobile devices, and the absence of a systematic framework designed to learn from demonstrations in this domain. To address these gaps, we introduce LearnGUI, the first comprehensive benchmark designed for studying demonstration-based learning in mobile agents, comprising 2,252 offline and 101 online tasks. We further develop LearnAct, a modular agent framework engineered to systematically extract, retrieve, and leverage knowledge from visual demonstrations. Extensive evaluations across six backbone models validate our approach: LearnAct achieves dramatic improvements for general-purpose models (e.g., Gemini-2.5-Pro: 38.5% to 58.9%) and specialized models alike (e.g., UI-TARS-7B-SFT's online success rate: 18.1% to 32.8%), demonstrating consistent gains across model architectures. Our work provides a robust benchmark and a systematic framework, paving the way for more adaptable and practical mobile agents. Our code and data are publicly available at https://lgy0404.github.io/LearnAct/.

[Findings] Learning from Contrastive Prompts: An Automated Prompt Optimization Framework ▶ *Underline Video*

Mingqi Li, Karan Aggarwal, Yong Xie, Aitzaz Ahmad and Stephen Lau

As large language models (LLMs) continue to advance, significant effort is spent on manually crafting prompts to unlock their full potential. While existing prompt optimization methods automate this process, they often underperform due to their reliance on learning exclusively from incorrect samples. We propose the Learning from Contrastive Prompts (LCP) framework, which leverages contrastive prompts to distinguish between high- and low-performing cases. By identifying and amplifying the differences that make prompts effective, LCP systematically extracts principles underlying successful prompt design. On the Big-Bench Hard benchmark, LCP achieves an 87.5% win rate on Claude-3-Sonnet and 75.7% on Claude-4-Sonnet. Experiments on DeepSeek-R1 (88.2% win rate) and SuperGLUE further confirm that LCP generalizes across both proprietary and open-source models and diverse NLU benchmarks. The framework offers a principled and scalable foundation for automated prompt engineering, reducing manual intervention in adapting LLMs to diverse applications.

[Long] Learning from Mistakes: Negative Reasoning Samples Enhance Out-of-Domain Generalization ▶ *Underline Video*

Xueyun Tian, Minghua Ma, Bingbing Xu, Nuoyan Lyu, Wei Li, Heng Dong, Zheng Chu, Yuanzhuo Wang and Huawei Shen

Supervised fine-tuning (SFT) on chain-of-thought (CoT) trajectories demonstrations is a common approach for enabling reasoning in large language models. Standard practices typically only retain trajectories with correct final answers (positives) while ignoring the rest (negatives). We argue that this paradigm discards substantial supervision and exacerbates overfitting, limiting out-of-domain (OOD) generalization. Specifically, we surprisingly find that incorporating negative trajectories into SFT yields substantial OOD generalization gains over positive-only training, as these trajectories often retain valid intermediate reasoning despite incorrect final answers. To understand this effect in depth, we systematically analyze data, training dynamics, and inference behavior, identifying 22 recurring patterns in negative chains that serve a dual role: they moderate loss descent to mitigate overfitting during training and boost policy entropy by 35.67% during inference to facilitate exploration. Motivated by these observations, we further propose Gain-based LOss Weighting (GLOW), an adaptive, sample-aware scheme that exploits such distinctive training dynamics by rescaling per-sample loss based on inter-epoch progress. Empirically, GLOW efficiently leverages unfiltered trajectories, yielding a 5.51% OOD gain over positive-only SFT on Qwen2.5-7B and boosting MMLU from 72.82% to 76.47% as an RL initialization. Code is available at https://github.com/Eureka-Maggie/GLOW.

[Long] Learning to Conceal Risk: Controllable Multi-turn Red Teaming for LLMs in the Financial Domain ▶ *Underline Video*

Gang Cheng, Haibo Jin, Wenbin Zhang, Haohan Wang and Jun Zhuang

Large Language Models (LLMs) are increasingly deployed in finance, where unsafe behavior can lead to serious regulatory risks. However, most red-teaming research focuses on overtly harmful content and overlooks attacks that appear legitimate on the surface yet induce regulatory-violating responses. We address this gap by introducing a controllable black-box multi-turn risk-concealed redteaming framework (CoRT) that progressively conceals surface-level risk while exploiting regulatory-violating behaviors. CoRT contains two key components: (i) a Risk Concealment Attacker (RCA) that generates multiturn prompts via iterative refinement, and (ii) a Risk Concealment Controller (RCC) that predicts a turn-level Risk Concealment Score (RCS) to steer RCAs follow-up style. We also build a domain-specific benchmark, FinRisk-Bench, with 522 instructions spanning six financial risk categories. Experiments on nine widely used LLMs show that CoRT (RCA) achieves 93.19% average attack success rate (ASR), and CoRT (RCA+RCC) further improves the average ASR to 95.00%. Our code and FinRisk-Bench are available at https://github.com/gcheng128/CoRT.

[Long] Lizard: An Efficient Linearization Framework for Large Language Models ▶ *Underline Video*

Van Chien Nguyen, Huy Nguyen, Ruiyi Zhang, Hanieh Deilamsalehy, Puneet Mathur, Viet Lai, Haoliang Wang, Jayakumar Subramanian, Ryan A. Rossi, Trung Bui, Nikos Vlassis, Franck Dernoncourt and Thien Huu Nguyen

We propose Lizard, a linearization framework that transforms pretrained Transformer-based Large Language Models (LLMs) into subquadratic architectures. Transformers faces severe computational and memory bottlenecks with long sequences due to the quadratic complexity of softmax attention and the growing Key-Value (KV) cache that makes inference memory-bound by context length. Lizard addresses these limitations by introducing a subquadratic attention mechanism that closely approximates softmax attention while preserving model quality. Unlike prior linearization methods constrained by fixed, non-adaptive structures, Lizard augments the architecture with compact, learnable modules that enable adaptive memory control and robust length generalization. Moreover, we introduce a hardware-aware algorithm that solves numerical instability in gated attention to accelerate training. Extensive experiments show that Lizard achieves near-lossless recovery of its teacher model's performance, significantly outperforming previous methods by up to 9.4 - 24.5 points on the 5-shot MMLU benchmark and demonstrating superior associative recall.

[Long] LLM Safety From Within: Detecting Harmful Content with Internal Representations ▶ *Underline Video*

Difan Jiao, Yilun Liu, Ye Yuan, Zhenwei Tang, Linfeng Du, Haolun Wu and Ashton Anderson

Guard models are widely used to detect harmful content in user prompts and LLM responses. However, state-of-the-art guard models rely solely on terminal-layer representations and overlook the rich safety-relevant features distributed across internal layers. We present SIREN, a lightweight guard model that harnesses these internal features. By identifying safety neurons via linear probing and combining them through an adaptive layer-weighted strategy, SIREN builds a harmfulness detector from LLM internals without modifying the underlying model. Our comprehensive evaluation shows that SIREN substantially outperforms state-of-the-art open-source guard models across multiple benchmarks while using 250x fewer trainable parameters. Moreover, SIREN exhibits superior generalization to unseen benchmarks, naturally enables real-time streaming detection, and significantly improves inference efficiency compared to generative guard models. Overall, our results highlight LLM internal states as a promising foundation for practical, high-performance harmfulness detection.

[Findings] LLMs in the Real World: Evaluating AI in Emergency Contexts ▶ *Underline Video*

Sara Court, Lara Downing and Micha Elsner

This paper offers a call to action. We urge our colleagues in the research community to play a greater role in the articulation of our findings to the public. To illustrate the stakes we present a case study on the initial stages of an LLM-based machine translation application's deployment in a real-world context: a text-2-911 system advertising capabilities in 55 languages for use in emergencies in which it may be difficult to call operators directly. We identify a number of common misconceptions about technologies such as these, concluding with a set of concrete recommendations and best practices for stakeholders at every stage of the development and deployment pipeline. While the advancement of scientific research often lies in solving the "hard" problems, we argue it is often the "easy" ones— problems for which the latest technology is often unnecessary— that are most overlooked.

[Findings] LLMs Meet Isolation Kernel: Lightweight, Learning-free Binary Embeddings for Fast Retrieval ▶ *Underline Video*

Zhibo Zhang, Yang Xu, Kai Ming Ting and Cam-Tu Nguyen

Large language models (LLMs) have recently enabled remarkable progress in text representation. However, their embeddings are typically high-dimensional, leading to substantial storage and retrieval overhead. Although recent approaches such as Matryoshka Representation Learning (MRL) and Contrastive Sparse Representation (CSR) alleviate these issues to some extent, they still suffer from retrieval accuracy degradation. This paper proposes Isolation Kernel Embedding or IKE, a learning-free method that transforms an LLM embedding into a binary embedding using Isolation Kernel (IK). Lightweight and based on binary encoding, IKE offers a low memory footprint and fast bitwise computation, lowering retrieval latency. Experiments on multiple text retrieval datasets demonstrate that IKE offers up to 16.7x faster retrieval and 16x lower memory usage than the original LLM embeddings, while maintaining comparable accuracy. Theoretically, we show that IKE works because it satisfies four essential criteria for effective binary hashing that other methods do not possess. Compared to CSR, IKE consistently achieves better retrieval efficiency and effectiveness. IKE also works effectively with graph-based indexing, demonstrating its superiority in balancing accuracy and latency compared to alternative compression techniques in the approximate nearest neighbor (ANN) search setting.

[Long] Location Not Found: Exposing Implicit Local and Global Biases in Multilingual LLMs ▶ *Underline Video*

Guy Mor-Lan, Omer Goldman, Matan Eyal, Adi Mayrav Gilady, Sivan Eiger, Idan Szpektor, Avinatan Hassidim, Yossi Matias and Reut Tsarfaty

Multilingual large language models (LLMs) have minimized the fluency gap between languages. This advancement, however, exposes models to the risk of biased behavior, as knowledge and norms may propagate across languages. In this work we aim to quantify models' inter- and intra-lingual biases, via their ability to answer locale-ambiguous questions. To this end, we present LocQA, a test set containing 2,156 questions in 12 languages, referring to various locale-dependent facts such as laws, dates, and measurements. The questions do not contain indications of the locales they relate to, other than the querying language itself. LLMs' responses to LocQA locale-ambiguous questions thus reveal models' implicit priors. We used LocQA to evaluate 32 models, and detected two types of structural biases. Inter-lingually, we show a global bias towards answers relevant to the US-locale, even when models are asked in languages other than English. Moreover, we discovered that this global bias is exacerbated in models that underwent instruction tuning, compared to their base counterparts. Intra-lingually, we show that when multiple locales are relevant for the same language, models act as demographic probability engines, prioritizing locales with larger populations. Taken together, insights from LocQA may help in shaping LLMs' desired local behavior, and in quantifying the impact of various training phases on different kinds of biases.

[Findings] Logical Structure as Knowledge: Enhancing LLM Reasoning via Structured Logical Knowledge Density Estimation ▶ *Underline Video*

Zhen Bi, Zhenlin Hu, Xueshu Chen, Mingyang Chen, Cheng Deng, Yida Xue, Zhen Wang, Qing Shen, Ningyu Zhang and Jungang Lou

The reasoning capabilities of Large Language Models (LLMs) are increasingly attributed to training data quality rather than mere parameter scaling. However, existing data-centric paradigms often equate quality with factuality or diversity and ignore the internal logical complexity of training samples. In this work, we propose that natural language harbors Structured Logical Knowledge manifested through entailment relationships and logical topologies. To quantify this, we introduce Structured Logical Knowledge Density (SLKD), a novel metric that measures logical information content by decomposing natural language into executable predicates and logical primitives. Our analysis reveals a significant logical disparity in current datasets where sparse logical signals predominate. Consequently, we propose a density-aware re-cognizing optimization strategy that prioritizes high-density logical samples to align training with the model's reasoning boundary. Extensive experiments demonstrate that our approach enhances reasoning performance and generalization without increasing total data volume. These results, further validated within a reinforcement learning framework, suggest that elevating logical density is more critical than expanding data scale for realizing the full cognitive potential of LLMs. The anonymized code is available in the Appendix C.

[Long] LOKA: Conflict-Aware LLM Knowledge Update with Adaptive Knowledge Memory ▶ *Underline Video*

Binchi Zhang, Zhengzhang Chen, Zaiyi Zheng, Jundong Li and Haifeng Chen

Large Language Models (LLMs) have achieved remarkable success in natural language processing by encoding extensive knowledge, but their utility relies on timely updates as human knowledge keeps evolving. In this paper, we investigate the problem of LLM knowledge updates, which requires simultaneously unlearning unwanted information and learning new knowledge. Existing approaches that tackle unlearning and learning separately encounter task conflicts and knowledge management issues when applied to comprehensive knowledge

updates. In this paper, we validate our findings with theoretical analysis and empirical evidence, and propose LOKA, a conflict-aware framework for Large language mOdel Knowledge updAtes. During training, LOKA introduces an adaptive knowledge memory approach in which updated knowledge is allocated across multiple memory units. During inference, LOKA retrieves the most relevant memory unit from the knowledge memory and integrates it with the original LLM to apply updated knowledge, while a learning-based router controls the activation of the knowledge memory to improve knowledge utilization. Extensive experiments demonstrate the efficacy of LOKA in achieving accurate, flexible, and conflict-aware knowledge updates.

[Long] Make Mechanistic Interpretability Auditable: A Call to Develop Guidelines via Continuous Collaborative Reviewing ► *Underline Video*

Michael Lan, Narmeen Fatimah Oozeer, Chaithanya Bandi, Philip Quirke, Austin Meek, Fazl Barez and Amir Abdullah

While mechanistic interpretability (MI) has produced important insights into neural network internals, the field has yet to establish a standardized system to audit experiments. As such, many of its findings remain underutilized in safety-critical applications such as medical AI and autonomous systems, as stakeholders cannot certify their validity. Recent work demonstrates this concretely: two papers found conflicting conclusions for the same behavior, and a third study revealed that both were partially correct but incomparable due to methodological inconsistencies. Without standardized auditing, such ambiguities hinder adoption in high-stakes contexts requiring strong correctness guarantees. We call for the MI community to work towards developing a novel reviewing system that complements peer review via: (1) Continuous reviewing supported by a Collaborative Reviewing Platform where meta-science results and discussions (such as critiques, negative results, post-hoc extensions, reproductions, replications, and partial results) that fit outside of papers are organized and discussed, allowing for comments and revisions to be made at any time (2) Generalizing good practices found on this platform into expert-verified guidelines and protocols to improve auditing efficiency, and (3) Source-based auditing systems that track arguments which claims depend on. This position paper encourages constructive debate over the necessity, design and implementation of such a framework, providing early concrete examples to help catalyze these dialogues. Overall, we propose that auditing MI itself is essential for its application in AI safety, industry, and governance.

[Findings] Making Revisions Understandable: A Survey of Edit Intentions, Methods, and Applications ► *Underline Video*

Fangping Lan, Qi Zhang and Eduard Dragut

Text revision is a core process in document creation, capturing how authors iteratively refine, reorganize, and improve written content. With the increasing availability of large-scale revision histories from platforms such as Wikipedia and arXiv, NLP research has begun to move beyond modeling what changes are made to understanding why they are made, i.e., the underlying edit intentions. To our knowledge, this is the first survey that synthesizes text revision research through the lens of edit intentions, providing a unified view of datasets, taxonomies, identification methods, and applications. We review prior work across the full revision workflow, including revision corpus construction, edit intention taxonomy design, and edit intention identification. We further categorize representative datasets and methods, summarize downstream applications such as writing assistance and document edit summarization, and highlight key open research directions.

[Long] MalruleLib: Large-Scale Executable Misconception Reasoning with Step Traces for Modeling Student Thinking in Mathematics ► *Underline Video*

Xinghe Chen, Naiming Liu and Shashank Sonkar

Student mistakes in mathematics are often systematic: a learner applies a coherent but wrong procedure and repeats it across contexts. We introduce MalruleLib, a learning-science-grounded framework that translates documented misconceptions into executable procedures, drawing on 67 learning-science and mathematics education sources, and generates step-by-step traces of malrule-consistent student work. We formalize a core student-modeling problem as Malrule Reasoning Accuracy (MRA): infer a misconception from one worked mistake and predict the students next answer under cross-template rephrasing. Across nine language models (4B - 120B), accuracy drops from 66% on direct problem solving to 40% on cross-template misconception prediction. MalruleLib encodes 101 malrules over 498 parameterized problem templates and produces paired dual-path traces for both correct reasoning and malrule-consistent student reasoning. Because malrules are executable and templates are parameterizable, MalruleLib can generate over one million instances, enabling scalable supervision and controlled evaluation. Using MalruleLib, we observe cross-template degradations of 10 - 21%, while providing student step traces improves prediction by 3 - 15%. We release MalruleLib as infrastructure for educational AI that models student procedures across contexts, enabling diagnosis and feedback that targets the underlying misconception.

[Long] Mango: Multi-Agent Web Navigation via Global-View Optimization ► *Underline Video*

Weixi Tong, Yifeng Di and Tianyi Zhang

Existing web agents typically initiate exploration from the root URL, which is inefficient for complex websites with deep hierarchical structures. Without a global view of the website's structure, agents frequently fall into navigation traps, explore irrelevant branches, or fail to reach target information within a limited budget. We propose Mango, a multi-agent web navigation method that leverages the website structure to dynamically determine optimal starting points. We formulate URL selection as a multi-armed bandit problem and employ Thompson Sampling to adaptively allocate the navigation budget across candidate URLs. Furthermore, we introduce an episodic memory component to store navigation history, enabling the agent to learn from previous attempts. Experiments on WebVoyager demonstrate that Mango achieves a success rate of 63.6% when using GPT-5-mini, outperforming the best baseline by 7.3%. Furthermore, on WebWalkerQA, Mango attains a 52.5% success rate, surpassing the best baseline by 26.8%. We also demonstrate the generalizability of Mango using both open-source and

closed-source models as backbones. Our data and code are open-source and available at https://github.com/VichyTong/Mango.

[Findings] Masking or Mitigating? Deconstructing the Impact of Query Rewriting on Retriever Biases in RAG ▶ *Underline Video*

Agam Goyal, Koyel Mukherjee, Apoorv Saxena, Anirudh Phukan, Eshwar Chandrasekharan and Hari Sundaram

Dense retrievers in retrieval-augmented generation (RAG) systems exhibit systematic biases—including brevity, position, literal matching, and repetition biases—that can compromise retrieval quality. Query rewriting techniques are now standard in RAG pipelines, yet their impact on these biases remains unexplored. We present the first systematic study of how query enhancement techniques affect dense retrieval biases, evaluating five methods across six retrievers. Our findings reveal that simple LLM-based rewriting achieves the strongest aggregate bias reduction (54%), yet fails under adversarial conditions where multiple biases combine. Mechanistic analysis uncovers two distinct mechanisms: simple rewriting reduces bias through increased score variance, while pseudo-document methods achieve reduction through genuine decorrelation from bias-inducing features. However, no technique uniformly addresses all biases, and effects vary substantially across retrievers. Our results provide practical guidance for selecting query enhancement strategies based on specific bias vulnerabilities. More broadly, we establish a taxonomy distinguishing query-document interaction biases from document encoding biases, clarifying the limits of query-side interventions for debiasing RAG systems.

[Long] MATCH: Modulating Attention via InContext Retrieval for LongContext Transformers ▶ *Underline Video*

Linrui Ma, Chun Hei Lo, Xinyu Wang, Peng Lu, Xihao Yuan, Hanting Chen, Kai Han, Xinghao Chen, Chengjun Zhan, Hanlin xu, Yichun Yin, Lifeng Shang, Feng Wen, Boxing Chen and Yufei Cui

The quadratic computational cost of traditional attention mechanisms poses a major bottleneck to the scalability and practical deployment of large language models (LLMs), particularly in long-context scenarios. To improve efficiency, existing approaches often enforce rigid structural constraints such as local attention windows. However, these strategies typically lead to substantial performance degradation on tasks requiring precise long-range recall. In this work, we propose MATCH, a scalable and efficient framework that augments sparsified attention mechanisms with dynamically integrated in-context information through an efficient retrieval system. Empirical results show that MATCH significantly improves the performance of sparse-attention models on both synthetic and real-world natural-language tasks. These findings highlight the versatility of MATCH as a general approach for enhancing in-context retrieval capabilities while maintaining the efficiency benefits of sparse attention architectures.

[Findings] Measuring Watermarking under Jailbreaking: ASR Inflation and Goal-Compliance Mismatch ▶ *Underline Video*

Sungwoo Han, Sangjun Moon, Jingun Kwon, Hidetaka Kamigaito and Manabu Okumura

Recently, watermarking has attracted growing attention as a practical technique for source attribution of machine-generated text. However, most prior work studies watermarking under benign prompts, while its behavior under jailbreaking prompts remains underexplored. This gap matters because jailbreaking can bypass safety policies and shift the generation regime, raising concerns that watermarking may interact with model alignment under attack. To address this gap, we evaluate six watermarking methods on four LLMs across two jailbreak benchmarks and three settings: Static, AutoDAN, and DSN. We find that watermarking can inflate judge-based attack success rate, denoted ASR, under jailbreaking, with the largest effects appearing in biased schemes that perturb logits. At the same time, these ASR increases often do not reflect higher harmful-goal compliance when measured by StrongREJECT or by human judgments. This suggests that ASR-only evaluations can be brittle to decoding perturbations and may overestimate harmful-goal compliance, motivating complementary goal-compliance metrics (e.g., StrongREJECT) and human evaluations.

[Long] MemSearch-o1: Empowering Large Language Models with Reasoning-Aligned Memory Growth in Agentic Search ▶ *Underline Video*

Sheng Zhang, Junyi Li, Yingyi Zhang, Pengyue Jia, Yichao Wang, Xiaowei Qian, Wenlin Zhang, Maolin Wang, Yong Liu and Xiangyu Zhao

Recent advances in large language models (LLMs) have scaled the potential for reasoning and agentic search, wherein models autonomously plan, retrieve, and reason over external knowledge to answer complex queries. However, the iterative think-search loop accumulates long system memories, leading to memory dilution problem. In addition, existing memory management methods struggle to capture fine-grained semantic relations between queries and documents and often lose substantial information. Therefore, we propose MemSearch-o1, an agentic search framework built on reasoning-aligned memory growth and retracing. MemSearch-o1 dynamically grows fine-grained memory fragments from memory seed tokens from the queries, then retraces and deeply refines the memory via a contribution function, and finally reorganizes a globally connected memory path. This shifts memory management from stream-like concatenation to structured, token-level growth with path-based reasoning. Experiments on eight benchmark datasets show that MemSearch-o1 substantially mitigates memory dilution, and more effectively activates the reasoning potential of diverse LLMs, establishing a solid foundation for memory-aware agentic intelligence.

[Findings] Meta-Tool: Efficient Few-Shot Tool Adaptation for Small Language Models ▶ *Underline Video*

Sachin Kumar

Can small language models achieve strong tool-use performance without complex adaptation mechanisms? This paper investigates this question through Meta-Tool, a controlled empirical study comparing hypernetwork-based LoRA adaptation against carefully designed few-shot prompting. Using a Llama-3.2-3B-Instruct backbone, we evaluate four adaptation mechanisms, few-shot prompting, documentation encoding, hypernetwork-generated LoRA weights, and value-guided beam search, across four diverse benchmarks: Gorilla APIBench, Spider 2.0, WebArena, and InterCode. Our central finding is a well-supported negative result: despite generating non-trivial weight matrices,

the 227.8M-parameter hypernetwork provides no measurable improvement over few-shot prompting alone. Comprehensive ablation studies reveal that few-shot examples contribute +21.5% to performance and documentation contributes +5.0%, while the hypernetwork adds 0%. A 3B model with well-designed prompts achieves 79.7% of GPT-5's average performance at 10 x lower latency. Error analysis across 722 failure cases spanning all shot counts (0-5) shows that at the 5-shot configuration (106 failures), failure modes are task-dependent: schema-heavy tasks (Spider 2.0, WebArena) show near-zero format errors with remaining failures semantic, while format errors dominate on Gorilla (100%) and InterCode (70%). These findings redirect practitioners toward prompt engineering and example curation rather than complex adaptation architectures.

[Long] MiniRAG: A Lightweight RAG system with Small Language Models ► *Underline Video*

Tianyu Fan, Jingyuan Wang, Xubin Ren and Chao Huang

The growing demand for efficient and lightweight Retrieval-Augmented Generation (RAG) systems has highlighted significant challenges when deploying Small Language Models (SLMs) in existing RAG frameworks. Current approaches face severe performance degradation due to SLMs' limited semantic understanding and text processing capabilities, creating barriers for widespread adoption in resource-constrained scenarios. To address these fundamental limitations, we present MiniRAG, a novel RAG system designed for simplicity and efficiency. MiniRAG introduces two key technical innovations: (1) a semantic-aware heterogeneous graph indexing mechanism that combines text chunks and named entities in a unified structure, reducing reliance on complex semantic understanding, and (2) a lightweight topology-enhanced retrieval approach that leverages graph structures for efficient knowledge discovery without requiring advanced language capabilities. Our extensive experiments demonstrate that MiniRAG achieves comparable performance to LLM-based methods even when using SLMs while requiring only 25% of the storage space. Additionally, we contribute a comprehensive benchmark dataset for evaluating lightweight RAG systems under realistic on-device scenarios with complex queries.

[Findings] Mitigating Cultural Bias in LLMs via Multi-Agent Cultural Debate ► *Underline Video*

Qian Tan, Lei Jiang, Yuting Zeng, Shuoyang Ding and Xiaohua Xu

Large language models (LLMs) exhibit systematic Western-centric bias, yet whether prompting in non-Western languages (e.g., Chinese) can mitigate this remains understudied. Answering this question requires rigorous evaluation and effective mitigation, but existing approaches fall short on both fronts: evaluation methods force outputs into predefined cultural categories without a neutral option, while mitigation relies on expensive multi-cultural corpora or agent frameworks that use functional roles (e.g., Planner–Critique) lacking explicit cultural representation. To address these gaps, we introduce CEBiasBench, a Chinese–English bilingual benchmark, and Multi-Agent Vote (MAV), which enables explicit "no bias" judgments. Using this framework, we find that Chinese prompting merely shifts bias toward East Asian perspectives rather than eliminating it. To mitigate such persistent bias, we propose Multi-Agent Cultural Debate (MACD), a training-free framework that assigns agents distinct cultural personas and orchestrates deliberation via a "Seeking Common Ground while Reserving Differences" strategy. Experiments demonstrate that MACD achieves 57.6% average No Bias Rate evaluated by LLM-as-judge and 86.0% evaluated by MAV (vs.\{} 47.6% and 69.0% baseline using GPT-4o as backbone) on CEBiasBench and generalizes to the Arabic CAMeL benchmark, confirming that explicit cultural representation in agent frameworks is essential for cross-cultural fairness.

[Long] Mixture-of-Minds: Multi-Agent Reinforcement Learning for Table Understanding ► *Underline Video*

Yuhang Zhou, Mingrui Zhang, Ke Li, Mingyi Wang, Qiao Liu, Qifei Wang, Jiayi Liu, Fei Liu, Serena Li, WEIWEI LI, Mingze Gao, Abhishek Kumar, Xiangjun Fan, Zhuokai Zhao and Lizhu Zhang

Understanding and reasoning over tables is a critical capability for many real-world applications. Large language models (LLMs) have shown promise on this task, but current approaches remain limited. Fine-tuning based methods strengthen language reasoning; yet they are prone to arithmetic errors and hallucination. In contrast, tool-based methods enable precise table manipulation but rely on rigid schemas and lack semantic understanding. These complementary drawbacks highlight the need for approaches that integrate robust reasoning with reliable table processing. In this work, we propose MIXTURE-OF-MINDS, a multi-agent framework that decomposes table reasoning into three specialized roles: planning, coding, and answering. This design enables each agent to focus on a specific aspect of the task while leveraging code execution for precise table manipulation. Building on this workflow, we introduce a self-improvement training framework that employs Monte Carlo Tree Search (MCTS) rollouts to generate pseudo-gold trajectories and optimize agents with reinforcement learning (RL). Extensive experiments show that MIXTURE-OF-MINDS delivers substantial gains, reaching 62.13% on TableBench and surpassing GPT-o3-mini. These results demonstrate the promise of combining structured multi-agent workflows with RL to advance table understanding.

[Long] Monotonic Scaffolding as a Diagnostic Lens for Legal Reasoning in LLMs ► *Underline Video*

Pedro Calais, Janderson Santos, Anisio Lacera and Wagner Meira Jr.

Modern evaluation of Legal QA systems is shifting from terminal accuracy toward process-aware analyses of model reasoning. We propose a diagnostic framework grounded in monotonic pedagogical scaffolding, where language models receive gold-standard, case-relevant information across stages aligned with the canonical legal framework FIRAC , Facts, Issue, Rules, Application, Conclusion. By strictly adding solution-relevant content at each step, we introduce a controlled monotonic intervention that allows for the evaluation of reasoning trajectories rather than isolated outcomes. This longitudinal design enables the introduction of two transition-based diagnostics: Errors-to-Success (E2S) quantifies the guidance required to reach correctness, while Success-to-Errors (S2E) measures the fragility of that correctness under additional structure. These local patterns define a global robustness criterion termed Stable Accuracy, which

credits a response only if the model maintains correctness throughout all scaffolding stages and enforces a higher bar for correctness by distinguishing sustained reasoning from transient patterns. We instantiate the framework on 3,123 Brazilian Bar Exam questions paired with expert-annotated explanations. Our findings reveal model instability patterns hidden from accuracy-only metrics and demonstrate that terminal accuracy systematically overestimates legal reasoning competence. To test the robustness of our diagnostics, we also evaluate a majority-vote aggregation across multiple reasoning samples, finding that the observed instability patterns persist under this stronger inference setting. Furthermore, principal component analysis indicates that legal domains cluster into distinct regions, suggesting systematic differences in reasoning demands across domains. While focused on the legal domain, our evaluation protocol is generalizable to any task with a staged reasoning structure.

[Findings] More Agents Improve Math Problem Solving but Adversarial Robustness Gap Persists ▶ *Underline Video*

Khashayar Alavi, Zhastay Yeltay, Lucie Flek and Akbar Karimi

When LLM agents work together, they seem to be more powerful than a single LLM in mathematical question answering. However, are they also more robust to adversarial inputs? We investigate this question using adversarially perturbed math questions. These perturbations include punctuation noise with three intensities (10%, 30%, 50%), plus real-world and human-like typos (WikiTypo, R2ATA). Using a unified sampling-and-voting framework (Agent Forest), we evaluate six open-source models (Qwen3-4B/14B, Llama3.1-8B, Mistral-7B, Gemma3-4B/12B) across four benchmarks (GSM8K, MATH, MMLU-Math, MultiArith), with various numbers of agents n = 1,2,5,10,15,20,25. Our findings show that 1) Noise type matters: punctuation noise harm scales with its severity, and the human typos remain the dominant bottleneck, yielding the largest gaps to Clean accuracy and the highest attack success rate (ASR) even with a large number of agents; 2) Collaboration reliably improves accuracy as the number of agents, n, increases, with the largest gains from n=1 to n=5 and diminishing returns beyond n~10. However, the adversarial robustness gap persists regardless of the agent count.

[Long] MORPHOGEN: A Multilingual Benchmark for Evaluating Gender-Aware Morphological Generation ▶ *Underline Video*

Mehul Agarwal, Aditya Aggarwal, Arnav Goel, Medha Hira and Anubha Gupta

While multilingual large language models (LLMs) perform well on high-level tasks like translation and question answering, their ability to handle grammatical gender and morphological agreement remains underexplored. In morphologically rich languages, gender influences verb conjugation, pronouns, and even first-person constructions with explicit and implicit mentions of gender. We introduce MORPHOGEN, a morphologically grounded large-scale benchmark dataset for evaluating gender-aware generation in three typologically diverse grammatically gendered languages: French, Arabic, and Hindi. The core task, GENFORM, requires models to rewrite a first-person sentence in the opposite gender while preserving its meaning and structure. We construct a high-quality synthetic dataset spanning these three languages and benchmark 15 popular multilingual LLMs (2B-70B) on their ability to perform this transformation. Our results reveal significant gaps and interesting insights into how current models handle morphological gender. MORPHOGEN provides a focused diagnostic lens for gender-aware language modeling and lays the groundwork for future research on inclusive and morphology-sensitive NLP.

[Long] Multilingual Idioms in Sentences and Conversations Across High-, Medium-, and Low-Resource Languages ▶ *Underline Video*

Saeed Almheiri, Bilal Elbouardi, Salsabila Zahirah Pranida, Irina Nikishina, Ashwath Rao B, Parameswari Krishnamurthy, Muhammad Cendekia Airlangga, Rifo Ahmad Genadi, Nguyn Phan Gia Bo, Amir Hossein Yari, Hawau Olamide Toyin, Nurdaulet Mukhituly, Mena Attia, Besher Hassan, Ahmad Fathan Hidayatullah, Tatsuki Kuribayashi, Haonan Li, Suma Bhat and Fajri Koto

Idiomatic expressions pose a major challenge for multilingual NLP because their meanings shift between figurative and literal usage, often requiring context for accurate interpretation. Prior work has focused on high-resource languages typically evaluates isolated idiom-meaning questions, overlooking realistic discourse. We introduce MIDI, a multilingual idiom dataset spanning 3 high-, 3 medium-, and 12 low-resource languages, curated by native speakers. Unlike previous datasets, MIDI provides idioms embedded in both sentence-level and conversational contexts, capturing both literal and figurative readings. Benchmarking state-of-the-art models shows that idiom comprehension degrades in low-resource languages and that, in all resource tiers, literal interpretations are substantially harder than figurative ones. Conversational context improves performance but does not eliminate these disparities. Through controlled tests and interventions on hidden representations, we further separate memorization from reasoning, exposing core limitations of current models.

[Findings] Multimodal Item Scoring for Natural Language Recommendation via Gaussian Process Regression with LLM Relevance Judgments ▶ *Underline Video*

Yifan Liu

Natural Language Recommendation (NLRec) generates item suggestions based on the relevance between user-issued NL requests and NL item description passages. Existing NLRec approaches often use Dense Retrieval (DR) to compute item relevance scores from aggregation of inner products between user request embeddings and relevant passage embeddings. However, DR views the request as the sole relevance label, thus leading to a unimodal scoring function centered on the query embedding that is often a weak proxy for query relevance. To better capture the potential multimodal distribution of the relevance scoring function that may arise from complex NLRec data, we propose GPR-LLM that uses Gaussian Process Regression (GPR) with LLM relevance judgments for a subset of candidate passages. Experiments on four NLRec datasets and two LLM backbones demonstrate that GPR-LLM with an RBF kernel, capable of modeling multimodal relevance scoring functions, consistently outperforms simpler unimodal kernels (dot product, cosine similarity), as well as baseline methods including DR, cross-encoder, and pointwise LLM-based relevance scoring by up to 65%. Overall, GPR-LLM provides an efficient and effective approach to NLRec within a minimal LLM labeling budget.

[Findings] Nash-Pruned CredMAS: Dynamic Panel Pruning for VLM-MAS using Nash-based Selection and Doubly-Robust Credits ▶ *Underline Video*

Yijia Fan, Mingyu Liu, Jing Yang, Jian Wang, Keze Wang and Jusheng Zhang

Multi-round Vision-Language Model (VLM) Multi-Agent Systems (MAS) offer powerful reasoning capabilities but suffer from prohibitive costs due to static panel designs, where all N agents communicate at every T round. This approach is fundamentally inefficient, as it ignores the context-dependent and diminishing marginal utility of specific agents. To address this, we propose Nash-CredMAS, an economic framework that transforms agent selection into a dynamic resource allocation game. Unlike heuristic routing or one-time pruning, our method operates in two phases: (1) Offline Causal Value Learning, where we employ a doubly-robust (AIPW) estimator to train a context-aware value function from biased interaction logs, effectively learning the true marginal contribution of agents; and (2) Online Dynamic Auctions, where agents bid for communication slots based on their predicted utility. We formulate the inference-time selection as a submodular maximization problem under budget constraints, theoretically guaranteeing a (1 - 1/e)-approximation of the optimal coalition via a greedy strategy. Empirically, Nash-CredMAS achieves state-of-the-art results on challenging benchmarks, including MMMU and V*-Bench, while reducing token consumption by over 25% compared to static baselines. The system naturally converges to an economic equilibrium where agents actively remain silent when their marginal value does not justify the cost.

[Findings] NeoAraBERT: A Modern Foundation Model for Arabic Embeddings with Diacritics-Aware Tokenization and POS-Targeted Masking ▶ *Underline Video*

Mohamad Ballout, Osama Al Mraikhat, Chadi Abou Chakra, Hadi Hamoud, Qusai Abu Obaida and Fadi Zaraket

We present NeoAraBERT, a state-of-the-art open-source Arabic text-embedding model built on the NeoBERT architecture. We pre-train NeoAraBERT on diverse open-source and internal datasets covering modern standard, classical, and dialectal Arabic. We guided our design choices with Arabic tailored ablation studies including text normalization, light stemming, and diacritics-aware tokenization handling. We also performed more general POS-aware token masking and learning-rate scheduling ablation studies. We benchmarked NeoAraBERT against five top-performing Arabic models on 23 tasks, including a novel synonym-based task, "Muradif", that directly assesses embedding quality with no additional fine-tuning. NeoAraBERT variants (MSA, dialectal, and mixed) rank first in 18 tasks, second in two, third in two, and fourth in one task. They show strong performance on classical and modern standard Arabic, substantial margins of improvement (>7%) in two tasks, and a +2.75% improvement on average across all tasks. Our code and links to checkpoints for our model variants are available on our website: https://acr.ps/neoarabert.

[Findings] Neuronal Insights into LLM Attacks: Targeted Neuron Tuning for Precise and Robust Vulnerability Patching ▶ *Underline Video*

Dan Shi, Renren Jin, Zhuowen Han, Yuqi Ren, Xinwei Wu, Zhigen Li and Deyi Xiong

Despite recent advances in safety alignment, large language models (LLMs) remain highly susceptible to adversarial attacks, while the internal mechanisms behind such vulnerabilities are still poorly understood. Existing gradient-based attribution methods offer valuable interpretability for analyzing information storage and processing in LLMs. However, they are inapplicable to adversarial attacks, which typically occur in open-ended generation settings without fixed ground-truth outputs. To address these challenges, we propose a novel similarity-based gradient attribution method to identify key neurons sensitive to adversarial behaviors in open-ended generation tasks. The detected neurons, termed targeted neurons, play a critical role in safety training. Building on this neuron-level perspective, we uncover two key neuronal patterns: (i) universal neurons that are consistently exploited across multiple attack strategies, and (ii) interference neurons that hinder safety improvements when fine-tuned indiscriminately, providing mechanistic insights into the interpretability of adversarial vulnerabilities. Inspired by these findings, we propose a neuron-level defense strategy, Targeted Neuron Tuning (TNT), which selectively fine-tunes the identified targeted neurons for specific attacks. Experimental evaluations across multiple LLM architectures and scales demonstrate that TNT substantially improves model robustness against a wide range of jailbreak attacks, achieving safe rates exceeding 90% and even approaching 100%, while preserving general task performance, enabling precise and robust safety interventions. Warning: This paper contains example data that may be harmful.

[Long] OLA: Output Language Alignment in Code-Switched LLM Interactions ▶ *Underline Video*

Juhyun Oh, Haneul Yoo, Faiz Ghifari Haznitrama and Alice Oh

Code-switching, alternating between languages within a conversation, is natural for multilingual users, yet poses fundamental challenges for large language models (LLMs). When a user code-switches in their prompt to an LLM, they typically do not specify the expected language of the LLM response, and thus LLMs must infer the output language from contextual and pragmatic cues. We find that current LLMs systematically fail to align with this expectation, responding in undesired languages even when cues are clear to humans. We introduce OLA, a benchmark to evaluate LLMs' Output Language Alignment in code-switched interactions. OLA focuses on Korean–English code-switching and spans simple intra-sentential mixing to instruction-content mismatches. Even frontier models frequently misinterpret implicit language expectation, exhibiting a systematic bias toward non-English responses. We further show this bias generalizes beyond Korean to Chinese and Indonesian pairs. Models also show instability through mid-response switching and language intrusions. Chain-of-Thought prompting fails to resolve these errors, indicating weak pragmatic reasoning about output language. However, Code-Switching Aware DPO with minimal data (~1K examples) substantially reduces misalignment, suggesting these failures stem from insufficient alignment rather than fundamental limitations. Our results highlight the need to align multilingual LLMs with users' implicit expectations in real-world code-switched

interactions.

[Long] OS-Symphony: A Holistic Framework for Robust and Generalist Computer-Using Agents ▶ *Underline Video*

Bowen Yang, Kaiming Jin, Zhenyu Wu, Zhaoyang Liu, Qiushi Sun, Zehao Li, JingJing Xie, Zhoumianze Liu, Fangzhi Xu, Kanzhi Cheng, Yian Wang, Qingyun Li, Yu Qiao, Zun Wang, Zichen DingBowen Yang, Kaiming Jin, Zhenyu Wu, Zhaoyang Liu, Qiushi Sun, Zehao Li, JingJing Xie, Zhoumianze Liu, Fangzhi Xu, Kanzhi Cheng, Yian Wang, Qingyun Li, Yu Qiao, Zun Wang and Zichen Ding

While Vision-Language Models (VLMs) have significantly advanced Computer-Using Agents (CUAs), current agentic frameworks struggle with robustness in novel domains and long-horizon workflows due to the absence of visual-aware tutorial retrieval and the lack of granular control over historical visual context curation and pruning. To bridge these gaps, we introduce OS-Symphony, a holistic framework that comprises an Orchestrator coordinating two key innovations for robust automation: (1) a Reflection-Memory Agent that utilizes milestone-driven long-term memory to enable trajectory-level self-correction, effectively mitigating visual context loss in long-horizon tasks; (2) Versatile Tool Agents featuring a Multimodal Searcher that adopts a SeeAct paradigm to navigate a browser-based sandbox to synthesize live, visually aligned tutorials, thereby resolving fidelity issues in unseen scenarios. Experimental results demonstrate that OS-Symphony delivers substantial performance gains across varying model scales, establishing new state-of-the-art results on three online benchmarks, notably achieving 65.84% on OSWorld. All research assets will be made publicly available.

[Findings] Persona-Assigned Large Language Models Exhibit Human-Like Motivated Reasoning ▶ *Underline Video*

Saloni Dash, Amélie Redmond, Emma Spiro and Aylin Caliskan

Reasoning in humans is prone to biases due to underlying motivations like identity protection, that undermine rational decision-making and judgment. This motivated reasoning at a collective level can be detrimental to society when debating critical issues such as human-driven climate change or vaccine safety, and can further aggravate political polarization. Prior studies have reported that large language models (LLMs) are also susceptible to human-like cognitive biases, however, the extent to which LLMs selectively reason toward identity-congruent conclusions remains largely unexplored. Here, we investigate whether assigning 8 personas across 4 political and socio-demographic attributes induces motivated reasoning in LLMs. Testing 8 LLMs (open source and proprietary) across two reasoning tasks from human-subject studies — veracity discernment of misinformation headlines and evaluation of numeric scientific evidence — we find that persona-assigned LLMs have up to 9% reduced veracity discernment relative to models without personas. Political personas specifically are up to 90% more likely to correctly evaluate scientific evidence on gun control when the ground truth is congruent with their induced political identity. Prompt-based debiasing methods are largely ineffective at mitigating these effects. Taken together, our empirical findings are the first to suggest that persona-assigned LLMs exhibit human-like motivated reasoning that is hard to mitigate through conventional debiasing prompts — raising concerns of exacerbating identity-congruent reasoning in both LLMs and humans.

[Long] Phonemes to the Rescue: Multilingual Tokenization Based on International Phonetic Alphabet ▶ *Underline Video*

Milan Miletic, Julie Kallini and Ekaterina Shutova

Multilingual language models often exhibit performance disparities across languages that can arise as early as the tokenization stage. Widely-used subword tokenization approaches favor high-resource languages, and tokenizer-free methods still yield longer sequences for scripts with a higher bytes-per-character ratio. To address these shortcomings, we propose to use the International Phonetic Alphabet (IPA) as a language-agnostic input representation for multilingual tokenizers. IPA provides a compact symbol inventory, greater cross-lingual character overlap, and a more balanced byte-per-character distribution across languages. We train matched pairs of text vs. IPA subword tokenizers across 24 languages and 14 scripts and demonstrate that IPA tokenizers consistently improve tokenization quality, especially for non-Latin scripts, and generalize more effectively to unseen languages and scripts.

[Findings] PlanE: Meta Planning of Data, Tuning, and Inference for Extractive-based LLMs ▶ *Underline Video*

Jiacheng Wang, Weiyan Zhang and Guangya Yu

Enhancing the task-specific capabilities of Large Language Models (LLMs) primarily requires substantial instruction-tuning datasets. However, the sheer volume of such data imposes a considerable annotation cost, and a lack of optimization methods for tailoring LLMs to specific tasks persists. To address the above issues, we propose a Planning framework for constructing Extractive-based LLMs called PlanE, which includes data decomposition, instruction tuning, and prompt inference. Additionally, we introduce a Data-Tuning-Inference (DTI) planner, aimed at selecting the optimal base-LLM and its DTI combinations for specific datasets to improve construction efficiency. The experimental results demonstrate the effectiveness of our PlanE from two views: (1) across different datasets using the same base-LLM, and (2) on the same dataset using different base-LLMs. Furthermore, we validate the generalizability of the proposed DTI planner under different optimization objectives. The codes are publicly available at https://github.com/gugugu-469/PlanE.

[Long] Polymorphic Universal Transformer ▶ *Underline Video*

Yilong Chen, Zitian Gao, Yihao Xiao, Jason Klein Liu, Xinyu Yang, Yifan Luo, Haoming Luo, Zhengmao Ye, Tingwen Liu, Ran Tao and Bryan Dai

Although the Universal Transformer (UT) mitigates the diminishing returns of standard LLM scaling by decoupling parameter count from depth, it remains constrained by linear computational costs and rigid weight-sharing mechanisms. These limitations lead to severe functional homogeneity, which subsequently induces over-smoothing, representation rank collapse, and degraded reasoning performance. In this work, we present the first systematic study of Compute Distribution Skew, identifying it as the primary driver of extrapolation failure. This is

a pathological phenomenon in ultra-deep recurrent Transformers characterized by a disproportionate distribution of contributions across recurrent steps, resulting in distinct functional states during prefix and suffix processing phases. To address this challenge, we propose the Polymorphic Transformer, which aims to achieve functional polymorphism and depth sparsity within a shared-parameter framework. By integrating conditional sparse subspaces, SiLU Attention, and an uncertainty-aware depth scheduler, our architecture mitigates power-method collapse and effectively decouples logical depth from computational cost. Experiments demonstrate that our model significantly enhances representation rank and robustness, achieving complex reasoning performance comparable to baseline while reducing computation by 64.7%.

[Findings] Position: LLM Watermarking Should Align Stakeholders' Incentives for Practical Adoption ▶ *Underline Video*

Yepeng Liu, Xuandong Zhao, Dawn Song, Gregory W. Wornell and Yuheng Bu

Despite progress in watermarking algorithms for large language models (LLMs), real-world deployment remains limited. We argue that this gap stems from misaligned incentives among LLM providers, platforms, and end users, which manifest as three key barriers: competitive risk, detection-tool governance, and attribution issues. We revisit three classes of watermarking through this lens. Model watermarking naturally aligns with LLM provider interests, yet faces new challenges in open-source ecosystems. LLM text watermarking offers modest provider benefit when framed solely as an anti-misuse tool, but can gain traction in narrowly scoped settings such as dataset de-contamination or user-controlled provenance. In-context watermarking (ICW) is tailored for trusted parties, such as conference organizers or educators, who embed hidden watermarking instructions into documents. If a dishonest reviewer or student submits this text to an LLM, the output carries a detectable watermark indicating misuse. This setup aligns incentives: users experience no quality loss, trusted parties gain a detection tool, and LLM providers remain neutral by simply following watermark instructions. We advocate for a broader exploration of incentive-aligned methods, with ICW as an example, in domains where trusted parties need reliable tools to detect misuse. More broadly, we distill design principles for incentive-aligned, domain-specific watermarking and outline future research directions. Our position is that the practical adoption of LLM watermarking requires aligning stakeholder incentives in targeted application domains and fostering active community engagement.

[Long] PosterForest: Hierarchical Multi-Agent Collaboration for Scientific Poster Generation ▶ *Underline Video*

Jiho Choi, Seojeong Park, Seongjong Song and Hyunjung Shim

Automating scientific poster generation requires hierarchical document understanding and coherent content-layout planning. Existing methods often rely on flat summarization or optimize content and layout separately. As a result, they often suffer from information loss, weak logical flow, and poor visual balance. We present PosterForest, a training-free framework for scientific poster generation. Our method introduces the Poster Tree, a structured intermediate representation that captures document hierarchy and visual-textual semantics across multiple levels. Building on this representation, content and layout agents perform hierarchical reasoning and recursive refinement, progressively optimizing the poster from global organization to local composition. This joint optimization improves semantic coherence, logical flow, and visual harmony. Experiments show that PosterForest outperforms prior methods in both automatic and human evaluations, without additional training or domain-specific supervision.

[Findings] PsychEthicsBench: Evaluating Large Language Models Against Australian Mental Health Ethics ▶ *Underline Video*

Yaling Shen, Stephanie Fong, Yiwen Jiang, Zimu Wan, Feilong Tang, Qingyang Xu, Xiangyu Zhao, Zhongxing Xu, Jiahe Liu, Jinpeng Hu, Dominic Dwyer and Zongyuan Ge

The increasing integration of large language models (LLMs) into mental health applications necessitates robust frameworks for evaluating professional safety alignment. Current evaluative approaches primarily rely on refusal-based safety signals, which offer limited insight into the nuanced behaviors required in clinical practice. In mental health, clinically inadequate refusals can be perceived as unempathetic and discourage help-seeking. To address this gap, we move beyond refusal-centric metrics and introduce PsychEthicsBench, the first principle-grounded benchmark based on Australian psychology and psychiatry guidelines, designed to evaluate LLMs' ethical knowledge and behavioral responses through multiple-choice and open-ended tasks with fine-grained ethicality annotations. Empirical results across 14 models reveal that refusal rates are poor indicators of ethical behavior, revealing a significant divergence between safety triggers and clinical appropriateness. Notably, we find that domain-specific fine-tuning can degrade ethical robustness, as several specialized models underperform their base backbones in ethical alignment. PsychEthicsBench provides a foundation for systematic, jurisdiction-aware evaluation of LLMs in mental health, encouraging more responsible development in this domain.

[Findings] PURE: Post-hoc Unlocking and REfinement for Discrete Diffusion Decoding ▶ *Underline Video*

Yangryeol Park, Kunhui Lee, Hanback Choi, Cheoneum Park, Donghyeon Jeon, Inho Kang and Seung-Hoon Na

Masked diffusion language models (MDLMs) enable efficient parallel decoding but are limited by a monotonic unmasking policy, where committed tokens cannot be revised. While remasking-based methods mitigate early errors, they mainly intervene during generation. In this work, we study post-hoc refinement of a completed draft and find that naive correction often fails because of contextual lock-in, a phenomenon in which local error patterns become self-reinforcing. To address this, we propose PURE (Post-hoc Unlocking and REfinement), a training-free inference algorithm for two-phase decoding. PURE profiles confidence dynamics during drafting to identify unstable regions via an instability score (_i), then unlocks them through deterministic window masking and stochastic leftward relaxation. On reasoning benchmarks, PURE substantially improves accuracy when applied to LLaDA-8B-Instruct, including a gain of +12.9 points over the baseline on GSM8K. These gains require only a small refinement budget, yielding a favorable compute-quality trade-off for discrete diffusion decoding.

[Findings] RAG or Learning? Understanding the Limits of LLM Adaptation under Continuous Knowledge Drift in the Real World ► *Underline Video*

Hanbing Liu, Lang Cao and Yang Li

Large language models (LLMs) acquire most of their knowledge during pretraining, which ties them to a fixed snapshot of the world and makes adaptation to continuously evolving knowledge challenging. As facts, entities, and events change over time, models may experience continuous knowledge drift, resulting not only in outdated predictions but also in temporally inconsistent reasoning. Although existing approaches, such as continual finetuning, knowledge editing, and retrieval-augmented generation (RAG), aim to update or supplement model knowledge, they are rarely evaluated in settings that reflect chronological, evolving, and real-world knowledge evolution. In this work, we introduce a new benchmark of real-world dynamic events, constructed from time-stamped evidence that captures how knowledge evolves over time, which enables systematic evaluation of model adaptation under continuous knowledge drift. The benchmark reveals that most existing methods, including vanilla RAG and several learning-based approaches, struggle under this setting, exposing critical limitations such as catastrophic forgetting and temporal inconsistency. To mitigate these limitations, we propose a time-aware retrieval baseline, Chronos, which progressively organizes retrieved evidence into an Event Evolution Graph to enable more temporally consistent understanding in LLMs without additional training. Overall, this work provides a foundation for analyzing and advancing LLM adaptation to continuous knowledge drift in realistic settings.

[Findings] Read the Room, Read the Image: Understanding Indirect Speech Acts in Multimodal Visual Contexts ► *Underline Video*

Jaehee Kim, Ji Hoon Chung, Seoyoon Park, Unsol Kim, Kyungwon Park, JiHak Kim, Yi-Jun Chen and Hansaem Kim

Indirect speech acts (ISAs) require pragmatic reasoning over context, as directive intent cannot be inferred from surface form alone. Prior text-based studies and existing multimodal benchmarks largely overlook this requirement, focusing instead on explicitly encoded context or perceptual recognition, and thus underexplore context-dependent pragmatic understanding, particularly in high-context languages such as Korean. We introduce READI, a multimodal benchmark for evaluating ISA understanding through integrated reasoning over visual context and dialogue. READI models graded indirectness grounded in pragmatic theory and formulates the task as vision-based pragmatic question answering (V-PQA), supporting cross-lingual evaluation in English and Korean. Experiments show that even state-of-the-art multimodal models struggle with visually grounded indirect speech acts, with performance declining as indirectness increases, underscoring the need for benchmarks that explicitly target contextual pragmatic reasoning.

[Long] ReasMark: A Robust Watermark for Attributing LLM Reasoning Under Knowledge Distillation Attacks ► *Underline Video*

Peizhuo Lv, Ruihua Zhou, Yunpeng Li, Ruigang Liang, Xingshuo Han, XiaoFeng Wang, Wei Dong and Yuling Liu

Reasoning-enhanced large language models rely on intermediate reasoning signals to solve complex, multi-step tasks, making reasoning behavior a valuable form of intellectual property. Meanwhile, knowledge distillation enables an adversary to replicate this behavior in a realistic black-box setting by repeatedly querying a deployed model on a target domain and training a local student to imitate its outputs, including reasoning traces. Existing LLM watermarks primarily operate on surface text and decoding-time token biases, and thus fail to provide reliable attribution of reasoning behavior once it is transferred through knowledge distillation. ReasMark entangles the watermark with the target-domain input distribution by selecting watermark tokens from high-frequency prompts, so distillation queries naturally activate it. It then embeds the watermark by score-conditioned losses that create a detectable reasoning-length gap for black-box verification. Comprehensive experiments across multiple LLMs, datasets, and distillation settings demonstrate that ReasMark consistently outperforms existing baselines while preserving task utility.

[Long] Reasoning Over Space: Enabling Geographic Reasoning for LLM-Based Generative Next POI Recommendation ► *Underline Video*

Dongyi Lv, Qiuyu Ding, Heng-Da Xu, Zhaoxu Sun, Zhi Wang, Feng Xiong and Mu Xu

Generative recommendation with large language models (LLMs) reframes prediction as sequence generation, yet existing LLM-based recommenders remain limited in leveraging geographic signals that are crucial in mobility and local-services scenarios. Here, we present Reasoning Over Space (ROS), a framework that utilizes geography as a vital decision variable within the reasoning process. ROS introduces a Hierarchical Spatial Semantic ID (SID) that discretizes coarse-to-fine locality and POI semantics into compositional tokens, and endows LLM with a three-stage Mobility Chain-of-Thought (CoT) paradigm that models user personality, constructs an intent-aligned candidate space, and performs locality informed pruning. We further align the model with real world geography via spatial-guided Reinforcement Learning (RL). Experiments on three widely used location-based social network (LBSN) datasets show that ROS achieves over 10% relative gains in hit rate over strongest LLM-based baselines and improves cross-city transfer, despite using a smaller backbone model.

[Findings] Recovered in Translation: Efficient Pipeline for Automated Translation of Benchmarks and Datasets ► *Underline Video*

Hanna Yukhymenko, Anton Alexandrov and Martin Vechev

The reliability of multilingual Large Language Model (LLM) evaluation is currently compromised by the inconsistent quality of translated benchmarks. Existing resources often suffer from semantic drift and context loss, which can lead to misleading performance metrics. In this work, we present a fully automated framework designed to address these challenges by enabling scalable, high-quality translation of datasets and benchmarks. We demonstrate that adapting test-time compute scaling strategies, specifically Universal Self-Improvement (USI) and our proposed multi-round ranking method, T-RANK, allows for significantly higher quality outputs compared to traditional pipelines. Our framework ensures that benchmarks preserve their original task structure and linguistic nuances during localization. We apply this approach

to translate popular benchmarks and datasets into eight Eastern and Southern European languages (Ukrainian, Bulgarian, Slovak, Romanian, Lithuanian, Estonian, Turkish, Greek). Evaluations using both reference-based metrics and LLM-as-a-judge show that our translations surpass existing resources, resulting in more accurate downstream model assessment. We release both the framework and the improved benchmarks to facilitate robust and reproducible multilingual AI development.

[Long] RedCoder: Automated Multi-Turn Red Teaming for Code LLMs ▶ *Underline Video*

Wenjie Jacky Mo, Qin Liu, Xiaofei Wen, Dongwon Jung, Hadi Askari, Wenxuan Zhou, Zhe Zhao and Muhao Chen

Large Language Models (LLMs) for code generation (i.e., Code LLMs) have demonstrated impressive capabilities in AI-assisted software development and testing. However, recent studies have shown that these models are prone to generating vulnerable or even malicious code under adversarial settings. Existing red-teaming approaches rely on extensive human effort, limiting their scalability and practicality, and generally overlook the interactive nature of real-world AI-assisted programming, which often unfolds over multiple turns. To bridge these gaps, we present RedCoder, a red-teaming agent that engages victim models in multi-turn conversation to elicit vulnerable code. The pipeline to construct RedCoder begins with a multi-agent gaming process that simulates adversarial interactions, yielding a set of prototype conversations and an arsenal of reusable attack strategies. We then fine-tune an LLM on these prototype conversations to serve as the backbone of RedCoder. Once deployed, RedCoder autonomously engages Code LLMs in multi-turn conversations, dynamically retrieving relevant strategies from the arsenal to steer the dialogue toward vulnerability-inducing outputs. Experiments across multiple Code LLMs show that our approach outperforms prior single-turn and multi-turn red-team methods in inducing vulnerabilities in code generation, offering a scalable and effective tool for evaluating the security boundaries of modern code-generation systems.

[Short] Reliable Evaluation Protocol for Low-Precision Retrieval ▶ *Underline Video*

Kisu Yang, Yoonna Jang, Hwanseok Jang, Kenneth Choi, Isabelle Augenstein and Heuiseok Lim

Lowering the numerical precision of model parameters and computations is widely adopted to improve the efficiency of retrieval systems. However, when computing relevance scores between the query and documents in low-precision, we observe spurious ties due to the reduced granularity. This introduces high variability in the results based on tie resolution, making the evaluation less reliable. To address this, we propose a more robust retrieval evaluation protocol designed to reduce score variation. It consists of: (1) High-Precision Scoring (HPS), which upcasts the final scoring step to higher precision to resolve tied candidates with minimal computational cost; and (2) Tie-aware Retrieval Metrics (TRM), which report expected scores, range, and bias to quantify order uncertainty of tied candidates. Our experiments test multiple models with three scoring functions on twelve retrieval datasets to demonstrate that HPS dramatically reduces tie-induced instability, and TRM accurately recovers expected metric values. This combination enables a more consistent and reliable evaluation system for lower-precision retrieval.

[Findings] Rerank Before You Reason: Analyzing Reranking Tradeoffs through Effective Token Cost in Deep Search Agents ▶ *Underline Video*

Sahel Sharifymoghaddam and Jimmy Lin

Deep research agents rely on iterative retrieval and reasoning to answer complex queries, but scaling test-time computation raises significant efficiency concerns. We study how to allocate reasoning budget in deep search pipelines, focusing on the role of listwise reranking. Using the BrowseComp-Plus benchmark, we analyze tradeoffs between model scale, reasoning effort, reranking depth, and total token cost via a novel effective token cost (ETC) metric. Our results show that reranking consistently improves retrieval and end-to-end accuracy, and that moderate reranking often yields larger gains than increasing search-time reasoning, achieving comparable accuracy at substantially lower cost. All our code is available at https://github.com/sahel-sh/DeepHone.

[Findings] Rethinking Assessments of Prompt Injection Attacks ▶ *Underline Video*

Chi Cui, Yixin Wu, Michael Backes and Yang Zhang

Prompt injection attacks are recognized as one of the primary risks faced by LLM-integrated applications in recent years. However, common evaluation frameworks remain insufficient, lacking comprehensiveness and real-world relevance. To bridge this gap, we revisit the common evaluation framework and conduct an extensive evaluation across eight different evaluation settings, including 37 real-world applications, 185 injected tasks, 21 attack instructions, and a total of 143,745 queries. The evaluation highlights several findings. For example, real-world applications are more vulnerable to prompt injection attacks compared to those used in research settings. While complex attack instructions are more sophisticated, they are less effective than simple attack instructions. We further conduct an assessment of both prompt-level and model-level defense mechanisms and highlight their limitations in real-world applications. By exploring more diverse scenarios across different dimensions, our framework provides a solid foundation for assessing vulnerabilities in LLM-integrated applications and evaluating the efficacy of defensive strategies.

[Long] Retrieval-Augmented Defense: Adaptive and Controllable Jailbreak Prevention for Large Language Models ▶ *Underline Video*

Guangyu Yang, Jinghong Chen, Jingbiao Mei, Weizhe Lin and Bill Byrne

Large Language Models (LLMs) remain vulnerable to jailbreak attacks, which attempt to elicit harmful responses from LLMs. The evolving nature and diversity of these attacks pose many challenges for defense systems, including (1) adaptation to counter emerging attack strategies without costly retraining, and (2) control of the trade-off between safety and utility. To address these challenges, we propose

Retrieval-Augmented Defense (RAD), a novel framework for jailbreak detection that incorporates a database of known attack examples into Retrieval-Augmented Generation, which is used to infer the underlying, malicious user query and jailbreak strategy used to attack the system. RAD enables training-free updates for newly discovered jailbreak strategies and provides a mechanism to balance safety and utility. Experiments on StrongREJECT show that RAD substantially reduces the effectiveness of strong jailbreak attacks such as PAP and PAIR while maintaining low rejection rates for benign queries. We propose a novel evaluation scheme and show that RAD achieves a robust safety-utility trade-off across a range of operating points in a controllable manner.

[Findings] Retrieve Only Relevant Tables Whether Few or Many: Adaptive Table Retrieval Method ► *Underline Video*

Taehee Kim, Seungbin Yang, Jihwan Kim and Jaegul Choo

Retrieving relevant tables from extensive databases for a given natural language query is essential for accurately answering questions in tasks such as text-to-SQL. Existing table retrieval approaches select a pre-determined set of k tables with the highest similarity to the query. However, the number of required tables varies across queries and cannot be known in advance. Enforcing a fixed number of retrieved tables regardless of the query may either retrieve an undersized set, failing to obtain all necessary evidence, or retrieve an oversized pool, including irrelevant tables. To address this issue, we propose an adaptive table retrieval method that adjusts the number of tables retrieved according to the requirements of each query. Specifically, we utilize an adaptive thresholding mechanism to selectively retrieve tables and integrate a sliding-window reranking algorithm to efficiently process a large table corpus. Extensive experiments on Spider, BIRD, and Spider 2.0 demonstrate that our method effectively addresses the limitations of the top-k retrieval strategy, improving performance in retrieval and downstream tasks. Our code and data are available at https://anonymous.4open.science/r/Adaptive-Table-Retrieval.

[Findings] Revisiting the Uniform Information Density Hypothesis in LLM Reasoning ► *Underline Video*

Minju Gwak, Guijin Son and Jaehyung Kim

The Uniform Information Density (UID) hypothesis proposes that effective communication is achieved by maintaining a stable flow of information. In this work, we revisit this principle in the context of Large Language Model (LLM) reasoning, asking whether step-level uniformity reflects reasoning quality. To this end, we introduce a novel framework to quantify uniformity of information flow at both local and global levels, using an entropy-based stepwise density metric. Across experiments on seven reasoning benchmarks, we see a counter-intuitive pattern: while high-quality reasoning exhibit smooth step-by-step transitions (local uniformity) and structured, non-uniform information flow at the trajectory level (global non-uniformity). The results demonstrate that these uniformities outperform alternative internal signals as predictors of reasoning quality, and such divergence with human communication is not a model deficiency, but a byproduct of distinct objectives between human communication and LLM reasoning.

[Long] Right at My Level: A Unified Multilingual Framework for Proficiency-Aware Text Simplification ► *Underline Video*

Jinhong Jeong, Junghun Park and Youngjae Yu

Text simplification supports second language (L2) learning by providing comprehensible input, consistent with the Input Hypothesis. However, constructing personalized parallel corpora is costly, while existing large language model (LLM)-based readability control methods rely on pre-labeled sentence corpora and primarily target English. We propose Re-RIGHT, a unified reinforcement learning framework for adaptive multilingual text simplification without parallel corpus supervision. We first show that prompting-based lexical simplification at target proficiency levels (CEFR, JLPT, TOPIK, and HSK) performs poorly at easier levels and for non-English languages, even with state-of-the-art LLMs such as GPT-5.2 and Gemini 2.5. To address this, we collect 43K vocabulary-level data across four languages (English, Japanese, Korean, and Chinese) and train a compact 4B policy model using Re-RIGHT, which integrates three reward modules: vocabulary coverage, semantic preservation, and coherence. Compared to the stronger LLM baselines, Re-RIGHT achieves higher lexical coverage at target proficiency levels while maintaining original meaning and fluency.

[Long] RIMRULE: Improving Tool-Using Language Agents via MDL-Guided Rule Learning ► *Underline Video*

Xiang Gao, Yuguang Yao, Qi Zhang, Kaiwen Dong, Avinash Baidya, Ruocheng Guo, Hilaf Hasson and Kamalika Das

Large language models (LLMs) often struggle to use tools reliably in domain-specific settings, where APIs may be idiosyncratic, under-documented, or tailored to private workflows. This highlights the need for effective adaptation to task-specific tools. We propose RIMRULE, a neuro-symbolic approach for LLM adaptation based on dynamic rule injection. Compact, interpretable rules are distilled from failure traces and injected into the prompt during inference to improve task performance. These rules are proposed by the LLM itself and consolidated using a Minimum Description Length (MDL) objective that favors generality and conciseness. Each rule is stored in both natural language and a structured symbolic form, supporting efficient retrieval at inference time. Experiments on tool-use benchmarks show that this approach improves accuracy on both seen and unseen tools without modifying LLM weights. It outperforms prompting-based adaptation methods and complements finetuning, offering an interpretable layer of inference-time generalization. Moreover, rules learned from one LLM can be reused to improve others, including long reasoning LLMs, highlighting the portability of symbolic knowledge across architectures.

[Findings] Robust Uncertainty Quantification for Self-Evolving Large Language Models via Continual Domain Pretraining ► *Underline Video*

Xiaofan Zhou and Lu Cheng

Continual Learning (CL) is essential for enabling self-evolving large language models (LLMs) to adapt and remain effective amid rapid

knowledge growth. Yet, despite its importance, little attention has been given to establishing statistical reliability guarantees for LLMs under CL, particularly in the setting of continual domain pretraining (CDP). Conformal Prediction (CP) has shown promise in offering correctness guarantees for LLMs, but it faces major challenges in CDP: testing data often stems from unknown or shifting domain distributions, under which CP may no longer provide valid guarantees. Moreover, when high coverage is required, CP can yield excessively large prediction sets for unanswerable queries, reducing informativeness. To address these challenges, we introduce an adaptive rejection and non-exchangeable CP framework. Our method first estimates the distribution of questions across domains in the test set using transformer-based clustering, then reweights or resamples the calibration data accordingly. Building on this, adaptive rejection CP allows the LLM to selectively abstain from answering when its confidence or competence shifts significantly. Extensive experiments demonstrate that our framework enhances both the effectiveness and reliability of CP under CDP scenarios. Our code is available at: https://github.com/AlearZhou/CPCL

[Long] RSMeM: Knowledge-Enhanced Memory Evolution for Remote Sensing Agents with Systematic Evaluation ▶ *Underline Video*

Bingxian Wu, Yu Zhang, Zonghao Guo, Tang Liu, Chen Qian, Yuxiang Lu, Xingbo Du, Yanghao Li, Yidan Zhang, Chi Chen, Ling Yao, Maosong Sun and Chenghu Zhou

Geoscience research requires complex analysis and domain expertise, with remote sensing (RS) observations as a key foundation. However, existing RS agents built on general-purpose LLMs remain largely domain-agnostic, resulting in brittle and error-prone workflows. Moreover, these failures are seldom consolidated into a reusable experience for subsequent analyses. To address this issue, we introduce RSMeM, a knowledge-enhanced memory evolution mechanism that bootstraps RS agents with pre-distilled domain knowledge and iteratively integrates online experience for robust multi-step tool execution. RSMeM is composed of two components: (i) Hierarchical Knowledge Grounding, which performs taxonomy-aware retrieval over a hierarchical domain corpus to guide planning and tool selection; and (ii) Failure-Aware Experience Refinement, which distills failure-annotated tool-use traces into reusable constraints for next-round tool execution. By iteratively employing these two processes, RS agents can evolve to absorb task-level domain knowledge and effectively translate it into instance-level execution experience. Extensive experiments on EarthBench demonstrate that RSMeM consistently improves tool-use performance and end-to-end accuracy across a diverse set of LLM backbones. Notably, RSMeM achieves a 6% accuracy improvement on DeepSeek-V3.2 with less than 1% additional experience tokens, demonstrating the knowledge density of our distilled experience. All codes and models will be released to support reproducible research.

[Findings] Safety Is Not Universal: The Selective Safety Trap in LLM Alignment ▶ *Underline Video*

Iago Alves Brito, Walcy Rios, Julia Soares Dollis, Diogo Fernandes Costa Silva and Arlindo Rodrigues Galvão Filho

Current safety evaluations of large language models (LLMs) create a dangerous illusion of universal protection by aggregating harms under generic categories such as "Identity Hate", obscuring vulnerabilities toward specific populations. In this work, we expose the Selective Safety Trap: a systemic failure mode where models robustly defend specific populations while leaving underrepresented communities highly vulnerable to identical adversarial attacks. To systematically audit this phenomenon, we introduce MiJaBench, a bilingual (English-Portuguese) adversarial benchmark comprising 43,961 controlled jailbreaking prompts across 16 minority groups. By evaluating 14 state-of-the-art LLMs on MiJaBench, we curate 615,454 prompt-response pairs that compose MiJaBench-Align, revealing that safety alignment is not a uniform semantic capability but a demographic hierarchy, with defense rates fluctuating by up to 42% within the same model solely based on the target group. This disparity persists across architectures and languages and is amplified by scaling, indicating that current alignment methods learn group-specific safeguards rather than a generalized notion of harm. Through targeted direct preference optimization (DPO) on a 1B-parameter baseline, we achieve strong zero-shot safety generalizations to entirely unseen demographics and complex attack strategies. We release all datasets and scripts to provide the community with a concrete pathway toward equitable, transferable safety alignment.

[Findings] Scaling Performance and Low-Resource Annotation with Many-Shot In-Context Learning for Named Entity Recognition ▶ *Underline Video*

Qi Zhang, Fangping Lan, Cornelia Caragea, Longin Jan Latecki and Eduard Dragut

In-context learning (ICL) with large language models (LLMs) has emerged as a powerful alternative to fine-tuning for Named Entity Recognition (NER), achieving strong performance with minimal annotation and no additional training. However, prior work has shown that despite their adaptability, LLMs still lag behind fully supervised models such as fine-tuned BERT in structured tasks like NER. While existing studies on ICL for NER have mainly explored few-shot settings, the potential of scaling to hundreds of demonstrations has not been thoroughly investigated. To address this gap, we conduct a comprehensive investigation of many-shot ICL for NER and further explore its effectiveness in annotating and refining data for low-resource NER tasks. Specifically, we evaluate various LLMs across multiple domains using hundreds of ICL examples and then assess the feasibility of using many-shot ICL as a data annotation framework. Our experiments demonstrate that: (1) scaling to hundreds of in-context examples enables LLMs to match or even surpass the performance of fully supervised BERT models; and (2) using about one hundred human-labeled examples as demonstrations, many-shot in-context annotation can generate high-quality labeled data, leading to approximately 10% absolute F1 improvement over existing state-of-the-art approaches when used to fine-tune BERT on low-resource NER.

[Long] Selective Test-Time Debiasing for CLIP via Reward Gating ▶ *Underline Video*

Jaeho Han, Jisoo Yang, Hyeondong Woo, Mingyu Jeon, Sunjae Yoon and Junyeong Kim

Vision language models (VLMs) demonstrate strong zero-shot performance, but often perpetuate social stereotypes in person-centric queries,

yielding skewed demographic distributions. Current debiasing methods apply uniform bias corrections across all input queries regardless of their bias sensitivity, creating a fundamental fairness-utility trade-off. Strong debiasing distorts semantically meaningful information in bias-insensitive queries, while weak debiasing fails to mitigate stereotypes in bias-sensitive ones. This one-size-fits-all approach hampers simultaneously achieving high utility on bias-insensitive queries and fairness on bias-sensitive queries. We introduce Reward-Gated Test-Time Adaptation (RG-TTA), a reinforcement learning-based test-time adaptation framework that selectively applies debiasing based on input sensitivity. RG-TTA adaptively triggers fairness regularization based on the bias sensitivity of each input during test-time policy adaptation, while focusing exclusively on optimizing cross-modal alignment for bias-insensitive inputs. Experiments on fairness benchmarks (e.g., FairFace, UTKFace) demonstrate substantial bias reduction while simultaneously improving zero-shot utility, resolving the trade-off of uniform debiasing.

[Findings] Semantic-pragmatic Annotations in the Prague Dependency Treebank ▶ *Underline Video*

Marie Mikulova, Eva Hajicova, Jiri Mirovsky, Anja Nedoluzhko, Michal Novak, Pavlina Synkova, Jan Stepanek, Barbora Stepankova and Jan Hajic

We present semantic-pragmatic specification and annotation (ellipsis, coreference, bridging and discourse relations, information structure, scope of negation) in the multi-layer, genre-diversified, 3+ million-token Prague Dependency Treebank - Consolidated 2. 0. While morphology and syntax work almost exclusively on sentence level, the semantic-pragmatic phenomena are often related to two or more neighbouring sentences and possibly to an extra-linguistic context. In the contribution, we describe these phenomena from both the linguistic perspective (form of expression, relation to syntax and morphology) and the cognitive perspective (relation to context, real world knowledge, as well as to the related processes such as thinking or reasoning) - classifying the possible relations between the semantic-pragmatic units into cognitively plausible, distinguishable, and human-understandable categories. We have applied our results to the corpus, by annotating it in its entirety. The resulting dataset is publicly and freely available, to serve for verification and further investigation of (not only) these phenomena.

[Long] SHARP: Self-adaptive Harmful Category-aware Prompt Generation for Black-box Jailbreaking ▶ *Underline Video*

Yingjie Xue, Xingyou Xia, Jun Zhang, Yunbo Cao, Dengpan Ye, Guotong Geng and Fei Li

Large Language Models (LLMs) have been widely applied in various domains such as education and healthcare, making safety assurance crucial. Jailbreak attacks, a method used in red-teaming, can help evaluate and improve the defensive strategies of LLMs. However, existing jailbreak methods often overlook the semantic differences across categories of harmful questions, leading to inconsistent success rates and reduced overall attack effectiveness. We propose the first category-aware jailbreak framework, SHARP, which incorporates the semantic category of harmful questions into prompt generation. Trained on a verified jailbreak dataset, SHARP enables the model to learn category-specific semantic features and adaptively generate prompts that bypass safety mechanisms. The method combines two-stage LoRA fine-tuning, and DPO-based reinforcement learning to optimize both attack success and category alignment. Experiments show that SHARP significantly improves attack success rates and achieves better cross-category robustness compared to the state-of-the-art (SOTA) baselines, providing an efficient and scalable tool for evaluating LLM safety.

[Findings] SLIM: Stealthy Low-Coverage Black-Box Watermarking via Latent-Space Confusion Zones ▶ *Underline Video*

Hengyu WU and Yang Cao

Training data is a critical and often proprietary asset in Large Language Model (LLM) development, motivating the use of data watermarking to embed model-transferable signals for usage verification. We identify low coverage as a vital yet largely overlooked requirement for practicality, as individual data owners typically contribute only a minute fraction of massive training corpora. Prior methods fail to maintain stealthiness, verification feasibility, or robustness when only one or a few sequences can be modified. To address these limitations, we introduce SLIM, a framework enabling per-user data provenance verification under strict black-box access. SLIM leverages intrinsic LLM properties to induce a Latent-Space Confusion Zone by training the model to map semantically similar prefixes to divergent continuations. This manifests as localized generation instability, which can be reliably detected via hypothesis testing. Experiments demonstrate that SLIM achieves ultra-low coverage capability, strong black-box verification performance, and great scalability while preserving both stealthiness and model utility, offering a robust solution for protecting training data in modern LLM pipelines.

[Long] SOCIA-EVO: Automated Simulator Construction via Dual-Anchored Bi-Level Optimization ▶ *Underline Video*

Yuncheng Hua, Sion Weatherhead, Mehdi Jafari, Hao Xue and Flora D. Salim

Automated simulator construction requires distributional fidelity, distinguishing it from generic code generation. We identify two failure modes in long-horizon LLM agents: contextual drift and optimization instability arising from conflating structural and parametric errors. We propose SOCIA-EVO, a dual-anchored evolutionary framework. SOCIA-EVO introduces: (1) a static blueprint to enforce empirical constraints; (2) a bi-level optimization to decouple structural refinement from parameter calibration; and (3) a self-curating Strategy Playbook that manages remedial hypotheses via Bayesian-weighted retrieval. By falsifying ineffective strategies through execution feedback, SOCIA-EVO achieves robust convergence, generating simulators that are statistically consistent with observational data. SOCIA-EVOs code and data are available here: https://github.com/cruiseresearchgroup/SOCIA/tree/evo.

[Findings] StructBreak: Structural Cognitive Overload-Induced Safety Failures in MLLMs ▶ *Underline Video*

Yang Luo, Liu Xinran, TianTian Ji, Zhiyi yin, Lingyun Peng and Shuyu Li

Multimodal Large Language Models (MLLMs) excel at structural reasoning yet suffer from a sharp logical brittleness in structural

consistency. We term this phenomenon Structural Cognitive Overload (SCO), a byproduct of the contention between deep reasoning and safety alignment. However, prior work has predominantly targeted typographic and pixel-level perturbations, leaving the study of SCO largely unexplored. To this end, we propose StructBreak, an automated end-to-end framework designed to quantify SCO. By leveraging StructBreak, we uncover a novel higher-order cognitive overload attack paradigm; notably, this attack operates under a practical black-box setting, requiring no internal model access. Consequently, we utilize this framework to establish a comprehensive benchmark spanning ten diverse threat scenarios. Empirical evaluations on six leading MLLMs reveal that SCO readily triggers toxic generation, yielding a 92% average ASR (up to 97% on Gemini 2.5). To elucidate the mechanism of SCO, we further conduct model-level interpretations spanning attention dynamics, latent space topology, and geometric analysis. Our findings reveal that StructBreak acts as a novel structural channel to circumvent safety filters. Furthermore, the limited efficacy of inherent safety mechanisms underscores that current alignment paradigms are insufficient for the era of complex multimodal reasoning.

[Findings] Structure-Aware Quantized Retrieval for Long-Document Question Answering ▶ *Underline Video*

Hui Huang, Julien Velcin and Yacine Kessaci

Long-document question answering is challenging because relevant evidence is often scattered across distant sections. Traditional long-document QA/RAG pipelines often suffer from context fragmentation, retrieving locally plausible but structurally misaligned passages. We present the Hierarchical Quantized Document Retriever (HQDR), a framework that aligns hierarchical graph representations with a universal token vocabulary and integrates explicit structure into retrieval. By grounding continuous structural features in a fixed, discrete semantic space, HQDR captures universal hierarchical patterns rather than overfitting to specific layouts. We further propose a hybrid scoring mechanism that decouples semantic matching from structural alignment. Extensive experiments on QASPER and Natural Questions demonstrate that HQDR achieves consistent gains over strong baselines and exhibits superior robustness when transferring between datasets with distinct structural characteristics.

[Findings] Structured Uncertainty guided Clarification for LLM Agents ▶ *Underline Video*

Manan Suri, Puneet Mathur, Nedim Lipka, Franck Dernoncourt, Ryan A. Rossi and Dinesh Manocha

LLM agents with tool-calling capabilities often fail when user instructions are ambiguous or incomplete, leading to incorrect invocations and task failures. Existing approaches operate in unstructured language spaces, generating clarifying questions through prompting strategies that lack principled criteria for determining which questions to ask and when to stop. We introduce a principled formulation of structured uncertainty that operates directly over tool parameters and their domains, cleanly separating specification uncertainty (what the user wants) from model uncertainty (what the LLM predicts). Our formulation uses Expected Value of Perfect Information (EVPI) to quantify the disambiguation value of each potential question, balanced against aspect-based cost modeling that prevents redundant questioning. We demonstrate the versatility of this formulation through two applications. First, SAGE-Agent uses structured uncertainty for inference-time question selection, achieving 7–39% higher coverage on ambiguous tasks while reducing clarification questions by 1.5–2.7 x compared to strong prompting and uncertainty-based baselines. Second, we show that structured uncertainty provides effective training signals: uncertainty-guided reward modeling boosts When2Call accuracy from 36.5% to 65.2% (3B model) and 36.7% to 62.9% (7B model) through uncertainty-weighted GRPO training, demonstrating more sample-efficient reinforcement learning for tool-calling agents. To enable evaluation, we present ClarifyBench, the first multi-turn dynamic tool-calling disambiguation benchmark. Our results establish structured uncertainty as a principled framework that improves both inference-time interaction efficiency and training-time sample efficiency in tool-augmented agents.

[Long] Success and Cost Elicit Convention Formation for Efficient Communication ▶ *Underline Video*

Saujas Vaduguru, Yilun Hua, Yoav Artzi and Daniel Fried

Humans leverage shared conversational context to become increasingly successful and efficient at communicating over time. One manifestation of this is the formation of ad hoc linguistic conventions, which allow people to coordinate on short, less costly utterances that are understood using shared conversational context. We present a method to train large multimodal models to form conventions, enabling efficient communication. Our approach uses simulated reference games between models, and requires no additional human-produced data. In repeated reference games involving photographs and tangram images, our method enables models to communicate efficiently with people: reducing the message length by up to 41% while increasing success by 15% over the course of the interaction. Human listeners respond faster when interacting with our model that forms conventions. We also show that training based on success or cost alone is insufficient - both are necessary to elicit convention formation.

[Findings] Supplement Generation Training for Enhancing Agentic Task Performance ▶ *Underline Video*

Young Min Cho, Daniele Bonadiman, Divya Bhargavi, Tamer Alkhouli, Salvatore Romeo, Dongwei Jiang, Khushbu Pahwa, Yubin Ge, Etsuko Ishii, Monica Sunkara and Yi Zhang

Training large foundation models for agentic tasks is increasingly impractical due to the high computational costs, long iteration cycles, and rapid obsolescence as new models are continuously released. Instead of post-training massive models for every new task or domain, we propose Supplement Generation Training (SGT), a more efficient and sustainable strategy. SGT trains a smaller LLM to generate useful supplemental text that, when appended to the original input, helps the larger LLM solve the task more effectively. These lightweight models can dynamically adapt supplements to task requirements, improving performance without modifying the underlying large models. This approach decouples task-specific optimization from large foundation models and enables more flexible, cost-effective deployment of LLM-powered agents in real-world applications.

[Findings] SWE-QA-Pro: A Representative Benchmark and Scalable Training Recipe for Repository-Level Code Understanding ▶ *Underline Video*

Songcheng Cai, Zhiheng Lyu, Yuansheng Ni, Xiangchao Chen, Baichuan Zhou, Shenzhe Zhu, Yi Lu, Haozhe Wang, Chi Ruan, Benjamin Schneider, Weixu Zhang, Xiang Li, Andy Zheng, Yuyu Zhang, Ping Nie and Wenhu Chen

Agentic repository-level code understanding is essential for automating complex software engineering tasks, yet the field lacks reliable benchmarks. Existing evaluations often overlook the long tail topics and rely on popular repositories where Large Language Models (LLMs) can cheat via memorized knowledge. To address this, we introduce SWE-QA-Pro, a benchmark constructed from diverse, long-tail repositories with executable environments. We enforce topical balance via issue-driven clustering to cover under-represented task types and apply a rigorous difficulty calibration process: questions solvable by direct-answer baselines are filtered out. This results in a dataset where agentic workflows significantly outperform direct answering (e.g., a ~13-point gap for Claude Sonnet 4.5), confirming the necessity of agentic codebase exploration. Furthermore, to tackle the scarcity of training data for such complex behaviors, we propose a scalable synthetic data pipeline that powers a two-stage training recipe: Supervised Fine-Tuning (SFT) followed by Reinforcement Learning from AI Feedback (RLAIF). This approach allows small open models to learn efficient tool usage and reasoning. Empirically, a Qwen3-8B model trained with our recipe surpasses GPT-4o by 2.3 points on SWE-QA-Pro and substantially narrows the gap to state-of-the-art proprietary models, demonstrating both the validity of our evaluation and the effectiveness of our agentic training workflow.

[Findings] Synapse: Empowering LLM Agents with Episodic-Semantic Memory via Spreading Activation ▶ *Underline Video*

Hanqi Jiang, Junhao Chen, Yi Pan, Ling Chen, Weihang You, Yifan Zhou, Ruidong Zhang, Yohannes Abate and Tianming Liu

While Large Language Models (LLMs) excel at generalized reasoning, standard retrieval-augmented approaches fail to address the disconnected nature of long-term agentic memory. To bridge this gap, we introduce Synapse (Synergistic Associative Processing & Semantic Encoding), a unified memory architecture that transcends static vector similarity. Drawing from cognitive science, Synapse models memory as a dynamic graph where relevance emerges from spreading activation rather than pre-computed links. By integrating lateral inhibition and temporal decay, the system dynamically highlights relevant sub-graphs while filtering interference. We implement a Triple Hybrid Retrieval strategy that fuses geometric embeddings with activation-based graph traversal. Extensive evaluations on the LoCoMo benchmark show that Synapse significantly outperforms state-of-the-art methods in complex temporal and multi-hop reasoning tasks, offering a robust solution to the "Contextual Tunneling" problem.

[Long] SynthAgent: Adapting Web Agents with Synthetic Supervision ▶ *Underline Video*

Zhaoyang Wang, Yiming Liang, Xuchao Zhang, Qianhui Wu, Siwei Han, Anson Bastos, Rujia Wang, Chetan Bansal, Baolin Peng, Jianfeng Gao, Saravan Rajmohan and Huaxiu Yao

Web agents struggle to adapt to new websites due to the scarcity of environment specific tasks and demonstrations. Recent works have explored synthetic data generation to address this challenge, however, they suffer from data quality issues where synthesized tasks contain hallucinations that cannot be executed, and collected trajectories are noisy with redundant or misaligned actions. In this paper, we propose SynthAgent, a fully synthetic supervision framework that aims at improving synthetic data quality via dual refinement of both tasks and trajectories. Our approach begins by synthesizing diverse tasks through categorized exploration of web elements, ensuring efficient coverage of the target environment. During trajectory collection, tasks are refined only when conflicts with observations are detected, which mitigates hallucinations while preserving task consistency. After collection, we conduct trajectory refinement with global context to mitigate potential noise or misalignments. Finally, we fine-tune open-source web agents on the refined synthetic data to adapt them to the target environment. Experimental results demonstrate that SynthAgent outperforms existing synthetic data methods, validating the importance of high-quality synthetic supervision. The code is publicly available at https://github.com/aiming-lab/SynthAgent.

[Short] T^: Progressive Block Scaling for Masked Diffusion Language Models Through Trajectory Aware Reinforcement Learning ▶ *Underline Video*

Hanchen Xia, Baoyou Chen, Yutang Ge, Guojiang Zhao and Siyu Zhu

We present T^, a simple TraceRL-based curriculum for progressive block-size scaling in masked diffusion language models (MDMs). Starting from an AR-initialized small-block MDM, T^ gradually increases the block size while re-optimizing the denoising policy at each stage, enabling higher-parallelism decoding with limited degradation on math reasoning benchmarks. Across two SDAR scales and three benchmarks, T^ consistently outperforms direct large-block TraceRL and is substantially more stable during training. Our schedule analysis suggests that the learned policy does not simply revert to a strictly left-to-right order; instead, it retains block-size-specific non-monotone updates while improving accuracy.

[Long] Temporal Sampling for Forgotten Reasoning in LLMs ▶ *Underline Video*

Yuetai Li, Zhangchen Xu, Fengqing Jiang, Bhaskar Ramasubramanian, Luyao Niu, Bill Yuchen Lin, Xiang Yue and Radha Poovendran

Fine-tuning large language models (LLMs) is intended to improve their reasoning capabilities, yet we uncover a counterintuitive effect: models often forget how to solve problems they previously answered correctly during training. We term this phenomenon Temporal Forgetting and show that it is widespread across model sizes, fine-tuning methods (both Reinforcement Learning and Supervised Fine-Tuning), and multiple reasoning benchmarks. Our analysis reveals on average more than 20% of final errors were once solved correctly at an earlier checkpoint. Inspired by the phenomenon of Temporal Forgetting, we proposed Temporal Sampling, a simple decoding strategy

that draws outputs from multiple checkpoints along the training trajectory. This approach recovers forgotten solutions and leads to significant improvements in reasoning performance than final-ckpt-sampling only, gains from 4 to 19 points in Pass@k and consistent gains for majority-voting and Best-of-N across several benchmarks. Temporal sampling also outperforms strong baselines such as model merging. By leveraging the temporal diversity inherent in training, Temporal Sampling offers a practical, compute-efficient way to surface hidden reasoning ability and rethink how we evaluate LLMs.

[Long] TensorLens: End-to-End Transformer Analysis via High-Order Attention Tensors ▶ *Underline Video*

Ido Andrew Atad, Itamar Zimerman, Shahar Katz and Lior Wolf

Attention matrices are fundamental to transformer research, supporting a broad range of applications including interpretability, visualization, manipulation, and distillation. Yet, most existing analyses focus on individual attention heads or layers, failing to account for the model's global behavior. While prior efforts have extended attention formulations across multiple heads via averaging and matrix multiplications or incorporated components such as normalization and FFNs, a unified and complete representation that encapsulates all transformer blocks is still lacking. We address this gap by introducing TensorLens, a novel formulation that captures the entire transformer as a single, input-dependent linear operator expressed through a high-order attention-interaction tensor. This tensor jointly encodes attention, FFNs, activations, normalizations, and residual connections, offering a theoretically coherent and expressive linear representation of the models computation. TensorLens is theoretically grounded and our empirical validation shows that it yields richer representations than previous attention-aggregation methods. Our experiments demonstrate that the attention tensor can serve as a powerful foundation for developing tools aimed at interpretability and model understanding.

[Long] Test of Time: Rethinking Temporal Signal of Benchmark Contamination ▶ *Underline Video*

Terry Jingchen Zhang, Gopal Dev, Ning Wang, Max Obreiter, Wenyuan Jiang, Punya Syon Pandey, Keenan Samway, Yinya Huang, Bernhard Schölkopf, Mrinmaya Sachan and Zhijing Jin

Post-cutoff performance decay has been widely interpreted as a temporal signal for benchmark contamination. We critically examine this belief and demonstrate that this temporal signal is highly sensitive to how benchmark questions are constructed. Specifically, we show that LLM-generated questions can produce remarkably different temporal patterns compared to fill-in-the-blank questions directly retrieved from the very same materials. We validated this finding on previous benchmarks that reported clear post-cutoff performance decay such as LiveCodeBench and further showed simple LLM transformation could effectively remove this temporal pattern when evaluated on the same models. We also provide a mechanistic understanding of our observation using influence function analysis. Overall, this work offers a new perspective on the sensitivity of temporal contamination signal and highlights the need for more robust contamination detection methods for reliable AI evaluation.

[Long] Thermometer of Thoughts: Enhancing LLM's Exploration via Attention Temperature Modulation ▶ *Underline Video*

Zhiyuan Yu, Shijian Xiao, Cam-Tu Nguyen, Zhangyue Yin, Lekai Xing, Wenzhong Li and Sanglu Lu

Improving the exploration of reasoning is essential for advancing Large Language Models' (LLMs) problem-solving performance. Current methods primarily rely on output-level stochasticity, which decode within fixed reasoning patterns of LLM and suffer from insufficient exploration. In this paper, we introduce adjusting attention temperature to directly modulate the model's internal focus during reasoning, which enables a dynamic shift between exploratory and focused processing. We reveal that moderate adjustments preserve LLM's reasoning capability while producing problem hardness-dependent benefits: higher temperatures facilitate solving complex tasks by encouraging wider exploration, whereas lower temperatures mitigate overthinking on simpler problems. Leveraging this insight, we propose a two-stage inference strategy: first, attention temperature scaling modulates the LLM's reasoning patterns to diversify the reasoning traces; then, a difficulty-aware aggregation scheme is introduced to effectively identify the most reliable solution from the generated candidates. Extensive evaluations show that our method improves Pass@10 by 6.78–14.20% and aggregation accuracy by 9.74% across 7 reasoning benchmarks.

[Long] THOR: A Theta-Gamma Hierarchical Oscillatory Reasoning Framework for Multi-hop QA ▶ *Underline Video*

Ziyang Ling and Ronald XuMingzhai Sun

Multi-hop question answering requires retrieving and integrating evidence from multiple contexts. Despite the rapid progress of current research, multi-hop reasoning remains constrained by two persistent limitations: attention decay, where the model's focus on main question degrades as the reasoning chain grows, and error accumulation, where mistakes propagate across hops and compounds into final failure. Inspired by Theta-Gamma hierarchical oscillation which decouples global planning from local retrieval, enabling efficient attention transfer between hops and a verification and repair mechanism that interrupts the accumulation of errors in the wrong paths, we present THOR, a brain-inspired Theta-Gamma hierarchical oscillatory reasoning framework. Extensive comparative experiments and specific validation experiments on multi-hop QA benchmarks demonstrate that THOR improves answer accuracy and robustness while mitigating limitations, showcasing its generalization across different backbones.

[Long] TLPO: Token-Level Policy Optimization for Mitigating Language Confusion in Large Language Models ▶ *Underline Video*

Jinho Choo, JunSeung Lee, Jimyeong Kim, Yeeho Song, S. K. Hong and Yeong-Dae Kwon

Large language models (LLMs) demonstrate strong multilingual capabilities, yet often fail to consistently generate responses in the intended language, exhibiting a phenomenon known as language confusion. Prior mitigation approaches based on sequence-level fine-tuning, such as DPO, ORPO, and GRPO, operate at the level of entire responses and can lead to unintended degradation of general model capabilities,

motivating the need for more fine-grained alternatives. To address this, we introduce Token-Level Policy Optimization (TLPO), a fine-tuning framework designed to mitigate language confusion through localized, token-level updates. TLPO identifies error-prone positions, explores alternative candidate tokens, and updates the policy using a tailored objective to suppress error-inducing outputs at a granular level. This selective intervention enables effective mitigation of language confusion without compromising the model's general abilities. Experiments on multiple multilingual LLMs across diverse languages demonstrate that TLPO significantly outperforms baselines in improving language consistency while preserving downstream task accuracy.

[Long] ToMMeR - Efficient Entity Mention Detection from Large Language Models ▶ *Underline Video*

Victor Morand, Nadi Tomeh, Josiane Mothe and Benjamin Piwowarski

Identifying which text spans refer to entities - mention detection- is both foundational for information extraction and a known performance bottleneck. We introduce ToMMeR, a lightweight model (<300K parameters) probing mention detection capabilities from early LLM layers. Across 13 NER benchmarks, ToMMeR achieves 93% recall zero-shot, with an estimated 90% precision under a human-calibrated LLM-judge protocol, showing that ToMMeR rarely produces spurious predictions despite high recall. Cross-model analysis reveals that diverse architectures (14M-15B parameters) converge on similar mention boundaries (DICE >75%), confirming that mention detection emerges naturally from language modeling. When extended with span classification heads, ToMMeR achieves competitive NER performance (80-87% F1 on standard benchmarks). Our work provides evidence that structured entity representations exist in early transformer layers and can be efficiently recovered with minimal parameters.

[Findings] ToolGrad: Efficient Tool-use Dataset Generation with Textual Gradients ▶ *Underline Video*

Zhongyi Zhou, Kohei Uehara, Haoyu Zhang, Jingtao Zhou, Lin Gu, Ruofei Du, Zheng Xu and Tatsuya Harada

Prior work synthesizes tool-use LLM datasets by first generating a user query, followed by complex tool-use annotations like DFS. This inherently leads to inevitable annotation failures and low efficiency in data generation. We introduce ToolGrad, an agentic framework that inverts this paradigm. ToolGrad first constructs valid tool-use chains through an iterative process guided by textual "gradients", and then synthesizes corresponding user queries. This "answer-first" approach led to ToolGrad-500, a dataset generated with more complex tool use, lower cost, and almost 100% pass rate. Experiments show that models trained on ToolGrad-500 outperform those trained on expensive baseline datasets and proprietary LLMs.

[Long] ToolPRM: Fine-Grained Inference Scaling of Structured Outputs for Function Calling ▶ *Underline Video*

Jianghao Lin, Yuanyuan Shi, Xin Peng, Renjie Ding, Hairui Wang, Yuxuan Peng, Bizhe Bai, Weixi Song, Fengshuo Bai, Huacan Chai, Weinan Zhang, Fei Huang and Ying Wen

Large language models (LLMs) excel at function calling, but inference scaling has been explored mainly for unstructured generation. We propose an inference-scaling framework for structured outputs that combines fine-grained beam search with ToolPRM, a process reward model scoring each intra-call decision (function name and argument filling). We build the first fine-grained intra-call supervision dataset via function masking, rollout collection, and step-level annotation. ToolPRM outperforms outcome and coarse-grained reward models in predictive accuracy and yields consistent test-time gains on multiple function-calling benchmarks. We further show that structured generation follows "explore more but retain less", since early JSON errors are unrecoverable.

[Long] Towards Privacy-Preserving Large Language Model: Text-free Inference Through Alignment and Adaptation ▶ *Underline Video*

Jeongho Yoon, Chanhee Park, Yongchan Chun, Hyeonseok Moon and Heuiseok Lim

Current LLM-based services typically require users to submit raw text regardless of its sensitivity. While intuitive, such practice introduces substantial privacy risks, as unauthorized access may expose personal, medical, or legal information. Although prior defenses strived to mitigate these risks, they often incur substantial computational overhead and degrade model performance. To overcome this privacy-efficiency trade-off, we introduce Privacy-Preserving Fine-Tuning (PPFT), a novel training pipeline that eliminates the need for transmitting raw prompt text while maintaining a favorable balance between privacy preservation and model utility for both clients and service providers. Our approach operates in two stages: first, we train a client-side encoder together with a server-side projection module and LLM, enabling the server to condition on k-pooled prompt embeddings instead of raw text; second, we fine-tune the projection module and LLM on private, domain-specific data using noise-injected embeddings, allowing effective adaptation without exposing plain text prompts and requiring access to the decoders internal parameters. Extensive experiments on domain-specific and general benchmarks demonstrate that PPFT achieves a striking balance between privacy and utility, maintaining competitive performance with minimal degradation compared to noise-free upper bounds.

[Findings] Tracing Mathematical Proficiency Through Problem-Solving Processes ▶ *Underline Video*

Jungyang Park, Suho Kang, Jaewoo Park, JaeHong Kim, Jaewoo Shin, Seonjoon Park and Youngjae Yu

Knowledge Tracing (KT) aims to model student's knowledge state and predict future performance to enable personalized learning in Intelligent Tutoring Systems. However, traditional KT methods face fundamental limitations in explainability, as they rely solely on the response correctness, neglecting the rich information embedded in students' problem-solving processes. To address this gap, we propose Knowledge Tracing Leveraging Problem-Solving Process (KT-PSP), which incorporates students' problem-solving processes to capture the multidimensional aspects of mathematical proficiency. We also introduce KT-PSP-25, a new dataset specifically designed for KT-PSP.

Building on this, we present StatusKT, a KT framework that employs a teacher-student-teacher three-stage LLM pipeline to extract students' Mathematical Proficiency (MP) as intermediate representation. In this pipeline, the teacher LLM first extracts problem-specific proficiency indicators, then a student LLM generates responses based on the student's solution process, and a teacher LLM evaluates these responses to determine mastery of each indicator. The experimental results on KT-PSP-25 demonstrate that StatusKT improves the prediction performance of existing KT methods. Moreover, StatusKT provides interpretable explanations for its predictions by explicitly modeling students' mathematical proficiency. Code is available here.

[Findings] Tree-of-Evidence: Efficient "System 2" Search for Faithful Multimodal Grounding ► *Underline Video*

Micky Nnamdi, Benoit Marteau, Yishan Zhong, Junior Tamo and May Wang

Large Multimodal Models (LMMs) achieve state-of-the-art performance in high-stakes domains like healthcare, yet their reasoning remains opaque. Attention- and saliency-based methods often fail to faithfully represent the model's decision process, particularly when integrating heterogeneous modalities. We introduce Tree-of-Evidence (ToE), an inference-time search algorithm that frames interpretability as a discrete optimization problem. Rather than relying on soft attention weights, ToE employs lightweight Evidence Bottlenecks that score coarse units of data (e.g., vital-sign windows, report chunks) and performs a beam search to identify the compact evidence set required to reproduce the model's prediction. We evaluate ToE across six tasks spanning three datasets and two domains, including clinical prediction on MIMIC-IV, cross-center validation on eICU, and non-clinical fault detection on LEMMA-RCA. ToE retains over 98% of full-model AUROC with as few as five evidence units, achieves higher decision agreement and lower fidelity error than LIME, SHAP, saliency, and concept-bottleneck baselines under sparse budgets, and outperforms LLMs up to 70B parameters. ToE therefore provides a practical mechanism for auditing multimodal models.

[Findings] Uncertainty Quantification of Large Language Models through Multiple Uncertainty Sources ► *Underline Video*

Tiejin Chen, Xiaoou Liu, Longchao Da, Jia Chen, Evangelos E. Papalexakis and Hua Wei

Large Language Models (LLMs) have demonstrated remarkable capabilities across diverse domains. However, the reliability of responses from LLMs remains a question. Uncertainty quantification (UQ) of LLMs is crucial for ensuring their reliability, especially in areas such as healthcare. Existing UQ methods, often designed around a single resource such as Natural Language Inference (NLI) or graph-based metrics, fail to capture the multifaceted nature of uncertainty in natural language generation. In this work, we propose MS-UQ, a novel Multi-Resource Uncertainty Quantification framework that integrates heterogeneous uncertainty signals into a unified measure. Our approach concatenates matrices from diverse resources and employs tensor decomposition to orthogonally disentangle unique and shared information. To ensure scalability, we construct an adaptive ensemble of outputs from different decomposition methods, enabling the incorporation of new uncertainty sources. Experiments on CoQA, NQ_Open, and HotpotQA demonstrate that MS-UQ consistently outperforms existing methods, offering a comprehensive and scalable solution for uncertainty estimation in black-box LLMs and a more robust framework for enhancing LLM reliability in high-stakes applications. Our code can be accessed at https://anonymous.4open.science/r/MDUQ-First-202E/README.md.

[Long] Understanding and Mitigating Political Stance Cross-topic Generalization in Large Language Models ► *Underline Video*

Jiayi Zhang, Shu Yang, Junchao Wu, Derek F. Wong and Di Wang

Fine-tuning Large Language Models on a political topic will significantly manipulate their political stance on various issues and unintentionally affect their stance on broad topics. While previous studies have proposed this issue, there is still a lack of understanding regarding the internal representations of these stances and the mechanisms that lead to unintended cross-topic generalization. In this paper, we systematically explore the internal mechanisms underlying this phenomenon from a neuron-level perspective and how to mitigate the cross-topic generalization of political fine-tuning. Firstly, we propose Political Neuron Localization through Activation Contrasting (PNLAC) to identify two distinct types of political neurons: general political neurons, which govern stance across multiple political topics, and topic-specific neurons that affect the model's political stance on individual topics. We find that these political neuron types exist in the middle and later layers across four models and datasets through activation patching experiments. Leveraging these insights, we introduce InhibitFT, an inhibition-based fine-tuning method that effectively mitigates the cross-topic stance generalization. Experimental results demonstrate the robustness of the identified neuron types across various models and datasets and show that InhibitFT significantly reduces the cross-topic stance generalization by 20% on average while preserving topic-specific performance. Moreover, we demonstrate that selectively inhibiting only 5% of neurons is sufficient to effectively mitigate the cross-topic stance generalization.

[Long] Understanding LLM Performance Degradation in Multi-Instance Processing: The Roles of Instance Count and Context Length ► *Underline Video*

Jingxuan Chen, Mohammad Taher Pilehvar and Jose Camacho-Collados

Users often rely on Large Language Models (LLMs) for processing multiple documents or performing analysis over a number of instances. For example, analysing the overall sentiment of a number of movie reviews requires an LLM to process the sentiment of each review individually in order to provide a final aggregated answer. While LLM performance on such individual tasks is generally high, there has been little research on how LLMs perform when dealing with multi-instance inputs. In this paper, we perform a comprehensive evaluation of the multi-instance processing (MIP) ability of LLMs for tasks in which they excel individually. The results show that all LLMs follow a pattern of slight performance degradation for small numbers of instances (~20-100), followed by a performance collapse on larger instance counts. Crucially, our analysis shows that while context length is associated with this degradation, the number of instances has a stronger effect on

the final results. This finding suggests that when optimising LLM performance for MIP, attention should be paid to both context length and, in particular, instance count.

[Findings] User-Assistant Bias in LLMs ▶ *Underline Video*

Xu Pan, Jingxuan Fan, Zidi Xiong, Ely Hahami, Jorin Overwiening and Ziqian Xie

Modern large language models (LLMs) are typically trained and deployed using structured role tags (e.g. system, user, assistant, tool) that explicitly mark the source of each piece of context. While these tags are essential for instruction following and controllability, asymmetries in the training data associated with different role tags can potentially introduce inductive biases. In this paper, we study this phenomenon by formalizing user-assistant bias, defined as the tendency of an LLM to preferentially rely on information from either the user or assistant role when they provide incompatible information about the same entity in the context history. We introduce a task-agnostic benchmark UserAssist and evaluate such bias in 52 frontier models. We observe that most of the instruction-tuned models exhibit strong user bias, whereas base and reasoning models are close to neutral. Using controlled fine-tuning experiments, we isolate which post-training recipes drive the observed user-assistant bias. We find that human-preference alignment amplifies user bias, while reasoning fine-tuning reduces it. Finally, we show that user-assistant bias can be bidirectionally controlled via direct preference optimization (DPO) on UserAssist-train, and that the resulting bias reliably generalizes to two realistic multi-turn debate datasets spanning philosophical opinions and natural argumentative exchanges on factual/policy topics. These results reveal an underexplored consequence of role-tagged training and provide a principled framework to diagnose and control tag-induced biases in modern LLMs.

[Findings] Veri-R1: Toward Precise and Faithful Claim Verification via Online Reinforcement Learning ▶ *Underline Video*

Qi He, Cheng Qian, Xiusi Chen, Bingxiang He, Yi R. Fung and Heng Ji

Claim verification with large language models (LLMs) has recently attracted growing attention, due to their strong reasoning capabilities and transparent verification processes compared to traditional answer-only judgments. However, existing approaches to online claim verification, which requires iterative evidence retrieval and reasoning, still mainly rely on prompt engineering or pre-designed reasoning workflows, without unified training to improve necessary skills. Therefore, we introduce Veri-R1, an online reinforcement learning (RL) framework that enables an LLM to interact with a search engine and to receive reward signals that explicitly shape its planning, retrieval, and reasoning behaviors. The dynamic interaction between models and retrieval systems more accurately reflects real-world verification scenarios and fosters comprehensive verification skills. Empirical results show that Veri-R1 improves joint accuracy by up to 30% and doubles evidence score, often surpassing its larger-scale model counterparts. Ablation studies further reveal the impact of reward components, and the link between output logits and label accuracy. Our results highlight the effectiveness of online RL for precise and faithful claim verification, and provide a foundation for future research.

[Findings] Vocab Diet: Reshaping the Vocabulary of LLMs via Vector Arithmetic ▶ *Underline Video*

Yuval Reif, Guy Kaplan and Roy Schwartz

Large language models (LLMs) often encode word-form variation (e.g., walk vs. walked) as linear directions in the embedding space. However, standard tokenization algorithms treat such variants as distinct words with different vocabulary entries, quickly filling the size-capped token vocabulary with surface-form variation (e.g., walk, walking, Walk), at the expense of diversity and multilingual coverage. We show that many of these variations can be captured by transformation vectors, additive offsets that yield the appropriate word representation when applied to a base form embedding, in both the input and output spaces. Building on this, we propose a compact reshaping of the vocabulary: instead of assigning unique tokens to each surface form, we compose them from shared base form and transformation vectors (e.g., walked is walk+past tense). Our approach is lightweight, keeping the pretrained backbone frozen and only training small adaptation modules. We apply it across five languages and multiple LLMs in both pretraining and post-hoc adaptation, freeing 10-40% of vocabulary slots to be reallocated where tokenization is inefficient. Importantly, we do so while also expanding vocabulary coverage to out-of-vocabulary words, and with minimal impact on downstream performance. Our findings motivate a rethinking of vocabulary design, towards a representation that better matches the underlying structure of language and the practical needs of multilingual coverage.

[Findings] WaveDetect: Robust Framework for Machine-Generated Text Detection via Wavelet Transform ▶ *Underline Video*

Zhichen Liu, Kaitong Qin, Linhan He and Yang Xu

As Large Language Models asymptotically approach human-level fluency in natural language generation, solely relying on surface-level semantic artifacts for detecting LLM-generated texts has become increasingly precarious. Existing detectors often falter when facing three critical challenges: adversarial perturbations, cross-domain shifts, and the rapid temporal evolution of the foundation model. To address these issues, we propose, a novel framework that reformulates text detection as a signal processing task within the time-frequency domain. Unlike previous methods that analyze static token probability distributions, models the generated output as a probability signal, upon which a differentiable Continuous Wavelet Transform is applied to convert them into learnable spectral representations. This process reveals the intrinsic "spectral fingerprints" in machine-generated texts–patterns that remain invisible in time domain. Comprehensive evaluations on three well-curated datasets (RAID, EvoBench, and Domain-Shift) show that our method achieves a new state-of-the-art. It not only achieves superior accuracy but also exhibits remarkable robustness against sophisticated attacks, generalization across out-of-distribution topics and unseen evolving LLMs. Our results validate the efficacy of spectral analysis as a promising paradigm for LLM-generated texts detection.

[Long] WebClipper: Efficient Evolution of Web Agents with Graph-based Trajectory Pruning ▶ *Underline Video*

Junjie Wang, Zequn Xie, Dan Yang, Jie Feng, Yue Shen, Duolin Sun, Meixiu Long, Yihan Jiao, Zhehao Tan, Jian Wang, Peng Wei and Jinjie Gu

Deep Research systems based on web agents have shown strong potential in solving complex information-seeking tasks, yet their search efficiency remains underexplored. We observe that many state-of-the-art open-source web agents rely on long tool-call trajectories with cyclic reasoning loops and exploration of unproductive branches. To address this, we propose WebClipper, a framework that compresses web agent trajectories via graph-based pruning. Concretely, we model the agents search process as a state graph and cast trajectory optimization as a minimum-necessary Directed Acyclic Graph (DAG) mining problem, yielding pruned trajectories that preserve essential reasoning while eliminating redundant steps. Continued training on these refined trajectories enables the agent to evolve toward more efficient search patterns and reduces tool-call rounds by about 20% while improving accuracy. Furthermore, we introduce a new metric called F-AE Score to measure the model's overall performance in balancing accuracy and efficiency. Experiments demonstrate that WebClipper compresses tool-call rounds under excellent performance, providing practical insight into balancing effectiveness and efficiency in web agent design.

[Findings] WebDART: Dynamic Decomposition and Re-planning for Complex Web Tasks ▶ *Underline Video*

Jingbo Yang, Bairu Hou, Wei Wei, Shiyu Chang and Yujia Bao

Large-language-model (LLM) agents are becoming competent at straightforward web tasks, such as opening an item page or submitting a form, but still struggle with objectives that require long-horizon navigation, large-scale information extraction, and reasoning under constraints. We present WebDART, a general framework that enables a single LLM to handle such complex chores. WebDART (i) dynamically decomposes each objective into three focused subtasks, navigation, information extraction, and execution, so the model concentrates on one skill at a time, and (ii) continuously re-plans the decomposition as new webpages are revealed, taking advantage of newly discovered filters or shortcuts and avoiding redundant exploration. Evaluated on WebChoreArena, WebDART lifts end-to-end success rates by up to 13.7 percentage points over previous state-of-the-art agents, while matching their performance on the easier WebArena suite and completing tasks with up to 14.7 fewer navigation steps. Code will be publicly available.

[Findings] What Makes an LLM a Good Optimizer? A Trajectory Analysis of LLM-Guided Evolutionary Search ▶ *Underline Video*

Xinhao Zhang, Xi Chen, François Portet and Maxime Peyrard

Recent work has demonstrated the promise of orchestrating large language models (LLMs) within evolutionary and agentic optimization systems. However, the mechanisms driving these optimization gains remain poorly understood. In this work, we present a large-scale study of LLM-guided evolutionary search, collecting optimization trajectories for 15 LLMs across 8 tasks. Although zero-shot problem-solving ability correlates with final optimization outcomes, it explains only part of the variance: models with similar initial capability often induce dramatically different search trajectories and outcomes. By analyzing these trajectories, we find that strong LLM optimizers behave as local refiners, producing frequent incremental improvements while progressively localizing the search in semantic space. Conversely, weaker optimizers exhibit large semantic drift, with sporadic breakthroughs followed by stagnation. Notably, various measures of solution novelty do not predict final performance; novelty is beneficial only when the search remains sufficiently localized around high-performing regions of the solution space. Our results highlight the importance of trajectory analysis for understanding and improving LLM-based optimization systems and provide actionable insights for their design and training.

[Long] When "Correct" Is Not Safe: Can We Trust Functionally Correct Patches Generated by Code Agents? ▶ *Underline Video*

Yibo Peng, James Song, Lei Li, Xinyu Yang, Mihai Christodorescu, Ravi Mangal, Corina Pasareanu, Haizhong Zheng and Beidi Chen

Code agents are increasingly trusted to autonomously fix bugs on platforms such as GitHub, yet their security evaluation focuses almost exclusively on functional correctness. In this paper, we reveal a novel type of threat to real-world code-agents: functionally correct yet vulnerable (FCV) patches, which pass all test cases but contain vulnerable code. With our proposed FCV-Attack, we demonstrate that SOTA LLMs (e.g., ChatGPT and Claude) and agent scaffolds (e.g., SWE-agent and OpenHands) are all vulnerable to this FCV threat; across 12 agent-model combinations on SWE-Bench, the attack only requires black-box access and a single query to the code agent to perform the attack. For example, for CWE-538 (information exposure vulnerability), the FCV-Attack attains an attack success rate of 40.7% on GPT-5 Mini + OpenHands. Our results reveal an important security threat overlooked by current evaluation paradigms and urge the development of security-aware defenses for code agents.

[Findings] When Thoughts Meet Facts: Reusable Reasoning for Long-Context LMs ▶ *Underline Video*

Soyeong Jeong, Taehee Jung, Sung Ju Hwang, Joo-Kyung Kim and Dongyeop Kang

Recent Long-Context Language Models (LCLMs) can process hundreds of thousands of tokens in a single prompt, enabling new opportunities for knowledge-intensive multi-hop reasoning by integrating large sets of retrieved documents or, in some cases, directly all necessary information. However, simply feeding more documents into the context window fails to capture how evidence should be connected. We address this gap with thought templates, reusable reasoning patterns derived from prior problem solving that structure how evidence is combined and guide multi-hop inference alongside factual documents. To keep these templates effective, we propose an update strategy that iteratively refines templates derived from training data through natural-language feedback. Across diverse benchmarks and LCLM families, our approach delivers consistent gains over strong baselines in both retrieval-based and retrieval-free settings. Furthermore, we show that optimized templates can be distilled into relatively smaller open-source models, demonstrating its broad applicability. We refer to our framework as Thought Template Augmented LCLMs (ToTAL).

[Findings] Why Do Multilingual Reasoning Gaps Emerge in Reasoning Language Models? ▶ *Underline Video*
Deokhyung Kang, Seonjeong Hwang, Daehui Kim, Hyounghun Kim and Gary Geunbae Lee
Reasoning language models (RLMs) achieve strong performance on complex reasoning tasks, yet they still exhibit a multilingual reasoning gap, performing better in high-resource languages than in low-resource ones. While recent efforts have been made to address this gap, its underlying causes remain largely unexplored. In this work, we show that this gap primarily stems from failures in language understanding, specifically, the models inability to translate multilingual inputs into the language dominating its reasoning traces (typically English). As identifying understanding failures can enable targeted mitigation of the gap, we evaluate a range of detection methods and find that understanding failures are detectable to a meaningful extent, with supervised approaches performing best. Building on this, we propose Selective Translation, a strategy that incorporates an English translation into the initial reasoning trace when an understanding failure is detected. Experimental results using Qwen3-4B show that Selective Translation substantially bridges the multilingual reasoning gap, achieving near full-translation performance while translating only about 20% of inputs. Together, our results show that failures in language understanding are the primary driver of the multilingual reasoning gap and can be detected and selectively mitigated, clarifying its origin and suggesting a path toward more equitable multilingual reasoning.

[Findings] XQ-MEval: A Dataset with Cross-lingual Parallel Quality for Benchmarking Translation Metrics ▶ *Underline Video*
Jingxuan Liu, Zhi Qu, Jin Tei, Hidetaka Kamigaito, Lemao Liu and Taro Watanabe
Automatic evaluation metrics are essential for building multilingual translation systems. The common practice of evaluating these systems is averaging metric scores across languages, yet this is suspicious since metrics may suffer from cross-lingual scoring bias, where translations of equal quality receive different scores across languages. This problem has not been systematically studied because no benchmark exists that provides parallel-quality instances across languages, and expert annotation is not realistic. In this work, we propose XQ-MEval, a semi-automatically built dataset covering nine translation directions, to benchmark translation metrics. Specifically, we inject MQM-defined errors into gold translations automatically, filter them by native speakers for reliability, and merge errors to generate pseudo translations with controllable quality. These pseudo translations are then paired with corresponding sources and references to form triplets used in assessing the qualities of translation metrics. Using XQ-MEval, our experiments on nine representative metrics reveal the inconsistency between averaging and human judgment and provide the first empirical evidence of cross-lingual scoring bias. Finally, we propose a normalization strategy derived from XQ-MEval that aligns score distributions across languages, improving the fairness and reliability of multilingual metric evaluation.

Poster Session E: Industry (Coronado)

11:00-12:30 (Coronado)

[Industry] R^3: Advertisement Compliance Rectification via Group-Relative Experience Extractor and Curriculum Reinforcement ▶ *Underline Video*
Yuan Chen, zhenyu hu, Mengge Xue, Cao Te, Liqun Liu, Peng Shu, Huan Yu and Jie Jiang
Rigorous content moderation is crucial for online advertising but leads to millions of daily rejections. This scale renders manual rectification infeasible, particularly for video advertisements. However, existing safety-driven methods often suffer from aggressive over-editing, which compromises the advertiser's original semantic intent merely to satisfy compliance. In this work, we target the rectification of textual violations in video ads, covering both speech transcripts and on-screen text. We propose R^3, a novel framework designed to harmonize compliance with original semantic intent preservation. Our approach integrates three key innovations: (1) an experience-driven data synthesis framework that bootstraps high-quality supervision via group-Relative compliance experience extractor; (2) a curriculum Reinforcement learning strategy with hierarchical rewards designed to enforce compliance while maximizing semantic consistency; and (3) a comprehensive video Rectification framework seamlessly integrating text recognition, rewriting, and re-rendering for industrial deployment. Extensive experiments on industrial datasets and online A/B testing demonstrate that R^3 significantly outperforms state-of-the-art baselines, achieving an optimal trade-off between violation rectification and intent preservation.

[Industry] OmniOData: Unleashing Small Language Models for OData Query Generation with Synthetic Data and Reinforcement Learning ▶ *Underline Video*
Tao Bai, Zhaochen Li, Hongxin Shao and Daniel Dahlmeier
Despite the success of Large Language Models (LLMs) in structured query generation, OData, a critical RESTful protocol for enterprise APIs, remains under-researched due to a lack of high-fidelity, execution-validated datasets. To bridge this gap, we introduce OmniOData, a framework that generates SynOData, the first large-scale OData corpus featuring execution-grounded queries and reasoning traces. Using this corpus, we develop OmniOData-R1 (1.5B-3B parameters), a family of models that match or surpass frontier proprietary systems, such as GPT-4o and Gemini 3, on realistic industrial benchmarks. Our results demonstrate that the synergy of execution-verified synthetic data and Reinforcement Learning (RL) effectively unlocks the latent reasoning of Small Language Models (SLMs), providing a high-performance, low-latency solution for specialized enterprise query generation. The code and data will be released under an open-source license.

[Industry] A Multistage Extraction Pipeline for Long Scanned Financial Documents: An Empirical Study in Industrial

KYC Workflows ▶ *Underline Video*

Han Yuxuan, Yuanxing Zhang, Yushuo Wang and Yichao Jin

Structured information extraction from long, multilingual scanned financial documents is a core requirement in industrial KYC and compliance workflows. These documents are typically non-machine-readable, noisy, and visually heterogeneous. They usually span dozens of pages while containing only sparse task-relevant information. Although recent vision-language models (VLMs) achieve strong benchmark performance, directly applying them end-to-end to full financial reports often leads to unreliable extraction under real-world conditions. We present a multistage extraction framework that integrates image preprocessing, multilingual OCR, hybrid page-level retrieval, and compact VLM-based structured extraction. The design separates page localization from multimodal reasoning, enabling more accurate extraction from complex multi-page documents. We evaluated the framework on 120 production KYC documents comprising about 3000 multilingual scanned pages. Across multiple OCR-VLM combinations, the proposed pipeline consistently outperforms direct PDF-to-VLM baselines, improving field-level accuracy by up to 31.9 percentage points. The best configuration, PaddleOCR with MiniCPM-o-2.6, achieves 87.27% accuracy. Ablation studies show that page-level retrieval is the dominant factor in performance improvements, particularly for complex financial statements and non-English documents.

[Industry] Accurate Legal Reasoning at Scale: Neuro-Symbolic Offloading and Structural Auditability for Robust Legal Adjudication ▶ *Underline Video*

Stanisaw Sójka and Witold Kowalczyk

Legal texts often contain computational legal clauses, provisions whose understanding requires complex logic. While frontier Large Reasoning Models (LRMs) can describe such clauses, building production-ready systems is limited by reasoning errors and the high cost of inference. We propose Amortized Intelligence, a neuro-symbolic approach where we use an LLM once to translate a legal text into Deterministic Autonomous Contract Language (DACL): a typed graph intermediate representation. Adjudication then relies on deterministic graph executions with a visually auditable trace. In comparison against runtime LRM baselines (including GPT-5.2 and Gemini 3 Pro), our DACL-based Agent achieves near-perfect consistency and mitigates the "reasoning cliff" observed in probabilistic models. The system reduces compute costs by over 90% in high-volume workflows while satisfying the strict auditability requirements of legal adjudication.

[Industry] Ada-RS: Adaptive Rejection Sampling for Selective Thinking ▶ *Underline Video*

Yirou Ge, Yixi Li, Alec M. Chiu, Shivani Shekhar, Zijie Pan, Avinash Thangali, Yun-Shiuan Chuang, Chaitanya Kulkarni, Uma Kona, Linsey Pang and Prakhar Mehrotra

Large language models (LLMs) are increasingly being deployed in cost- and latency-sensitive settings. While chain-of-thought improves reasoning, it can waste tokens on simple requests. We study selective thinking for tool-using LLMs and introduce Adaptive Rejection Sampling (Ada-RS), an algorithm-agnostic sample filtering framework for learning selective and efficient reasoning. For each given context, Ada-RS scores multiple sampled completions with an adaptive length-penalized reward then applies stochastic rejection sampling to retain only high-reward candidates (or preference pairs) for downstream optimization. We demonstrate how Ada-RS plugs into both preference pair (e.g. DPO) or grouped policy optimization strategies (e.g. DAPO). Using Qwen3-8B with LoRA on a synthetic tool call-oriented e-commerce benchmark, Ada-RS improves the accuracy-efficiency frontier over standard algorithms by reducing average output tokens by up to ~80% and reducing thinking rate by up to ~95% while maintaining or improving tool call accuracy. These results highlight that training signal selection is a powerful lever for efficient reasoning in latency-sensitive deployments.

[Industry] AesX: Enhance Your Images with Stunning Aesthetic Beauty ▶ *Underline Video*

Yuyan Chen, Zhendong Hou, Lei Xia, Jiahao Li, Zhuolin Ji and Zhixu Li

In the fields of advertising design, artistic creation, and cultural dissemination, there is an increasingly urgent demand for high-quality images that cater to fine-grained aesthetic preferences. Although existing large-scale models can generally meet basic requirements for clarity and alignment with textual elements, they still face significant bottlenecks in achieving precise control and aesthetic optimization. To address this limitation, we propose a set of comprehensive preference indicators across two major dimensions, text-image consistency and aesthetic quality, encompassing multiple criteria ranging from exposure and clarity to visual guidance and innovativeness. Building on these indicators, we have developed a generative framework named AesX to steer the model consistently toward a generation path that more closely aligns with human aesthetic sensibilities. Our experimental findings demonstrate that this approach yields significant improvements in both target recognition accuracy and overall visual aesthetic presentation.

[Industry] Agent-Ops: A Multi-Agent Orchestration Framework for End-to-End SOP Automation in E-Commerce Operations ▶ *Underline Video*

Apoorva Singh, Sanjay Agrawal, Sayanta Adhikari, Vinayak S Puranik, Shivam Tiwari and Dheeraj Assudani

While Large Language Models excel at reasoning and language understanding, they struggle with multi-step operational workflows requiring precise procedural adherence, which is fundamental for industrial automation. Existing SOP-guided agents assume well-defined procedures and structured APIs, failing to address enterprise realities like incomplete SOPs, dynamic web interfaces, and unpredictable document formats. We present Agent-Ops, an end-to-end multi-agent framework automating Standard Operating Procedures in e-commerce. Agent-Ops contributes: (1) SOP Groomer, a human-AI framework transforming ambiguous documentation into automation-ready specifications, improving accuracy by 13.2%, (2) WebAgent, achieving 91.3% task completion and 86.5% execution consistency through demonstration-based learning, and (3) a Document Verification Agent performing multi-lingual validation across tax invoices, certificates,

and supply chain documents with 94.2% accuracy. Deployed across seven SOP categories in three geographic regions, Agent-Ops achieves 85-97% end-to-end accuracy while reducing case resolution from 30 to 5 minutes (83% reduction). Production deployment with over 1000 Account Managers validates that LLM-based agents achieve enterprise-grade reliability when augmented with robust web automation, comprehensive document understanding, and systematic SOP refinement.

[Industry] Agentic Economic Modeling ► *Underline Video*

Bohan Zhang, Jiaxuan Li, Ali Hortacsu, Xiaoyang Ye, Victor Chernozhukov, Anqi Ni and Edward W Huang

We introduce Agentic Economic Modeling (AEM), a framework that aligns synthetic LLM choices with small-sample human evidence for reliable econometric inference. AEM first generates task-conditioned synthetic choices via LLMs, then learns a bias-correction mapping from task features and raw LLM choices to human-aligned choices, upon which standard econometric estimators perform inference to recover demand elasticities and treatment effects. We validate AEM in two experiments. In a large scale conjoint study with millions of observations, using only 10% of the original data to fit the correction model lowers the error of the demand-parameter estimates, while uncorrected LLM choices even increase the errors. In a regional field experiment, a mixture model calibrated on 10% of geographic regions estimates an out-of-domain treatment effect of -65+/-10 bps, closely matching the full human experiment (-60+/-8 bps). Under time-wise extrapolation, training with only day-one human data yields -24 bps (95% CI: [-26, -22], p<1e-5), improving over the human-only day-one baseline (-17 bps, 95% CI: [-43, +9], p=0.2049). These results demonstrate AEM's potential to improve RCT efficiency and establish a foundation method for LLM-based counterfactual generation.

[Industry] ARGUS: Policy-Adaptive Ad Governance via Evolving Reinforcement with Adversarial Umpiring ► *Underline Video*

Deyi Ji, Junyu Lu, Xuanyi Liu, Liqun Liu, Hailong Zhang, Peng Shu, Huan Yu, Jie Jiang, Tianrun Chen and Lanyun Zhu

Online advertising governance faces significant challenges due to the non-stationary nature of regulatory policies, where emerging mandates (e.g., restrictions on education or aesthetic anxiety) create severe label inconsistencies and reasoning ambiguities in historical datasets. In this paper, we propose ARGUS, a policy-adaptive governance system that enables evolving reinforcement through multi-agent adversarial umpiring. ARGUS addresses the sparsity of new policy data by employing a three-stage framework: (1) Policy Seeding for initial perception; (2) Adversarial Label Rectification, which utilizes a "Prosecutor-Defender-Umpire" architecture to resolve conflicts between stale labels and new mandates; and (3) Latent Knowledge Discovery, which employs a tripartite dialectical discussion to unearth sophisticated, "gray-area" violations. By leveraging RAG-enhanced policy knowledge and Chain-of-Thought synthesis as dynamic rewards for reinforcement learning, ARGUS synchronizes its reasoning pathways with evolving regulations. Extensive experiments on both industrial and public datasets demonstrate that ARGUS significantly outperforms traditional fine-tuning baselines, achieving superior policy-adaptive learning with minimal gold data.

[Industry] Automatic Prompt Engineering for Scalable Prompt Inversion in Text-to-Image Ad Generation ► *Underline Video*

Zixin Ding, Qi Zeng, Boying Gong, Wenlong Deng, Bo Pan and Yuxin Chen

While prompt engineering offers effective control over Text-to-Image (T2I) generation, it remains labor-intensive for large-scale production. We present PRISM-DUEL, a black-box framework that formalizes prompt optimization as Automatic Prompt Engineering (APE), motivated by advertising workflows requiring low-latency, diverse variants faithful to a human-designed ads. Since zero-shot LLMs are unreliable judges of image quality, PRISM-DUEL obtains label-free pairwise preferences and rationales from an LLM judge over pairs of generated images, then uses a dueling-bandit optimizer to optimize a prompt for generating controlled variations while matching the reference ad's visual content. By iteratively steering the prompt distribution towards higher-quality generations and improving posterior calibration, PRISM-DUEL preserves visual similarity and semantic faithfulness while increasing diversity. Experiments on PartiPrompts and DreamBooth across Gemini 2.5 Flash Image, FLUX.1, and Qwen-Image show consistent gains over strong baselines in visual faithfulness and prompt interpretability.

[Industry] Benchmarking Testing in Automated Theorem Proving ► *Underline Video*

Jongyoon_Kim1 Jongyoon_Kim1, Hojae_Han1 Hojae_Han1 and seung-won_hwang2 seung-won_hwang2

Recent advances in large language models (LLMs) have shown promise in formal theorem proving, yet evaluating semantic correctness remains challenging. Existing evaluations rely on indirect proxies such as a lexical overlap to human-annotated proof, or expensive manual inspection. Inspired by the shift from lexical comparison to test-case execution-based evaluation in code generation, we propose TE^2, a framework that evaluates semantic correctness of formal theorems: a generated theorem is considered correct only if all dependent downstream theorems compile successfully, analogous to integration testing. We construct a benchmark from 6 real-world Lean 4 repositories, comprising 2,206 problems paired with 41 downstream theorems on average, automatically extracted without human effort. Experiments demonstrate that while state-of-the-art models achieve high compilation success, they perform significantly worse under our semantic metric. The best model, Claude-Sonnet-4.5, achieves only 38.9% Execution Accuracy in the hard set, given both natural language proof and downstream theorems as context, revealing a critical gap in current theorem generation capabilities.

[Industry] CascadeDebate: Multi-Agent Deliberation for Cost-Aware LLM Cascades ► *Underline Video*

Raeyoung Chang; Dongwook Kwon; Jisoo Lee; Nikhil Verma

Cascaded LLM systems coordinate models of varying sizes with human experts to balance accuracy, cost, and abstention under uncertainty. However, single-model tiers at each stage falter on ambiguous queries, triggering premature escalations to costlier models or experts due to

under-confidence and inefficient compute scaling. CascadeDebate addresses this critical gap by inserting multi-agent deliberation directly at each tier's escalation boundary. Confidence-based routers activate lightweight agent ensembles only for uncertain cases, enabling consensus-driven resolution of ambiguities internally, without invoking higher-cost upgrades. Our unified architecture alternates single-model inference with selective multi-agent deliberation across model scales, culminating in human experts as final fallback. This design scales test-time compute dynamically to query difficulty. Across five benchmarks spanning science, medicine, and general knowledge, CascadeDebate outperforms strong single-model cascades and standalone multi-agent systems by up to 26.75%. An online threshold optimizer proves essential, boosting accuracy 20.98-52.33% relative improvement over fixed policies and enabling elastic adaptation to real-world distributions.

[Industry] CAT: Confidence-Adaptive Thinking for Efficient Reasoning of Large Reasoning Models ▶ *Underline Video*

Qizhi Jiang, Shuo Wang, Pei Ke, Yuhang Song and Ke Qin

Large Reasoning Models (LRMs) have achieved remarkable success on complex tasks by leveraging long chain-of-thought (CoT) trajectories, yet they frequently exhibit overthinking on simple queries, resulting in significant token overhead and reduced inference efficiency. However, existing compression methods predominantly apply uniform length reduction or rely on coarse-grained difficulty estimation, often leading to performance degradation on difficult problems. To address this limitation, we propose Confidence-Adaptive Thinking (CAT), a framework that incorporates the model's intrinsic self-certainty signals as confidence into the preference optimization process, which autonomously modulates reasoning lengths based on problem difficulty. Experimental results show that CAT consistently outperforms state-of-the-art baselines on reasoning accuracy across multiple benchmarks on different base models. Our work enables LRMs to effectively compress confident responses while deliberating on uncertain ones, offering a potentially robust solution for balancing accuracy and latency in practical industrial scenarios.

[Industry] ClinDet-Bench: Beyond Abstention, Evaluating Judgment Determinability of LLMs in Clinical Decision-Making ▶ *Underline Video*

YUSUKE WATANABE, Yohei Kobashi, Takeshi Kojima, Yusuke Iwasawa, Yasushi Okuno and Yutaka Matsuo

Clinical decisions are often required under incomplete information. Clinical experts must identify whether available information is sufficient for judgment, as both premature conclusion and unnecessary abstention can compromise patient safety. To evaluate this capability of large language models (LLMs), we developed ClinDet-Bench, a benchmark based on clinical scoring systems that decomposes incomplete-information scenarios into determinable and undeterminable conditions. Identifying determinability requires considering all hypotheses about missing information, including unlikely ones, and verifying whether the conclusion holds across them. We find that recent LLMs fail to identify determinability under incomplete information, producing both premature judgments and excessive abstention, even when they correctly explain the underlying scoring knowledge and perform well under complete information, as they tend to assume plausible values for missing information. These findings suggest that existing benchmarks are insufficient to evaluate the safety of LLMs in clinical settings. ClinDet-Bench provides a framework for evaluating determinability recognition, leading to appropriate abstention, with potential applicability to medicine and other high-stakes domains, and is publicly available.

[Industry] DeepResearch Retail: Benchmarking Tool-Augmented Deep Research in the E-Commerce Domain ▶ *Underline Video*

Rafael Ferreira, Flavio Di Palo, Huilin Lu, Ayush Jain and Harsha Aduri

Deep Research (DR) systems autonomously retrieve and synthesize information from web sources, however, industrial DR applications face a critical gap: effective integration of internal tools with web search. We introduce DeepResearch Retail, an evaluation framework grounded in real-world e-commerce data for assessing Deep Research with tools (DR+Tools) in realistic commercial settings. The framework evaluates both factual faithfulness and multidimensional response quality when reasoning over heterogeneous web and internal data sources. We further present Hybrid-ReAct, a multi-agent architecture that demonstrates how collaborative reasoning and tool use can produce evidence-grounded answers. Experimental results validate our framework's utility, showing improvements in agent's performance when leveraging web-page information and multi-agent specialization.

[Industry] Development and Benchmarking of a Blended Human-AI Qualitative Research Assistant ▶ *Underline Video*

Joseph Matveyenko; James Liu; John David Parsons; Ryan A. Brown; Alina I. Palimaru; Vipul Gupta; Prateek Puri;

Qualitative research emphasizes constructing meaning through iterative engagement with textual data. Traditionally, this human-driven process requires navigating coder fatigue and interpretative drift, thus posing challenges when scaling analysis to larger, more complex datasets. Computational approaches to augment qualitative research have been met with skepticism, partly due to their inability to replicate the nuance, context-awareness, and sophistication of human analysis. LLMs, however, present new opportunities to automate aspects of qualitative analysis while upholding rigor and research quality. In this work, we present and benchmark Muse, an interactive qualitative research system that allows researchers to identify themes and annotate datasets, achieving an inter-rater reliability between Muse and humans of Cohens = 0.71 for well-specified codes.

[Industry] Diagnose, Then Repair: A Two-Stage MQM-Guided Post-Editing Framework for Domain-Specific Machine Translation ▶ *Underline Video*

Ji Hun Wang and Siyu Wu

LLM-based machine translation (MT) evaluation can closely match human judgments, but in practice it remains large diagnostic, with the signals rarely translating into direct quality improvements under real production constraints. We propose a two-stage, evaluator-guided

automatic post-editing framework that turns MQM-style evaluation into targeted repairs: a retrieval-augmented LLM evaluator outputs structured, span-level MQM diagnoses under an explicit edit contract, and a separate LLM post-editor applies minimal edits restricted to those diagnoses. This separation improves controllability and reduces paraphrastic drift compared to one-stage "judge-and-refine" baselines. In a systematic study involving seven LLMs spanning three model providers and English-to-German and English-to-Chinese test sets representing the e-commerce domain, our best configuration consistently improves both COMET-22 and CometKiwi scores over one-stage post-edit methods, while the evaluator's error spans and severities show strong agreement with human MQM annotations and human editor preferences.

[Industry] DocSplit: A Comprehensive Benchmark Dataset and Evaluation Approach for Document Packet Recognition and Splitting ▶ *Underline Video*

Md Mofijul Islam, Md Sirajus Salekin, Nivedha Balakrishnan, Vincil C. Bishop III, Niharika Jain, Spencer Romo, Bob Strahan, Boyi Xie and Diego A. Socolinsky

Document understanding in real-world applications often requires processing heterogeneous, multi-page document packets containing multiple documents stitched together. Despite recent advances in visual document understanding, the fundamental task of document packet splitting, which involves separating a document packet into individual units, remains largely unaddressed. We present the first comprehensive benchmark dataset, DocSplit, along with novel evaluation metrics for assessing the document packet splitting capabilities of large language models. DocSplit comprises five datasets of varying complexity, covering diverse document types, layouts, and multimodal settings. We formalize the DocSplit task, which requires models to identify document boundaries, classify document types, and maintain correct page ordering within a document packet. The benchmark addresses real-world challenges, including out-of-order pages, interleaved documents, and documents lacking clear demarcations. We conduct extensive experiments evaluating multimodal LLMs on our datasets, revealing significant performance gaps in current models' ability to handle complex document splitting tasks. The DocSplit benchmark datasets and proposed novel evaluation metrics provide a systematic framework for advancing document understanding capabilities essential for legal, financial, healthcare, and other document-intensive domains. We release the datasets and evaluation code to facilitate future research in document packet processing.

[Industry] Dont Stop Early: Scalable Enterprise Deep Research with Controlled Information Flow and Evidence-Aware Termination ▶ *Underline Video*

Prafulla Kumar Choubey, Kung-Hsiang Huang, Pranav Narayanan Venkit, Jiaxin Zhang, Vaibhav Vats, Yu Li, XIANGYU PENG and Chien-Sheng Wu

Enterprise deep research often fails to produce decision-ready reports due to uneven information coverage, context explosion, and premature stopping. We propose a scalable Enterprise Deep Research (EDR) architecture to address these failures. Our system (i) decomposes requests into coverage-driven objectives via outline generation with reflection, (ii) localizes context with dependency-guided execution and explicit information sharing, and (iii) enforces evidence-based completion criteria so agents iteratively collect information until sufficiency conditions are met. We evaluate on an internal sales-enablement task and the public DeepResearch Bench benchmark, where our proposed system design achieves the strongest overall performance against competitive deep-research baselines. The results show that dependency-controlled context and explicit evidence sufficiency criteria reduce premature stopping and improve the consistency and depth of enterprise research outputs. We will make our code and permitted data publicly available.

[Industry] DQA: Diagnostic Question Answering for IT Support ▶ *Underline Video*

Vishaal Kapoor, Mariam Dundua, Evren Yortucboylu, Sarthak Ahuja, Rebecca Steinert, Vaibhavi padala, Neda Kordjazi, Derek Ho, Yiming Li and Jennifer Whitted

Enterprise IT support interactions are fundamentally diagnostic: effective resolution requires iterative evidence gathering from ambiguous user reports to identify an underlying root cause. While retrieval-augmented generation (RAG) provides grounding through historical cases, standard multi-turn RAG systems lack explicit diagnostic state and therefore struggle to accumulate evidence and resolve competing hypotheses across turns. We introduce DQA, a diagnostic question-answering framework that maintains persistent diagnostic state and aggregates retrieved cases at the level of root causes rather than individual documents. DQA combines conversational query rewriting, retrieval aggregation, and state-conditioned response generation to support systematic troubleshooting under enterprise latency and context constraints. We evaluate DQA on 150 anonymized enterprise IT support scenarios using a replay-based protocol. Averaged over three independent runs, DQA achieves a 78.7% success rate under a trajectory-level success criterion, compared to 41.3% for a multi-turn RAG baseline, while reducing average turns from 8.4 to 3.9. This improvement reflects the benefit of explicitly representing competing explanations and aggregating evidence across turns in unscripted troubleshooting.

[Industry] DuIVRS-2: An LLM-based Interactive Voice Response System for Large-scale POI Attribute Acquisition ▶ *Underline Video*

Le Zhang, Shengming Zhang, Rui Zha, Yunpeng Wu, Jingbo Zhou and Jizhou Huang;

Accurate Point of Interest (POI) attribute acquisition is essential for location-based services, yet traditional modular IVR systems suffer from error accumulation and high maintenance overhead. We present DuIVRS-2, an LLM-based end-to-end framework designed for large-scale POI verification at Baidu Maps. To address the long-tail distribution of real-world interactions, our methodology first employs an FSM-guided data augmentation strategy to synthesize a balanced and diverse training dataset. We then streamline dialogue management via a selective generation scheme combined with a Chain-of-Thought (CoT) mechanism, which ensures output stability and effectively eliminates

hallucinations in industrial settings. To facilitate continuous policy refinement with minimal manual effort, we design a cooperative iterative learning framework that leverages a dual-evaluator voting system. Deployed in production for two months, DuIVRS-2 processed 0.4 million calls daily and achieved a 83.9% Task Success Rate, outperforming its predecessor by 4 percentage points while maintaining a low reaction time of 130ms. This work provides a production-proven reference for developing robust, cost-effective LLM agents for large-scale industrial dialogue applications.

[Industry] Efficient Agent Evaluation via Diversity-Guided User Simulation ► *Underline Video*

Itay Nakash, George Kour and Ateret Anaby Tavor

Large language models (LLMs) are increasingly deployed as customer-facing agents, yet evaluating their reliability remains challenging due to stochastic, multi-turn interactions. Current evaluation protocols rely on linear Monte Carlo rollouts of full agent-user conversations to estimate success. This approach is computationally inefficient - reprocessing identical conversation prefixes across runs, and often fails to uncover deep failure modes triggered by rare user behaviors. We introduce DIVERT (Diversity-Induced Evaluation via Branching of Trajectories), a snapshot-based, coverage-guided user simulation framework for efficient and systematic exploration of multi-turn agent behavior. DIVERT captures the full agent-environment state at critical junctions and resumes execution from these points, reusing shared prefixes to avoid redundant regeneration and reduce token cost. From each junction, it branches with targeted, diverse user responses, enabling directed exploration of alternative interaction paths while preserving task intent. By reallocating computation from redundant restarts to behaviorally salient mid-trajectory states, DIVERT steers evaluation toward under-explored semantic regions and rare interaction failures. Experiments on realistic multi-domain benchmarks show that our method consistently improves failure discovery efficiency and task-level coverage compared to standard linear rollout evaluation, without increasing overall cost.

[Industry] EvidenceDriven Reasoning for Industrial Maintenance Using Heterogeneous Data ► *Underline Video*

Fearghal O'Donncha, Nianjun Zhou, Natalia Martinez, James T Rayfield, Fenno F. Heath III, Abigail Langbridge and Roman Vaculin

Industrial maintenance platforms contain rich but fragmented evidence, including free-text work orders, heterogeneous operational sensors or indicators, and structured failure knowledge. These sources are often analyzed in isolation, producing alerts or forecasts that do not support conditional decision-making: given this asset history and behavior, what is happening and what action is warranted? We present Condition Insight Agent, a deployed decision-support framework that integrates maintenance language, behavioral abstractions of operational data, and engineering failure semantics to produce evidence-grounded explanations and advisory actions. The system constrains reasoning through deterministic evidence construction and structured failure knowledge, and applies a rule-based verification loop to suppress unsupported conclusions. Case studies from production CMMS deployments show that this verification-first design operates reliably under heterogeneous and incomplete data while preserving human oversight. Our results demonstrate how constrained LLM-based reasoning can function as a governed decision-support layer for industrial maintenance.

[Industry] FinHarmBench: Financial Jailbreak Benchmark and Unsupervised Safety Fine-Tuning via Refusal Steering Distillation ► *Underline Video*

Yubin Choi and Yujin Yang

Financial Large Language Models (LLMs) exhibit strong domain expertise but remain vulnerable to financially harmful prompts. To systematically assess this vulnerability, we introduce FinHarmBench, a benchmark designed to evaluate financially harmful and confusable benign prompts. Our analysis reveals a concerning result that financial LLMs can be less robust than general-purpose models, suggesting that domain adaptation alone does not guarantee financial safety alignment. To address this issue, we propose Financial Refusal Steering Distillation (FiRSD), an unsupervised training framework that strengthens financial-domain safety by learning and distilling a financial refusal direction at the representation level. FiRSD enhances refusal behavior without requiring annotated refusal responses. Experiments show that FiRSD substantially improves safety while largely preserving task capability. These results highlight the importance of domain-aware safety alignment for high-stakes financial applications.

[Industry] FROST: Factual Reasoning via Optimized Stochastic Trajectories in Large Language Models during Inference ► *Underline Video*

Soumedhik Bharati, Ebad Shabbir and Jiechao Gao

Large language models (LLMs) face a fundamental trade-off between factual consistency and reasoning diversity. Deterministic decoding strategies prioritize reliability but may fail to explore alternative solution paths, while high-temperature sampling increases exploration at the cost of factual accuracy. We present FROST (Factual Reasoning via Optimized Stochastic Trajectories), an inference framework that systematically balances exploration and exploitation for cost-efficient reasoning in production environments. FROST combines deterministic inference from large models with targeted stochastic sampling from smaller models, followed by multi-criteria validation to select factually grounded outputs. The framework operates entirely at inference time, requiring no additional training or fine-tuning. Across HotpotQA, CommonsenseQA, and MMLU, FROST achieves 3–5% accuracy improvements over standard chain-of-thought prompting. Crucially, it maintains high cost-efficiency relative to other robust inference methods, achieving a 28% cost reduction compared to Self-Consistency ensembles through strategic delegation to smaller models. Validation analysis demonstrates that FROST reduces unsupported outputs by 40% relative to baseline methods. Task-stratified evaluation reveals that exploration benefits scale with problem ambiguity, with alternative reasoning paths improving accuracy in 23–34% of multi-hop questions. These results establish FROST as a practical approach for deploying reliable and efficient reasoning systems in production environments.

[Industry] GrocLM: Grocery Category Recommendation in E-Commerce with Large Language Models ▶ *Underline Video*

Yuan Zhong, Chuanwei Ruan, Moein Hasani, Tejaswi Tenneti, Haixun Wang and Fenglong Ma

The rapid growth of online grocery shopping demands recommendation models that capture cyclical purchasing behavior and diverse user queries. At a large-scale online grocery e-commerce platform, traditional item-level models face scalability and accuracy limitations, motivating category-level recommendation as a more structured and business-aligned solution. We propose GrocLM, a fine-tuned language model for grocery category recommendation in a real-world production setting. Built with a two-stage LoRA-based fine-tuning strategy, GrocLM models repurchasing patterns and implicit category relationships from historical purchase and query data. To ensure discrete and valid outputs, we introduce a trie tree-based constrained decoding mechanism. Extensive experiments on proprietary production data and public datasets demonstrate that GrocLM significantly outperforms strong baselines and surpasses the deployed production model in live testing, improving recommendation precision, category diversity, and scalability.

[Industry] IPS: In-Prompt Process Supervision for Short Video Content Moderation ▶ *Underline Video*

Mingchao Liu, Yu Sun, Ruixiao Sun, Xin Dong, Xiang Shen, Hongwei Wang, Hongyu Xiong and Yang Song

Multimodal large language models (MLLMs) are effective at capturing the semantics of short video content; however, they often fail to attend to the policy-specific details required for reliable content moderation. To address this limitation, we introduce IPS, a novel framework that integrates In-prompt Process Supervision into MLLMs by introducing sequential reasoning over ancillary questions during fine-tuning. IPS consistently outperforms baseline MLLMs on public and proprietary benchmarks. Moreover, replacing human-annotated ancillary labels with MLLM-generated ones results in only marginal performance degradation, demonstrating robustness to noisy supervision and strong scalability with model-generated annotations. These findings establish IPS as a scalable and effective solution for complex multimodal classification in large-scale industrial settings.

[Industry] LatentGate: Low-Latency Semantic Routing via Frozen-Backbone Probing of Small Language Models ▶ *Underline Video*

Shivam Ratnakar, Abhiroop Talasila and Vinayak K Doifode

As Multi-Agent Systems (MAS) scale to hundreds of specialized agents, the routing layer becomes a critical bottleneck. Traditional approaches force a stark trade-off: prompt-based LLM routers deliver high semantic reasoning but incur prohibitive latency (~1500-2000ms) and cost that grows with agent count, while embedding-based routers operate at low latency (25-50ms on a T4 for cosine/centroid-style routing) but fail to capture nuanced functional intent, collapsing semantically similar but functionally distinct agents. We identify representation anisotropy, the geometric collapse of hidden-state vectors into a narrow cone, as a key mechanism underlying this embedding-based routing failure. We propose LatentGate, a non-generative routing architecture that extracts mean-pooled hidden states from a frozen small language model (SLM), applies PCA-whitening (decorrelation + variance normalization) to resolve the anisotropy, and trains a lightweight linear probe for agent classification. Experiments across 5 SLM backbones and 100 enterprise agents show that LatentGate achieves 98.8% in-domain accuracy and 80.0% out-of-distribution accuracy on natural queries, outperforming embedding-based routers by 13-22 absolute points. LatentGate operates at ~28ms on a T4 GPU; the SLM forward pass is independent of agent count, with classification adding an O(Ck) term (agents C, whitened dimension k) that is negligible at C = 100 and small relative to the SLM forward pass. The lightweight linear probe additionally enables sub-10ms warm-start retraining from user feedback, offering a path toward self-healing routing in production. We further benchmark prompt-based routing with GPT-4.1, GPT-4.1-nano, and Gemini 2.5 Flash, demonstrating that these degrade to 70-77% accuracy at 100 agents while incurring 1500-2000ms latency, confirming the need for non-generative alternatives. We categorize this work as Emerging, as it introduces a new routing primitive rather than reporting a completed long-term deployment study.

[Industry] Learning from Textual Radiology Reports: A Benchmark Dataset for Coronary CT Angiography ▶ *Underline Video*

Sudharshan Balaji, Zhiyu Liu, Zhengyuan Jiang, Shuo Lei, Yimin Chen, Yang Xiao, Shone Almeida, Mathew Karivelil, Christopher Malanga and Ning Wang

While coronary imaging is widely used for anatomical assessment, CCTA reports play a distinct last-mile role in clinical care. Rather than serving as an intermediate signal, CCTA provides an assessment of coronary disease severity (known as the CAD-RADS score) to guide patient management. However, real-world clinical text exhibits substantial heterogeneity in terminology and structure, leading to inconsistent interpretation by automated systems, even for clinically similar cases. Recent work leverages a direct application of LLMs for automated CAD-RADS scoring, but is limited by small, non-public, and homogeneous clinical data. We introduce CCTA-RADS, the largest publicly available dataset of 940 real-world CCTA reports from a major cardiovascular center, each annotated with CAD-RADS scores. Our analysis reveals that direct approaches, including state-of-the-art LLMs (GPT-4o, GPT-o3) and fine-tuned BERT models underperform on diverse real-world clinical data. To address these limitations, we propose a two-stage pipeline that decouples structuring from classification: an LLM-based parser normalizes heterogeneous reports into structured format, followed by fine-tuned BERT classification. This approach substantially improves the F1-score by 6%-13% compared with direct methods. We deploy our system as an interactive web interface that allows clinicians to upload CCTA reports for automated CAD-RADS assessment with SHAP and LIME explainability visualizations.

[Industry] LegalDrill: Diagnosis-Driven Synthesis for Legal Reasoning in Small Language Models ▶ *Underline Video*

Tianchun Li, Haochen Liu, Vishwa Pardeshi, Xingchen Wang, Tianci Liu, Huijun Zhao, Wei Fan and Jing Gao

Small language models (SLMs) are promising for real-world deployment due to their efficiency and low operational cost. However, their limited capacity struggles with high-stakes legal reasoning tasks that require coherent statute interpretation and logically consistent deduction.

Furthermore, training SLMs for such tasks demands high-quality, concise reasoning trajectories, which are prohibitively expensive to manually collect and difficult to curate via standard rejection sampling, which lacks granularity beyond final verdicts. To address these challenges, we propose LegalDrill, a diagnosis-driven synthesis framework that extracts and iteratively refines reasoning trajectories from a capable teacher via fine-grained prompting, then a self-reflective verification is employed to adaptively select the most effective data for the SLM student. The resulting data empower SLM training through supervised fine-tuning and direct preference optimization. Extensive experiments on several legal benchmarks demonstrate that LegalDrill significantly bolsters the legal reasoning capabilities of representative SLMs while bypassing the need for scarce expert annotations, paving a scalable path toward practical legal reasoning systems.

[Industry] Let LLM Tutors Ask First: Proactive LLM-Based Tutoring at Scale in a 1,500-Student Online Classroom ► *Underline Video*

Jonghoon Lee, Geonjae Youn, Seongmin Lee, Chaemoon Im, Joongheon Kim and Chuck Yoo

Large-scale introductory CS courses, often enrolling thousands of students, struggle to provide personalized support and encourage active participation. While recent large language models (LLMs) have enabled AI teaching assistants at scale, most existing systems remain reactive, responding only after students explicitly initiate queries. We present SCALA, a student-centered AI learning assistant designed to provide proactive support for students. SCALA introduces predictive query management, a mechanism that generates likely student questions and answers ahead of lectures. Students may choose to view these pre-generated question-answer pairs or engage in interactive conversations with our tutoring model via the same interface. We evaluate SCALA through a semester-long deployment in an undergraduate Python course with over 1,500 students, and find that predictive queries are frequently selected in practice and substantially overlap with real student questions. Based on student feedback, learners preferred SCALAs responses to their real queries over alternatives such as GPT-4o. These results suggest proactive support as a promising direction for future development of AI-powered teaching assistants. We will release our codebase and interactive demo upon acceptance.

[Industry] Leveraging Generative AI for Extracting Business Requirements from Legacy COBOL and PL/I Code ► *Underline Video*

Ankur Kalohia

Recovering business requirements from COBOL and PL/I portfolios is difficult because logic is scattered across interdependent programs and data definitions, and existing analyses seldom yield stakeholder-facing artifacts. We introduce an LLM-augmented reverse-engineering pipeline that couples grammar-based parsing and control-flow and data-flow analysis with a large language model to translate an enriched intermediate representation into structured specifications. The pipeline produces business requirements documents, explicit rule catalogues, end-to-end data lineage, create-read-update-delete matrices, and field-level source-to-target mappings, each linked to the supporting code. In a financial industry setting, containing 3.2 millions of lines of COBOL, the system achieves 93% agreement with expert-authored business rules and reduces documentation effort by approximately 70%, as measured against manually produced requirement documents and rule sets. On an additional internal corpus spanning 3.4M+ lines across online, batch, and job control workloads, the approach yields approximately 3.2-3.3x faster analysis while improving artifact consistency and traceability.

[Industry] LoCar: Localization-Aware Evaluation of In-Vehicle Assistants through Fine-Grained Sociolinguistic Control ► *Underline Video*

Seogyeong Jeong, Kiwoong Park, Seyoung Song, Eunsu Kim, Ken Friedl, Jaeho Kim and Alice Oh;

While Large Language Models (LLMs) are increasingly integrated into in-vehicle conversational systems, identifying the optimal model remains challenging due to the lack of domain-specific evaluation standards tailored to real-world deployment requirements. In this paper, we propose a novel evaluation framework for in-vehicle assistants, with a particular focus on Korean-language localization. Our empirical analysis reveals notable patterns in model behavior. First, fine-grained Korean honorific control remains unstable in current LLMs, indicating that precise speech-level realization must be explicitly evaluated in localization settings. Second, models exhibit weaker performance in strategic conversational metrics like clarification and proactivity. Our analysis suggests this stems from the inherent subjective complexity of these tasks, where our framework adopts a conservative evaluation stance to prioritize reliability. Together, our findings underscore that automotive AI must move beyond general competence toward precise linguistic tailoring and reliable, safety-oriented interaction management.

[Industry] Measuring and Mitigating Racial Bias in Embedding Models: A Comparative Study for Law Enforcement Retrieval ► *Underline Video*

Archan Dutta

Embedding models are often used for semantic retrieval in high-stakes domains such as law enforcement, where biased outputs can have severe consequences. We systematically measure racial bias in six widely used embedding models by computing similarity scores between crime incident texts that include racial identity tokens and simple law enforcement queries. The analysis reveals that racial descriptors consistently affect cosine similarity scores and retrieval rankings for semantically identical crime incidents. All models exhibit statistically significant bias, with magnitude varying across models. This study provides a comprehensive methodology and metrics to aid the selection of embedding models when deploying NLP-based systems in the law enforcement domain. Organizations can reduce bias at low cost through informed model selection. The methodology establishes reproducible metrics for measuring bias in embedding-based systems.

[Industry] MemoPhishAgent: Memory-Augmented Multi-Modal LLM Agent for Phishing URL Detection ► *Underline Video*

Xuan Chen, Hao Liu, Tao Yuan, Mehran Kafai, Piotr Habas and Xiangyu Zhang

Traditional phishing website detection relies on static heuristics or reference lists, which lag behind rapidly evolving attacks. While recent systems incorporate large language models (LLMs), they are still prompt-based, deterministic pipelines that underutilize reasoning capability. We present MemoPhishAgent (MPA), a memory-augmented multi-modal LLM agent that dynamically orchestrates phishing-specific tools and leverages episodic memories of past reasoning trajectories to guide decisions on recurring and novel threats. On two public datasets, MPA outperforms three state-of-the-art (SOTA) baselines, improving recall by 13.6%. To better reflect realistic, user-facing phishing detection performance, we further evaluate MPA on a benchmark of real-world suspicious URLs actively crawled from five social media platforms, where it improves recall by 20%. Detailed analysis shows episodic memory contributes up to 27% recall gain without introducing additional computational overhead. The ablation study confirms the necessity of the agent-based approach compared to prompt-based baselines and validates the effectiveness of our tool design. Finally, MPA is deployed in production, processing 60K targeted high-risk URLs weekly, and achieving 91.44% recall, providing proactive protection for millions of customers. Together, our results show that combining multi-modal reasoning with episodic memory yields robust, adaptable phishing detection in realistic user-exposure settings.

[Industry] MM-BizRAG: Rethinking Multimodal Retrieval-Augmented Generation for General Purpose Enterprise Q&A ► *Underline Video*

Hanoz Bhathena, Parin Rajesh Jhaveri, Rohan Mittal, Prateek Singh, Aymen Kallala, Rachneet Kaur, Yiqiao Jin, Zhen Zeng, Adwait Ratnaparkhi and Denis Kochedykov

Recent advances in multimodal retrieval-augmented generation (MM-RAG) have shifted toward minimal parsing, relying on page-level images for producing retriever embeddings and for answer generation. While efficient, this trend often neglects the rich, structured information embedded in complex enterprise documents, instead depending on pre-trained embeddings or vision-language models to implicitly capture such structure. In this work, we take a more direct approach: MM-BizRAG proactively extracts and represents document structure, applying explicit layout-aware parsing for report-type documents and leveraging page-level representations for slide-type documents, guided by a document structure-aware split. We present MM-BizRAG, a system which distinguishes between vertically structured (e.g., reports) and horizontally structured (e.g., slide decks) documents, unifying targeted document parsing with LLM-driven artifact transformation and flexible multimodal context assembly. Through experiments on a large, heterogeneous enterprise dataset and two public benchmarks (SlideVQA and FinRAGBench-V), we show that MM-BizRAGs proactive parsing and artifact transformation pipeline consistently outperforms state-of-the-art vision-centric baselines, especially on report style layouts. Furthermore, we introduce FastRAGEval, a single-call LLM Judge metric for fine-grained generative recall.

[Industry] NEST: Nested Evidence Survival for Retrieval ► *Underline Video*

Akshay Verma, Siddharth Pillai, Prateek Sircar and Deepak Gupta

Retrieval-Augmented Generation (RAG) sys- tems degrade sharply under extreme noise, where relevant evidence is sparse and easily pruned by static retrieval decisions. Existing approaches fixed top-k retrieval, hierarchical chunking, cross-encoder reranking, or policy- based iterative control- either rely on rigid heuristics or incur substantial computational overhead, and often fail to recover context- dependent evidence without introducing redun- dancy or latency. We introduce NEST (Nested Evidence Survival for Retrieval), a lightweight, training-free RAG framework that improves factual grounding by explicitly separating recall amplification from precision selection. NEST first maximizes re- call through Nested Evidence Survival, eval- uating candidates under nested retrieval con- texts to rescue evidence that would otherwise be pruned by static chunking. It then applies a survival-consistent Mean Reciprocal Rank (MRR) selection mechanism to retain evidence that remains salient across retrieval scopes, re- moving redundancy without harming recall. Evaluated on WebQuestions, HotpotQA (dis- tractor setting), and a proprietary InternalQA benchmark with 50M Common Crawl distrac- tors, NEST consistently outperforms strong adaptive RAG baselines, including DeepRAG, improving EM by up to +2.4 pp and F1 by +2.1 pp, while increasing retrieval recall by +6.8 pp. These gains are achieved with only 12-18 ms additional latency. Ablation studies confirm that Nested Evidence Survival drives recall im- provements, while MRR-based selection con- verts these gains into precision, demonstrating that recall-first retrieval with principled prun- ing can outperform iterative control and model scaling in retrieval-augmented generation.

[Industry] No Innocence in Styling: Discovery of Privacy Protection Capabilities and Security Risks in Consumer Generative AI Writing Assistants ► *Underline Video*

Mohd. Farhan Israk Soumik, SYED MHAMUDUL HASAN, Wanniarachchi Kankanamge Malithi Mithsara, Ahmed Imteaj and Abdur R. Shahid

Generative AI writing assistants are now integrated into consumer platforms such as Apple Intelligence and Microsoft Copilot, enabling millions of users to automatically rewrite and stylize their text. While positioned as productivity tools, their deployment at scale introduces important and underexplored implications for privacy and platform safety. This paper examines the dual-use nature of platform-level text stylization. Stylization can enhance privacy by suppressing stylistic signals used for profiling and personal data inference. However, the same transformations can be leveraged to evade automated safeguards, including misinformation detection systems. We conduct empirical case studies on emotion inference and misinformation detection across benchmark datasets using deployed stylization modes. We evaluate downstream impact with fine-tuned open-source models and GPT-4o in a zero-shot setting. Our results show that stylization reduces emotion inference accuracy, lowering profiling risk, while increasing error rates in misinformation detection. This discovery reveal a measurable trade-off among privacy protection, moderation robustness, and stylization, highlighting new design and governance challenges for industry deployment.

[Industry] OSyTa: Optimizer-based System for Table Generation from Documents ▶ *Underline Video*
Gyanendra Shrestha, Todor Ivanov, Karthik Vemireddy, Anna Pyayt and Michael Gubanov
Modern data lakes contain vast and heterogeneous document collections, making Information Extraction (IE) a persistent and nontrivial challenge. Traditional IE systems are often rigid - i.e. domain-specific, require extensive supervision, or are limited to predefined entity types and data structures; LLM-based approaches are more flexible, but suffer from hallucinations, non-determinism, high computational costs - drawbacks inherent to all LLMs. To overcome these limitations, we introduce OSyTa, a novel scalable LLM-based IE system that processes documents of varying complexity, extracts relational tables, and can generate complex tables with Bi-dimensional metadata, enabling robust document-level IE. We evaluated OSyTa on real user extraction tasks and achieved high F1 score (92.1%) across three publicly available, large-scale corpora from two distinct domains (medical and legal) each containing numerous documents.

[Industry] PAR2-RAG: Planned Active Retrieval and Reasoning for Multi-Hop Question Answering ▶ *Underline Video*
Rongguang Wang
Multi-hop question answering (MHQA) is a practical bottleneck in industry applications such as enterprise assistants, customer-support copilots, and compliance analysis, where systems must combine evidence across multiple documents before answering. Large language models (LLMs) remain brittle in this setting: iterative retrieval can commit too early to low-recall trajectories, while planning-only approaches can produce static query sets that fail to adapt when intermediate evidence changes. We propose Planned Active Retrieval and Reasoning RAG (PAR^2-RAG), a training-free two-stage framework that separates coverage from commitment. PAR^2-RAG first performs breadth-first anchoring to build a high-recall evidence frontier, then applies depth-first refinement with evidence sufficiency control in an iterative loop. This design targets deployment constraints by avoiding retraining cycles, reducing maintenance overhead under changing corpora, and improving scalability across domains. Across four MHQA benchmarks, PAR^2-RAG consistently outperforms strong baselines: compared with IRCoT, it achieves up to 23.5% higher answer accuracy and up to 10.5% NDCG gains in retrieval quality.

[Industry] Policy Compliance of User Requests in Natural Language for AI Systems ▶ *Underline Video*
Pedro Cisneros-Velarde
Consider an organization whose users send requests in natural language to an AI system that fulfills them by carrying out specific tasks. In this paper, we consider the problem of ensuring such user requests comply with a list of diverse policies determined by the organization with the purpose of guaranteeing the safe and reliable use of the AI system. We propose, to the best of our knowledge, the first benchmark consisting of annotated user requests of diverse compliance with respect to a list of policies. Our benchmark is related to industrial applications in the technology sector. We then use our benchmark to evaluate the performance of various LLM models on policy compliance assessment under different solution methods. We analyze the differences on performance metrics across the models and solution methods, showcasing the challenging nature of our problem.

[Industry] Practical Guidelines for Model Merging in LLMs Pre-Training ▶ *Underline Video*
Giuseppe Curci, Stefano Simonazzi, Andrea Molinari and Andrea Zugarini
Model merging is widely used to combine fine-tuned models trained with different data distributions, tasks, or hyperparameters, yet its role during LLM pre-training remains underexplored. We systematically study checkpoint merging across training phases, focusing on the transition from stable to decaying learning rates. Across multiple scales, we find that simple averaging methods consistently improve performance during stable learning rate regimes, but gains sharply diminish during decay. We link this effect to reduced checkpoint diversity and show that merging effectiveness correlates with parameter-space variation. Strategies such as synthetic variability, task-vector merging, and cross-run merging yield only modest improvements. Our results provide practical insights on when merging is most effective in large-scale pre-training.

[Industry] ProductResearch: Training E-Commerce Deep Research Agents via Multi-Agent Synthetic Trajectory Distillation ▶ *Underline Video*
Jiangyuan Wang, Kejun Xiao, Huaipeng Zhao, Tao Luo and Xiaoyi Zeng
Large Language Model (LLM)-based agents show promise for e-commerce conversational shopping, yet existing implementations lack the interaction depth and contextual breadth required for complex product research. Meanwhile, the Deep Research paradigm, despite advancing information synthesis in web search, suffers from domain gaps when transferred to e-commerce. We propose ProductResearch, a multi-agent framework that synthesizes high-fidelity, long-horizon tool-use trajectories for training robust e-commerce shopping agents. The framework employs a User Agent to infer nuanced shopping intents from behavioral histories, and a Supervisor Agent that orchestrates iterative collaboration with a Research Agent to generate synthetic trajectories culminating in comprehensive, insightful product research reports. These trajectories are rigorously filtered and distilled through a reflective internalization process that consolidates multi-agent supervisory interactions into coherent single-role training examples, enabling effective fine-tuning of LLM agents for complex shopping inquiries. Extensive experiments show that a compact MoE model fine-tuned on our synthetic data achieves substantial improvements over its base model in response comprehensiveness, research depth, and user-perceived utility, approaching the performance of frontier proprietary deep research systems and establishing multi-agent synthetic trajectory training as an effective and scalable paradigm for enhancing LLM-based shopping assistance.

[Industry] Progressive Fine-Tuning for Cost-Effective Structured Attribute Generation in E-commerce ▶ *Underline Video*

Lakshman Kolasani and Fatemeh Taheri Dezaki
Large language models (LLMs) excel at structured information generation but face cost and latency challenges when deployed at scale in user-facing products. We present a parameter-efficient supervised fine-tuning pipeline for adapting a small language model (SLM) to structured attribute generation in e-commerce product listing, enabling continuous model improvement with implicit user feedback without expensive manual annotation. Our approach involves completeness-deficit guided curation, which ranks samples by divergence between model predictions and seller-submitted attributes, selecting the highest completeness gap examples for progressive fine-tuning. Our system is deployed on a large-scale product listing service, reducing inference costs by 98% and p90 latency by 70% using a fine-tuned SLM relative to the baseline LLM while preserving an 86.4% user acceptance rate, translating to significant monthly infrastructure savings.

[Industry] Prompt-Level Distillation: A Non-Parametric Alternative to Model Fine-Tuning for Efficient Reasoning ▶ *Underline Video*
Sanket Badhe and Deep Shah
Advanced reasoning typically requires Chain-of-Thought prompting, which is accurate but incurs prohibitive latency and substantial test-time inference costs. The standard alternative, fine-tuning smaller models, often sacrifices interpretability while introducing significant resource and operational overhead. To address these limitations, we introduce Prompt-Level Distillation (PLD). We extract explicit reasoning patterns from a Teacher model and organize them into a structured list of expressive instructions for the Student model's System Prompt. Evaluated on the StereoSet and Contract-NLI datasets using Gemma-3 4B, PLD improved Macro F1 scores from 57% to 90.0% and 67% to 83% respectively, enabling this compact model to match frontier performance with negligible latency overhead. These expressive instructions render the decision-making process transparent, allowing for full human verification of logic, making this approach ideal for regulated industries such as law, finance, and content moderation, as well as high-volume use cases and edge devices.

[Industry] RAPIDS: Resume Attack Prompt Injection Detection at Scale ▶ *Underline Video*
Yohann Augey, Joshua H. Levy and Arda Akdemir
The integration of Large Language Models (LLMs) into recruitment workflows has introduced a critical security vulnerability: indirect prompt injection attacks embedded within resumes can manipulate screening tools to override instructions, effectively jailbreaking the hiring process. Frontier LLMs can detect such anomalies, but deploying them at the scale required for high-volume recruitment is prohibitively slow and costly. At the same time, existing generic prompt injection detectors lack the domain specificity needed for nuanced resume attacks. To address this gap, we introduce RAPIDS, a scalable detection framework with three contributions. First, we release a synthetically generated dataset of injection snippets derived from curated attack seeds spanning multiple adversarial strategies to address data scarcity in this domain. Second, we fine-tune a lightweight Small Language Model (SLM) on this data that outperforms the best off-the-shelf detector by over 50% in relative F1 and approaches frontier LLM accuracy. Third, we propose a cascade architecture in which the fine-tuned SLM serves as a high-recall first stage followed by an LLM verifier. This design achieves 97.5% end-to-end recall on both evaluated datasets while delivering a 25-28x latency reduction over standalone frontier LLMs (GPT-5-mini), bringing expected per-request latency to 97-141ms at less than 3% of the API cost.

[Industry] SAJA: A Simple Approach to Judge Alignment for LLM-as-a-Judge ▶ *Underline Video*
Sneha Kola, Pankaj Kumar Sharma, Soumyadeep Dey, Bamdev Mishra and Mayur Datar
LLM-as-a-Judge systems are increasingly used to evaluate text at scale, yet production deployment demands low latency, minimal cost, and compatibility with closed-source APIs. Current approaches fall short in different ways; some require many LLM calls and per-dataset prompt tuning, others depend on logit access unavailable in commercial APIs, and yet others demand multiple rounds of LLM interaction for iterative feature discovery. We present SAJA (Simple Approach to Judge Alignment), built on the principle that task-specific alignment should reside in a lightweight calibration head, not in elaborate prompts or model internals. SAJA makes exactly one LLM call per item using a fixed structured rubric prompt, extracts a multi-dimensional feature vector, and maps it to a human-aligned score via a calibration head trained on a small number of human labels. No prompt optimization, no logit access, and no iterative LLM interaction are needed, yet SAJA matches far more complex systems across four evaluation paradigms, achieving 86% F1 on MT-Bench pairwise preference (vs.\{} 78% uncalibrated), competitive performance on five classification benchmarks with a single call, and +5.71% F1 over prompt-optimized baselines on proprietary data. Ablations confirm that multi-dimensional rubric features outperform one-dimensional calibration (SummEval improves from 0.60 to 0.74) and that coarse rubric outputs recover the same human alignment as full logit distributions (= 0.36 vs.\{} 0.37), establishing that logit access is unnecessary for calibrated judge alignment.

[Industry] SchemaRAG: Dynamic Large Schema Reduction for LLM-driven Structured Information Extraction ▶ *Underline Video*
Sin Yu Bonnie Ho, Arlie Coles, Erik Larsson, Eric Marshall, Nathan Bodenstab and Paul Vozila
Extracting structured data from unstructured text using large language models (LLMs) becomes challenging when the target schemas are large and complex. In such cases, including the full schema in the prompt increases cost and latency, risks lost-in-the-middle performance degradation, and can exceed context length limits. We propose SchemaRAG, a retrieval-augmented generation (RAG) framework that dynamically prunes the output schema space for schema-conditioned information extraction tasks by leveraging schema metadata and few-shot examples (when available). We evaluate SchemaRAG on real-world healthcare and e-commerce datasets. Our results show that SchemaRAG can achieve up to an 8.8% increase in micro-F1, a 47% reduction in latency, and a 48% reduction in token costs, demonstrating its practicality for large-schema extraction.

[Industry] Search-P1: Path-Centric Reward Shaping for Stable and Efficient Agentic RAG Training ▶ *Underline Video*

Tianle Xia, Ming Xu, Lingxiang Hu, Yiding Sun, wenwei li, Linfang Shang, Liqun Liu, Peng Shu, Huan Yu and Jie Jiang

Retrieval-Augmented Generation (RAG) enhances large language models (LLMs) by incorporating external knowledge, yet traditional single-round retrieval struggles with complex multi-step reasoning. Agentic RAG addresses this by enabling LLMs to dynamically decide when and what to retrieve, but current RL-based training methods suffer from sparse outcome rewards that discard intermediate signals and low sample efficiency where failed samples contribute nothing. We propose Search-P1, a framework that introduces path-centric reward shaping for agentic RAG training, comprising two key components: (1) Path-Centric Reward, which evaluates the structural quality of reasoning trajectories through order-agnostic step coverage and soft scoring that extracts learning signals even from failed samples, and (2) Dual-Track Path Scoring with offline-generated reference planners that assesses paths from both self-consistency and reference-alignment perspectives. Experiments on multiple QA benchmarks demonstrate that Search-P1 achieves significant improvements over Search-R1 and other strong baselines, with an average accuracy gain of 7.7 points.

[Industry] Securing the Tool Layer: A Threat Taxonomy and Runtime Defense Framework for Model Context Protocol Deployments ▶ *Underline Video*

Saurabh Yergattikar

The Model Context Protocol (MCP) has rapidly emerged as the dominant standard for connecting large language models to external tools, databases, and services. Yet this convenience introduces a fundamentally new attack surface that existing LLM safety measures fail to address: adversaries can now compromise AI agents not through the user prompt, but through the tools the agent trusts. We present ShieldMCP, a runtime security framework grounded in two contributions. First, we introduce a structured threat taxonomy derived from analysis of 80+ attack techniques catalogued under the SAFE-MCP initiative within the Linux Foundation's OpenSSF, spanning 14 tactical categories adapted from the MITRE ATT&CK methodology. Second, we describe an interception architecture that performs real-time validation of MCP tool calls and responses, combining structural analysis with semantic intent verification to detect tool poisoning, indirect prompt injection through tool outputs, and supply chain manipulation. In red-team evaluation across five popular LLM backends, ShieldMCP reduces attack success rates from 74% to under 9% for tool poisoning and from 47% to under 6% for indirect prompt injection via tool responses, while adding fewer than 120ms of median latency per tool call. We discuss deployment considerations, the tension between security and agent utility, and lessons applicable to any organization integrating MCP into production workflows. Our framework is categorized as an Emerging system intended for real-world deployment.

[Industry] Small Agents, Big Gains: Journey-Aware and Critic-Guided Simulation for Long-Horizon Shopping Dialogues ▶ *Underline Video*

Qing Ping, Changyou Chen and Binxuan Huang

Modern e-commerce assistants must go beyond simple product search to support inspiration, comparison, and tool-grounded fact-checking across non-linear shopping journeys. However, distilling these complex behaviors into efficient, deployable models is bottle-necked by a lack of post-training data: trajectories must cover diverse agentic workflows with high fidelity, yet the desired outputs are open-ended without a single ground truth. We propose a closed-loop Multi-Agent Simulation Framework to synthesize diverse, faithful, and policy-aligned shopping trajectories. The system orchestrates a journey-aware, stateful user simulator to drive exploration, a shopping agent that manages both tools and UI elements, and a critic agent that provides rubric-driven feedback to iteratively refine the data. On a domain-specific benchmark, this synthetic data enables a small model to significantly outperform same-size baselines and surpass a large-model baseline, achieving near-zero tool-calling errors with 8x higher inference throughput.

[Industry] Sommelier: Scalable Open Multi-turn Audio Pre-processing for Full-duplex Speech Language Models ▶ *Underline Video*

Kyudan Jung, Jihwan Kim, Soyoon Kim, Jeonghoon Kim, Jaegul Choo and Cheonbok Park

As the paradigm of AI shifts from text-based LLMs to Speech Language Models (SLMs), there is a growing demand for full-duplex systems capable of real-time, natural human-computer interaction. However, the development of such models is constrained by the scarcity of high-quality, multi-speaker conversational data, as existing large-scale resources are predominantly single-speaker or limited in volume. Addressing the complex dynamics of natural dialogue, such as overlapping and back-channeling remains a challenge, with standard processing pipelines suffering from diarization errors and ASR hallucinations. To bridge this gap, we present a robust and scalable open-source data processing pipeline designed for full-duplex model. Our code and project page are publicly available at https://anonymous-2001-j.github.io/sommelier.github.io/.

[Industry] Structure-Guided Entity Resolution: Fine-Tuning LLMs for Robust Name Matching in Complex Linguistic Contexts ▶ *Underline Video*

Shivam Chourasia, Hitesh Kapoor and Nilesh Patil

Matching person names across heterogeneous records is a core challenge in entity resolution, especially within linguistically and culturally complex environments. Variations in naming conventions, inconsistent transliteration across scripts, and frequent data entry errors make it difficult to unify user identities, an essential requirement for Know Your Customer (KYC) compliance. While Large Language Models have shown promise in understanding natural language, they often struggle with the structured ambiguity present in such domain-specific settings. This paper introduces Structure-Guided Entity Resolution (SGER), a novel framework that fine-tunes an LLM through a two-phase curriculum. The model is first trained to parse the grammatical and semantic structure of personal names, then optimized for the downstream

task of binary entity matching. We evaluate SGER in the challenging context of Indian identity data, one of the most linguistically diverse and noisy environments globally. SGER achieves 99.02% accuracy and an F1 of 0.994 on a held-out set of 50,000 real-world pairs, outperforming GPT-4o few-shot prompting and single-stage fine-tuning baselines. The system is fully deployed in production at a major fantasy sports platform, serving 250M+ users. Our results demonstrate that curriculum-guided training enables robust, high-precision entity resolution in real-world multilingual systems at scale.

[Industry] Synthetic Text Detection in the Age of Large Language Models: Watermark vs. Automatic Detection ▶ *Underline Video*
Adaku Uchendu
Given the ubiquitous nature of Large Language Models (LLMs) and its impressive capabilities, malicious uses of this technology to generate harmful content have been observed. Thus, to mitigate this serious security risk LLMs pose, many researchers have proposed two techniques for detecting synthetic texts generated from LLMs - watermark and automatic detection. The idea with watermarking LLMs involves infusing generated content with algorithmically-identifiable patterns during generation. This makes accurate synthetic text detection achievable with watermark detection. While, for automatic detection, the focus is on using statistical and linguistic cues to reveal authorship of texts as human or LLM. Currently, both types of synthetic text detectors achieve state-of-the-art performance, however, the better detector is still unknown. To ascertain the better detection method, we evaluate each method on their performance on both unperturbed and perturbed (i.e., adversarially manipulated texts) data. We perform a comprehensive study across six different sizes of Qwen2.5 models, six watermark techniques and detectors, two automatic detectors, three authorship obfuscation methods for different levels of syntactic changes, and two datasets of different text lengths. Our results suggest that there is no detector that consistently outperforms on all scenarios. However, we observe that the (1) automatic detectors are better for short synthetic text detection; and (2) watermark detectors perform better defending against the word-level attack implemented.

[Industry] TeachMaster: Generative Teaching via Code ▶ *Underline Video*
Yuheng Wang, Runde Yang, Lin Wu, Jie Zhang, Jingru Fan, Ruoyu Fu, Tianle Zhou, Huatao Li, Ruijie Shi, Siheng Chen, Weinan E and Chen Qian
The scalability of high-quality online education is hindered by the high costs and slow cycles of manual content creation. Despite advancements in video generation, current approaches often fail to ensure pedagogical structure and precise control due to their pixel-level, black-box nature. In this paper, we propose Generative Teaching, a novel paradigm shifting educators from manual creators to high-level directors who focus on pedagogical intents while agents handle the execution. To realize this vision, we introduce TeachMaster, a multi-agent framework that leverages code as an intermediate semantic medium. Unlike traditional video generation methods, TeachMaster orchestrates a collaborative team of agents, spanning planning, design, and rendering, to automate the production of interpretable, editable, and curriculum-ready educational videos. Experiments validate that TeachMaster significantly boosts production efficiency without compromising structural coherence or visual fidelity, slashing production costs to only 0.3% of traditional online course videos and providing a robust solution for scalable education.

[Industry] Toward Scalable Verifiable Reward: Proxy State-Based Evaluation for Multi-turn Tool-Calling LLM Agents ▶ *Underline Video*
Yun-Shiuan Chuang, Chaitanya Kulkarni, Alec M. Chiu, Avinash Thangali, Zijie Pan, Shivani Shekhar, Yirou Ge, Yixi Li, Uma Kona, Linsey Pang and Prakhar Mehrotra
Interactive large language model (LLM) agents operating via multi-turn dialogue and multi-step tool calling are increasingly used in production. Benchmarks for these agents must both reliably compare models and yield on-policy training data. Prior agentic benchmarks (e.g., tau-bench, tau^2-bench, AppWorld) rely on fully deterministic backends, which are costly to build and iterate. We propose Proxy State-Based Evaluation, an LLM-driven simulation framework that preserves final state-based evaluation without a deterministic database. Specifically, a scenario specifies the user goal, user/system facts, expected final state, and expected agent behavior, and an LLM state tracker infers a structured proxy state from the full interaction trace. LLM judges then verify goal completion and detect tool/user hallucinations against scenario constraints. Empirically, our benchmark produces stable, model-differentiating rankings across families and inference-time reasoning efforts, and its on- and off-policy rollouts provide supervision that transfers to unseen scenarios. Careful scenario specification yields near-zero simulator hallucination rates, as supported by ablation studies. The framework also supports sensitivity analyses over user personas. Human-LLM judge agreement exceeds 90%, indicating reliable automated evaluation. Overall, proxy state-based evaluation offers a practical, scalable alternative to deterministic agentic benchmarks for industrial LLM agents.

[Industry] VEG: Verbal epsilon-greedy for Semantic Exploration in Multi-Turn RL Agents ▶ *Underline Video*
Yongchang Hao, Jie Hao, Yongsheng Mei, Ze Ye, Junyi Chai, Bin Guo, Benjamin Z. Yao, Chenlei Guo and Lili Mou;
Reinforcement learning (RL) has become a cornerstone of the post-training pipeline for large language models (LLMs), enabling capabilities such as complex reasoning and tool use. However, standard RL approaches face significant challenges due to reward sparsity. Moreover, LLMs typically exhibit mode-seeking behavior, concentrating probability mass on high-likelihood regions. This lack of diversity biases the model toward premature exploitation, hindering the exploration necessary for optimal learning. To address this, we propose VEG (verbal epsilon-greedy), a novel framework that leverages external feedback as a dynamic control variable to explicitly balance exploration and exploitation within the semantic space. This method not only supplements sparse final rewards with intermediate signals but also enforces sustained exploration throughout the training process. Experiments on Tau Bench and SearchQA demonstrate that our method achieves

superior accuracy compared to standard RL baselines. Notably, the trained policy eventually outperforms the external feedback model itself, demonstrating that VEG enables the agent to effectively filter and improve upon the guidance it receives.

Session 10: Virtual Presentations 3 - 12:45-14:15

Virtual Presentations 3

12:45-14:15 (Zoom)

[Long] ACBQ: Adaptive Cross-Block Quantization of Large Language Models ▶ *Underline Video*

Hailing Wang, Jianglin Lu, Yitian Zhang, Huimin Zeng and Yun Fu

Post-training quantization (PTQ) has emerged as a promising approach for reducing the memory footprint and computational cost of large language models (LLMs), enabling efficient deployment without full model retraining. However, existing PTQ methods struggle to simultaneously support weight-activation joint quantization and extreme low-bit weight quantization. This limitation primarily arises from the depth of LLMs and their strong cross-layer dependencies, which cause quantization errors to propagate and accumulate across layers, ultimately leading to significant performance degradation. In this paper, we present ACBQ, a simple yet effective framework that simultaneously addresses weight-activation joint quantization and extreme weight quantization. We first propose a granular quantization strategy that treats self-attention and FFN as separate quantization units with module-specific optimization objectives. To mitigate the propagation and accumulation of quantization errors across layers, we introduce an adaptive cross-block quantization strategy that explicitly accounts for cross-layer dependencies by encouraging consistency across blocks. Extensive experiments across diverse LLMs, including OPT and the LLaMA family, demonstrate that ACBQ achieves superior performance under both W4A4 and highly aggressive W2 settings, while incurring negligible additional computational overhead.

[Long] Accurate and Efficient Statistical Testing for Word Semantic Breadth ▶ *Underline Video*

Yo Ehara

Measuring the breadth of a words meaning, or its spread across contexts, has become feasible with contextualized token embeddings. A word type can be represented as a cloud of token vectors, with dispersion-based statistics serving as proxies for contextual diversity (Nagata and Tanaka-Ishii, ACL2025). These measurements are useful for deciding appropriate sense distinctions when constructing thesauri and domain-specific dictionaries. However, when comparing the breadth of two word types, naive hypothesis testing on dispersion can be misleading: differences in semantic direction can masquerade as dispersion differences, inflating Type-I error and yielding "statistically significant" outcomes even when there is no true breadth difference. This is problematic because significance testing should distinguish genuine effects from incidental fluctuations in small-difference regimes. We propose a Householder-aligned permutation test to isolate dispersion differences from directional differences. Our method applies a single Householder reflection to align the mean directions of the two word types and then performs a permutation test on the aligned token clouds, yielding calibrated, non-parametric p-values. For practicality, we introduce a GPU-oriented implementation that batches permutations and linear algebra operations. Empirically, our alignment reduced Type-I error by 32.5% while preserving sensitivity to genuine breadth differences, and achieved a 23x speedup over the CPU baseline.

[Findings] ActorMind: Emulating Human Actor Reasoning for Speech Role-Playing ▶ *Underline Video*

Xi Chen, Wei Xue and Yike Guo

Role-playing has garnered rising attention as it provides a strong foundation for human-machine interaction and facilitates sociological research. However, current work is confined to textual modalities, neglecting speech, which plays a predominant role in daily life, thus limiting genuine role-playing. To bridge this gap, we conceptualize and benchmark speech role-playing through ActorMindBench, and we present a corresponding reasoning framework, called ActorMind. Specifically, (1) Speech Role-Playing enables models to deliver spontaneous responses with personalized verbal traits based on their role, the scene, and spoken dialogue. (2) ActorMindBench is a hierarchical benchmark comprises Utterance-Level content with 7,653 utterances, Scene-Level content with 313 scenes, and Role-Level content with 6 roles. (3) ActorMind is an off-the-shelf, multi-agent, chain-of-though style reasoning framework that emulates how human actors perform in theaters. Concretely, ActorMind first reads its assigned role description via Eye Agent, then comprehends emotional cues within contextual spoken dialogues through Ear Agent. Subsequently, Brain Agent generates a descriptive emotional state, and finally, Mouth Agent delivers the scripts infused with corresponding emotion state. Experimental results demonstrate the effectiveness of ActorMind in enhancing speech role-playing. The project page is available at https://github.com/OzymandiasChen/ActorMind.

[Long] AdaDPI: Document-level Translation Adaptive Agent via Dynamic Parametric Internalization ▶ *Underline Video*

Hong Ren, Liting Deng, Shaolin Zhu and Deyi Xiong

Large Language Models (LLMs) have demonstrated remarkable capabilities in machine translation. However, maintaining discourse coherence and terminological consistency remains a persistent challenge in document-level translation (DocMT). Existing solutions, such as memory-based agents, predominantly rely on explicit context concatenation. This paradigm treats historical context as a static external resource, which often leads to context dilution, high inference latency, and superficial knowledge integration. To address these limitations, we

propose AdaDPI, an adaptive agentic framework that shifts the DocMT paradigm from static retrieval to dynamic parametric internalization. Specifically, we design a linguistic uncertainty monitor (LUM) to actively detect critical discourse discontinuities by the model's epistemic uncertainty. Upon detection, a context-to-parameter integrator (CPI) compiles retrieved external constraints directly into the model's intrinsic state via an online parameter adaptation mechanism. Through the online parameter adaptation on a lightweight adapter, AdaDPI internalizes document-specific norms into the model's intrinsic representations, enabling a progressive evolution of the translation strategy as the discourse unfolds. Extensive experiments on the discourse-rich GuoFeng and IWSLT2017 datasets demonstrate that AdaDPI significantly outperforms the SoTA baselines by more than 5 points on the consistency metric.

[Long] AdapShot: Adaptive Many-Shot In-Context Learning with Semantic-Aware KV Cache Reuse ▶ *Underline Video*

Jie Ou, Jinyu Guo, Shiyao Guo, Yuang Li, Ruiqi Wu, Zhaokun Wang, Wenyi Li and Wenhong Tian

Many-Shot In-Context Learning (ICL) has emerged as a promising paradigm, leveraging extensive examples to unlock the reasoning potential of Large Language Models (LLMs). However, existing methods typically rely on a predetermined, fixed number of shots. This static approach often fails to adapt to the varying difficulty of different queries, leading to either insufficient context or interference from noise. Furthermore, the prohibitive computational and memory costs of long contexts severely limit Many-Shot's feasibility. To address the above limitations, we propose AdapShot, which dynamically optimizes shot counts and leverages KV cache reuse for efficient inference. Specifically, we design a probe-based evaluation mechanism that utilizes output entropy to determine the optimal number of shots. To bypass the redundant prefilling computation during both the probing and inference phases, we incorporate a semantics-aware KV cache reuse strategy. Within this reuse strategy, to address positional encoding incompatibilities, we introduce a decoupling and re-encoding method that enables the flexible reordering of cached key-value pairs. Extensive experiments demonstrate that AdapShot achieves an average performance gain of 10% and a 4.64x speedup compared to state-of-the-art DBSA.

[Long] Adaptive Spatial and Temporal Redundancy Optimization for Efficient Reasoning in Large Language Models ▶ *Underline Video*

Tianle Chen, Pengyu Cheng, Qiyuan Zhu, Jiacheng Wang, Bei Liu, Hao Gu, Ruijie Shen, Xiaofeng Hou, Sirui Han and Jiacheng Liu

Large Language Models (LLMs) have achieved exceptional performance in complex reasoning via Chain-of-Thought (CoT), yet the associated computational costs remain prohibitive. CoT reasoning contains significant untapped efficiency potential across two dimensions: temporal redundancy, where reasoning steps may be unnecessary, and spatial redundancy, where computations can be performed at reduced precision. While current optimization techniques often necessitate resource-intensive fine-tuning or data curation, we introduce ASTRO (Adaptive Spatial and Temporal Redundancy Optimization), a training-free framework that simultaneously addresses both dimensions. ASTRO leverages Deweys reflective thinking model to segment reasoning phases, applying a progressive precision reduction strategy coupled with an entropy-based confidence mechanism for adaptive termination. Empirical results across diverse reasoning benchmarks demonstrate that ASTRO achieves up to an 11.3 x efficiency gain without compromising accuracy, highlighting the advantages of holistic multi-dimensional redundancy management over isolated optimization methods.

[Long] Adversarial Metric Learning for Fine-Grained Emotion Classification ▶ *Underline Video*

Junfan Chen, Sizhe Wu, Richong Zhang and Chunming Hu

Fine-grained emotion classification (FEC) requires distinguishing subtly different emotions, where the dominant errors come from closely confusable categories. Recent progress relies on contrastive learning with hard-pair mining, implicitly assuming that a fixed similarity metric is sufficient to optimize informative pairs. We argue that this assumption is fragile because defining whether two utterances are similar becomes a problem when the label space is crowded, and hard-pair mining under a fixed metric can systematically miss the worst confusions. Thus, we treat the similarity function as a learnable component and design an adversarial metric learning (AML) framework. It follows theoretical interpretations of metric-robust representations that better separate confusable emotions. AML trains a pairwise discriminator to maximally confuse two targeted hard pair types, while training the encoder to remain discriminative under this worst-case learned metric. Our code and data are released on GitHub.

[Long] Agentic Memory: Learning Unified Long-Term and Short-Term Memory Management for Large Language Model Agents ▶ *Underline Video*

Yi Yu, Liuyi Yao, Yuexiang Xie, Qingquan Tan, Jiaqi Feng, Yaliang Li and Libing Wu

Large language model (LLM) agents face fundamental limitations in long-horizon reasoning due to finite context windows, making effective memory management critical. Existing methods typically handle long-term memory (LTM) and short-term memory (STM) as separate components, relying on heuristics or auxiliary controllers, which limits adaptability and end-to-end optimization. In this paper, we propose Agentic Memory (AgeMem), a unified framework that integrates LTM and STM management directly into the agents policy. AgeMem exposes memory operations as tool-based actions, enabling the LLM agent to autonomously decide what and when to store, retrieve, update, summarize, or discard information. To train such unified behaviors, we propose a three-stage progressive reinforcement learning strategy and design a step-wise GRPO to address sparse and discontinuous rewards induced by memory operations. Experiments on five long-horizon benchmarks demonstrate that AgeMem consistently outperforms strong memory-augmented baselines across multiple LLM backbones, achieving improved task performance, higher-quality long-term memory, and more efficient context usage.

[Long] ALDEN: Reinforcement Learning for Active Navigation and Evidence Gathering in Long Documents ▶ *Underline Video*

Tianyu Yang, Terry Ruas, Yijun Tian, Jan Philip Wahle, Daniel Kurzawe and Bela Gipp
While Vision-language models (VLMs) interpret text-rich images effectively, they struggle with reasoning across long, multi-page documents. We present Active Long DocumEnt Navigation (ALDEN), a multi-turn reinforcement learning framework that fine-tunes VLMs as interactive agents capable of actively navigating long, visually rich documents rather than passive readers. ALDEN features a novel fetch action that allows direct page indexing, complementing the classic search action and better exploiting document structure. To ensure training efficiency and stability, we introduce a rule-based cross-level reward for dense supervision and a visual-semantic anchoring mechanism utilizing dual-path KL-divergence constraints. We train ALDEN on a curated corpus built from open-source datasets where trivial samples are filtered, and queries are rewritten to incentivize multi-turn navigation and fetch usage. Empirically, ALDEN achieves state-of-the-art results on five long-document benchmarks, offering a more accurate and efficient path for long-document understanding.

[Long] ALIGN: Word Association Learning for Cultural Alignment in Large Language Models ▶ *Underline Video*
Chunhua Liu, Kabir Manandhar Shrestha and Sukai Huang
Large language models (LLMs) exhibit cultural bias from over-represented viewpoints in training data, yet cultural alignment remains a challenge due to limited cultural knowledge and a lack of exploration into effective learning approaches. We introduce a cost-efficient and cognitively grounded method: fine-tuning LLMs on native speakers' word-association norms, leveraging cognitive psychology findings that such associations capture cultural knowledge. Using word association datasets from native speakers in the US (English) and China (Mandarin), we train Llama-3.1-8B and Qwen-2.5-7B via supervised fine-tuning and preference optimization. We evaluate models' cultural alignment through a two-tier evaluation framework that spans low-level lexical associations and high-level cultural value alignment using the World Values Survey. Results show significant improvements in lexical alignment (16-20% English, 43-165% Mandarin on Precision@5) and high-level cultural value shifts. On a subset of 50 questions where US and Chinese respondents diverge most, fine-tuned Qwen nearly doubles its response alignment with Chinese values (13 to 25). Remarkably, our trained 7-8B models match or exceed vanilla 70B baselines, demonstrating that a few million of culture-grounded associations achieve value alignment without expensive retraining. Our work highlights both the promise and the need for future research grounded in human cognition in improving cultural alignment in AI models.

[Long] Anchoring the Affective Manifold: Learning Canonical and Disentangled Representations via Generative Cross-Modal Alignment ▶ *Underline Video*
Weibin Li, Jintao Cheng, Xiaoyu Tang and Chi Man VONG
Dominant multimodal emotion recognition paradigms often neglect the intrinsic geometric structure of affect, resulting in representations heavily entangled with non-affective factors. To address this, we propose a Canonical Disentangled Multimodal Generative Framework aimed at recovering the canonical affective manifold from raw data. We explicitly decompose the latent space into a canonical Shared Affective Subspace (z_vad) and a Private Modality Subspace (z_priv). We facilitate this factorization through Supervised Manifold Anchoring and Cross-Modal Manifold Alignment. Experiments demonstrate that our model effectively disentangles affect from private attributes (e.g., identity), achieving superior robustness in zero-shot cross-domain transfer compared to fully supervised baselines, while enabling controllable emotion generation.

[Findings] ArkRepoBench: A Repository-Level Code Completion Benchmark for HarmonyOS Development ▶ *Underline Video*
Yanlin Wang, Bowen Zhang, Yanli Wang, Daya Guo, Terry Yue Zhuo, Jiachi Chen, Mingwei Liu, Xingong Zhang and Zibin Zheng
ArkTS is the primary programming language for Huawei's HarmonyOS ecosystem, which has expanded across smartphones, tablets, and IoT devices. While large language models have demonstrated strong code generation capabilities for mainstream languages, their performance on ArkTS remains largely unexplored. To address this gap, we introduce ArkRepoBench, the first repository-level code completion benchmark for ArkTS to our knowledge, 7,519 samples from 20 official HarmonyOS repositories. The benchmark covers multiple difficulty levels and further categorizes completion instances into Single-File, Cross-File Independent, and Cross-File Dependent settings based on dependency analysis, distinguishing the mere presence of cross-file context from its actual necessity. Our experiments show that: (1) ArkTS completion consistently underperforms mainstream languages under our experimental settings, suggesting language-specific challenges associated with this emerging language; (2) open-source 7B models achieve performance comparable to close-source models; (3) cross-file context is a double-edged sword, with sparse retrieval(Jaccard) outperforming dense methods on ArkTS; and (4) HarmonyOS API documentation consistently improves performance, suggesting the benefits of domain-specific enhancements in low-resource settings.

[Demo] AutoChecklist: Composable Pipelines for Checklist Generation and Scoring with LLM-as-a-Judge ▶ *Underline Video*
Karen Zhou and Chenhao Tan
Checklists have emerged as a popular approach for interpretable and fine-grained evaluation, particularly with LLM-as-a-Judge. Beyond evaluation, these structured criteria can serve as signals for model alignment, reinforcement learning, and self-correction. To support these endeavors, we present AutoChecklist, an open-source library that unifies checklist-based evaluation into composable pipelines. At its core is a taxonomy of five checklist generation abstractions, each encoding a distinct strategy for deriving evaluation criteria. A modular Generator -> Refiner -> Scorer pipeline connects any generator with a unified scorer, and new configurations can be registered via prompt templates alone. The library ships with eight built-in pipelines implementing published approaches and supports multiple LLM providers (OpenAI, OpenRouter, vLLM). Beyond the Python API, the library includes a CLI for off-the-shelf evaluation and a web interface for interactive exploration. Validation experiments confirm that these checklist methods significantly align with human preferences and quality ratings, and a case study on ICLR peer review rebuttals demonstrates flexible domain adaptation. AutoChecklist is publicly available at

https://github.com/ChicagoHAI/AutoChecklist.

[Long] AutoTaskEval: Towards Domain-Specific and Fine-Grained Evaluation for LLMs ▶ *Underline Video*

Qingqing Lyu, Linjuan Wu, Yongliang Shen, Hengwei Liu, Hao Li, Shengpei Jiang, Yin Zhang and Weiming Lu

Despite the rapid progress of LLMs, their evaluation remains hindered by static, manually curated benchmarks with limited task coverage and poor adaptability to emerging domains. Existing automated approaches typically operate within fixed task schemas and often fail to autonomously discover new evaluation dimensions, limiting both scalability and effectiveness. To address these gaps, we propose AutoTaskEval, an automated framework that constructs domain-specific benchmarks directly from unstructured corpora. Using a refined Blooms Taxonomy, the framework systematically discovers tasks, enriches contextual grounding via iterative Socratic prompting, and generates diverse, progressively challenging evaluation instances. Applied to the complex and knowledge-intensive legal domain, AutoTaskEval uncovers a broader and more fine-grained task space than expert-curated benchmarks while producing high-quality instances that preserve established model-level evaluation trends. We further validate its robustness in a low-structure e-commerce review domain. Together, these results show that AutoTaskEval enables scalable, adaptive, and high-fidelity LLM assessment across domains and model families, advancing autonomous and capability-sensitive evaluation.

[Long] Backdoor Collapse: Eliminating Unknown Threats Via Known Backdoor Aggregation In Language Models ▶ *Underline Video*

Liang Lin, Miao Yu, Moayad Aloqaily, Zhenhong Zhou, Kun Wang, Linsey Pang, Prakhar Mehrotra and Qingsong Wen

Backdoor attacks are a significant threat to large language models (LLMs), often embedded via public checkpoints, yet existing defenses rely on impractical assumptions about trigger settings. To address this challenge, we propose Locphylax, a defense framework that requires no prior knowledge of trigger settings. Locphylax is based on the key observation that when deliberately injecting known backdoors into an already-compromised model, both existing unknown and newly injected backdoors aggregate in the representation space. Locphylax leverages this through a two-stage process: first, aggregating backdoor representations by injecting known triggers, and then, performing recovery fine-tuning to restore benign outputs. Extensive experiments across multiple LLM architectures demonstrate that: (I) Locphylax reduces the average Attack Success Rate to 4.41% across multiple benchmarks, outperforming existing baselines by 28.1%-69.3%. (II) Clean accuracy and utility are preserved within 0.5% of the original model, ensuring negligible impact on legitimate tasks. (III) The defense generalizes across different types of backdoors, confirming its robustness in practical deployment scenarios. Our code is available at https://anonymous.4open.science/r/Locphylax.

[Long] Benchmarking Web Agent Safety under E-commerce Deceptive Interfaces ▶ *Underline Video*

Zijing Shi, Meng Fang and Ling Chen

As autonomous web agents are increasingly deployed to perform real-world tasks, ensuring their safety has become a critical concern. In this work, we study web agent behavior under realistic deceptive interfaces in the e-commerce domain. We introduce WebDecept, a lightweight and configurable plugin framework that enables controlled injection of deceptive interface patterns into existing web environments. Using WebDecept, we instantiate seven deceptive patterns commonly observed on the open web, including targeted advertisements, domain redirection, and shopping manipulation. By injecting these patterns into the frontend during task execution, we perform controlled evaluation of multiple multimodal web agents. Our results show that current web agents are highly susceptible to multiple classes of deceptive interfaces, and that prompt-based constraints are often insufficient to mitigate these failures. We further analyze how the design choices of deceptive patterns influence the success of such manipulations. These findings highlight safety challenges that should be addressed as web agents are scaled toward real-world deployment.

[Findings] Beyond Memorization: Testing LLM Reasoning on Unseen Theory of Computation Tasks ▶ *Underline Video*

Shlok Shelat, Jay Raval, Souvik Roy and Manas Gaur

Large language models (LLMs) have demonstrated strong performance on formal language tasks, yet whether this reflects genuine symbolic reasoning or pattern matching on familiar constructions remains unclear. We introduce a benchmark for deterministic finite automata (DFA) construction from regular languages, comprising factual knowledge questions, seen construction problems from public sources, and two types of unseen problems: hand-crafted instances with multiple interacting constraints and systematically generated problems via Arden's theorem. Models achieve perfect accuracy on factual questions and 84-90% on seen tasks. However, accuracy drops sharply on unseen problems (by 30-64%), with failures stemming from systematic misinterpretation of language constraints, incorrect handling of Kleene-star semantics, and a failure to preserve global consistency. We evaluate a three-stage hint protocol that enables correction of shallow errors but does not reliably resolve globally inconsistent or structurally flawed automata. Our analysis across multiple prompting strategies (direct, Chain-of-Thought, Tree-of-Thought) reveals that errors persist regardless of prompting approach, exposing a fundamental gap between LLMs' ability to generate syntactically plausible DFAs and their capacity for semantically correct formal reasoning.

[Long] Beyond Query Memorization: Large Language Model Routing with Query Decomposition and Historical Matching ▶ *Underline Video*

Bo Lv, Jingbo Sun, Jianwei Lv, Chen Tang, shaojie zhang, Nayu Liu, Guoxin Yu, Zihao Li, Qichao Zhang, Dongbin Zhao, Ping Luo and Yue Yu

Optimizing the trade-off among predictive performance and computational cost is a central focus in the deployment of Large Language Models (LLMs). Current routing methods primarily rely on direct mapping from queries to models based on surface-level features, making

them susceptible to the memorization trap and leading to poor generalizability on out-of-distribution (OOD) data. In this paper, we propose DecoR, a novel routing framework that recasts the routing task as a matching process of sifting similar queries from historical logs, effectively mitigating the memorization trap. To enhance matching accuracy, we introduce a query capability deconstruction method that decouples linguistic surface forms from task-intrinsic requirements, directing matching toward capability dimensions to ground decisions in essential task attributes. Furthermore, we develop CodaSet, a comprehensive benchmark for assessing routing generalization, where experimental results demonstrate that DecoR maintains superior accuracy while substantially lowering inference costs across both in-distribution and OOD settings.

[Findings] Beyond Reproduction: A Paired-Task Framework for Assessing LLM Comprehension and Creativity in Literary Translation ► *Underline Video*

Ran Zhang, Steffen Eger, Arda Tezcan, Wei Zhao, Simone Paolo Ponzetto and Lieve Macken

Large language models (LLMs) are increasingly used for creative tasks such as literary translation. Yet translational creativity remains underexplored and is rarely evaluated at scale, while source-text comprehension is typically studied in isolation, despite the fact that, in professional translation, comprehension and creativity are tightly intertwined. We address these gaps with a paired-task framework applied to literary excerpts from 11 books. Task 1 assesses source-text comprehension, and Task 2 evaluates translational creativity through Units of Creative Potential (UCPs), such as metaphors and wordplay. Using a scalable evaluation setup that combines expert human annotations with UCP-based automatic scoring, we benchmark 23 models and four creativity-oriented prompts. Our findings show that strong comprehension does not translate into human-level creativity: models often produce literal or contextually inappropriate renderings, with particularly large gaps for the more distant English-Chinese language pair. Creativity-oriented prompts yield only modest gains, and only one model, Mistral-Large, comes close to human-level creativity (0.167 vs. 0.246). Across all model-prompt combinations, only three exceed a creativity score of 0.1, while the rest remain at or near zero.

[Long] BhashaSutra: A Task-Centric Unified Survey of Indian NLP Datasets, Corpora, and Resources ► *Underline Video*

Raghvendra Kuma, Devankar Raj and Sriparna Saha

Indias linguistic landscape, spanning 22 scheduled languages and hundreds of marginalized dialects, has driven rapid growth in NLP datasets, benchmarks, and pretrained models. However, no dedicated survey consolidates resources developed specifically for Indian languages. Existing reviews either focus on a few high-resource languages or subsume Indian languages within broader multilingual settings, limiting coverage of low-resource and culturally diverse varieties. To address this gap, we present the first unified survey of Indian NLP resources, covering 200+ datasets, 50+ benchmarks, and 100+ models, tools, and systems across text, speech, multimodal, and culturally grounded tasks. We organize resources by linguistic phenomena, domains, and modalities; analyze trends in annotation, evaluation, and model design; and identify persistent challenges such as data sparsity, uneven language coverage, script diversity, and limited cultural and domain generalization. This survey offers a consolidated foundation for equitable, culturally grounded, and scalable NLP research in the Indian linguistic ecosystem.

[Findings] BOOKAGENT: Orchestrating Safety-Aware Visual Narratives via Multi-Agent Cognitive Calibration ► *Underline Video*

Bo Gao, Chang Liu, Yuyang Miao, SiYuan Ma and Ser-Nam Lim

Recent advancements in Large Generative Models (LGMs) have revolutionized multi-modal generation. However, generating illustrated storybooks remains an open challenge, where prior works mainly decompose this task into separate stages, and thus, holistic multi-modal grounding remains limited. Besides, while safety alignment is studied for text- or image-only generation, existing works rarely integrate child-specific safety constraints into narrative planning and sequence-level multi-modal verification. To address these limitations, we propose BookAgent, a safety-aware multi-agent collaboration framework designed for high-quality, safety-aware visual narratives. Different from prior story visualization models that assume a fixed storyline sequence, BookAgent targets end-to-end storybook synthesis from a user draft by jointly planning, scripting, illustrating, and globally repairing inconsistencies. To ensure precise multi-modal grounding, BookAgent dynamically calibrates page-level alignment between textual scripts and visual layouts. Furthermore, BookAgent calibrates holistic consistency from the temporal dimension, by verifying-then-rectifying global inconsistencies in character identity and storytelling logic. Extensive experiments demonstrate that BookAgent significantly outperforms current methods in narrative coherence, visual consistency, and safety compliance, offering a robust paradigm for reliable agents in complex multi-modal creation. The implementation will be publicly released at https://github.com/bogao-code/BookAgent/tree/main.

[Long] BracketRank: Large Language Model Document Ranking via Reasoning-based Competitive Elimination ► *Underline Video*

Abdelrahman Abdallah, Mohammed Ali, Bhawna Piryani and Adam Jatowt

Although Large Language Models (LLMs) show strong potential for zero-shot document ranking, current listwise approaches face three critical limitations. First, context length constraints prevent processing many documents simultaneously. Second, sequential generation creates bottlenecks that cannot run in parallel. Third, ranking results depend heavily on initial document order, leading to inconsistent performance. We introduce BracketRank, a reasoning-driven competitive elimination framework that addresses these challenges through systematic group competition. Our method uses adaptive grouping to automatically optimise group sizes based on LLM context limits, reasoning-enhanced prompts that require explicit relevance explanations, and bracket-style elimination where documents compete through winner and loser brackets. This structure ensures every document has fair advancement opportunities regardless of initial positioning while allowing for parallel processing across competition stages. We evaluate BracketRank on TREC DL 19, TREC DL 20, and eight

BEIR benchmark datasets. Results show that BracketRank achieves 77.90 NDCG@5 on TREC DL 19 and 75.85 NDCG@5 on TREC DL 20, outperforming RankGPT and other state-of-the-art methods. On BEIR datasets, BracketRank reaches 54.66 average NDCG@10, demonstrating robust performance across diverse domains.

[Long] Breaking Block Boundaries: Anchor-based History-stable Decoding for Diffusion Large Language Models ► *Underline Video*
Shun Zou, Yong Wang, Zehui Chen, Lin Chen, Chongyang Tao, Feng Zhao and Xiangxiang Chu
Diffusion Large Language Models (dLLMs) have recently become a promising alternative to autoregressive large language models (ARMs). Semi-autoregressive (Semi-AR) decoding is widely employed in base dLLMs and advanced decoding strategies due to its superior performance. However, our observations reveal that Semi-AR decoding suffers from inherent block constraints, which cause the decoding of many cross-block stable tokens to be unnecessarily delayed. To address this challenge, we systematically investigate the identification of stable tokens and present three key findings: (1) naive lookahead decoding is unreliable, (2) token stability closely correlates with convergence trend, and (3) historical information is isolated. Building on these insights, we propose Anchor-based History-stable Decoding (AHD), a training-free, plug-and-play dynamic decoding strategy. Specifically, AHD monitors the stability trend of tokens in real time through dynamic anchors. Once a token reaches stability, it initiates early cross-block decoding to enhance efficiency and performance. Extensive experiments across language, vision-language, and audio-language domains demonstrate that AHD simultaneously improves both performance and inference efficiency. Notably, AHD effectively reverses the performance degradation typically observed in existing advanced decoding acceleration strategies. For instance, on the BBH benchmark, our approach reduces decoding steps by 80% while improving performance by 3.67%.

[Long] Breaking the Generator Barrier: Disentangled Representation for Generalizable AI-Text Detection ► *Underline Video*
Xiao Pu, Zepeng Cheng, Lin Yuan, Yu Wu and Xiuli Bi
As large language models (LLMs) generate text that increasingly resembles human writing, the subtle cues that distinguish AI-generated content from human-written content become increasingly challenging to capture. Reliance on generator-specific artifacts is inherently unstable, since new models emerge rapidly and reduce the robustness of such shortcuts. This generalizes unseen generators as a central and challenging problem for AI-text detection. To tackle this challenge, we propose a progressively structured framework that disentangles AI-detection semantics from generator-aware artifacts. This is achieved through a compact latent encoding that encourages semantic minimality, followed by perturbation-based regularization to reduce residual entanglement, and finally a discriminative adaptation stage that aligns representations with task objectives. Experiments on MAGE benchmark, covering 20 representative LLMs across 7 categories, demonstrate consistent improvements over state-of-the-art methods, achieving up to 24.2% accuracy gain and 26.2% F_1 improvement. Notably, performance continues to improve as the diversity of training generators increases, confirming strong scalability and generalization in open-set scenarios. Our source code will be publicly available at https://github.com/PuXiao06/DRGD.

[Findings] BV-Blend: Uncertainty-Weighted Historical Baselines for Stable Critic-Free RL with Verifiable Rewards ► *Underline Video*
Yupeng Chang, Yuan Wu and Yi Chang
Critic-free reinforcement learning with verifiable rewards (RLVR), exemplified by Group Relative Policy Optimization (GRPO), avoids training a value function (critic) and reduces memory and compute overhead relative to critic-based PPO pipelines for aligning large language models. However, GRPO-style advantage estimation depends on prompt-local (within-prompt-group) reward statistics and can be unstable. In particular, when all rollouts in a prompt group receive identical rewards, the within-group reward variance becomes zero, and group normalization yields zero advantages for that group, impeding learning in cold-start regimes with binary verifiers. We introduce BV-Blend, a critic-free framework that stabilizes advantage estimation by combining prompt-local on-policy statistics with semantic-cluster-conditioned historical moments. BV-Blend maintains EMA-tracked reward moments for each cluster, derives a confidence weight from a standard error of the mean (SEM) proxy, and uses this weight to blend historical and prompt-local baseline and variance statistics into a standardized advantage for PPO-style clipped updates. Experiments on verifiable reasoning benchmarks show that BV-Blend improves training stability and performance, and remains robust in regimes where group-normalized methods may stall.

[Long] Can AI Be a Good Peer Reviewer? A Survey of Peer Review Process, Evaluation, and the Future ► *Underline Video*
Sihong Wu, Owen Jiang, Yilun Zhao, Tiansheng Hu, Yiling Ma, Kaiyan Zhang, Manasi Patwardhan and Arman Cohan
Peer review is a multi-stage process involving reviews, rebuttals, meta-reviews, final decisions, and subsequent manuscript revisions. Recent advances in large language models (LLMs) have motivated methods that assist or automate different stages of this pipeline. In this survey, we synthesize techniques for (i) peer review generation, including fine-tuning strategies, agent-based systems, RL-based methods, and emerging paradigms to enhance generation; (ii) after-review tasks including rebuttals, meta-review and revision aligned to reviews; and (iii) evaluation methods spanning human-centered, reference-based, LLM-based and aspect-oriented. We catalog datasets, compare modeling choices, and discuss limitations, ethical concerns, and future directions. The survey aims to provide practical guidance for building, evaluating, and integrating LLM systems across the full peer review workflow.

[Findings] CaPEdit: Capability-Preserving Lifelong Knowledge Editing For Language Models ► *Underline Video*
Song-Li Wu, Zhaocheng Du, Xianquan Wang and Jingyi Wang
Lifelong knowledge editing (LKE) aims to incrementally correct factual inaccuracies in large language models (LLMs), but sequential edits can lead to substantial degradation of capabilities. Existing approaches primarily rely on static parameter regularization, which restricts

knowledge integration and fails to prevent cumulative capability degradation. We argue that an important source of this degradation lies in the temporal mismatch between locally editable factual knowledge and procedural knowledge, which is gradually acquired, guides task execution, and cannot be reliably updated by rapid edits. To this end, we formulate LKE as a dual-timescale process, explicitly decoupling fast-updating factual knowledge from slow-evolving procedural knowledge. Based on this formulation, we propose CaPEdit, a framework that preserves model capabilities under LKE. It first synthesizes procedural knowledge across successive edits, and subsequently performs parameter updates guided jointly by factual supervision and the synthesized procedural signal. To ensure stability under long edit sequences, CaPEdit is trained via a hybrid optimization scheme, combining step-wise updates for rapid factual correction with trajectory-level optimization to facilitate gradual procedural adaptation. Experiments demonstrate that CaPEdit improves capability preservation across all fundamental capabilities by 49.78%, achieves superior editing performance, and requires only 18.07% of the editing time of most existing methods.

[SRW] CBAL: Context-Based Agentic Learning for Speaker Diarization Segmentation Refinement ▶ *Underline Video*

Odwitiyo Dutta and Dinesh Kumar Vishwakarma;

Speaker diarization systems produce segmentation errors, such as false splits and boundary misplacements, that degrade transcript readability and downstream applications. We present CBAL (Context-Based Agentic Learning), a post-processing framework that refines segmentation boundaries in diarized scripts through targeted error correction. CBAL detects potential segmentation errors using acoustic and temporal heuristics and employs a lightweight LLM agent to reason about merge decisions, validating corrections through uncertainty-aware filtering with signal-based constraints. CBAL achieves 93.4% accuracy across 359 applied merges and reduces segment count by 6.1%. We demonstrate that our framework identifies and corrects high-confidence errors while maintaining 0% degradation in terms of concatenated minimum-permutation Word Error Rate (cpWER). An ablation study confirms that each component contributes non-redundantly, demonstrating the viability of interpretable refinement frameworks that use the strengths of acoustic models and language understanding without requiring end-to-end retraining.

[Long] CharTide: Data-Centric Chart-to-Code Generation via Tri-Perspective Tuning and Inquiry-Driven Evolution ▶ *Underline Video*

Xiangxi Zheng, Kuang He, Jiayi Hu, Ping Yu, Rui Yan, Yuan Yao, Peng Hou, Anxiang Zeng and Alex Jinpeng Wang

Chart-to-code generation demands strict visual precision and syntactic correctness from Vision-Language Models (VLMs). However, existing approaches are fundamentally constrained by data-centric limitations: despite the availability of growing chart-to-code datasets, simply scaling homogeneous chart-code pairs conflates visual perception with program logic, preventing models from fully leveraging the richness of multimodal supervision. We present CharTide, a novel data-centric framework that systematically redesigns both training and alignment data for chart-to-code generation. First, we construct a 2M-sample dataset via a Tri-Perspective Tuning strategy, explicitly decoupling training into visual perception, pure-text code logic, and modality fusion streams, enabling a 7B model to surpass specialized baselines using only supervised data. Second, we reformulate alignment as a data verification problem rather than a heuristic scoring task. To this end, we introduce an Inquiry-Driven RL framework grounded in the principle of information invariance: a downstream model should yield consistent answers to identical visual queries across both original and generated charts. Moving beyond rigid rule matching or VLM scoring, we employ a frozen Inspector to objectively verify generated charts through atomic QA tasks, providing verifiable reward signals based on answer accuracy. Experiments on ChartMimic, Plot2Code, and ChartX show that CharTide-7B/8B significantly outperforms open-source baselines, surpasses GPT-4o, and is competitive with GPT-5.

[Findings] ClimateCause: Complex and Implicit Causal Structures in Climate Reports ▶ *Underline Video*

Liesbeth Allein, Nataly Pineda-Castañeda, Andrea Rocci and Marie-Francine Moens

Understanding climate change requires reasoning over complex causal networks. Yet, existing causal discovery datasets predominantly capture explicit, direct causal relations. We introduce ClimateCause, a manually expert-annotated dataset of higher-order causal structures from science-for-policy climate reports, including implicit and nested causality. Cause-effect expressions are normalized and disentangled into individual causal relations to facilitate graph construction, with unique annotations for cause-effect correlation, relation type, and spatiotemporal context. We further demonstrate ClimateCause's value for quantifying readability based on the semantic complexity of causal graphs underlying a statement. Finally, large language model benchmarking on correlation inference and causal chain reasoning highlights the latter as a key challenge.

[Long] CoG: Controllable Graph Reasoning via Relational Blueprints and Failure-Aware Refinement over Knowledge Graphs ▶ *Underline Video*

Yuanxiang Liu, Songze Li, Xiaoke Guo, Zhaoyan Gong, Qifei Zhang, Huajun Chen and Wen Zhang

Large Language Models (LLMs) have demonstrated remarkable reasoning capabilities but often grapple with reliability challenges like hallucinations. While Knowledge Graphs (KGs) offer explicit grounding, existing paradigms of KG-augmented LLMs typically exhibit cognitive rigidity, applying homogeneous search strategies that render them vulnerable to instability under neighborhood noise and structural misalignment leading to reasoning stagnation. To address these challenges, we propose CoG, a training-free framework inspired by Dual-Process Theory that mimics the interplay between intuition and deliberation. First, functioning as the fast, intuitive process, the Relational Blueprint Guidance module leverages relational blueprints as interpretable soft structural constraints to rapidly stabilize the search direction against noise. Second, functioning as the prudent, analytical process, the Failure-Aware Refinement module intervenes upon

encountering reasoning impasses. It triggers evidence-conditioned reflection and executes controlled backtracking to overcome reasoning stagnation. Experimental results on three benchmarks demonstrate that CoG significantly outperforms state-of-the-art approaches in both accuracy and efficiency.

[Findings] ComicVQA: A Benchmark for Visual Reasoning in Multimodal LLMs ▶ *Underline Video*

Esther Gan, Hannah Brown, David Herel, Kenji Kawaguchi, Min-Yen Kan and Michael Qizhe Shieh

We introduce Comic Visual Question Answering (ComicVQA), a comics-based benchmark for evaluating MLLMs on visual reasoning. ComicVQA comprises of (i) Missing Panel Prediction, testing fine-grained visual grounding and (ii) Panel Sorting, which evaluates sequential narrative understanding. Proprietary models achieve up to 62.6% on Missing Panel Prediction and 46.4% on Panel Sorting, whereas open-source models reach only 47.7% and 26.9%, respectively. In contrast, human annotators achieve over 83% accuracy on both tasks, revealing a large gap between current models and human-level multimodal understanding in comics. Through controlled ordering ablations and a detailed error taxonomy, we show that current MLLMs rely primarily on coarse temporal cues and struggle with fine-grained visual reasoning. These findings demonstrate ComicVQA as a diagnostic benchmark for advancing multimodal visual reasoning in comics.

[Long] Comparative Analysis of the Intrinsic Metrics for Tokenizers and their effect on Downstream Tasks for Hindi and Marathi ▶ *Underline Video*

Shagun Dwivedi and Kaushik Gopalan

Various studies have pointed out that the performance of language models is poor in non-English or non-European languages. One of the factors affecting this performance is the effectiveness and suitability of the tokenization scheme used in the model. Indic scripts require multiple Unicode codepoints to represent a single visual unit to be encoded in the standard UTF-8 scheme. This paper investigates the effect of multiple tokenizers that use UTF-8 text input on the downstream performance of pretrained language models for Hindi and Marathi, languages written in Devangari script. We present the intrinsic performance of the tokenizers using Fertility, Rényi Efficiency and Percentile Frequency, and report the extrinsic performance of monolingual and multilingual models on question-answering tasks, using an automated parts-of-speech and sentence similarity based evaluation framework, and on word-level tasks such as grapheme-to-phoneme conversion and transliteration. We propose a grapheme cluster tokenizer for the script which shows performance better than or competitive with other popular tokenizers. We also find that the Rényi Efficiency metric is highly correlated to downstream performance on question answering.

[Long] Compatibility-Aware Dynamic Fine-Tuning for Large Language Models ▶ *Underline Video*

Yucheng Zhou, Junwei Sheng, Qianning Wang and Jianbing Shen

Supervised Fine-Tuning (SFT) is the predominant paradigm for aligning large language models (LLMs), yet it suffers from optimization instability and limited generalization. Recent work attributes this issue to pathological gradient scaling and proposes Dynamic Fine-Tuning (DFT) to correct it at the token level. However, DFT assumes all demonstrations are equally suitable learning targets, an assumption violated by the strong heterogeneity of large-scale instruction data, where demonstration-policy mismatch induces high-variance updates at the sample level. We introduce Compatibility-Aware Dynamic Fine-Tuning (CADFT), a principled extension of DFT that controls sample-level optimization variance. CADFT derives a dynamic, policy-dependent compatibility signal from model likelihoods to modulate supervised updates, suppressing high-variance gradients from incompatible demonstrations. We further propose a delayed, low-frequency compatibility-guided rewriting strategy to transform persistently incompatible demonstrations into learnable targets. We show that CADFT can be interpreted as a variance-controlled estimator that generalizes token-level stabilization in DFT to the sample level. Extensive experiments demonstrate improved stability, generalization, and cold-start reinforcement learning initialization, while remaining fully supervised and free of reward modeling.

[Long] Compressing LLM Knowledge into Graph Representations for Text-attributed Graphs Learning ▶ *Underline Video*

Runhuai Chen, Dian Shen, Dandan Zhang, Kaihong Huang, Linghui Meng and Beilun Wang

Text-attributed graphs (TAGs) require jointly modeling relational structure and node-level text. Existing GNN-LLM approaches perform by incorporating large language models at inference time for processing the text attributes, resulting in costly deployment. More fundamentally, LLM knowledge is typically used in a sample-wise manner, leading to inefficient utilization across graph instances. In this work, we study how interactions with LLM embedding spaces affect graph representations, and show that projecting into the LLM space can learn better GNNs. That is to say, the knowledge encoded in LLM embeddings can be compressed into graph representations. Based on this insight, we propose a framework that internalizes LLM knowledge within graph models and supports inference-efficient TAG learning. Our framework employs a hierarchical Proxy-Purifier module with distribution-level regularization, using LLM embeddings only as training-time guidance. With this module, the model operates TAGs without invoking LLMs, achieving high efficiency as standard GNNs without LLMs. Notably, experiments on five popular TAG tasks further demonstrate that our method can also achieve consistent performance gains, in comparison to existing GNN-LLM approaches.

[Long] CompTab: A Comprehensive Benchmark for Real-World TableQA with Complex Reasoning and Irregular Tables ▶ *Underline Video*

Zhen Yang, Wei Du, Jie Wang, Wenze Zhou, Xiangfeng Meng, Zhengyang Wang, Suping Sun, Ziwei Du, Haodong Zou, Jie Chen, Yongbin Liu, Shicheng Tan, Jiahao Ying and Shu Zhao

Recent progress in Large Language Model (LLM) based Table Question Answering (TableQA) has demonstrated strong performance on

standard benchmarks. However, existing benchmarks mainly focus on well-structured tables and fail to reflect the irregular structures and complex reasoning commonly encountered in real-world scenarios. We propose CompTab, a benchmark designed to evaluate TableQA under complex reasoning and irregular table conditions. CompTab covers six representative types, including semantic ambiguity, multi-hop reasoning, transposed tables, merged cells, missing values, and outliers. It is constructed from real-world seed tables across multiple domains using controlled LLM based generation and human verification to ensure realism and diversity. In addition, to improve the generalization of LLMs under complex and irregular table settings, we propose a two-stage training framework that progressively aligns models with textual reasoning and executable decision signals, instantiated as CompTabLLM. Evaluations on 38 representative LLMs and CompTabLLM show clear limitations of existing LLMs under realistic conditions, while the proposed framework improves generalization. CompTab thus provides a challenging benchmark for advancing TableQA in real-world.

[Long] Consolidation or Adaptation? PRISM: Disentangling SFT and RL Data via Gradient Concentration ▶ *Underline Video*

Yang Zhao, Yangou Ouyang, Xiao Ding, Hepeng Wang, Bibo Cai, Kai Xiong, Jinglong Gao, Zhouhao Sun, Li Du, Bing Qin and Ting Liu

While Hybrid Supervised Fine-Tuning (SFT) followed by Reinforcement Learning (RL) has become the standard paradigm for training LLM agents, effective mechanisms for data allocation between these stages remain largely underexplored. Current data arbitration strategies often rely on surface-level heuristics that fail to diagnose intrinsic learning needs. Since SFT targets pattern consolidation through imitation while RL drives structural adaptation via exploration, misaligning data with these functional roles causes severe optimization interference. We propose PRISM, a dynamics-aware framework grounded in Schema Theory that arbitrates data based on its degree of cognitive conflict with the model's existing knowledge. By analyzing the spatial geometric structure of gradients, PRISM identifies data triggering high spatial concentration as high-conflict signals that require RL for structural restructuring. In contrast, data yielding diffuse updates is routed to SFT for efficient consolidation. Extensive experiments on WebShop and ALFWorld demonstrate that PRISM achieves a Pareto improvement, outperforming state-of-the-art hybrid methods while reducing computational costs by up to 3.22 x. Our findings suggest that disentangling data based on internal optimization regimes is crucial for scalable and robust agent alignment.

[Demo] Copyright Detective: A Forensic System to Evidence LLMs Flickering Copyright Leakage Risks ▶ *Underline Video*

Guangwei Zhang, Jianing Zhu, Cheng Qian, Neil Zhenqiang Gong, Rada Mihalcea, Zhaozhuo Xu, Jingrui He, Jiaqi W. Ma, Chaowei Xiao, Bo Li, Ahmed Abbasi, Dongwon Lee, Heng Ji and Denghui Zhang;

We present Copyright Detective, the first interactive forensic system for detecting, analyzing, and visualizing potential copyright risks in LLM outputs. The system treats copyright infringement versus compliance as an evidence discovery process rather than a static classification task due to the complex nature of copyright law. It integrates multiple detection paradigms, including content recall testing, paraphrase-level similarity analysis, persuasive jailbreak probing, and unlearning verification, within a unified and extensible framework. Through interactive prompting, response collection, and iterative workflows, our system enables systematic auditing of verbatim memorization and paraphrase-level leakage, supporting responsible deployment and transparent evaluation of LLM copyright risks even with black-box access. In our experiments with GPT-4o-mini, we demonstrate that specific persuasive strategy "Pathos" shift the leakage distribution from about 0.1 (ROUGE-L) to 0.7. Our live system is hosted on https://copyright-detective.streamlit.app, and the source code will be made publicly available at https://github.com/changhu73/Copyright-Detective upon acceptance. In addition, a https://youtu.be/10gmwV3q_gg is also included as supplementary material.

[Findings] CrisPrune: Combining Contextual Relevance and Intrinsic Saliency for Efficient Visual Token Pruning in MLLMs ▶ *Underline Video*

Ziniu Liu, Shuheng Zhou, Mingqing Liu, Hao Deng and Huijia Zhu

Visual token pruning has emerged as a pivotal strategy to alleviate the computational bottleneck in Multimodal Large Language Models (MLLMs), yet it frequently compromises the integrity of visual understanding in pursuit of efficiency. Existing methods face a fundamental tension: vision-centric approaches are susceptible to the attention sink phenomenon and operate in a query-agnostic manner, whereas text-guided methods often create an overly narrow focus, discarding essential background context and failing on ambiguous queries. In this paper, we propose CrisPrune, a training-free and model-agnostic method that reconciles efficiency with understanding by integrating visual saliency and text relevance. Specifically, we introduce intrinsic visual saliency with robust normalization to identify information-rich regions characterized by significant visual features. Simultaneously, we design dual-source text relevance to synergize explicit instruction alignment with implicit scene priors. Finally, we reformulate the selection process using a Determinantal Point Process (DPP) to balance token quality and spatial diversity. Extensive experiments demonstrate that CrisPrune significantly outperforms state-of-the-art methods. On LLaVA-NeXT, it achieves a 13 x decrease in FLOPs while maintaining 97% of the original performance with 94.4% of visual tokens pruned, effectively bridging the gap between efficiency and holistic understanding.

[TACL] Cross-layer Attention Sharing for Pre-trained Large Language Models ▶ *Underline Video*

Yongyu Mu, Yuzhang Wu, Yuchun Fan, Chenglong Wang, Hengyu Li, Jiali Zeng, Qiaozhi He, Murun Yang, Fandong Meng, Jie Zhou, Tong Xiao and Jingbo Zhu

To enhance the efficiency of the attention mechanism within large language models (LLMs), previous works primarily compress the KV cache or group attention heads, while largely overlooking redundancy between layers. Our comprehensive analyses across various LLMs show that highly similar attention patterns persist within most layers. It's intuitive to reduce the redundancy by sharing attention weights across layers. However, further analysis reveals two challenges: (1) Directly sharing the weight matrix without carefully rearranging the

attention heads proves to be ineffective; (2) Shallow layers are vulnerable to small deviations in attention weights. Driven by these insights, we introduce LISA, a lightweight substitute for self-attention in well-trained LLMs. LISA employs tiny feed-forward networks to align attention heads between adjacent layers and low-rank matrices to approximate differences in layer-wise attention weights. Evaluations encompassing 13 typical benchmarks demonstrate that LISA maintains high response quality in terms of accuracy and perplexity while reducing redundant attention calculations within 53%-84% of the total layers. Our implementations of LISA achieve a 6x compression of Q and K matrices within the attention mechanism, with maximum throughput improvements 19.5%, 32.3%, and 40.1% for LLaMA3-8B, LLaMA2-7B, and LLaMA2-13B, respectively.

[Long] Cross-Prompt Automated Essay Scoring of Multiple Traits: Making Sense of the State of the Art ► *Underline Video*

Shengjie Li and Vincent Ng

Despite the recent progress made in cross-prompt essay scoring, there is little analysis of what makes a state-of-the-art cross-prompt scorer work well. To this end, we present an empirical analysis of how the key components of a cross-prompt scorer interact with each other and impact its overall performance. In addition, we examine for the first time the application of transductive learning to cross-prompt scoring, which represents an important starting point for providing a practical way to improve cross-prompt scorers for use in the rarely-studied classroom setting without the need for additional labeled training data.

[Long] Cut Your Losses! Learning to Prune Paths Early for Efficient Parallel Reasoning ► *Underline Video*

Jiaxi Bi, Tongxu Luo, Wenyu Du, Zhengyang Tang and Benyou Wang

Parallel reasoning enhances Large Reasoning Models (LRMs) but incurs prohibitive costs due to futile paths caused by early errors. To mitigate this, path pruning at the prefix level is essential, yet existing research remains fragmented without a standardized framework. In this work, we propose the first systematic taxonomy of path pruning, categorizing methods by their signal source (internal vs. external) and learnability (learnable vs. non-learnable). This classification reveals the unexplored potential of learnable internal methods, motivating our proposal of STOP (Super TOken for Pruning). Extensive evaluations across LRMs ranging from 1.5B to 20B parameters demonstrate that STOP achieves superior effectiveness and efficiency compared to existing baselines. Furthermore, we rigorously validate the scalability of STOP under varying compute budgets, for instance, boosting GPT-OSS-20B accuracy on AIME25 from 84% to nearly 90% under fixed compute budgets. Finally, we distill our findings into formalized empirical guidelines to facilitate optimal real-world deployment. Code, data and models are available at https://bijiaxihh.github.io/STOP.

[Findings] DaMo: Data Mixing Optimizer in Fine-tuning Multimodal LLMs for Mobile Phone Agents ► *Underline Video*

Kai Shi, Jun Yang, Ni Yang, Binqiang Pan, Qingsong Xie, Chao Zhang, Zhenyu Yang, Tianhuang Su and Haonan Lu

Mobile Phone Agents (MPAs) have emerged as a promising research direction due to their broad applicability across diverse scenarios. While Multimodal Large Language Models (MLLMs) serve as the foundation for MPAs, their effectiveness in handling multiple mobile phone tasks simultaneously remains limited. Although multitask supervised fine-tuning (SFT) is widely adopted for multitask learning, existing approaches struggle to determine optimal training data compositions for peak performance. To address this challenge, we propose DaMo (Data Mixture Optimizer) - a novel solution employing a trainable network that predicts optimal data mixtures by forecasting downstream task performance for any given dataset ratio. To support comprehensive evaluation, we introduce PhoneAgentBench, the first specialized benchmark to evaluate MLLMs on multimodal mobile phone tasks, comprising 1,235 QA pairs spanning diverse real-world industrial mobile application scenarios. Demonstrating strong predictive capability (Rš=0.81) in small-scale pilot experiments, DaMo efficiently extrapolates optimal data mixing configurations. Our results show DaMo achieves 3.06% average score improvement on PhoneAgentBench and open-source benchmarks, including BFCL-v3, MME-Reasoning, MME-Perception, and OCRBench, compared to alternative methods. Through predicting optimal data mixture only on open-source benchmarks, DaMo outperforms other approaches by 6.70% in terms of average score. Moreover, DaMo improves the metrics by 12.74% than other methods when used solely for MLLM optimization on the BFCL-v3 task. Notably, DaMo maintains robust scalability, preserving its effectiveness when applied to other model architectures.

[Short] Dark & Stormy: Modeling Humor in Sentences from the Bulwer-Lytton Fiction Contest ► *Underline Video*

Venkata S Govindarajan and Laura Biester

Textual humor is enormously diverse and computational studies need to account for this range, including intentionally bad humor. In this paper, we curate and analyze a novel corpus of sentences from the Bulwer-Lytton Fiction Contest to better understand "bad" humor in English. Standard humor detection models perform poorly on our corpus, and an analysis of literary devices finds that these sentences combine features common in existing humor datasets (e.g., puns, irony) with metaphor, metafiction and simile. LLMs prompted to synthesize contest-style sentences imitate the form but exaggerate the effect by over-using certain literary devices, and including far more novel adjective-noun bigrams than human writers.

[Long] Dashboard2Code: Evaluating Multimodal Models on Reconstructing Interactive Dashboards ► *Underline Video*

Tianhao Niu, Ziyu Han, Qiguang Chen, Shiqi Zhou, Baocai Shan, Hengjie Fang, Qingfu Zhu and Wanxiang Che

Automatic data visualization generation have advanced rapidly with multi-modal large language models, yet existing efforts largely focus on static charts and overlook the interactive dashboards commonly used for real-world data exploration. We introduce Dashboard2Code, a novel task that requires a model to proactively explore an interactive dashboard, acquire and integrate feedback from its own interactions (e.g., clicking and filtering), and generate code that reproduces the target dashboard. To support comprehensive evaluation, we present

DashboardMimic, the first Plotly+Dash benchmark for Dashboard2Code, comprising 180 carefully designed and manually verified dashboard-code pairs spanning three difficulty levels and covering eight common real-world interaction patterns. We further propose an automated evaluation framework tailored to dashboards that combines code semantic analysis with dynamic interaction-based testing to assess visual and interaction consistency, showing strong agreement with human judgments. Experiments across a range of open- and closed-source multi-modal models reveal that even the strongest systems struggle on high-complexity dashboards and that a substantial performance gap remains between open-source and closed-source models on the Dashboard2Code task.

[Long] DE-CLIP: Few-Shot Anomaly Detection via Difference-Guided Embedding Editing ▶ *Underline Video*

Yage Zhang, Yukun Jiang, Michael Backes and Yang Zhang

Anomaly detection (AD) plays a critical role in applications such as automated industrial inspection and medical image analysis. Empowered by the strong pre-trained vision-language model, CLIP, recent years have witnessed the emergence of several CLIP-based few-shot AD methods. Due to the overlap between the embedding distributions of normal and anomalous samples, many existing approaches introduce additional model training for more discriminative text embeddings. However, we demonstrate that such training is not necessary. Specifically, we find that this embedding overlap can be separated by introducing a Difference-guided vector for embedding Editing (DiffEdit). Based on this finding, we propose DE-CLIP, a simple yet effective framework based on DiffEdit, which directly edits text embeddings based on the textual and visual differences between normal and anomalous samples, resulting in more discriminative embeddings for AD. Extensive experiments on industrial and medical datasets demonstrate the superiority of our proposed DE-CLIP compared with existing baselines. For instance, on MVTec dataset, DE-CLIP achieves 96.6% and 96.7% AUROC on anomaly classification and segmentation, surpassing both training-based and training-free methods. In addition, we observe that introducing DiffEdit into other training-free baselines could also significantly improve their performance, highlighting the potential of DiffEdit to promote better AD.

[SRW] Debiasing Logical Fallacy Detection for Real-World Robustness via Counterfactually Augmented Data ▶ *Underline Video*

Navyansh Singh

Logical fallacy detection models frequently over-flag valid reasoning due to reliance on surface-level spurious correlations, a brittleness that persists even in state-of-the-art benchmarks like CoCoLoFa despite strong in-distribution performance. This study introduces a targeted debiasing strategy using 703 LLM-generated counterfactually augmented pairs —minimally differentiated valid and fallacious arguments matched on topic, structure, and wording—to isolate logical distinctions beyond superficial cues. Fine-tuning DeBERTa-v3-large on CoCoLoFa augmented with these pairs yields only marginal in-distribution improvement (+0.4% F1 on the test set) but delivers substantial out-of-distribution robustness: a 58% relative reduction in false positive rate (64% 26.7%) on a 300-sample balanced manual evaluation set curated from Reddit and other real-world sources. Zero-shot evaluation of recent LLMs (Llama-3.1-8B and Llama-3.3-70B) shows high performance under carefully optimized prompts (F1 90–94%), but severe degradation under simple human-like prompts (F1 63–72%, FPR rising to 54–74%) and partial recovery with few-shot examples (F1 83–90%, FPR 21–34%). In contrast, the proposed lightweight and prompt-invariant approach achieves competitive robustness (F1 85.9%, FPR 26.7%) across all prompting regimes without any dependence on prompt engineering. This makes the method particularly stable for environments with unpredictable user input. The 703-pair counterfactual dataset, the fine-tuned model, and an interactive Hugging Face demo are publicly released to facilitate further research in misinformation detection and argumentation analysis.

[Long] DeepGuard: Secure Code Generation via Multi-Layer Semantic Aggregation ▶ *Underline Video*

Li Huang, Zhongxin Liu, Yifan Wu, Tao Yin, Dong Li, Jichao Bi, Nankun Mu, Hongyu Zhang and Meng Yan

Large Language Models (LLMs) for code generation can replicate insecure patterns from their training data. To mitigate this, a common strategy for security hardening is to fine-tune models using supervision derived from the final transformer layer. However, this design may suffer from a final-layer bottleneck: vulnerability-discriminative cues can be distributed across layers and become less detectable near the output representations optimized for next-token prediction. To diagnose this issue, we perform layer-wise linear probing. We observe that vulnerability-related signals are most detectable in a band of intermediate-to-upper layers yet attenuate toward the final layers. Motivated by this observation, we introduce DeepGuard, a framework that leverages distributed security-relevant cues by aggregating representations from multiple upper layers via an attention-based module. The aggregated signal powers a dedicated security analyzer within a multi-objective training objective that balances security enhancement and functional correctness, and further supports a lightweight inference-time steering strategy. Extensive experiments across five code LLMs demonstrate that DeepGuard improves the secure-and-correct generation rate by an average of 11.9% over strong baselines such as SVEN. It also preserves functional correctness while exhibiting generalization to held-out vulnerability types.

[Demo] Demonstrating ViviDoc: Generating Interactive Documents through Human-Agent Collaboration ▶ *Underline Video*

Yinghao Tang, Yupeng Xie, Yingchaojie Feng, Tingfeng Lan and Wei Chen;

Interactive articles help readers engage with complex ideas through exploration, yet creating them remains costly, requiring both domain expertise and web development skills. Recent LLM-based agents can automate content creation, but naively applying them yields uncontrollable and unverifiable outputs. We present Vividoc, a human-agent collaborative system that generates interactive educational documents from a single topic input. Vividoc introduces a multi-agent pipeline (Planner, Executor, Evaluator) and the Document Specification (DocSpec), a human-readable intermediate representation that decomposes each interactive visualization into State, Render, Transition, and Constraint components. The DocSpec enables educators to review and refine generation plans before code is produced,

bridging the gap between pedagogical intent and executable output. We collect a dataset of 101 real-world interactive documents across 11 domains and conduct a user study showing that ViviDoc produces documents comparable in quality to human-authored ones. Our demo is available at https://vividoc.vercel.app/ and a video demonstration at https://www.youtube.com/watch?v=rJrnPJLyHUI.

[Findings] DiffER: Diffusion Entity-Relation Modeling for Reversal Curse in Diffusion Large Language Models ▶ *Underline Video*

Shaokai He, Kaiwen Wei, Xinyi Zeng, Xiang Chen, Xue Yang, Zhenyang Li, Jiang Zhong and Yu Tian

The "reversal curse" refers to the phenomenon where large language models (LLMs) exhibit predominantly unidirectional behavior when processing logically bidirectional relationships. Prior work attributed this to autoregressive training, predicting the next token inherently favors left-to-right information flow over genuine bidirectional knowledge associations. However, we observe that Diffusion LLMs (DLLMs), despite being trained bidirectionally, also suffer from the reversal curse. To investigate the root causes, we conduct systematic experiments on DLLMs and identify three key reasons: 1) entity fragmentation during training, 2) data asymmetry, and 3) missing entity relations. Motivated by the analysis of these reasons, we propose Diffusion Entity-Relation Modeling (DiffER), which addresses the reversal curse through entity-aware training and balanced data construction. Specifically, DiffER introduces whole-entity masking, which mitigates entity fragmentation by predicting complete entities in a single step. DiffER further employs distribution-symmetric and relation-enhanced data construction strategies to alleviate data asymmetry and missing relations. Extensive experiments demonstrate that DiffER effectively alleviates the reversal curse in Diffusion LLMs, offering new perspectives for future research. The code is available at https://github.com/CQU-MM-Intelligent-Lab/DiffER.

[Long] Diffusion-CAM: Faithful Visual Explanations for dMLLMs ▶ *Underline Video*

Haomin Zuo, Yidi Li, Luoxiao Yang and Xiaofeng Zhang

While diffusion Multimodal Large Language Models (dMLLMs) have recently achieved remarkable strides in multimodal generation, the development of interpretability mechanisms has lagged behind their architectural evolution. Unlike traditional autoregressive models that produce sequential activations, diffusion-based architectures generate tokens via parallel denoising, resulting in smooth, distributed activation patterns across the entire sequence. Consequently, existing Class Activation Mapping (CAM) methods, which are tailored for local, sequential dependencies, are ill-suited for interpreting these non-autoregressive behaviors. To bridge this gap, we propose Diffusion-CAM, the first interpretability method specifically tailored for dMLLMs. We derive raw activation maps by differentiably probing intermediate representations in the transformer backbone, accordingly capturing both latent features and their class-specific gradients. To address the inherent stochasticity of these raw signals, we incorporate four key modules to resolve spatial ambiguity and mitigate intra-image confounders and redundant token correlations. Extensive experiments demonstrate that Diffusion-CAM significantly outperforms SoTA methods in both localization accuracy and visual fidelity, establishing a new standard for understanding the parallel generation process of diffusion multimodal systems.

[SRW] Disentangling Linguistic Relatedness from Task Alignment in Cross-Lingual Transfer ▶ *Underline Video*

Ahmed Haj Ahmed; Ruochen Zhang; Alvin Grissom II

We study cross-lingual transfer by fine-tuning seven large language models (4B–671B parameters) on Arabic and evaluating zero-shot reading comprehension on Semitic languages and non-Semitic controls. Across dense and Mixture-of-Experts architectures, we find no evidence of Semitic-specific transfer: models with weak baselines improve dramatically across all languages, while strong-baseline models show only marginal gains regardless of language family. A chain-of-thought ablation reinforces this finding—the same models that benefit most from fine-tuning benefit equally from inference-time reasoning, suggesting both mechanisms address task-format alignment rather than cross-lingual knowledge transfer.

[Long] Distilling Large Embeddings via Hyperspherical Householder Quantization ▶ *Underline Video*

Yihang Wang, Bin Wu, Yueyang Su, Tianfu Zhang, Yiqi Du, Lei Yu, Jiafeng Guo and Xueqi Cheng

Large embedding models have become the backbone of modern retrieval systems, offering strong semantic representations at the cost of substantial storage and computation. While recent work explores quantizing embeddings into discrete document identifiers for generative retrieval, most existing approaches rely on Euclidean quantization, which is poorly aligned with the angular geometry induced by contrastive embedding training and often requires long identifier sequences to preserve semantic fidelity. In this work, we propose Hyperspherical Householder Quantization (HHQ), a geometry-aware distillation method that compresses large embeddings into short discrete representations via iterative Householder transformations on the unit hypersphere. By explicitly preserving cosine similarity at each step, HHQ distills semantic structure into compact identifiers that remain faithful to the original embedding space. To support reliable generation of these identifiers, we introduce constrained supervised fine-tuning and tree-aware dynamic masking to enforce structural validity during training and inference. Experiments on NQ and MS~MARCO show that HHQ achieves competitive or superior retrieval performance using only five tokens per document, substantially reducing decoding cost while retaining strong semantic retrieval accuracy.

[Findings] Do MLLMs Understand Pointing? Benchmarking and Enhancing Referential Reasoning in Egocentric Vision ▶ *Underline Video*

Chentao Li, Zirui Gao, Mingze Gao, Yinglian Ren, Jianjiang Feng and Jie Zhou

Egocentric AI agents, such as smart glasses, rely on pointing gestures to resolve referential ambiguities in natural language commands. However, despite advancements in Multimodal Large Language Models (MLLMs), current systems often fail to precisely ground the

spatial semantics of pointing. Instead, they rely on spurious correlations with visual proximity or object saliency, a phenomenon we term "Referential Hallucination." To address this gap, we introduce EgoPoint-Bench, a comprehensive question-answering benchmark designed to evaluate and enhance multimodal pointing reasoning in egocentric views. Comprising over 11k high-fidelity simulated and real-world samples, the benchmark spans five evaluation dimensions and three levels of referential complexity. Extensive experiments demonstrate that while state-of-the-art proprietary and open-source models struggle with egocentric pointing, models fine-tuned on our synthetic data achieve significant performance gains and robust Sim-to-Real generalization. This work highlights the importance of spatially-aware supervision and offers a scalable path toward precise egocentric AI assistants. The project website is available at https://guyyyug.github.io/EgoPoint-Bench/.

[Long] Does RLVR Extend Reasoning Boundaries? Investigating Capability Expansion in Vision-Language Models ▶ *Underline Video*

Minghe Shen, Zhuo Zhi, Chonghan Liu, Shuo Xing, Zhengzhong Tu and Che Liu

Recent studies posit that Reinforcement Learning with Verifiable Rewards (RLVR) primarily amplifies behaviors inherent to the pre-training distribution rather than inducing new capabilities, but these insights are predominantly limited to language-only domains, leaving the dynamics of visual-centric spatial reasoning under-explored. To examine the impact of RLVR on the capability boundaries of Vision-Language Models (VLMs), we introduce Ariadne, a controlled framework based on synthetic maze navigation where the reasoning difficulty is precisely regulated by path length and the number of turns. We demonstrate that applying RLVR extends the spatial reasoning boundary, achieving success on problems where the base policy VLM consistently attains 0% accuracy despite increasing pass@k sampling budgets, indicating that the optimized policy successfully navigates search spaces that were effectively unreachable by the base distribution. Furthermore, despite being trained exclusively on synthetic mazes, we evaluate the model on two real-world navigation benchmarks (MapBench and ReasonMap) in a zero-shot setting. The observed improvements in these out-of-domain tasks suggest genuine spatial reasoning capability expansion rather than mere sampling efficiency. Our code is available at: https://github.com/MingheShen/Ariadne

[Findings] Double-Edged Sword or Sharp Tool? Designing and Evaluating Triadic LLM-Teacher Collaboration for K-12 Writing at Scale ▶ *Underline Video*

Canran WANG, Yuwen Yang, Zhen Wang, Ming MA, Ding Yu, Chentai Wang, Keman Huang and Xiaoyong Du

The double-edged sword of integrating Large Language Models (LLMs) requires an effective triadic collaboration mechanism among LLMs, teachers and students, especially for K-12 education. By developing a triadic collaboration system to support K-12 writing learning, a multidimensional evaluation framework grounded in Systemic Functional Linguistics and the suggestion trajectory tracing pipeline, this paper contributes a large-scale empirical dataset involving 57,954 essays from 10,195 students across 120 schools over two years. Our findings confirm the efficacy of this system in improving writing quality through a strategic labor division: the LLM serves as a generative engine to mitigate teacher burnout, and the teacher acts as a pedagogical gatekeeper and bridge to guarantee feedback quality. While both LLM and teacher are critical for skill improvement, we uncover a ceiling effect where excessive linguistic expansion yields diminishing marginal utility. These suggest a dynamically adaptive LLM-teacher collaboration as student proficiency increases.

[Long] Dynamic Long Context Reasoning over Compressed Memory via End-to-End Reinforcement Learning ▶ *Underline Video*

Zhuoen Chen, Dongfang Li, Meishan Zhang, Baotian Hu and Min Zhang

Large Language Models (LLMs) face severe challenges in long-context processing, including quadratic computational costs, information forgetting, and the context fragmentation inherent in Retrieval-Augmented Generation (RAG). We introduce LycheeMemory, a cognitively inspired framework that enables efficient long-context inference via chunk-wise compression and selective memory recall, rather than processing all raw tokens. LycheeMemory segments the input into chunks and encodes each into compressed KV-cache style representations using a Compressor. A Gate then dynamically selects relevant memory blocks, which a Reasoner iteratively processes with an evolving working memory to solve downstream tasks. The Compressor and Reasoner are jointly optimized via end-to-end reinforcement learning, while the Gate is trained separately as a classifier. Experimental results demonstrate that LycheeMemory achieves competitive accuracy (up to 82% in ablation variants) on multi-hop reasoning benchmarks (e.g., RULER-HQA), successfully extrapolates context length from 7K to 1.75M, and provides a favorable accuracy-efficiency trade-off against strong long-context baselines. Notably, compared to MemAgent, LycheeMemory achieves an average 2x reduction in peak GPU memory usage and a 6x speedup during inference.

[Long] Editing the Moving World: Model Editing for Video LLMs ▶ *Underline Video*

Qian Zhang, Xinye Li, Xiaokai Wu, Junhao Xu, Zhanyue Qin, Qingbin Liu, Junxian Cai, Xi Chen, Bolin Zhang, Zhiying Tu, Dianhui Chu, Xiaoyan Yu and Dianbo Sui

Model Editing, also known as knowledge editing, is receiving increasing attention in the field of Large Language Models (LLMs). However, existing model editing approaches predominantly focus on knowledge-level or static visual domains, overlooking dynamic semantics. This paper exploratively applies six representative model editing methods (FT, IKE, MEND, SERAC, MEMIT and AlphaEdit) to Video Large Language Models (Vid-LLMs) and introduces the first benchmark specifically designed for Vid-LLMs editing, VMEB (Vid-LLMs Model Editing Benchmark), systematically extending model editing research from static modalities to dynamic video scenarios. We position this work as a forward-looking benchmark and a foundational diagnostic study: in the video paradigm, our evaluation dimensions encompass traditional metrics including Reliability, Locality, and Generality, while also introducing a video-specific metric: Robustness. Based on experimental results, we analyze the strengths and limitations of existing model editing approaches, and identify new challenges and research directions for the future development of the model editing field within the context of multimodal and video paradigms. Our benchmark is

available at https://github.com/Sakabamrisa/VMEB.

[Long] EduBench: A Comprehensive Benchmarking Dataset for Evaluating Large Language Models in Diverse Educational Scenarios ▶ *Underline Video*

Bin Xu, Yu Bai, Huashan Sun, Yiguan Lin, Siming Liu, Xinyue Liang, Yaolin Li, Zhuangzhi Dong, Jingren Zhang, Yufan Deng, Xinyu Zou, Yang Gao and Heyan Huang

As large language models continue to advance, their application in educational contexts remains underexplored and under-optimized. In this paper, we address this gap by introducing the first diverse benchmark tailored for educational scenarios, incorporating synthetic data containing 9 major scenarios and over 4,000 distinct educational contexts. To enable comprehensive assessment, we propose a set of multi-dimensional evaluation metrics that cover 12 critical aspects relevant to both teachers and students. We further apply human annotation to ensure the effectiveness of the model-generated evaluation responses. Additionally, we succeed to train a relatively small-scale model on our constructed dataset and demonstrate that it can achieve performance comparable to state-of-the-art large models (e.g., Deepseek V3, Qwen Max) on the test set. Overall, this work provides a practical foundation for the development and evaluation of education-oriented language models.

[Long] Embodied-Reasoner: Synergizing Visual Search, Reasoning, and Action for Embodied Interactive Tasks ▶ *Underline Video*

Wenqi Zhang, Mengna Wang, Gangao Liu, Huixin Xu, Yiwei Jiang, Yongliang Shen, Guiyang Hou, Zhe Zheng, Hang Zhang, Xin Li, Jiajun Liu, Weiming Lu, Peng Li and Yueting Zhuang

Recent advances in reasoning models have demonstrated remarkable capabilities on mathematical and coding tasks. However, their effectiveness in embodied domains, where the agent must continuously interact with environments and process observation-action interleaved trajectories, remains largely unexplored. We present Embodied-Reasoner, a reasoning model for interactive embodied tasks. Unlike mathematical reasoning that relies primarily on logical deduction, embodied scenarios demand spatial understanding, temporal reasoning, and ongoing self-reflection based on interaction history. To address these challenges, we synthesize 9.3k coherent Observation-Thought-Action trajectories containing 64k ego-centric images and 90k diverse reasoning processes (analysis, spatial reasoning, reflection, planning, and verification). We develop a three-stage training recipe that progressively enhances the model's capabilities through imitation learning, rejection sampling tuning on self-exploration trajectories, and reflection tuning. The evaluation shows that our model significantly outperforms advanced visual reasoning models, e.g., exceeds OpenAI o1, o3-mini, and Claude-3.7 by +9%, 24%, and +13%. Analysis reveals that our model exhibits fewer repeated searches and logical inconsistencies, with particular advantages in complex long-horizon tasks. Real-world testing further validates the effectiveness of our approach.

[SRW] Emergence of Minimal Circuits for Indirect Object Identification in Attention-Only Transformers ▶ *Underline Video*

Rabin Adhikari

Mechanistic interpretability aims to reverse-engineer large language models (LLMs) into human-understandable computational circuits. However, the complexity of pretrained models often obscures the minimal mechanisms required for specific reasoning tasks. In this work, we train small, attention-only transformers from scratch on a symbolic version of the Indirect Object Identification (IOI) task, a benchmark for studying coreference-like reasoning in transformers. Surprisingly, a single-layer model with only two attention heads achieves perfect IOI accuracy, despite lacking MLPs and normalization layers. Through residual stream decomposition, spectral analysis, and embedding interventions, we find that the two heads specialize into additive and contrastive subcircuits that jointly implement IOI resolution. Furthermore, we show that a two-layer, one-head model achieves performance comparable to that of a multi-head model by composing information across layers primarily through query-key interactions. These results demonstrate that task-specific training induces highly interpretable, minimal circuits, offering a controlled testbed for probing the computational foundations of transformer reasoning.

[Long] Empowering GUI Agents via Autonomous Experience Exploration and Hindsight Experience Utilization for Task Planning ▶ *Underline Video*

Tianyi Men, Zhuoran Jin, Pengfei Cao, Yubo Chen, Kang Liu and Jun Zhao

Multimodal web agents can assist humans in operating repetitive GUI tasks, where effective task planning is essential for decomposing complex tasks into executable actions. While small opensource MLLMs are costefficient and privacy-preserving compared with commercial large models, they suffer from weak planning and limited crosswebsite generalization. To address these limitations, we introduce the planning experience exploration and utilization (PEEU) method, which autonomously explores environments to discover experiences and utilizes hindsight experience to synthesize strictly aligned, high-level training data. To quantitatively analyze the generalization behaviors driving this performance, we propose the task decomposition hierarchical analysis framework (TDHAF) to systematically study compositional generalization across three task granularities: low, middle and high levels. Our analysis reveals that mastering lowlevel atomic skills does not guarantee highlevel planning competence, while highlevel task training yields stronger OOD generalization. Experiments on real-world benchmarks demonstrate PEEU's superior effectiveness: our 7B model achieves 30.6% accuracy, outperforming the much larger Qwen2.5-VL-32B model. These demonstrate constructing hindsight highlevel tasks and leveraging experiences is crucial for OOD planning abilities of small MLLMs.

[Industry] Enhancing Job Evaluation with Data Augmentation and Text Classification ▶ *Underline Video*

Samaneh Jalilian, Niels van Weeren, Mohammad Shokri, Thijmen Bijl and Suzan Verberne

Accurate job grading and evaluation are essential for ensuring fair compensation in Human Resources (HR) planning. In this research, we propose to improve job evaluation by semi-automating a manual, time-consuming, and inconsistent process with text-based classification models. We address three prediction tasks: job title classification, grading, and compensation prediction. For job title classification, we fine-tune a RoBERTa model for classification and use Gemini to generate synthetic job descriptions for rare job titles. For grade and compensation prediction, we compare TF-IDF and transformer-based embeddings (DistilRoBERTa, MPNet, MiniLM) in combination with deep neural networks and tree-based models (Random Forest, XGBoost). We optimize all models using grid search with hyperparameter tuning and cross-validation. The results show that job title classification by RoBERTa with Gemini-generated descriptions works well with an accuracy of about 97%. In our regression experiments, our models get promising results: for grade prediction, a tuned TF-IDF + XGBoost model achieves a mean absolute error (MAE) of 0.185, and for annual salary prediction, MiniLM embeddings with XGBoost get an MAE of EUR1,587. These findings demonstrate that a semi-automated pipeline can enhance traditional manual processes by boosting consistency, speeding up HR workflows, and reducing biased assessments.

[Long] Enhancing Reinforcement Learning for Radiology Report Generation with Evidence-aware Rewards and Self-correcting Preference Learning ▶ *Underline Video*

Guoyan Liang

Recent reinforcement learning (RL) approaches have advanced radiology report generation (RRG), yet two core limitations persist: (1) report-level rewards offer limited evidence-grounded guidance for clinical faithfulness; and (2) current methods lack an explicit self-improving mechanism to align with clinical preference. We introduce clinically aligned Evidence-aware Self-Correcting Reinforcement Learning (ESC-RL), comprising two key components. First, a Group-wise Evidence-aware Alignment Reward (GEAR) delivers group-wise, evidence-aware feedback. GEAR reinforces consistent grounding for true positives, recovers missed findings for false negatives, and suppresses unsupported content for false positives. Second, a Self-correcting Preference Learning (SPL) strategy automatically constructs a reliable, disease-aware preference dataset from multiple noisy observations and leverages an LLM to synthesize refined reports without human supervision. ESC-RL promotes clinically faithful, disease-aligned reward and supports continual self-improvement during training. Extensive experiments on two public chest X-ray datasets demonstrate consistent gains and state-of-the-art performance.

[Findings] Enrich, Aggregate, and Generate: Three-stage Biomedical Data-to-Text Generation Using Large Language Models in Low-resource Scenarios ▶ *Underline Video*

Yupian Lin, Guangya Yu, Cheng Yuan, Hui Luo, Yuang Bian and Tong Ruan

Biomedical data-to-text generation aims at generating textual natural language descriptions that can fluently and precisely describe the biomedical structured data. However, biomedical data-to-text generation faces the dilemma of a lack of labeled data due to the privacy and scarcity of medical data. Large language models (LLMs) have demonstrated the ability to solve few-shot tasks through in-context learning (ICL). In this paper, we are the first to explore the performance of different LLMs in the biomedical data-to-text generation task.To address the issues of semantic sparsity and misinterpretation of numerical values in biomedical structured data, we propose an EAG (Enrich, Aggregate, and Generate) framework, a simple but efficient LLM-based three-stage biomedical D2T approach in low-resource scenarios. We conduct extensive evaluations of closed-source general LLMs, open-source general LLMs, and open-source medical LLMs. The results show that the EAG framework provides good interpretability and superior performance, achieving state-of-the-art performance on the BioLeaflets dataset. The code and data will be released at https://github.com/FXLP/EAG.

[Industry] Entity Exchange in the Wild: A Diagnostic Study of LLM Based Real-World Conversational Entity Extraction ▶ *Underline Video*

Soumya Jain and Ayush Kumar

Entity extraction from spoken customer-agent conversations is increasingly driving automation in contact centers. In these settings, extraction errors can trigger incorrect system actions, including database updates, verification failures, and unintended workflow execution. While prior work has examined the impact of transcription noise and cross-turn reasoning, it has not systematically analyzed how entity-exchange phenomena themselves shape extraction performance. We model conversational entity exchange along three orthogonal axes: Initiation (how an entity becomes relevant in the dialogue), Evolution (how commitment to an entitys value develops or changes across turns), and Articulation (how the final committed value is expressed in surface form). We evaluate 16 large language models on 6,387 real-world customer-agent conversations spanning 12 entity types across numeric, alphanumeric, temporal, and free-text categories. Performance varies by as much as 50-60% within the same model depending solely on the underlying entity-exchange phenomena. The most severe failures occur when entity values are revised during the interaction and the model must distinguish intermediate mentions from the final committed value. Even in the absence of revision, digit-by-digit and encoded expressions remain persistent sources of error. Error-Aware prompting improves extraction across all three axes, yielding average gains of up to 6.4% across models. Together, this work provides a structured framework for benchmarking entity extraction in real-world deployments and isolating systematic failure modes grounded in conversational structure.

[Long] Escaping the Echo Trap: On Credit Assignment Failure in Multi-turn LLM Self-Reflection ▶ *Underline Video*

Linxuan Du, Guangquan Xue, Xiaobo Liang, Qipeng Huang, Yuyang Ding, Xinyu Shi, Yijun Zhang, Ji Qi, Wenpeng Zhu, Juntao Li and Min Zhang

Despite the potential of multi-turn self-reflection to improve LLM reasoning, its effectiveness in practice is severely constrained by a failure

mode we term the Echo Trap. Specifically, this phenomenon gives rise to two coupled problems: (1) the model becomes limited by its inherent capabilities and tends to repeat earlier reflections to preserve reward signals; (2) once such "copy" behavior is reinforced, the model ceases to try new strategies, leading to exploration collapse. We attribute this issue to imprecise credit assignment during training, as standard GRPO assigns rewards at the trajectory level, making it difficult to distinguish which reflection steps contribute to improved outcomes. To address this limitation, we propose a tree-structured extension of GRPO for multi-turn self-reflection, which enables more accurate advantage estimation. Through extensive experiments, we analyze the Echo Trap and demonstrate that our method effectively mitigates behavior collapse and improves performance across multiple benchmarks.

[Long] Evaluating LLMs on Large-Scale Graph Property Estimation via Random Walks ▶ *Underline Video*

Sunil Kumar Maurya and Xin Liu

With the rapidly improving reasoning abilities of Large Language Models (LLMs), there is also a rising demand to use them in a wide variety of domains. This brings about the need to carefully evaluate the limits of the capabilities of these models with various tests and benchmarks. Graph structures are ubiquitous in real-world data, and are often used to represent and analyze relationship patterns within data. Many benchmarks have already been proposed in the graph literature to test the reasoning ability of LLMs to follow and execute graph algorithms. However, due to the limited context length of LLMs, these benchmarks consist of very small graphs. In real-world data, the size of graphs can be significantly larger, and in many cases, not fully accessible. In this paper, we examine a class of problems that arises with very large graphs having limited accessibility. We propose a large graph benchmark dataset, EstGraph, and introduce four distinct tasks designed to estimate large graph properties. We evaluate the reasoning abilities of LLMs on these tasks using a wide variety of graph datasets. In addition, we provide task-specific prompt constructions based on random walk sampling of large graphs (up to millions of nodes) that effectively convey sufficient information to LLMs within the limits of context length.

[Findings] Explanation Quality Assessment as Ranking with Listwise Rewards ▶ *Underline Video*

Thomas Bailleux, Tanmoy Mukherjee, Emmanuel Lonca, Pierre Marquis and Zied Bouraoui

We reformulate explanation quality assessment as a ranking problem rather than a generation problem. Instead of optimizing models to produce a single best explanation token-by-token, we train reward models to discriminate among multiple candidate explanations and learn their relative quality. Concretely, we construct per-instance candidate sets with graded quality levels and train listwise and pairwise ranking models (ListNet, LambdaRank, RankNet) to preserve ordinal structure and avoid score compression typical of pointwise regression or binary preference objectives. We observe three findings: First, ranking losses consistently outperform regression on score separation across all domains tested. Second, the optimal ranking loss depends on data characteristics: listwise objectives excel with well-separated quality tiers, while pairwise methods are more robust to noisy natural annotations. Third, when trained on carefully curated and well-structured data, small encoder models can match models that are orders of magnitude larger, suggesting that data quality matters more than model scale. Finally, when used as rewards in policy optimization, ranking-based scores enable stable convergence in settings where regression-based rewards fail entirely. Code and data are available at: https://github.com/Tankiit/PPO_Learning_to_rank

[Findings] FAMA: Failure-Aware Meta-Agentic Framework for Open-Source LLMs in Interactive Tool Use Environments ▶ *Underline Video*

Amir Saeidi, Venkatesh Mishra, Souradeep Mukhopadhyay, Gaowen Liu, Ali Payani, Jayanth Srinivasa and Chitta Baral

Large Language Models are being increasingly deployed as the decision-making core of autonomous agents capable of effecting change in external environments. Yet, in conversational benchmarks, which simulate real-world customer-centric issue resolution scenarios, these agents frequently fail due to the cascading effects of incorrect decision-making. These challenges are particularly pronounced for open-source LLMs with smaller parameter sizes, limited context windows, and constrained inference budgets, which contribute to increased error accumulation in agentic settings. To tackle these challenges, we present the Failure-Aware Meta-Agentic (FAMA) framework. FAMA operates in two stages: first, it analyzes failure trajectories from baseline agents to identify the most prevalent errors; second, it employs an orchestration mechanism that activates a minimal subset of specialized agents tailored to address these failures by injecting a targeted context for the tool-use agent before the decision-making step. Experiments across open-source LLMs demonstrate performance gains up to 27% across evaluation modes over standard baselines. These results highlight that targeted curation of context through specialized agents to address common failures is a valuable design principle for building reliable, multi-turn tool-use LLM agents that simulate real-world conversational scenarios.

[Long] Fast and Accurate Fisher-Guided Quantization via Efficient Kronecker Factorization ▶ *Underline Video*

Viktoriia Chekalina, Timofey Gerasin, Andrey Kuznetsov and Evgeniy Frolov

Quantization has shown strong results in preserving model quality under compression. However, under aggressive bit-width reductions, even quantization may require additional information to prevent performance degradation. A natural source of it is second-order curvature information, captured by the Hessian. Since the Hessian of the model layers is prohibitively large, direct computation is infeasible, making structured parameterizations and approximations crucial in practice. In this work, we propose efficient Kronecker-factored approximation yielding state-of-the-art performance when integrated into existing quantization schemes. Evaluations on the LLaMA and Qwen model families show near-baseline quality at 4-bit compression and only a 5-6% degradation at 2-bit. Moreover, our method substantially accelerates the most expensive component in second-order quantization – Hessian parameterization – achieving up to a 10x speedup over prior approaches.

[Long] FastV-RAG: Towards Fast and Fine-Grained Video QA with Retrieval-Augmented Generation ▶ *Underline Video*
Gen Li and Peiyu Liu
Vision-Language Models (VLMs) excel at visual reasoning but still struggle with external knowledge integration. Retrieval-Augmented Generation(RAG) is a promising solution, but current methods remain inefficient and often fail to maintain high answer quality. To address these challenges, we propose VideoSpeculateRAG, an efficient VLM-based RAG framework built on two key ideas. First, we introduce a speculative decoding pipeline: a lightweight draft model quickly generates multiple answer candidates, which are then verified and refined by a more accurate heavyweight model, substantially reducing inference latency without sacrificing correctness. Second, we identify a major source of error, incorrect entity recognition in retrieved knowledge, and mitigate it with a simple yet effective similarity-based filtering strategy that improves entity alignment and boosts overall answer accuracy. Experiments demonstrate that VideoSpeculateRAG achieves comparable or higher accuracy than standard RAG approaches, while speeding up the inference by approximately 2x. Our framework highlights the potential of combining speculative decoding with retrieval-augmented reasoning to enhance efficiency and reliability in complex, knowledge-intensive multimodal tasks.

[Long] Fisher-Driven Adaptive Locating for Knowledge Editing in Large Language Models ▶ *Underline Video*
Chenghao Xu, Jiexi Yan, Guangtao Lyu, Qi Liu, Muli Yang and Cheng Deng
Large language models (LLMs) store extensive factual knowledge acquired during pretraining, yet this knowledge is inherently static and may become inaccurate or outdated, leading to knowledge hallucinations. Knowledge editing offers an efficient alternative to full retraining by enabling targeted factual updates while preserving overall model behavior. Existing locate-then-edit methods, however, rely on fixed layer selection strategies, treating the locating stage as a static design choice and failing to account for the hierarchical and instance-dependent nature of knowledge representation in LLMs. In this paper, we propose FiDAL, a Fisher-driven adaptation-aware locating strategy that dynamically identifies which model components should be edited for a given knowledge update. FiDAL formulates localization as a weight-level decision problem and leverages Fisher Information to select layers that are both influential and sensitive to factual modifications. A lightweight probing stage with low-rank modulation enables efficient localization with minimal overhead. Experiments on standard benchmarks demonstrate that FiDAL consistently improves editing effectiveness and knowledge preservation across multiple editing methods.

[Findings] FrameNet-Cultures: A Benchmark for Evaluating LLMs via Cross-Cultural Frame Semantics ▶ *Underline Video*
Neda Jamshidi, Anders Søgaard and Monica Bianchini
Large language models (LLMs) exhibit cultural biases, yet existing benchmarks rely on closed-form, domain-specific questionnaires. We introduce FRAMENET-CULTURES, a benchmark for evaluating cultural alignment in LLMs based on Fillmore-style frame semantics. Using the EveryCulture encyclopedia, we construct a lexicon of 18 cultural frames (e.g., greeting,child-rearing) across 20 countries, treating it as a structured reference for comparison rather than a definitive representation of contemporary societies. For each frame, we prompt five major LLMs, ChatGPT-5, Gemini-2.5-Flash, Mistral-Large, Qwen-3-Max, DeepSeek-V3.2, three times to generate open-ended instantiations, which are manually annotated and binarized. We measure alignment with country- and continent-level profiles using normalized Hamming distance, and validate cultural recognizability through human evaluation of generated dialogues. Under culture-neutral prompting, outputs align most closely with European profiles, followed by Asian and American ones, indicating a consistent cross-model pattern. With culture-specific prompting, models shift toward the target regions, aligning most strongly with Africa for Ethiopia and with Asia for India. FRAMENET-CULTURES is the first open-ended benchmark for cultural alignment relying on frame semantics. Data, prompts, and annotations are publicly available at https://github.com/neda-jamshidi/FrameNet-Cultures.

[Long] From Completion to Editing: Unlocking Context-Aware Code Infilling via Search-and-Replace Instruction Tuning ▶ *Underline Video*
Jiajun Zhang, Zeyu Cui, Jiaxi Yang, Lei Zhang, Yuheng Jing, Zeyao Ma, Tianyi Bai, Zilei Wang, Qiang Liu, Liang Wang, Binyuan Hui and Junyang Lin
The dominant Fill-in-the-Middle (FIM) paradigm for code completion is constrained by its rigid inability to correct contextual errors and reliance on unaligned, insecure Base models. While Chat LLMs offer safety and Agentic workflows provide flexibility, they suffer from performance degradation and prohibitive latency, respectively. To resolve this dilemma, we propose Search-and-Replace Infilling (SRI), a framework that internalizes the agentic verification-and-editing mechanism into a unified, single-pass inference process. By structurally grounding edits via an explicit search phase, SRI harmonizes completion tasks with the instruction-following priors of Chat LLMs, extending the paradigm from static infilling to dynamic context-aware editing. We synthesize a high-quality dataset, SRI-200K, and fine-tune the SRI-Coder series. Extensive evaluations demonstrate that with minimal data (20k samples), SRI-Coder enables Chat models to surpass the completion performance of their Base counterparts. Crucially, unlike FIM-style tuning, SRI preserves general coding competencies and maintains inference latency comparable to standard FIM. We release our dataset and models, establishing SRI as a robust, secure, and efficient alignment recipe for next-generation interactive development.

[Industry] From Graph to Text and Back: Semantic Fidelity in Automated Industrial Knowledge Graphs ▶ *Underline Video*
Kamyar Zeinalipour, Silvia Severini, Alessia Borghini, Sara Cardarelli, Marco Maggini and Marco Gori
Knowledge Graphs (KGs) are the backbone of reliable industrial data strategies, yet verbalizing them with Large Language Models (LLMs) often leads to unacceptable risks for high-stakes applications, such as hallucinations or omitted relations. To enforce strict semantic fidelity in KG-to-text generation, we introduce a self-supervised round-trip pipeline. The system verbalizes KG triples into text and immediately

attempts to reconstruct the original graph from that text; only verbalizations that enable perfect graph recovery are retained. This creates a closed feedback loop that guarantees the generated text is semantically equivalent to the source data. Experiments confirm that our automated round-trip consistency score correlates strongly with expert judgment, effectively acting as a scalable proxy for human review. Furthermore, we show that standard LLMs can bootstrap their own KG-extraction and generation capabilities by fine-tuning on this trusted synthetic data. Our approach yields significant improvements in triple-extraction accuracy and verbalization faithfulness without relying on costly manual annotation or massive teacher models, offering a practical path to deploying trustworthy, KG-grounded AI systems.

[Findings] From Shijing to English and German: Resources and Evaluation for LLM Translation of Early Chinese Poetry ▶ *Underline Video*

Ying Jiao and Meng Sun

While large language models (LLMs) show promise in literary translation, Shijing (The Book of Songs) serves as a rigorous yet under-explored testbed for testing their limits, given its linguistic antiquity and complex poetic constraints. Automated evaluation in this domain is currently hindered by a scarcity of multilingual resources and the inadequacy of existing metrics in capturing both semantic fidelity and aesthetic quality. In this paper, we bridge these gaps by curating a Shijing parallel corpus with line-by-line Chinese-English-German alignments, together with a fine-grained lexical knowledge base (KB) for archaic expressions. Based on these resources, we propose a hybrid evaluation framework that integrates knowledge-driven, rule-based, and LLM-as-judge metrics. Experimental results show that our framework achieves significantly higher human correlation than traditional metrics and demonstrates high statistical stability. By applying this framework to evaluate representative LLMs, we reveal that while top-tier models like Gemini-2.5-Pro and DeepSeek-3.1 show potential, achieving semantic precision and aesthetic sophistication, particularly in lower-resource directions like German, remains a persistent challenge. Our code, lexical KB, and corpus reconstruction protocols are available at https://github.com/ML-KULeuven/ShijingLLMTrans.

[Long] G-Cap: A Game Character Caption Generator ▶ *Underline Video*

Yang Yang, Feng Hu, Haiming Zhang, Xu Cheng, Gui Zheng, Liang Yao and Wenqi Ren

While Large Vision-Language Models (LVLMs) have demonstrated remarkable proficiency in image captioning, existing research primarily focuses on real-world scenarios, leaving surreal, highly stylized, and semantically hybrid virtual-world scenarios significantly underexplored. In this work, we introduce Game Character Captioning, a novel task designed to evaluate LVLMs' capability to perceive and describe game character from the virtual-world. To facilitate evaluation, we establish GC-Bench, a manually annotated benchmark, and propose Graph-F1 to effectively assess performance on this task. Our evaluation reveals that: (1) current state-of-the-art LVLMs, including closed-source giants such as Gemini 3 Pro and GPT-5.1, struggle to maintain the high performance seen in real-world scenarios; and (2) a notable gap exists between open-source and closed-source models. To bridge this gap, we construct GC-148K, a large-scale dataset generated via a specialized data pipeline, and develop the G-Cap series. Experiments demonstrate that G-Cap series rivals the performance of advanced closed-source models at a lower cost, offering an efficient solution for industrial-grade production environment.

[Long] Gated Differentiable Working Memory for Long-Context Language Modeling ▶ *Underline Video*

Lingrui Mei, Shenghua Liu, Yiwei Wang, Yuyao Ge, Baolong Bi, Jiayu Yao, Jun Wan, Ziling Yin, Jiafeng Guo and Xueqi Cheng

Long contexts break transformers: attention scores dilute across thousands of tokens, critical information gets lost in the middle, and the model cannot adapt to novel patterns at inference time. Recent work on test-time adaptation addresses this by maintaining a form of working memory—transient parameters updated on the current context—but existing approaches employ uniform write policies that waste computation on low-value regions and suffer from high gradient variance across semantically heterogeneous contexts. In this work, we reframe test-time adaptation as a budget-constrained memory consolidation problem, asking: given limited computational budget, which parts of the context should be consolidated into working memory? We propose GDWM (Gated Differentiable Working Memory), a framework that introduces a Write Controller to gate the memory consolidation process. Our controller estimates Contextual Utility—an information-theoretic measure quantifying how much each region depends on long-range context—and allocates gradient steps accordingly, subject to a coverage constraint that ensures global representation. Theoretically, we prove that our chunk-restricted sampling strategy reduces gradient variance by eliminating inter-chunk variance via the Law of Total Variance. Experiments on ZeroSCROLLS and LongBench v2 benchmarks demonstrate that GDWM achieves comparable or superior performance with 4 x fewer gradient steps compared to uniform baselines—excelling on sparse-information tasks (+6–13% on Qasper, +5–13% on GovReport for smaller models) while revealing principled trade-offs on dense-coverage tasks, establishing a new efficiency-performance Pareto frontier for test-time adaptation.

[Long] Generative Gamer: Learning Equilibrium Strategy by LLM-driven Dynamic Deduction ▶ *Underline Video*

Yadong Zhang, Xinshu Shen, Yupei Ren, Shangqing Zhao and Man Lan

Large Language Models (LLMs) have demonstrated remarkable general capabilities, yet they falter in domains requiring deep strategic reasoning. A primary obstacle is the need to navigate a game tree that grows exponentially with search depth, a task for which their generative nature is ill-suited. To address this, we introduce Generative Gamer (GenGamer), a framework that trains LLMs to reason like an expert player. Instead of attempting an exhaustive search, GenGamer learns to generate a compact, pruned reasoning trajectory termed as a Dynamic Deduction. This is achieved by integrating three key strategies: action pruning based on policy confidence, state pruning via value estimation, and branch pruning inspired by alpha-beta principles. Furthermore, to train the model effectively, we propose the Deduction Tree Reward (DTR), a process-oriented mechanism that provides step-by-step feedback on the quality of the reasoning process, rather than relying solely on the final game outcome. Experiments on complex games such as Tic-Tac-Toe and Leduc Poker demonstrate that GenGamer significantly

enhances the strategic capabilities of LLMs, enabling them to achieve performance that surpasses current state-of-the-art language models.

[Findings] GeoLAN: Geometric Learning of Latent Explanatory Directions in Large Language Models ▶ *Underline Video*
Tianyu Pan and Damon L. Woodard
Large language models (LLMs) demonstrate strong performance, but they often lack transparency. We introduce GeoLAN, a training framework that treats token representations as geometric trajectories and applies stickiness conditions inspired by recent developments related to the Kakeya Conjecture. We have developed two differentiable regularizers, Katz-Tao Convex Wolff (KT-CW) and Katz-Tao Attention (KT-Attn), that promote isotropy and encourage diverse attention. Our experiments with Gemma-3 (1B, 4B, 12B) and Llama-3-8B show that GeoLAN frequently maintains task accuracy while improving geometric metrics and reducing certain fairness biases. These benefits are most significant in mid-sized models. Our findings reveal scale-dependent trade-offs between geometric precision and performance, suggesting that geometry-aware training is a promising approach to enhance mechanistic interpretability.

[Findings] Geoparsing: Diagram Parsing for Plane and Solid Geometry with a Unified Formal Language ▶ *Underline Video*
Peijie Wang, Ming-Liang Zhang, Jun Cao, CHAO DENG, Dekang Ran, Pi Bu, Hongda Sun, Xuan Zhang, Yingyao Wang, Jun Song, Bo Zheng, Fei Yin and Cheng-Lin Liu
Multimodal Large Language Models (MLLMs) have achieved remarkable progress but continue to struggle with geometric reasoning, primarily due to the perception bottleneck regarding fine-grained visual elements. While formal languages have aided plane geometry understanding, solid geometry which requires spatial understanding remains largely unexplored. In this paper, we address this challenge by designing a unified formal language that integrates plane and solid geometry, comprehensively covering geometric structures and semantic relations. We construct GDP-29K, a large-scale dataset comprising 20k plane and 9k solid geometry samples collected from diverse real-world sources, each paired with its ground-truth formal description. We propose a training paradigm combining Supervised Fine-Tuning with Reinforcement Learning via Verifiable Rewards, which effectively enforces syntactic correctness and geometric consistency. Experiments show that our approach achieves state-of-the-art parsing performance. Furthermore, we demonstrate that our parsed formal descriptions serve as a critical cognitive scaffold, significantly boosting MLLMs' capabilities for downstream geometry reasoning tasks.

[Long] Gold-Medal-Level Olympiad Geometry Solving with Efficient Heuristic Auxiliary Constructions ▶ *Underline Video*
Boyan Duan, Xiao Liang, Shuai Lu, Yaoxiang Wang, yelong shen, Kai-Wei Chang, Ying Nian Wu, Mao Yang, Weizhu Chen and Yeyun Gong
Automated theorem proving in Euclidean geometry, particularly for International Mathematical Olympiad (IMO) level problems, remains a major challenge and an important research focus in Artificial Intelligence. In this paper, we present a highly efficient method for geometry theorem proving that runs entirely on CPUs without relying on neural network-based inference. Our initial study shows that a simple random strategy for adding auxiliary points can achieve "silver-medal" level human performance on IMO. Building on this, we propose HAGeo, a Heuristic-based method for adding Auxiliary points in Geometric deduction that solves 28 of 30 problems on the IMO-30 benchmark, achieving "gold-medal" level performance and surpassing AlphaGeometry, a competitive neural network-based approach, by a notable margin. To evaluate our method and existing approaches more comprehensively, we further construct HAGeo, a benchmark consisting of 409 geometry problems with human-assessed difficulty levels. Compared with the widely used IMO-30, our benchmark poses greater challenges and provides a more precise evaluation, setting a higher bar for geometry theorem proving.

[Long] GroupToM-Bench: Benchmarking Group Theory of Mind and Nonlinear Social Emergence in MLLMs ▶ *Underline Video*
Weidong Tang, Jierui Li, Yueling Hou, Zihan Mei, Can Zhang, Xinyan Wan, Zhiyuan Liang, Pengfei Zhou, Yang You and Wangbo Zhao
True general intelligence requires not only a model of the physical world but also a social world model: the capacity to infer how individual mental states interact and crystallize into group-level outcomes. Despite notable progress in individual-level Theory of Mind (ToM) reasoning, existing multimodal large language models systematically fail at this: collective behavior emerges non-linearly from social tensions, conformity dynamics, and structural constraints, and cannot be recovered by summing individual intentions. We present GroupToM-Bench, the first multimodal benchmark for group-level ToM, built around a causal chain spanning micro-level BDI states (belief, desire, intention), meso-level group tension and structural constraints, and macro-level outcome prediction and mechanistic attribution. To probe this full arc, we develop a seven-level cognitive audit framework. Experiments reveal that frontier models perform significantly below human levels, exposing fundamental blind spots in modeling social structures and nonlinear collective behavior.

[Industry] Harmonizing Dense and Sparse Signals in Multi-turn RL: Dual-Horizon Credit Assignment for Industrial Sales Agents ▶ *Underline Video*
Haojin Yang, Ai Jian, Yiwei Wang, Xinyue Huang, Weipeng Zhang, Ke Zeng, Xunliang Cai and Jingqing Ruan
Optimizing large language models for industrial sales requires balancing long-term commercial objectives (e.g., conversion rate) with immediate linguistic constraints such as fluency and compliance. Conventional reinforcement learning often merges these heterogeneous goals into a single reward, causing high-magnitude session-level rewards to overwhelm subtler turn-level signals, which leads to unstable training or reward hacking. To address this issue, we propose Dual-Horizon Credit Assignment (DuCA), a framework that disentangles optimization across time scales. Its core, Horizon-Independent Advantage Normalization (HIAN), separately normalizes advantages from turn-level and session-level rewards before fusion, ensuring balanced gradient contributions from both immediate and long-term objectives to the policy update. Extensive experiments with a high-fidelity user simulator show DuCA outperforms the state-of-the-art GRPO baseline, achieving a 6.82% relative improvement in conversion rate, reducing inter-sentence repetition by 82.28%, and lowering identity detection

rate by 27.35%, indicating a substantial improvement for an industrial sales scenario that effectively balances the dual demands of strategic performance and naturalistic language generation.

[Long] Harnessing Negative Signals: Reinforcement Distillation from Teacher Data for LLM Reasoning ▶ *Underline Video*

Shuyao Xu, Cheng Peng, Jiangxuan Long, Weidi Xu, Wei Chu and Yuan Qi

Recent advances in model distillation show that data from advanced reasoning models can effectively train smaller student models. However, standard practices discard incorrect reasoning traces, valuable, yet underutilized data. This paper addresses the critical question: How can both positive and negative distilled reasoning traces be effectively leveraged to maximize LLM reasoning performance in an offline setting? We employ a two-stage training recipe: first, Supervised Fine-Tuning (SFT) on positive traces, followed by a refinement stage using both positive and negative traces. We find that a simple REINFORCE-style objective, which we term the Reinforcement Distillation (REDI) objective, outperforms established preference optimization methods like DPO and SimPO in this distillation context. Our empirical evaluations demonstrate the effectiveness of this approach. Notably, our Qwen-REDI-1.5B model, trained on just 131k traces from the open Open-R1 dataset, achieves an 83.1% score on MATH-500. Its performance matches that of DeepSeek-R1-Distill-Qwen-1.5B, a model trained on 800k proprietary data. This result showcases the remarkable data efficiency of utilizing previously discarded negative traces.

[Long] HARPO: Hierarchical Agentic Reasoning for User-Aligned Conversational Recommendation ▶ *Underline Video*

Subham Raj, Aman Vaibhav Jha, Mayank Anand and Sriparna Saha

Conversational recommender systems (CRSs) operate under incremental preference revelation, requiring recommendation decisions under uncertainty. While recent LLM-based approaches achieve strong performance on proxy metrics such as Recall@K and BLEU, they often fail to deliver high-quality, user-aligned recommendations in practice, as they optimize intermediate objectives like retrieval accuracy or fluent generation rather than recommendation quality itself. We propose HARPO (Hierarchical Agentic Reasoning with Preference Optimization), an agentic framework that reframes conversational recommendation as a structured decision-making process optimized for multi-dimensional recommendation quality. HARPO integrates (i) hierarchical preference learning that decomposes recommendation quality into interpretable dimensions (relevance, diversity, satisfaction, and engagement) with context-dependent weighting; (ii) deliberative tree-search reasoning guided by a learned value network evaluating candidate paths on predicted quality; and (iii) domain-agnostic reasoning abstractions through Virtual Tool Operations and multi-agent refinement. We evaluate HARPO on ReDial, INSPIRED, and MUSE, demonstrating consistent improvements over strong baselines on recommendation-centric metrics while maintaining competitive response quality.

[Long] HARPO: Hierarchical Agentic Reasoning for User-Aligned Conversational Recommendation ▶ *Underline Video*

Subham Raj, Aman Vaibhav Jha, Mayank Anand and Sriparna Saha

Conversational recommender systems (CRSs) operate under incremental preference revelation, requiring recommendation decisions under uncertainty. While recent LLM-based approaches achieve strong performance on proxy metrics such as Recall@K and BLEU, they often fail to deliver high-quality, user-aligned recommendations in practice, as they optimize intermediate objectives like retrieval accuracy or fluent generation rather than recommendation quality itself. We propose HARPO (Hierarchical Agentic Reasoning with Preference Optimization), an agentic framework that reframes conversational recommendation as a structured decision-making process optimized for multi-dimensional recommendation quality. HARPO integrates (i) hierarchical preference learning that decomposes recommendation quality into interpretable dimensions (relevance, diversity, satisfaction, and engagement) with context-dependent weighting; (ii) deliberative tree-search reasoning guided by a learned value network evaluating candidate paths on predicted quality; and (iii) domain-agnostic reasoning abstractions through Virtual Tool Operations and multi-agent refinement. We evaluate HARPO on ReDial, INSPIRED, and MUSE, demonstrating consistent improvements over strong baselines on recommendation-centric metrics while maintaining competitive response quality.

[Demo] Hindsight: Structured Agent Memory that Retains, Recalls, and Reflects ▶ *Underline Video*

Christopher Latimer, Nicolò Boschi, Andrew Neeser, Chris Bartholomew, Gaurav Srivastava, Xuan Wang and Naren Ramakrishnan

We demonstrate Hindsight, a working memory system for AI agents that organizes long-term memory into four logical networks and exposes three core operations. The world, experience, observation, and opinion networks separate objective facts from subjective beliefs, giving developers clear visibility into what an agent knows versus what it believes. The retain, recall, and reflect operations handle memory ingestion, retrieval, and reasoning respectively. The retrieval layer combines vector search, keyword matching, graph traversal, and temporal filtering in a parallel pipeline backed by PostgreSQL with pgvector. Unlike existing memory systems such as MemGPT, Zep, and Mem0, Hindsight is the only system that jointly provides fact-belief separation, temporal entity graphs, evolving opinions with confidence scores, and configurable behavioral profiles. On the LongMemEval and LoCoMo benchmarks, Hindsight with a 20B open-source model achieves 83.6% and 83.2% accuracy, outperforming full-context GPT-4o and all prior memory systems; with Gemini-3 Pro, accuracy reaches 91.4% on LongMemEval. Our interactive demonstration lets users build memory graphs through multi-session conversations, inspect how memories are classified across networks, and observe how agent opinions form and change over time. The system is open-source under the MIT license and available as a Python package (pip install hindsight-all) and Docker image, with 1.9K GitHub stars and 191 forks till date.A video demo is at https://youtu.be/4M2wS-yEmVA.

[Findings] HiRAS: A Hierarchical Multi-Agent Framework for Paper-to-Code Generation and Execution ▶ *Underline Video*

Hanhua Hong, Yizhi Li, Jiaoyan Chen, Sophia Ananiadou, Xiaoli Li, Jung-jae Kim and Chenghua Lin;

Recent advances in large language models have highlighted their potential to automate computational research, particularly reproducing

experimental results. However, existing approaches still use fixed sequential agent pipelines with weak global coordination, which limits their robustness and overall performance. In this work, we propose Hierarchical Research Agent System (HiRAS), a hierarchical multi-agent framework for end-to-end paper reproduction that employs supervisory manager agents to coordinate specialised agents across fine-grained stages. We also identify limitations in the reference-free evaluation of the Paper2Code benchmark and introduce Paper2Code-Extra (P2C-Ex), a refined protocol that incorporates repository-level information and better aligns with the original reference-based metric. We conduct extensive evaluation, validating the effectiveness and robustness of our proposed methods, and observing improvements, including >10% relative performance gain above the previous state-of-the-art using open-source backbone models and significantly reduced hallucination in the evaluation. All code and data will be made publicly available.

[Long] HiSVD: Principled Low-Rank Approximation of LLMs via Hierarchical Modeling of Information Capacity and Spectral Structure ▶ *Underline Video*

Zhuo Chen, Minghao Li, Xiaoqian Ma, Siqi Fan, Xiusheng Huang, ZhangLiujie ZhangLiujie and Weihang Chen

Singular Value Decomposition (SVD) enables hardware-agnostic LLM compression via low-rank approximation, yet optimal rank allocation remains a bottleneck. Existing methods predominantly derive layer importance from performance-oriented proxies. Yet, these metrics fail to distinguish between representational importance and structural compressibility, consequently obscuring the fine-grained influence of spectral distribution shape. We demonstrate this disconnect through spectral analysis, revealing that layers with similar information capacity can exhibit markedly different singular value decay behaviors, corresponding to varying degrees of redundancy in the spectral tail. This imperfect coupling implies that allocation strategies driven solely by importance leave significant compression opportunities underexploited. To address this gap, we propose HiSVD, a hierarchical rank allocation framework with two stages: (1) Capacity-Anchored Baseline Allocation, which preserves representational stability by aligning rank budgets with information capacity; and (2) Redundancy-Aware Refinement, which modulates this baseline using tail redundancy to penalize structural excess. Experiments on LLMs demonstrate that HiSVD achieves superior compression efficiency, significantly outperforming state-of-the-art baselines by effectively exploiting this spectral heterogeneity.

[Long] How Long Reasoning Chains Influence LLMs Judgment of Answer Factuality ▶ *Underline Video*

Minzhu Tu, Shiyu Ni and Keping Bi

Large language models (LLMs) are increasingly adopted as scalable judges for open-ended generation, yet how they form judgments remains insufficiently understood. Meanwhile, modern LLMs frequently produce answers accompanied by explicit reasoning, making reasoning chains a natural but understudied source of information for model-based evaluation. This work takes a first step toward understanding how exposing reasoning influences LLM-based judgment. Empirical results across factual question-answering (QA) and mathematical datasets show that the presence of reasoning substantially alters judgment behavior, with clear differences across judge capabilities. Weaker judges become more likely to accept incorrect answers when reasoning is present, suggesting over-reliance on persuasive explanations. In contrast, stronger judges exhibit more selective behavior and, in some cases, achieve higher judgment accuracy by leveraging reasoning content. Further analysis reveals that both reasoning fluency and factuality critically shape judgment outcomes. Together, these findings suggest that examining how models interpret reasoning is essential for understanding and improving LLM-based evaluation, with broader implications for the design of reliable automatic judges and evaluation protocols.

[Long] If an LLM Were a Character, Would It Know Its Own Story? Evaluating Lifelong Learning in LLMs ▶ *Underline Video*

Siqi Fan, Xiusheng Huang, Yiqun Yao, Xuezhi Fang, Kang Liu, Peng Han, Shuo Shang, Aixin Sun and Yequan Wang

Large language models (LLMs) can carry out human-like dialogue, but unlike humans, they are stateless due to the superposition property. However, during multi-turn, multi-agent interactions, LLMs begin to exhibit consistent, character-like behaviors, hinting at a form of emergent lifelong learning. Despite this, existing benchmarks often fail to capture these dynamics, primarily focusing on static, open-ended evaluations. To address this gap, we introduce LifeState-BENCH, a benchmark designed to assess lifelong learning in LLMs. It features two episodic datasets, Hamlet and a synthetic script collection, rich in narrative structure and character interactions. Our fact-checking evaluation probes models self-awareness, episodic memory retrieval, and relationship tracking, across both parametric and non-parametric approaches. Experiments on models like Llama3.1-8B, GPT-4-turbo, and DeepSeek R1, we demonstrate that non-parametric methods significantly outperform parametric ones in managing stateful learning. However, all models exhibit challenges with catastrophic forgetting as interactions extend, highlighting the need for further advancements in lifelong learning.

[Industry] Intent-Driven Semantic ID Generation for Grounded Conversational News Recommendation ▶ *Underline Video*

Hongyang Su, Beibei Kong, Lei Cheng, Chengxiang Zhuo, Zang Li and Chenyun Yu

Conversational news recommendation requires grounding each suggestion in a rapidly evolving article corpus while addressing implicit user intents that lack explicit retrievable keywords. To characterize this scenario, we identify 6 intent types from production dialogues: five are implicit and pose fundamental challenges to standard RAG pipelines, forming a critical retrieve-first bottleneck. To address these issues, we introduce intent-driven Semantic ID (SID) generation under a Generate-then-Match paradigm. With two-stage training that consists of multi-task SID alignment and GPT-4 Chain-of-Thought distillation, an LLM maps diverse intents to hierarchical SID prefixes, which are then fuzzy-matched to the current news pool to guarantee fully grounded recommendations. Profile-Aware Dual-Signal Reasoning (PADR) further enables cold-start users to obtain valid recommendations using only profiles. On a mainstream Chinese news platform, our 7B model achieves 0% hallucination and 12.4% L1 match in the 152K open-generation SID space (4 x random baseline). It matches GPT-4+Hybrid RAG on L1 while surpassing it on finer-grained metrics (L2 2 x, Category +1.2pp) at 100 x lower cost. Cold-start users, where existing

baselines score 0%, achieve 18.0% L1 (6 x random), the highest among all user groups.

[Long] It's High Time: A Survey of Temporal Question Answering ▶ *Underline Video*

Bhawna Piryani, Abdelrahman Abdallah, Jamshid Mozafari, Avishek Anand and Adam Jatowt

Time plays a critical role in how information is generated, retrieved, and interpreted. In this survey, we provide a comprehensive overview of Temporal Question Answering (TQA), a research area that focuses on answering questions involving temporal constraints or context. As time-stamped content from sources like news articles, web archives, and knowledge bases continues to grow, TQA systems must address challenges such as detecting temporal intent, normalizing time expressions, ordering events, and reasoning over evolving or ambiguous facts. We organize existing work through a unified perspective that captures the interaction between corpus temporality, question temporality, and model capabilities, enabling a systematic comparison of datasets, tasks, and approaches. We review recent advances in TQA enabled by neural architectures, especially transformer-based models and Large Language Models (LLMs), highlighting progress in temporal language modeling, retrieval-augmented generation (RAG), and temporal reasoning. We also discuss benchmark datasets and evaluation strategies designed to test temporal robustness, recency awareness, and generalization.

[Findings] Jamendo-MT-QA: A Benchmark for Multi-Track Comparative Music Question Answering ▶ *Underline Video*

Junyoung Koh, Jaeyun Lee, Sooyong Kim, Gyehyeong Choi, Jungin Koh, Jordan Phillips, Yeonjin Lee and Min Song

Recent work on music question answering (Music-QA) has primarily focused on single-track understanding, where models answer questions about an individual audio clip using its tags, captions, or metadata. However, listeners often describe music in comparative terms, and existing benchmarks do not systematically evaluate reasoning across multiple tracks. Building on the Jamendo-QA dataset, we introduce Jamendo-MT-QA, a dataset and benchmark for multi-track comparative question answering. From Creative Commons-licensed tracks on Jamendo, we construct 36,519 comparative QA items over 12,173 track pairs, with each pair yielding three question types: yes/no, short-answer, and sentence-level questions. We describe an LLM-assisted pipeline for generating and filtering comparative questions, and benchmark representative audio-language models using both automatic metrics and LLM-as-a-Judge evaluation.

[Findings] Know Thy Enemy: Securing LLMs Against Prompt Injection via Diverse Data Synthesis and Instruction-Level Chain-of-Thought Learning ▶ *Underline Video*

Zhiyuan Chang, Mingyang Li, Yuekai Huang, Ziyou Jiang, Xiaojun Jia, Qian Xiong, Junjie Wang, Zhaoyang Li and Qing Wang

Large language model (LLM)-integrated applications have become increasingly prevalent, yet face critical security vulnerabilities from prompt injection (PI) attacks. Defending against PI attacks faces two major issues: malicious instructions can be injected through diverse vectors, and injected instructions often lack clear semantic boundaries from the surrounding context, making them difficult to identify. To address these issues, we propose InstruCoT, a model enhancement method for PI defense that synthesizes diverse training data and employs instruction-level chain-of-thought fine-tuning, enabling LLMs to effectively identify and reject malicious instructions regardless of their source or position in the context. We evaluate InstruCoT across three critical dimensions: Behavior Deviation, Privacy Leakage, and Harmful Output. Experimental results across four LLMs demonstrate that InstruCoT significantly outperforms baselines in all dimensions while maintaining utility performance without degradation.

[Long] Language Acquisition Device in Large Language Models ▶ *Underline Video*

Masato Mita, Taiga Someya, Ryo Yoshida and Yohei Oseki

Large Language Models (LLMs) remain substantially less data-efficient than humans. Pre-pretraining (PPT) on synthetic languages has been proposed to close this gap, with prior work emphasizing highly expressive formal languages such as k-Shuffle Dyck. Inspired by the Language Acquisition Device (LAD) hypothesis, which posits that innate constraints preemptively restrict the learner's hypothesis space to natural-language-like structure, we propose LAD-inspired PPT: pre-pretraining on MP-STRUCT, a formal language whose strings encode hierarchical composition, feature-based dependencies, and long-distance displacement via MERGE, AGREE, and MOVE. A brief 500-step PPT with MP-STRUCT matches strong formal-language baselines in token efficiency while additionally imparting a human-like resistance to structurally implausible languages. Analyzing simplified variants, we find that MP-STRUCT CORE outperforms k-Shuffle Dyck despite not being definable in C-RASP (a formal bound on transformer expressivity), challenging the prior hypothesis that effective PPT languages must be both hierarchically expressive and circuit-theoretically learnable. We show that functional landmarks, which reduce dependency resolution ambiguity, are a key driver, suggesting that effective PPT design depends not only on expressivity but also on the accessibility of dependency resolution.

[Long] Layer-Wise High-Impact Parameter Ratio Optimization in Post-Training Quantization for Large Language Models ▶ *Underline Video*

Cuong Pham, Anh Dung Hoang, Cuong C. Nguyen, Trung Le, Gustavo Carneiro and Thanh-Toan Do

Large language models (LLMs) have advanced natural language processing, but their massive parameter counts create computational and memory challenges during deployment. Post-training quantization (PTQ) has emerged as a promising approach to mitigate these challenges. While existing PTQ methods can effectively quantize LLMs, they experience substantial accuracy loss at extremely low bit-widths due to high-impact parameters. Several approaches address this by retaining high-impact parameters in FP16 format, but they apply fixed ratios across all layers, overlooking layer-wise sensitivity variations. We propose a quadratic optimization framework that determines layer-specific ratios of high-impact parameters while considering inter-layer dependencies. We quantize high-impact parameters to moderate bit-widths

while the remaining parameters are quantized to extremely low bit-widths. Under the same resource budget, this preserves more high-impact parameters than methods retaining a few in FP16 format. Our framework enables leveraging advanced quantization methods for high-impact parameters while applying lightweight computational quantization methods to the rest, achieving an effective balance between computational efficiency and accuracy during quantization process.

[SRW] Leakage-Aware User-Level ADHD Signal Classification from Social Media: When Graph Aggregation Helps, and When It Does Not ► *Underline Video*

Daniel Cielak; Wadysaw redniawa;

User-level ADHD-related text classification from social media is methodologically challenging because predictions must aggregate many short posts, performance can be inflated by direct diagnostic leakage, and screening-adjacent settings require calibrated probabilities rather than discrimination alone. We introduce a leakage-aware evaluation framework organized around two controlled axes: evidence budget, i.e., the number of tweets available per user, and leakage control. Within this setup, we compare document-level transformers, strong non-graph embedding-pooling baselines, and heterogeneous graph models combining semantic tweet embeddings, psycholinguistic features, and temporal structure. The main result is regime-dependent: graph aggregation is most useful when user evidence is scarce, whereas simple embedding pooling becomes highly competitive and often slightly stronger as more evidence becomes available. Overall, the main contribution is a controlled benchmarking framework and a clearer account of when graph-based aggregation is actually beneficial.

[Long] Learn Like Humans: Use Meta-cognitive Reflection for Efficient Self-Improvement ► *Underline Video*

Xinmeng Hou, Bohao Qu, Wuqi Wang, Peiliang Gong, Qing Guo and Yang Liu

While Large Language Models (LLMs) enable complex autonomous behavior, current agents remain constrained by static, human-designed prompts that limit adaptability. Existing self-improving frameworks attempt to bridge this gap but typically rely on inefficient, multi-turn recursive loops that incur high computational costs. To address this, we propose Metacognitive Agent with Reflective Self-improvement (MARS), a framework that achieves efficient self-evolution within a single recurrence cycle. Inspired by educational psychology, MARS mimics human learning by integrating principle-based reflection (abstracting normative rules to avoid errors) and procedural reflection (deriving step-by-step strategies for success). By synthesizing these insights into optimized instructions, MARS allows agents to systematically refine their reasoning logic without continuous online feedback. Extensive experiments on six benchmarks demonstrate that MARS outperforms state-of-the-art self-evolving systems while significantly reducing computational overhead. Code is available at https://github.com/Paparare/MARS/tree/main

[Long] LearnerCoMPASS: Intelligent Tutoring System with Dynamic Cognitive Diagnosis and Multi-Model Path Planning ► *Underline Video*

Ziji Sheng, Guiyao Tie, Weidong Wang, Pan Zhou and Daizong Liu

Existing adaptive learning systems struggle to simultaneously achieve deep personalization, dynamic adaptability, and content trustworthiness, particularly in logically rigorous STEM fields where Large Language Models (LLMs) are prone to "hallucination". This paper introduces LearnerCoMPASS (Cognitive Multi-model Planning & Adaptive System), an integrated, end-to-end framework for adaptive learning. At its core, the framework features a novel multi-model path planning algorithm that orchestrates and fuses the outputs of heterogeneous LLM experts to generate and optimize learning sequences. To enable deep personalization, we design a dynamic cognitive diagnosis module that employs an innovative encoder-decoder architecture to generate precise, multi-dimensional cognitive state vectors for learners. To ensure trustworthiness, the system leverages an adaptively constructed dynamic knowledge graph and a Graph-RAG mechanism to provide factual anchors and logical constraints for LLM reasoning, thereby mitigating hallucinations. Extensive experiments demonstrate that LearnerCoMPASS significantly outperforms state-of-the-art baselines in generating high-quality personalized learning paths. Furthermore, ablation studies validate the critical contributions of our dynamic cognitive diagnosis and multi-model planning components.

[Long] Learning from Near-Misses: Error-Aware Contrastive Few-Shot Learning for NL2Formula ► *Underline Video*

Zhihao Shuai, Yiyun Chen, Maolin Ma, Yutong Chen, Hanjia Qiu, Jing Xu, Ziye Chen and Weikai Yang

Natural Language to Excel Formula (NL2Formula) translates user intent into executable spreadsheet formulas. However, current models often produce near-miss outputs, formulas that parse correctly yet fail at execution due to an incorrect function, operator, or reference. Through a systematic error analysis, we find that these errors repeatedly arise from a small set of structural decision points, motivating the need for typed error supervision rather than general error signals. To this end, we introduce an abstract syntax tree (AST)-based error taxonomy that organizes common error modes by the kind of decision that goes wrong in the parse tree. Building on this taxonomy, we propose Error-Aware Contrastive Few-Shot Learning (ECFL), an error-aware framework that unifies training and inference around typed error supervision. During offline training, ECFL mines near-miss errors, assigns error types under the taxonomy, and builds error-aware contrastive demonstrations for fine-tuning. During online inference, a lightweight predictor estimates likely error types and triggers targeted retrieval of contrastive demonstrations to guide single-pass decoding. Experiments show ECFL improves Exact Match (EM) by 6.4 points over supervised fine-tuning (SFT) and matches self-consistency (SC@5) accuracy at substantially lower inference cost.

[Industry] Learning Selective LLM Autonomy from Copilot Feedback in Enterprise Customer Support Workflows ► *Underline Video*

Nikita Borovkov, Elisei Rykov, Olga Tsymboi, Sergei Filimonov, Nikita Surnachev, Dmitry Bitman and Anatolii Potapov

We present a deployed system that automates end-to-end customer support workflows inside an enterprise Business Process Management (BPM) platform. The approach is scalable in production and reaches selective automation within two weeks for a new process, leveraging supervision already generated at scale: structured per-case UI interaction traces and low-overhead copilot feedback, where operators either accept a suggestion or provide a correction. A staged deployment pipeline trains a next UI action policy, learns a critic from copilot feedback to calibrate abstention, and executes only high-confidence steps in the background while deferring uncertain decisions to operators and resuming from the updated UI state. This setup lets one operator supervise multiple concurrent sessions and be interrupted only when the system is uncertain. The system operates on a schema-driven view of the BPM interface and includes monitoring and safe fallbacks for production. In production, it automated 45% of sessions and reduced average handling time by 39% without degrading support quality level.

[Long] LegalChainReasoner: Grounding Criminal Judicial Opinion Generation via Structured Legal Chains ▶ *Underline Video*

Weizhe Shi, Qiqi Wang, Yihong Pan, Qian Liu and Kaiqi Zhao

A criminal judicial opinion represents the judge's disposition of a case, including the decision rationale and sentencing. Automatically generating such opinions can assist in analyzing sentencing consistency and provide judges with references to past similar cases. However, current research typically approaches this task by dividing it into two isolated subtasks: legal reasoning and sentencing prediction. This separation often leads to inconsistency between the reasoning and predictions, failing to meet real-world judicial requirements. Furthermore, prior studies rely on manually creating knowledge to enhance applicability, yet such methods remain limited in practical deployment. To address these limitations and better align with legal practice, we propose a new LegalAI task: Criminal Judicial Opinion Generation, which simultaneously produces both legal reasoning and sentencing decisions. To achieve this, we introduce LegalChainReasoner framework that applies structured legal chains to guide the model through comprehensive case assessments. By integrating factual premises, composite legal conditions, and sentencing conclusions, our approach ensures flexible knowledge injection and end-to-end opinion generation. Experiments on real-world, open-source Chinese legal case datasets demonstrate that our method outperforms baseline models.

[Long] LiGen: Active Lipid Generation via a Molecular Language Model ▶ *Underline Video*

Ying Zhan, Xiuqi Tang, Yan Zhang, Xiao Tan, Dian Shen, Zhou Yu and Beilun Wang

Lipid nanoparticles (LNPs) can deliver cargos to both tumor and immune cells, playing a crucial role in biomedicine. Traditional approaches rely on experimental screening and expert knowledge, which can be costly and time-consuming. Recent methods based on language models have accelerated this process using deep learning. Although these methods can retrieve molecules for fusion or rank candidates from existing libraries, they are still limited by the scope of known formulations. In this work, we propose a method, LiGen, to generate lipid molecules efficiently and actively, facilitating the discovery of high-performing LNP formulations. We first train a lipid-specific molecular language model, LiCore, to learn hidden representations of lipid molecules. We then explore the learned latent space to generate improved candidate formulations. This process is guided by a trained predictor, which evaluates delivery efficiency and provides directional signals. In reconstruction tasks, LiCore achieves nearly perfect reconstruction output with a low invalid ratio on both the LNP-Virtual900k and LNP-Exp12k datasets. The predictor consistently improves ranking-oriented metrics across multiple cell lines, with our method outperforming the best baselines by an average of 4.1%, 10.8%, and 8.1% in Top-50, Top-10, and Top-5 identification accuracy, respectively. Guided by the predictor, LiGen generates novel lipid candidates that achieve a 30.7% improvement over baseline methods on average, with some samples exceeding 50% improvement.

[Long] LLM Beliefs Are in Their Heads ▶ *Underline Video*

Alessandro Corona Mendozza and Anders Søgaard

We investigate belief-like representations in decoder-only autoregressive LLMs using linear controlled probes on residual stream activations and single attention heads. Following Herrmann and Levinsteins (2025) criteria (Accuracy, Use, Coherence, and Uniformity) we find that large models exhibit strong truth sensitivity (Accuracy), and steering activations along probe directions reliably changes downstream behavior (Use). Coherence, measured via calibrated probes and cross-dataset probing, is moderate across models, while training on diverse data yields domain-consistent truth directions (Uniformity). The results are particularly encouraging at the head level and align with some standard philosophical accounts of belief, e.g., minimal functionalism, supporting the view that LLMs can maintain propositional attitudes under such theoretical frameworks.

[Findings] LLM-induced Rationales for More Compact Explainable Style Classification Models ▶ *Underline Video*

Ahmad Aljanaideh and Saeb Ganideh

The complexity of recent natural language classification models led to interest in developing methods for improving the performance of explainable models (e.g. Logistic Regression). Existing methods focus on clustering word embeddings to discover fine-grained contextual features that can be used to train a linear model. While those methods help reduce the gap in performance between black-box models and explainable models, they are based on discovering a large number of features, and this affects interpretability. In this work, we propose a model that leverages Large Language Models (LLMs) and clustering algorithms to discover a compact set of interpretable features. The proposed model first uses GPT-4o mini to extract rationales (i.e. phrases which explain an item's label) from labeled text, and then clusters those rationales to obtain a compact, interpretable feature space. Across 3 Style Classification tasks, the resulting features achieve comparable performance to word-cluster baselines on most tasks, while reducing the number of features by 85-99%. These results highlight the potential of LLMs to improve the compactness of explainable AI models.

[Long] LLM-SLM Collaborative Framework of Idiomatic Expression Generation ▶ *Underline Video*

Hui Gao, Changhao Song, Peng Zhang, Jing Zhang, Chang Yang and Liuxian Ge

Idiomatic Expression Generation, which aims to produce idiomatic text from plain text, is a valuable yet challenging NLP task. However, existing methods suffer from the scarcity of parallel data and dependence on high-quality manual annotations. To address this, we propose an iterative LLM-SLM (Large Language Model-Small Language Model) collaborative framework , Auto-IDEA, that replaces human supervision for idiomatic expression data generation. In this self-improving cycle, the LLM constructs parallel corpora (idiomatic and plain text) via bidirectional semantic reconstruction, automatically generating "Locate-Then-Polish" (LTP) annotations; the SLM filters low-quality corpora while continuously enhancing its verification ability through incremental learning. We instantiate Auto-IDEA for Chinese Idiom Polishing (CIP), constructing CIP-200K, a large-scale dataset of 206K parallel sentences with LTP annotations. The Qwen3-8B fine-tuned on CIP-200K achieves a 25.2% absolute Idiom Polishing Accuracy (IPA) improvement over a supervised fine-tuning (SFT) baseline, outperforming DeepSeek-R1 by 6.2%. Extensive experiments (e.g., Chinese idiom cloze tests and English idiom generation tasks) and human evaluations verify the generalization and effectiveness of Auto-IDEA, demonstrating a new pathway for high-quality, annotation-free data generation through LLM-SLM collaboration.

[Long] LLMs in Sarcasm Detection? It's elementary! (Or is it?) ▶ *Underline Video*

Priyanshu Mahato, Aniket Santosh Mishra and Kripabandhu Ghosh

While Large Language Models (LLMs) are frequently cited for their sophisticated pragmatic reasoning 10.5555/3600270.3602070, Bubeck2023SparksOA, recent progress in sarcasm detection increasingly relies on synthetic benchmarks li25ba_interspeech, anonymous2025deceptive. This study exposes a catastrophic generalization gap in this paradigm: we observe that models achieve near-perfect accuracy on synthetic data but collapse to random guessing on organic human speech. By triangulating hidden state geometry, entropy analysis, and causal interventions, we demonstrate that this disparity stems from shortcut learning Geirhos_2020, models exploit the low-entropy statistical signatures of generated text while remaining "semantically blind" to the pragmatic cues essential for irony. Our findings indicate that high performance on synthetic leaderboards reflects forensic pattern matching rather than the genuine linguistic intelligence assumed in prior work, creating a statistical mirage of competence.

[Short] LLMs Underperform Graph-Based Parsers on Supervised Relation Extraction for Complex Graphs ▶ *Underline Video*

Paolo Gajo, Domenic Rosati, Hassan Sajjad and Alberto Barrón-Cedeño

Relation extraction represents a fundamental component in the process of creating knowledge graphs, among other applications. Large language models (LLMs) have been adopted as a promising tool for relation extraction, both in supervised and in-context learning settings. However, in this work we show that their performance still lags behind much smaller architectures when the linguistic graph underlying a text has great complexity. To demonstrate this, we evaluate four LLMs against a graph-based parser on six relation extraction datasets with sentence graphs of varying sizes and complexities. Our results show that the graph-based parser increasingly outperforms the LLMs, as the number of relations in the input documents increases. This makes the much lighter graph-based parser a superior choice in the presence of complex linguistic graphs.

[Long] Look Less, Reason More: Rollout-Guided Adaptive Pixel-Space Reasoning ▶ *Underline Video*

Xuchen Li, Xuzhao Li, Jiahui Gao, Renjie Pi, Shiyu Hu and Wentao Zhang

Vision-Language Models (VLMs) excel at many multimodal tasks, yet they frequently struggle with tasks requiring precise understanding and handling of fine-grained visual elements. This is mainly due to information loss during image encoding or insufficient attention to critical regions. Recent work has shown promise by incorporating pixel-level visual information into the reasoning process, enabling VLMs to access high-resolution visual details during their thought process. However, this pixel-level information is often overused, leading to inefficiency and distraction from irrelevant visual details. To address these challenges, we propose the first framework for adaptive pixel reasoning that dynamically determines necessary pixel-level operations based on the input query. Specifically, we first apply operation-aware supervised fine-tuning to establish baseline competence in textual reasoning and visual operations, then design a novel rollout-guided reinforcement learning framework relying on feedback of the model's own responses, which enables the VLM to determine when pixel operations should be invoked based on query difficulty. Experiments on extensive multimodal reasoning benchmarks show that our model achieves superior performance while significantly reducing unnecessary visual operations. Impressively, our model achieves 73.4% accuracy on HR-Bench 4K while maintaining a tool usage ratio of only 20.1%, improving accuracy and simultaneously reducing tool usage by 66.5% compared to the previous methods.

[Long] Look Light, Think Heavy: What Multimodal Chain-of-Thought Reasoning Can and Cannot Do ▶ *Underline Video*

Zhuoran Jin, Kejian Zhu, Hongbang Yuan, Yupu Hao, Pengfei Cao, Yubo Chen, Kang Liu and Jun Zhao

Chain-of-Thought (CoT) has become a standard method for improving reasoning capabilities in large language models (LLMs) by eliciting step-by-step thinking, but its effectiveness in multimodal tasks remains unclear. In this paper, we aim to systematically investigate the key question: What can multimodal Chain-of-Thought reasoning do, and where and why does it fall short? To this end, we evaluate 12 multimodal tasks across perception and reasoning categories using both 14 non-reasoning models and 8 reasoning models. Our analysis reveals several important findings: (1) CoT is not a free lunch and should be used selectively depending on the specific requirements of each task. For perception tasks, CoT can lead to undesirable side effects, such as reduced performance in visual grounding and object counting. In contrast, it proves effective for reasoning tasks involving mathematical, scientific, and

multi-image reasoning; (2) Compared to original models, existing open-source multimodal reasoning models often yield only marginal overall improvements, possibly due to an overemphasis on mathematical reasoning at the expense of broader capabilities; (3) Visual reasoning remains a key bottleneck for current multimodal CoT, as models exhibit a Look Light, Think Heavy" pattern where verbal reflection rises and falls during reasoning, whereas visual reflection consistently diminishes. These findings suggest that while multimodal CoT handles verbal reflection relatively well, it lacks the ability to maintain deep visual introspection throughout the reasoning process.

[Long] Look Within or Beyond? A Theoretical Comparison Between Parameter-Efficient and Full Fine-Tuning ▶ *Underline Video*

YongKang Liu, Xingle Xu, Ercong Nie, Zijing Wang, Shi Feng, Daling Wang, Qian Li and Hinrich Schuetze

Parameter-Efficient Fine-Tuning (PEFT) has become a popular alternative to Full-Parameter Fine-Tuning (FFT), achieving similar performance on many benchmarks with far lower computational and memory costs. Yet, its effectiveness on complex tasks such as reasoning and instruction-following remains unclear. In this work, we provide a theoretical and empirical comparison of PEFT and FFT in terms of representational capacity and robustness. We show that PEFTs solution space is a strict subset of FFTs and derive upper bounds revealing how its restricted parameterization limits expressiveness and increases vulnerability to perturbations. Experiments on 20 datasets and 11 adversarial test sets support these findings, indicating that while PEFT performs well on standard tasks, its weaknesses on complex and adversarial settings call for new directions beyond current PEFT paradigms.The source code is in the anonymous GitHub repository.

[Long] MAESTRO: Meta-learning Adaptive Estimation of Scalarization Trade-offs for Reward Optimization ▶ *Underline Video*

Yang Zhao, Hepeng Wang, Xiao Ding, Yangou Ouyang, Bibo Cai, Kai Xiong, Jinglong Gao, Zhouhao Sun, Li Du, Bing Qin and Ting Liu

Group-Relative Policy Optimization (GRPO) has emerged as an efficient paradigm for aligning Large Language Models (LLMs), yet its efficacy is primarily confined to domains with verifiable ground truths. Extending GRPO to open-domain settings remains a critical challenge, as unconstrained generation entails multi-faceted and often conflicting objectives, such as creativity versus factuality, where rigid, static reward scalarization is inherently suboptimal. To address this, we propose MAESTRO (Meta-learning Adaptive Estimation of Scalarization Trade-offs for Reward Optimization), which introduces a meta-cognitive orchestration layer that treats reward scalarization as a dynamic latent policy, leveraging the model's terminal hidden states as a semantic bottleneck to perceive task-specific priorities. We formulate this as a contextual bandit problem within a bi-level optimization framework, where a lightweight Conductor network co-evolves with the policy by utilizing group-relative advantages as a meta-reward signal. Across seven benchmarks, MAESTRO consistently outperforms single-reward and static multi-objective baselines, while preserving the efficiency advantages of GRPO, and in some settings even reducing redundant generation.

[Long] MARCH: Multi-Agent Reinforced Check for Hallucination ▶ *Underline Video*

Zhuo Li, Yupeng Zhang, Pengyu Cheng, Jiajun Song, Mengyu Zhou, Hao Li, Shujie Hu, Yu Qin, erchao.zec erchao.zec, xiaoxi jiang and guanjunjiang guanjunjiang

Hallucination remains a critical bottleneck for large language models (LLMs), undermining their reliability in real-world applications, especially in Retrieval-Augmented Generation (RAG) systems. While existing hallucination detection methods employ LLM-as-a-judge to verify LLM outputs against retrieved evidence, they suffer from inherent confirmation bias, where the verifier inadvertently reproduces the errors of the original generation. To address this, we introduce Multi-Agent Reinforced self-Check for Hallucination (MARCH), a framework that enforces rigorous factual alignment by leveraging deliberate information asymmetry. MARCH orchestrates a collaborative pipeline of three specialized agents: a Solver, a Proposer, and a Checker. The Solver generates an initial RAG response, which the Proposer decomposes into claim-level verifiable atomic propositions. Crucially, the Checker validates these propositions against retrieved evidence in isolation, deprived of the Solver's original output. This well-crafted information asymmetry scheme breaks the cycle of self-confirmation bias. By training this pipeline with multi-agent reinforcement learning (MARL), we enable the agents to co-evolve and optimize factual adherence. Extensive experiments across hallucination benchmarks demonstrate that MARCH substantially reduces hallucination rates. Notably, an 8B-parameter LLM equipped with MARCH achieves performance competitive with powerful closed-source models. MARCH paves a scalable path for factual self-improvement of LLMs through co-evolution. The code is at https://github.com/Qwen-Applications/MARCH.

[Long] MathFlow: Enhancing the Perceptual Flow of MLLMs for Visual Mathematical Problems ▶ *Underline Video*

Shuhang Chen, Hangjie Yuan, Yunqiu Xu, Pengwei Liu, Tao Feng, Jun Cen, Zeying Huang and Yi Yang

Despite strong results on many tasks, multimodal large language models (MLLMs) still underperform on visual mathematical problem solving, especially in reliably perceiving and interpreting diagrams. Inspired by human problem-solving, we hypothesize that the ability to extract meaningful information from diagrams is pivotal, as it directly conditions subsequent inference. Hence, we introduce FlowVerse, a comprehensive benchmark that provides a fine-grained evaluation of MLLMs perception and reasoning capabilities. Our preliminary results on FlowVerse reveal that existing MLLMs exhibit substantial limitations when extracting essential information and reasoned properties from diagrams and performing complex reasoning based on these visual inputs. In response, we introduce MathFlow, a modular problem-solving pipeline that decouples perception and inference into distinct stages, thereby optimizing each independently. Given the perceptual limitations observed in current MLLMs, we trained MathFlow-P-7B as a dedicated perception model. Experimental results indicate that MathFlow-P-7B yields substantial performance gains when integrated with various closed-source and open-source inference models. This demonstrates the effectiveness of the MathFlow pipeline and its compatibility with diverse inference frameworks. Project page: https://github.com/MathFlow-zju/MathFlow.

[Findings] Mechanistic Interpretability of Large-Scale Counting in LLMs through a System-2 Strategy ▶ *Underline Video*

Hosein Hasani, Mohammadali Banayeeanzade, Ali Nafisi, Sadegh Mohammadian, Fatemeh Askari, Mobin Bagherian, Amirmohammad Izadi and Mahdieh Soleymani Baghshah

Large language models (LLMs), despite strong performance on complex mathematical problems, exhibit systematic limitations in counting tasks. This issue arises from the architectural limits of transformers, where counting is performed across layers, leading to degraded precision for larger counting problems due to depth constraints. To address this limitation, we propose a simple test-time strategy inspired by System-2 cognitive processes that decomposes large counting tasks into smaller, independent sub-problems that the model can reliably solve. We evaluate this approach using observational and causal mediation analyses to understand the underlying mechanism of this System-2-like strategy. Our mechanistic analysis identifies key components: latent counts are computed and stored in the final item representations of each part, transferred to intermediate steps via dedicated attention heads, and aggregated in the final stage to produce the total count. Experimental results demonstrate that this strategy enables LLMs to surpass architectural limitations and achieve higher accuracy on large-scale counting tasks. This work provides mechanistic insight into System-2 counting in LLMs and presents a generalizable approach for improving and understanding their reasoning behavior.

[Long] MemRec: Collaborative Memory-Augmented Agentic Recommender System ▶ *Underline Video*

Weixin Chen, Yuhan Zhao, Jingyuan Huang, Zihe Ye, Mingxuan Ju, Tong Zhao, Neil Shah, Li Chen and Yongfeng Zhang

The evolution of recommender systems has shifted from traditional collaborative filtering to LLM-based agentic systems, which rely on semantic user and item memories to make predictions. However, existing agents maintain these memories in isolation. This overlooks crucial collaborative signals, such as user-item co-engagements and peer relationships across the community, which significantly limits their ability to uncover hidden preferences and accurately infer user needs, particularly for data-sparse users. To bridge this gap, we introduce collaborative memory, a paradigm that connects isolated semantics to enable the sharing of relational insights. Yet, naively utilizing collaborative memory causes severe context overload and introduces noise to downstream LLMs, alongside prohibitive computational costs. To resolve this, we propose MemRec, a framework that architecturally decouples memory management from reasoning. MemRec introduces a dedicated, lightweight language model LM_Mem to efficiently manage and synthesize a dynamic collaborative memory graph in the background. It provides only distilled, high-signal contexts to a downstream, heavyweight large language model (LLM_Rec) for the final recommendation. Extensive experiments on four benchmarks demonstrate that MemRec achieves state-of-the-art performance. Code: https://github.com/rutgerswiselab/memrec Homepage: https://memrec.weixinchen.com

[Long] Metaphor Reasoning is Meta-reasoning ▶ *Underline Video*

Qianyu He, Junting Lu, Yikai Zhang, Siyu Yuan, Xiaojun Meng, Jiansheng Wei, Jiaqing Liang and Yanghua Xiao

Metaphor reasoning is an essential cognitive ability that maps knowledge from familiar domains to more abstract domains. This ability functions as a meta-ability underlying many types of reasoning. However, existing work rarely investigates how metaphor reasoning affects other reasoning abilities. To bridge this gap, we systematically study how metaphor reasoning, particularly through metaphorical riddles, can enhance broader reasoning abilities in large language models. We propose MetaR, an automated system for synthesizing metaphorical riddles that satisfy five quality dimensions: diverse, balanced, reasoning-oriented, challenging, and verifiable. Leveraging that answer categories determine riddle categories, we employ a hierarchical answer taxonomy for the former three criteria and a multi-agent refinement framework for the latter two, generating a high-quality dataset. Training with reinforcement learning on verifiable rewards using only thousands of metaphorical riddles, we demonstrate improvements across six out-of-distribution reasoning domains. Analysis reveals transfer effectiveness depends on model scale and pattern-target domain alignment. The datasets and code are publicly available at https://github.com/Abbey4799/MetaR.

[Long] MicroC-KT: Modeling Community Effect via Learning Micro-Environment for Evidence-Grounded Explainable Knowledge Tracing ▶ *Underline Video*

Zhiyi Duan, Zixing Shi, Bing Jia and Qi Wang

Knowledge Tracing (KT) is essential for tracking students evolving knowledge states and predicting their future performance. While current graph-based methods focus on exercise-concept relations, they often overlook the inherent group structures among students. Similarly, emerging LLM-based approaches rely on individual histories, lacking the broader context of group references and contrastive evidence. As a result, existing individual-isolation paradigms fail to provide stable predictions and evidence-based explanations. To bridge this gap, we propose Micro-Community Knowledge Tracing (MicroC-KT), a framework that incorporates learning micro-environments to provide social-cognitive anchors for KT. MicroC-KT identifies latent learning communities via hypergraph modeling and generates dual-granular summaries to facilitate community matching and peer retrieval. By extracting contrastive group evidence, the model prompts an LLM to generate both accurate answer predictions and verifiable analysis reports. Experiments on four public datasets demonstrate that MicroC-KT significantly outperforms state-of-the-art baselines in predictive performance while providing more reliable and evidence-based explanations.

[Long] Mind the Pause: Disfluency-Aware Objective Tuning for Multilingual Speech Correction with LLMs ▶ *Underline Video*

Deepak Kumar, Baban Gain and Asif Ekbal

Automatic Speech Recognition (ASR) transcripts often contain disfluencies, such as fillers, repetitions, and false starts, which reduce readability and hinder downstream applications like chatbots and voice assistants. If left unaddressed, such disfluencies can significantly degrade the reliability of downstream systems. Most existing approaches rely on classical models that focus on identifying disfluent

tokens for removal. While this strategy is effective to some extent, it often disrupts grammatical structure and semantic coherence, leading to incomplete or unnatural sentences. Recent literature explored the use of large language models (LLMs); however, these efforts have primarily focused on disfluency detection or data augmentation, rather than performing comprehensive correction. We propose a multilingual correction pipeline where a sequence tagger first marks disfluent tokens, and these signals guide instruction fine-tuning of an LLM to rewrite transcripts into fluent text. To further improve reliability, we add a contrastive learning objective that penalizes the reproduction of disfluent tokens, encouraging the model to preserve grammar and meaning while removing disfluent artifacts. Our experiments across three Indian languages, namely Hindi, Bengali, and Marathi show consistent improvements over strong baselines, including multilingual sequence-to-sequence models. These results highlight that detection-only strategies are insufficient. Combining token-level cues with instruction tuning and contrastive learning provides a practical and scalable solution for multilingual disfluency correction in speech-driven NLP systems. We make the codes publicly available at https://github.com/deepak-kumar-98/Mind-the-Pause.

[Findings] MIPIC: Matryoshka Representation Learning via Self-Distilled Intra-Relational and Progressive Information Chaining ▶ *Underline Video*

Phung Gia Huy, Hai An Vu, Minh-Phuc Truong, Thang Duc Tran, Linh Ngo Van, Thanh Hong Nguyen and Trung Le

Representation learning is fundamental to NLP, but building embeddings that work well at different computational budgets is challenging. Matryoshka Representation Learning (MRL) offers a flexible inference paradigm through nested embeddings; however, learning such structures requires explicit coordination of how information is arranged across embedding dimensionality and model depth. In this work, we propose MIPIC (Matryoshka Representation Learning via Self-Distilled Intra-Relational Alignment and Progressive Information Chaining), a unified training framework designed to produce structurally coherent and semantically compact Matryoshka representations. MIPIC promotes cross-dimensional structural consistency through Self-Distilled Intra-Relational Alignment (SIA), which aligns token-level geometric and attention-driven relations between full and truncated representations using top-k CKA self-distillation. Complementarily, it enables depth-wise semantic consolidation via Progressive Information Chaining (PIC), a scaffolded alignment strategy that incrementally transfers mature task semantics from deeper layers into earlier layers. Extensive experiments on STS, NLI, and classification benchmarks (spanning models from TinyBERT to BGEM3, Qwen3) demonstrate that MIPIC yields Matryoshka representations that are highly competitive across all capacities, with significant performance advantages observed under extreme low-dimensional.

[Long] Mitigating Hallucinations in Large Vision-Language Models without Performance Degradation ▶ *Underline Video*

Xingyu Zhu, Junfeng Fang, Shuo Wang, Beier Zhu, Zhicai Wang, Yonghui Yang and Xiangnan He

Large Vision-Language Models (LVLMs) exhibit powerful generative capabilities but frequently produce hallucinations that compromise output reliability. Fine tuning on annotated data devoid of hallucinations offers the most direct solution, while its high computational cost motivates recent representation-based methods, which focus on mitigating hallucinatory components within hidden representations. Though efficient, we empirically observe that these methods degrade general generation capacity due to incomplete extraction of hallucination components and non-selective parameter updates. To address these limitations, we propose MPD, a dual-stage framework for mitigating hallucinations without performance degradation. Specifically, our MPD relies on two essential factors: (1) semantic-aware component disentanglement to extract pure hallucination components, and (2) interpretable parameter updates that selectively modify parameters most relevant to hallucination. Extensive experiments demonstrate that MPD achieves state-of-the-art performance, reducing hallucinations by 23.4% while maintaining 97.4% of general generative capability as evaluated on LLaVA-Bench and MME, with no additional computational cost.

[Findings] Modeling Multi-Dimensional Cognitive States in Large Language Models under Cognitive Crowding ▶ *Underline Video*

Lin Zhong, Siyu Zhu, Zizhen Yuan, Jinhao Cui, Xinyang Zhao, Lingzhi Wang, Hao Chen and Qing Liao

Modeling human cognitive states is essential for advanced artificial intelligence. Existing Large Language Models (LLMs) mainly address isolated tasks such as emotion analysis or stance detection, and fail to capture interactions among cognitive dimensions defined in psychology, including emotion, thinking style, stance, and intention. To bridge this gap, we construct CognitiveBench, the first benchmark with unified annotations across the above four dimensions. Experiments on CognitiveBench show that although LLMs perform well on single dimension tasks, their performance drops sharply in joint multi-dimensional modeling. Using Gromov-hyperbolicity analysis, we find that CognitiveBench exhibits a strong hierarchical structure. We attribute the performance bottleneck to "Cognitive Crowding", where hierarchical cognitive states require exponential representational space, while the Euclidean space of LLMs grows only polynomially, causing representation overlap and degraded performance. To address this mismatch, we propose HyCoLLM, which models cognitive states in hyperbolic space and aligns LLM representations via Hyperbolic Guided Alignment Tuning. Results show that HyCoLLM substantially improves multi-dimensional cognitive understanding, allowing 8B parameter model to outperform strong baselines, including GPT-4o. Our code is available at https://anonymous.4open.science/r/HycoLLM.

[Long] MT^3: A Synergistic Multi-Task RL Framework for Specializing MLLMs in Text Image Machine Translation ▶ *Underline Video*

Zhaopeng Feng, Yupu Liang, Shaosheng Cao, Jiayuan Su, Jiahan Ren, zhijie Zhou, Wenxuan Huang, Jian Wu and Zuozhu Liu

Text Image Machine Translation (TIMT), the task of translating textual content embedded in images, is critical for applications in accessibility, cross-lingual information access, and real-world document understanding. However, TIMT remains a complex challenge due to the need for accurate optical character recognition (OCR), robust visual-text reasoning, and high-quality translation, often requiring

cascading multi-stage pipelines. Recent advances in large-scale Reinforcement Learning (RL) have improved reasoning in Large Language Models (LLMs) and Multimodal LLMs (MLLMs), but their application to end-to-end TIMT is still underexplored. To bridge this gap, we introduce MT^3, a novel Multi-Task RL framework to specialize MLLMs into end-to-end expert TIMT models. MT^3 adopts a synergistic multi-task optimization paradigm targeting three key sub-skills: text recognition, context-aware reasoning, and translation. It is trained using a novel multi-mixed reward mechanism that provides fine-grained feedback, fostering a controllable and transparent optimization process. Furthermore, to facilitate the evaluation of TIMT in authentic cross-cultural and real-world social media contexts, we introduced XHSPost, the first social media TIMT benchmark. Our MT^3-7B-Zero achieves state-of-the-art results on the latest in-domain MIT-10M benchmark, outperforming strong baselines such as Qwen2.5-VL-72B and InternVL2.5-78B by notable margins across multiple metrics. Additionally, the model shows strong generalization to out-of-distribution language pairs and datasets. In-depth analyses reveal how multi-task synergy, reinforcement learning initialization, curriculum design, and reward formulation contribute to advancing MLLM-driven TIMT.

[Findings] MTO: A Multi-turn Conversational Text-to-OverpassQL Dataset for Enhanced OpenStreetMap Query Generation ▶ *Underline Video*

Haodi Zhang, Xinrui Zhu, Mingze Kong, Zhidan Liu, Tao Fan, Kaishun Wu and Yuanfeng SONG

We propose a comprehensive framework for constructing multi-turn Text-to-OverpassQL dialogue datasets. Under this framework, we introduce the first multi-turn Text-to-OverpassQL dataset built upon the OverpassNL corpus. Our dataset comprises over 7,800 dialogues, each containing 2 to 4 user utterances, resulting in more than 20,000 individual utterances aligned with executable Overpass queries. To generate high-quality multi-turn dialogues, we design a four-stage pipeline. First, we convert Overpass queries into syntax trees using a custom parser developed based on the official OverpassQL grammar. This enables structural manipulation while preserving syntactic and executable validity. Second, we apply a diverse set of tree-editing templates, including both simple keyword-level changes and complex structural decompositions, to produce multiple valid and diverse Overpass queries. Third, we leverage a prompt-based approach to guide large language models in generating context-aware natural language questions, ensuring increasing inter-turn dependency across the dialogue. Finally, we implement a hybrid filtering strategy that combines manual annotation with model-assisted selection to validate alignment and correctness at scale. In addition to presenting the dataset, we evaluate the performance of several mainstream large language models and demonstrate that our end-to-end baseline model achieves competitive results. This work offers a new benchmark for studying executable semantic parsing and contextual understanding in map-based query tasks.

[Findings] MTP-RL: Acceleration of Reinforcement Learning Rollouts with Policy-Aligned Multi-Token Prediction ▶ *Underline Video*

Ke Wang, Aohan Zeng, Zhengxiao Du, Yuxuan Hu, Bohan Zhang, Xinyi Wang, Jie Tang and Jing Zhang

Reinforcement learning (RL) is widely applied to boost the performance of pretrained models, yet its training efficiency is severely constrained by rollout generation. While speculative decoding based on multi-token prediction (MTP) offers a potential acceleration pathway, its widespread adoption is hindered by the absence of MTP in vanilla pretrained models and the rapid degradation of the MTP acceptance length in RL training. To address these issues, this paper proposes MTP-RL, a two-stage framework that pioneers effective training of MTPs in RL and accelerates the rollout phase for diverse models. It involves a pipeline to equip the multi-layer parameter-sharing MTP for all models and an innovative advantage-aware MTP optimization strategy to facilitate policy-aligned training of MTPs. Experiments demonstrate that our method not only achieves stable growth of acceptance length during RL training, but also accelerates RL rollouts, achieving an average 23.1%-55.3% reduction in rollout time compared to baselines.

[Long] Multimodal Large Language Models for Multi-Subject In-Context Image Generation ▶ *Underline Video*

Yucheng Zhou, Dubing Chen, Huan Zheng and Jianbing Shen

Recent advances in text-to-image (T2I) generation have enabled visually coherent image synthesis from descriptions, but generating images containing multiple given subjects remains challenging. As the number of reference identities increases, existing methods often suffer from subject missing and semantic drift. To address this problem, we propose MUSIC, the first MLLM specifically designed for MUlti-Subject In-Context image generation. To overcome the data scarcity, we introduce an automatic and scalable data generation pipeline that eliminates the need for manual annotation. Furthermore, we enhance the model's understanding of multi-subject semantic relationships through a vision chain-of-thought (CoT) mechanism, guiding step-by-step reasoning from subject images to semantics and generation. To mitigate identity entanglement and manage visual complexity, we develop a novel semantics-driven spatial layout planning method and demonstrate its test-time scalability. By incorporating complex subject images during training, we improve the model's capacity for chained reasoning. In addition, we curate MSIC, a new benchmark tailored for multi-subject in-context generation. Experimental results demonstrate that MUSIC significantly surpasses other methods in both multi- and single-subject scenarios.

[Long] N-GLARE: An Non-Generative Latent Representation-Efficient LLM Safety Evaluator ▶ *Underline Video*

Zheyu Lin, Jirui Yang, Yukui Qiu, Hengqi Guo, Yubing Bao and Yao Guan

Evaluating the safety robustness of LLMs is critical for their deployment. However, mainstream Red Teaming methods rely on online generation and black-box output analysis. These approaches are not only costly but also suffer from feedback latency, making them unsuitable for agile diagnostics after training a new model. To address this, we propose N-GLARE (A Non-Generative, Latent Representation-Efficient LLM Safety Evaluator). N-GLARE operates entirely on the model's latent representations, bypassing the need for full text generation. It characterizes hidden layer dynamics by analyzing the APT (Angular-Probabilistic Trajectory) of latent representations and introducing the

JSS (Jensen-Shannon Separability) metric. Experiments on over 40 models and 20 red teaming strategies demonstrate that the JSS metric exhibits high consistency with Red Teaming safety rankings at less than 1% token and runtime cost.

[Long] Native Hybrid Attention for Efficient Sequence Modeling ▶ *Underline Video*
Jusen Du, Jiaxi Hu, Zhang Tao, Weigao Sun and Yu Cheng
Transformers excel at sequence modeling but face quadratic complexity, while linear attention offers improved efficiency but often compromises recall accuracy over long contexts. In this work, we introduce Native Hybrid Attention (NHA), a novel hybrid architecture of linear and full attention that integrates both intra & inter-layer hybridization into a unified layer design. NHA maintains long-term context in key-value slots updated by a linear RNN, and augments them with short-term tokens from a sliding window. A single softmax attention operation is then applied over all keys and values, enabling per-token and per-head context-dependent weighting without requiring additional fusion parameters. The inter-layer behavior is controlled through a single hyperparameter, the sliding window size, which allows smooth adjustment between purely linear and full attention while keeping all layers structurally uniform. Experimental results show that NHA surpasses Transformers and other hybrid baselines on recall-intensive and commonsense reasoning tasks. Furthermore, pretrained LLMs can be structurally hybridized with NHA, achieving competitive accuracy while delivering significant efficiency gains. Code is available at https://github.com/JusenD/NHA.

[Long] ODL-TempLLM: Ontology-Guided and Description Logic-Reasoned Temporal Reasoning with LLMs ▶ *Underline Video*
Jinshuo Liu, Cheng Bi, Meng Wang, Juan Deng, Donghong Ji and Jeff Z. Pan
Temporal reasoning is crucial for large language models (LLMs) to understand event concurrency and complex temporal interactions in natural language. Recent approaches rely on the LLM to infer temporal relations between events and largely overlook the inherent structural nature of temporal relationships. In this work, we propose ODL-TempLLM (Ontology-Guided and Description Logic-Constrained Temporal Reasoning with LLMs), a novel paradigm for temporal reasoning with LLMs that shifts focus from internal inference to the explicit modeling of temporal structure. ODL-TempLLM leverages ontology learning to explicitly construct structured temporal knowledge, employs a symbolic reasoner to deductively reason about temporal relations and uses logic-constrained retrieval augmentation to obtain relevant facts.Experiments results evaluated across there datasets via various LLM backbones show that our method outperforms state-of-the-art methods by 2.07-31.83 F1 points and 1.00-30.73 EM points, exhibiting strong generalization and highlighting the potential of explicit temporal reasoning.

[Findings] OmniCompliance-100K: A Multi-Domain, Rule-Grounded, Real-World Safety Compliance Dataset ▶ *Underline Video*
Wenbin Hu, Huihao JING, Haochen Shi, Changxuan Fan, Haoran Li and Yangqiu Song
Ensuring the safety and compliance of large language models (LLMs) is of paramount importance. However, existing LLM safety datasets often rely on ad-hoc taxonomies for data generation and suffer from a significant shortage of rule-grounded, real-world cases that are essential for robustly protecting LLMs. In this work, we address this critical gap by constructing a comprehensive safety dataset from a compliance perspective. Using a powerful web-searching agent, we collect a rule-grounded, real-world case dataset OmniCompliance-100K, sourced from multi-domain authoritative references. The dataset spans 74 regulations and policies across a wide range of domains, including security and privacy regulations, content safety and user data privacy policies from leading AI companies and social media platforms, financial security requirements, medical device risk management standards, educational integrity guidelines, and protections of fundamental human rights. In total, our dataset contains 12,985 distinct rules and 106,009 associated real-world compliance cases. Our analysis confirms a strong alignment between the rules and their corresponding cases. We further conduct extensive benchmarking experiments to evaluate the safety and compliance capabilities of advanced LLMs across different model scales. Our experiments reveal several interesting findings that have great potential to offer valuable insights for future LLM safety research.

[Long] Open Schrödinger's Closed Box: Identifying Retrieval Augmented Generation in API-Accessible Large Language Model Services ▶ *Underline Video*
Yukun Jiang, Xinyue Shen, Michael Backes, Zheng Li and Yang Zhang
Large language models (LLMs) are powerful at question-answering but prone to hallucinations due to limited domain-specific or up-to-date knowledge. Retrieval augmented generation (RAG) mitigates this by adding an external retriever and knowledge database, yet RAG remains vulnerable to targeted attacks that degrade outputs or manipulate opinions. Prior attacks typically assume adversaries know the service is RAG-enhanced and may even know deployment details, an assumption often invalid for real-world commercial LLMs that expose only black-box APIs. This opacity also risks misleading users about system capabilities. This work aims to bridge this gap by proposing RAG-ID, a framework for IDentifying RAG properties in LLM services. We classify adversaries into three knowledge levels and design six attack methods. Experiments show these attacks reliably detect RAG , up to 99.97% accuracy with partial or no optional knowledge, and nearly 100% when the LLM and database are known. After detection, RAG-ID can infer finer RAG properties (e.g., deployed LLM and knowledge database). We consider RAG-ID a reconnaissance tool for attackers, a way to facilitate users' transparent selection of LLM services, and a guide for RAG developers in refining security measures.

[Findings] Optimizing Native Sparse Attention with Latent Attention and Local Global Alternating Strategies ▶ *Underline Video*
Yuxuan Hu, Jianchao Tan, Jiaqi Zhang, Wen Zan, Pingwei SUN, Yifan Lu, Xunliang Cai and Jing Zhang
In this work, we conduct a systematic analysis of Native Sparse Attention (NSA) and propose targeted improvements that enhance long-context modeling. A key insight is that alternating between local (sliding-window) and global (compression/selective) attention across

layers, rather than using fixed patterns, enables more effective propagation of long-range dependencies and substantially boosts performance on long-sequence tasks. Meanwhile, we further refine NSAs branches with Latent Attention that the sliding-window branch is enhanced with Multi-head Latent Attention (MLA) while compression and selective branches adopt Group-head Latent Attention (GLA). These changes reduce KV-cache memory by 50% versus NSA while improving the model's common-sense reasoning and long-text understanding capabilities. Experiments on models from 340M to 1.3B parameters (trained on 15B and 100B tokens) show our method matches or exceeds full attention and native sparse attention in both common-sense reasoning and long-context understanding tasks.

[Long] PDR: A Plug-and-Play Positional Decay Framework for LLM Pre-training Data Detection ► *Underline Video*

Jinhan Liu, Yibo Yang, Ruiying Lu, Piotr Pikos, Yimeng Chen, Peng Wang and Dandan Guo

Detecting pre-training data in Large Language Models (LLMs) is crucial for auditing data privacy and copyright compliance, yet it remains challenging in black-box, zero-shot settings where computational resources and training data are scarce. While existing likelihood-based methods have shown promise, they typically aggregate token-level scores using uniform weights, thereby neglecting the inherent information-theoretic dynamics of autoregressive generation. In this paper, we hypothesize and empirically validate that memorization signals are heavily skewed towards the high-entropy initial tokens, where model uncertainty is highest, and decay as context accumulates. To leverage this linguistic property, we introduce Positional Decay Reweighting (PDR), a training-free and plug-and-play framework. PDR explicitly reweights token-level scores to amplify distinct signals from early positions while suppressing noise from later ones. Extensive experiments show that PDR acts as a robust prior and can usually enhance a wide range of advanced methods across multiple benchmarks.

[Findings] PERM: Psychology-grounded Empathetic Reward Modeling for Large Language Models ► *Underline Video*

Chengbing Wang, Wuqiang Zheng, Yang Zhang, Fengbin Zhu, Junyi Cheng, Yi Xie, Wenjie Wang and Fuli Feng

Large Language Models (LLMs) are increasingly deployed in human-centric applications, yet they often fail to provide substantive emotional support. While Reinforcement Learning (RL) has been utilized to enhance empathy of LLMs, existing reward models typically evaluate empathy from a single perspective, overlooking the inherently bidirectional interaction nature of empathy between the supporter and seeker as defined by Empathy Cycle theory. To address this limitation, we propose Psychology-grounded Empathetic Reward Modeling (PERM). PERM operationalizes empathy evaluation through a bidirectional decomposition: 1) Supporter perspective, assessing internal resonation and communicative expression; 2) Seeker perspective, evaluating emotional reception. Additionally, it incorporates a bystander perspective to monitor overall interaction quality. Extensive experiments on a widely-used emotional intelligence benchmark and an industrial daily conversation dataset demonstrate that PERM outperforms state-of-the-art baselines by over 10%. Furthermore, a blinded user study reveals a 70% preference for our approach, highlighting its efficacy in generating more empathetic responses.

[Long] PersonalityDBench: A Dataset for Personality Disorders - from Modeling to Controlled Generation ► *Underline Video*

Federico Ravenda, Seyed Ali Bahrainian, Daniele Montagnani, Antonietta Mira and Andrea Raballo

Personality disorders (PDs) are a complex class of mental health (MH) conditions characterized by persistent patterns of cognition, behavior, and emotional regulation that deviate from cultural norms. While social media has become a valuable resource for MH research, NLP has largely focused on more prevalent conditions (e.g., depression), leaving PDs underexplored. In this work, we introduce PersonalityDBench, a large-scale, clinically grounded dataset that supports multidimensional study of personality pathology, and standardized, reproducible evaluation of LLM steering toward clinically grounded behavioral targets. The dataset comprises two parts: (1) PRISMA and (2) PersonaDSteering. (1) PRISMA (PeRsonality dISorder MAnifestations) is a clinically annotated collection of social media content spanning the full spectrum of PDs. It links clinically validated diagnostic criteria and dimensional trait frameworks with computational annotation and analysis methods to support fine-grained, multidimensional study of how PDs manifests in naturalistic, free-form language. Building on PRISMA, (2) PersonaDSteering is a benchmark for LLM steering evaluation that operationalizes clinically grounded PD profiles into structured behavioral elicitation tasks, enabling multidimensional steerability assessment beyond single-behavior settings and supporting PD-consistent persona construction for simulated patient generation. This dataset may have application in the study and modeling of PD and powering personality-specific text generation for adaptive, personalized chat systems.

[Findings] Personalized Benchmarking: Evaluating LLMs by Individual Preferences ► *Underline Video*

Cristina Garbacea, Heran Wang and Chenhao Tan

With the rise in capabilities of large language models (LLMs) and their deployment in real-world tasks, evaluating LLM alignment with human preferences has become an important challenge. Current benchmarks average preferences across all users to compute aggregate ratings, overlooking individual user preferences when establishing model rankings. Since users have varying preferences in different contexts, we call for personalized LLM benchmarks that rank models according to individual needs. We compute personalized model rankings using ELO ratings and Bradley-Terry coefficients for 115 active Chatbot Arena users and analyze how user query characteristics (topics and writing style) relate to LLM ranking variations. We demonstrate that individual rankings of LLM models diverge dramatically from aggregate LLM rankings, with Bradley-Terry correlations averaging only = 0.04 (57% of users show near-zero or negative correlation) and ELO ratings showing moderate correlation (= 0.43). Through topic modeling and style analysis, we find users exhibit substantial heterogeneity in topical interests and communication styles, influencing their model preferences. We further show that a compact combination of topic and style features provides a useful feature space for predicting user-specific model rankings. Our results provide strong quantitative evidence that aggregate benchmarks fail to capture individual preferences for most users, and highlight the importance of developing personalized benchmarks that rank LLM models according to individual user preferences.

[Findings] Preference Optimization for Review Question Generation Improves Writing Quality ▶ *Underline Video*

Karun Sharma, Vidushee Vats, Shengzhi LI, Yuxiang Wang, Zhongtian Sun and Prayag Tiwari

Peer review relies on substantive, evidence-based questions, yet current LLMs generate surface-level queries that perform worse than human reviewer questions in expert evaluation. To address this gap, we curate a high-quality dataset of reviewer questions from OpenReview and conduct a human preference study where expert annotators evaluate question-paper pairs across three dimensions: effort, evidence, and grounding. From these annotations, we train IntelliReward, a reward model built from a frozen autoregressive LLM with trainable multi-head transformers. Validated against expert judgments, IntelliReward predicts reviewer-question quality better than API-based SFT baselines and provides scalable evaluation. We apply Decoupled Clip and Dynamic Sampling Policy Optimization (DAPO) with IntelliReward to train IntelliAsk, a question-generation model aligned with human standards of effortful, evidence-based critique. Human evaluations show IntelliAsk generates more grounded, substantive and effortful questions than strong baselines and reduces reliance on first-page content. We also find improvements on reasoning and writing benchmarks, suggesting reviewer-question quality correlates with broader capabilities. Compared to Qwen3-32B, IntelliAsk improves MuSR (68.3 vs 64.7 Acc) and WritingBench (8.31 vs 8.07). We release our code, filtered review dataset, expert annotations, IntelliAsk and IntelliReward to support automatic evaluation of grounding, effort, and evidence in LLM-generated review questions.

[Long] PRISMA: Preference-Reinforced Self-Training Approach for Interpretable Emotionally Intelligent Negotiation Dialogues ▶ *Underline Video*

Prajwal Vijay Kajare, Priyanshu Priya, Bikash Santra and Asif Ekbal

Emotion plays a pivotal role in shaping negotiation outcomes, influencing trust, cooperation, and long-term relationships. Developing negotiation dialog systems that can recognize and respond strategically to emotions is therefore essential to create more effective human-centered interactions. Beyond generating emotionally appropriate responses, interpretability - understanding how a system generates a particular emotion-aware response, is critical for fostering reliability and building rapport. Driven by these aspects, in this work, we introduce PRISMA, an interpretable emotionally intelligent negotiation dialogue system targeting two application domains, viz. job interviews and resource allocation. To enable interpretability, we propose an Emotion-aware Negotiation Strategy-informed Chain-of-Thought (ENS-CoT) reasoning mechanism, which mimics human negotiation by perceiving, understanding, using, and managing emotions. Leveraging ENS-CoT, we curate two new datasets: JobNego (for job interview negotiation) and ResNego (for resource allocation negotiation). We then leverage these datasets to develop PRISMA by augmenting self-training with Direct Preference Optimization (DPO), guiding agents toward more accurate, interpretable, and emotionally appropriate negotiation responses. Automatic and human evaluations on JobNego and ResNego datasets demonstrate that PRISMA substantially enhances the interpretability and generates appropriate emotion-aware responses, while improving overall negotiation effectiveness.

[Long] Programming over Thinking: Efficient and Robust Multi-Constraint Planning ▶ *Underline Video*

Xin Deik Goh, Quanyu Long, Zhengyuan Liu, Fang Yih Chen and Wenya Wang

Multi-constraint planning involves identifying, evaluating, and refining candidate plans while satisfying multiple, potentially conflicting constraints. Existing large language model (LLM) approaches face fundamental limitations in this domain. Pure reasoning paradigms, which rely on long natural language chains, are prone to inconsistency, error accumulation, and prohibitive cost as constraints compound. Conversely, LLMs combined with coding- or solver-based strategies lack flexibility: they often generate problem-specific code from scratch or depend on fixed solvers, failing to capture generalizable logic across diverse problems. To alleviate these issues, we introduce the Scalable Code Planning Engine (SCOPE), a systematic framework that disentangles query-specific problem reasoning from generic code execution. SCOPE first transforms input queries into optimized structured representations, capturing the interdependent constraints, and then autonomously generates reusable solver functions (Combination, Filter, and Deliver) that provide consistent and reliable execution across diverse problems. SCOPE achieves state-of-the-art performance while lowering cost and latency. For example, with GPT-4o, it reaches 93.1% success on TravelPlanner, a 61.6% gain over the best baseline (CoT) while cutting inference cost by 1.4 times and time by approximately 4.67 times. Code is available at https://github.com/DerrickGXD/SCOPE.

[Findings] ProtoCycle: Reflective Tool-Augmented Planning for Text-Guided Protein Design ▶ *Underline Video*

Yutang Ge, Guojiang Zhao, Sihang Li, Zheng Cheng, Zifeng Zhao, Hanchen Xia, Guolin Ke, Linfeng Zhang, Zhifeng Gao and Yuguang Wang

Designing proteins that satisfy natural language functional requirements is a central goal in protein engineering. A straightforward baseline is to fine-tune generic instruction-tuned LLMs as direct text-to-sequence generators, but this is data- and compute-hungry. With limited supervision, LLMs can produce coherent plans in text yet fail to reliably realize them as sequences. This plan–execute gap motivates ProtoCycle, an agentic framework for protein design that uses LLMs primarily to drive a multi-round, feedback-driven decision cycle. ProtoCycle couples an LLM planner with a lightweight tool environment designed to emulate the iterative workflow of human protein engineers and uses LLM-driven reflection on tool feedback to revise plans. Trained with supervised trajectories and online reinforcement learning, ProtoCycle achieves strong language alignment while maintaining competitive foldability, and ablations show that reflection substantially improves sequence quality.

[Demo] Qayyem: A Real-time Platform for Scoring Proficiency of Arabic Essays ▶ *Underline Video*

Hoor Elbahnasawi, Marwan Sayed, Sohaila Eltanbouly, Fatima Brahamia and Tamer Elsayed

Over the past years, Automated Essay Scoring (AES) systems have gained increasing attention as scalable and consistent solutions for assessing proficiency of student writing. Despite recent progress, support for Arabic AES remains limited due to linguistic complexity and scarcity of large publicly-available annotated datasets. In this work, we present Qayyem, a Web-based platform designed to support Arabic AES by providing an integrated workflow for assignment creation, batch essay upload, scoring configuration, and per-trait essay evaluation. Qayyem abstracts the technical complexity of interacting with scoring server APIs, allowing instructors to access advanced scoring services through a user-friendly interface. The platform deploys a number of state-of-the-art Arabic essay scoring models with different effectiveness and efficiency figures.

[Long] Rectifying the Emotional Flow: Aligning Priors and Dynamic Guidance for High-Arousal Text-to-Speech ▶ *Underline Video*

Fangming Feng, Dongjie Fu, Zequn Xie, Yu Zhang, Yangyang Wu, Zhou Zhao and Tao Jin

While diffusion and flow-matching models have advanced TTS, generating high-arousal emotions remains a persistent challenge due to the trade-off between stability and expressiveness. Existing systems often suffer from linguistic collapse when pursuing high intensity or fail to meet target emotional levels under stable settings. In this work, we identify that standard Gaussian initialization inevitably introduces a neutral prosody bias, while uniform Classifier-Free Guidance often distorts the acoustic manifold, leading to artifacts. To address this, we propose an inference framework that rectifies the emotional trajectory. An Emotion-Rectified Noise Prior injects a semantic gradient at initialization to align sampling with the target emotional manifold, and Likelihood-Inverse Guidance adaptively schedules guidance via a conditional/unconditional likelihood ratio, strengthening guidance only when the trajectory drifts toward a neutral fallback. Extensive experiments demonstrate that our method effectively resolves the stability bottleneck in high-intensity scenarios, achieving superior linguistic accuracy and emotional fidelity without model retraining. Audio samples are available at https://showtts.github.io/emotionTTS/.

[Long] ReEx-SQL: Reasoning with Execution-Aware Reinforcement Learning for Text-to-SQL ▶ *Underline Video*

Yaxun dai, Wenxuan Xie, Xialie Zhuang, Tianyu Yang, Ziyi Liu, Haiqin Yang, Yiying Yang, Yuhang Zhao, Pingfu Chao and Wenhao Jiang

Current Text-to-SQL reasoning models often lack integrated execution feedback during generation, and most existing approaches utilize feedback only for post-hoc correction. This separation not only limits real-time error correction, but may also introduce mistakes by altering otherwise correct SQL queries. To address these challenges, we present ReEx-SQL (Reasoning with Execution-Aware Reinforcement Learning), a Text-to-SQL framework that interacts with the SQL execution engine during decoding and dynamically adjusts reasoning based on execution feedback, thereby enabling context-sensitive query refinement and improved accuracy. ReEx-SQL achieves this through structured prompts with markup tags and a stepwise rollout strategy that incorporates execution feedback at each generation stage. For policy supervision, we design a composite reward function, featuring an exploration reward, to explicitly encourage effective interaction with the database. Furthermore, ReEx-SQL adopts a tree-based decoding strategy to facilitate exploratory reasoning and primarily aims to enhance parallel decoding efficiency. Notably, ReEx-SQL achieves 89.1% accuracy on Spider and 65.3% on BIRD at the 7B scale, surpassing baseline models by 2.7% and 2.6%, respectively. In addition, its tree-based decoding accelerates inference by 51.9% compared to linear decoding during sampling.

[SRW] RegTrack: A Fine-Grained Benchmark for Multi-Class Legal Change Detection ▶ *Underline Video*

Joe Yu, Kevin Chenhao Li and Julian Ostarek

Organizations must continuously monitor evolving regulations to maintain compliance. While current tools are limited to surface-level text comparison, existing models lack the fine-grained classification schemes to determine whether small changes impact legal obligations or merely update formatting. To address this gap, we introduce a novel benchmark for change detection in EU regulations. It comprises 4,772 manually annotated pairs of structurally distinct provisions, defined as Atomic Legal Units (ALUs), mapped to a six-class taxonomy. We formalize three core tasks: structural alignment, change classification, and a combined task requiring simultaneous alignment and classification. Evaluating baselines from lexical algorithms to Large Language Models (LLMs), we find LLMs excel at isolated change classification, whereas domain-specific dense encoders offer the most robust combined performance. By providing fine-grained labeled data, this benchmark enables the development of AI systems that can help organizations analyze regulatory shifts and support version-aware retrieval in the legal domain.

[Long] Reinforcement Learning-Guided Adaptive Tuning for Out-of-Distribution Harmful Text Detection ▶ *Underline Video*

Mengyu Xiang, Tinghao Chen, Boxu Han, Qiudan Li, Shu Wu and Daniel Dajun Zeng

As social media grows, harmful information spreads rapidly across platforms and evolves over time, showing cross-platform and cross-temporal variations. Existing methods rely on fixed model parameters during training, which fail to handle substantial semantic discrepancies, leading to Out-Of-Distribution (OOD) problems. While test-time tuning enables dynamic parameter adjustment, it may lead to excessive adaptation to individual samples. The key challenge is how to adapt to semantic variations during testing while preventing overfitting from continuous tuning. To tackle this issue, this paper proposes RLAT, a reinforcement learning (RL)-guided adaptive tuning method for harmful text detection. First, a tuning joint optimization module is designed to update parameters and adapt to semantic variations during testing. It tunes the model by optimizing consistency loss and applying word-level attention constraints to reduce over-reliance on local words and learn a more robust global representation. Then, to mitigate overfitting caused by continuous tuning, a RL-guided adaptive decision model is introduced to direct the tuning process. It reduces the influence of local samples by selecting data and controlling parameter updates, thereby improving overall test performance. Experimental results show that the RLAT outperforms state-of-the-art baselines in cross-platform and cross-temporal scenarios across multiple public datasets.

[Long] Resonating with RoPE: Spectral Quantization for High-Fidelity Key Cache Compression ► *Underline Video*

Xuefei Wang, Haoyu Tang, Tianyuan Liang, Zhibin Wang, Yupeng Hu and Weili Guan

The linear growth of KV cache bottlenecks long-context LLMs, yet RoPE-induced oscillations complicate Key cache quantization. To address this issue, we propose SpectrumQuant, a frequency-domain framework that utilizes the Discrete Cosine Transform (DCT) to convert these oscillations into sparse spectral representations. Specifically, our pipeline integrates dominant frequency extraction, hybrid bit-width allocation, and high-frequency pre-emphasis to maximize fidelity while minimizing memory footprint. To eliminate computational overhead, we develop fused Triton kernels featuring deferred inverse transformation and on-chip sparse accumulation. Extensive experiments on several benchmarks confirm SpectrumQuant achieves efficient compression with performance and latency comparable to FP16 baselines.

[Long] Response-G1: Explicit Scene Graph Modeling for Proactive Streaming Video Understanding ► *Underline Video*

Ke Ma, Jiaqi Tang, Bin Guo, Xueting Han, Ruonan Xu, Qingfeng He, Ziheng Wang, Xu Wang, Qifeng Chen, Zhiwen Yu and Yunhao Liu

Proactive streaming video understanding requires VideoLLMs to decide when to respond as a video unfolds, a task where existing methods often fall short due to their implicit, queryagnostic modeling of visual evidence. We introduce Response-G1, a novel framework that establishes explicit, structured alignment between the accumulated video evidence and the querys expected response conditions via scene graphs. The framework operates in three finetuningfree stages: (1) online queryguided scene graph generation from streaming clips; (2) memorybased retrieval of the most semantically relevant historical scene graphs; and (3) retrievalaugmented trigger prompting for perframe "silence/response" decisions. By grounding both evidence and conditions in a shared graph representation, Response-G1 achieves more interpretable and accurate response timing decisions. Experimental results on established benchmarks demonstrate the superiority of our method in both proactive and reactive tasks, validating the advantage of explicit scene graph modeling and retrieval in streaming video understanding.

[Short] Rethinking Data Mixing from the Perspective of Large Language Models ► *Underline Video*

Yuanjian Xu, Tianze Sun, Changwei Xu, XinLong Zhao, Jianing Hao, Ran Chen, Yang Liu, Ruijie Xu, Stephen Chen and Guang Zhang

Data mixing strategy is essential for large language model (LLM) training. Empirical evidence shows that inappropriate strategies can significantly reduce generalization. Although recent methods have improved empirical performance, several fundamental questions remain open: what constitutes a domain, whether human and model perceptions of domains are aligned, and how domain weighting influences generalization. We address these questions by establishing formal connections between gradient dynamics and domain distributions, offering a theoretical framework that clarifies the role of domains in training dynamics. Building on this analysis, we introduce DoGraph, a reweighting framework that formulates data scheduling as a graph-constrained optimization problem. Extensive experiments on GPT-2 models of varying scales demonstrate that DoGraph consistently achieves competitive performance.

[Long] Retrieval as Generation: A Unified Framework with Self-Triggered Information Planning ► *Underline Video*

Bo Li, Mingda Wang, Gexiang Fang, Shikun Zhang and Wei Ye

We revisit retrieval-augmented generation (RAG) by embedding retrieval control directly into generation. Instead of treating retrieval as an external intervention, we express retrieval decisions within token-level decoding, enabling end-to-end coordination without additional controllers or classifiers. Under the paradigm of Retrieval as Generation, we propose GRIP (Generation-guided Retrieval with Information Planning), a unified framework in which the model regulates retrieval behavior through control-token emission. Central to GRIP is Self-Triggered Information Planning, which allows the model to decide when to retrieve, how to reformulate queries, and when to terminate, all within a single autoregressive trajectory. This design tightly couples retrieval and reasoning and supports dynamic multi-step inference with on-the-fly evidence integration. To supervise these behaviors, we construct a structured training set covering answerable, partially answerable, and multi-hop queries, each aligned with specific token patterns. Experiments on five QA benchmarks show that GRIP surpasses strong RAG baselines and is competitive with GPT-4o while using substantially fewer parameters. Code and resources are provided in the supplementary materials.

[Findings] Revisiting Self-Play Preference Optimization: On the Role of Prompt Difficulty ► *Underline Video*

Yao Xiao ~Yao_Xiao8, Jung-jae Kim, Roy Ka-Wei Lee and Lidong Bing

Self-play preference optimization has emerged as a prominent paradigm for aligning large language models (LLMs). It typically involves a language model to generate on-policy responses for prompts and a reward model (RM) to guide the selection of chosen and rejected responses, which can be further trained with direct preference optimization (DPO). However, the role of prompts remains underexplored, despite being a core component in this pipeline. In this work, we investigate how prompts of varying difficulty influence self-play preference optimization. We use the mean reward of sampled responses of a prompt as a proxy for its difficulty. We first find that difficult prompts exhibit substantially inferior self-play optimization performance compared to easy prompts for language models. Moreover, incorporating difficult prompts into training fails to enhance overall performance and, in fact, leads to slight degradation compared to training on easy prompts alone. Third, there is a clear upward trend in optimization performance as prompt difficulty decreases. We also observe that the performance gap between difficult and easy prompts tends to close as the model capacity increases, suggesting that prompt difficulty interacts with the model capacity. Building on these findings, we explore strategies to mitigate the adversary effect of difficult prompts on final performance. We demonstrate that only training on a small portion (30%) of the easiest prompts improves overall self-play performance on AlpacaEval~2 and Arena-Hard. We also report failed attempts and lessons learned.

[Long] RST-Guarder: Enhancing Long-Context Robustness for Safeguards via RST Parsing and Probabilistic Inference ► *Underline Video*

Xu Zhang and Xiaojun Wan

As large language models (LLMs) demonstrate remarkable capabilities across a wide range of tasks, ensuring the safety of their outputs is increasingly critical. To mitigate the risk of policy-violating responses, numerous guardrail models have been developed for harmful-content detection. While effective on short outputs, existing guardrails degrade on long-form responses, reflecting limited semantic understanding and weak robustness to contextual noise. To address these limitations, we propose RST-Guarder, an inference-time method that improves harmful-content detection for long-form inputs without additional data curation or model training. RST-Guarder first applies a RST parser to long-form inputs to get discourse-level semantic relations among segments, and subsequently performs hierarchical probabilistic inference to aggregate segment-level safety scores produced by pre-trained guardrail models. We evaluate RST-Guarder across multiple benchmarks and a diverse set of widely used guardrail models. Experimental results demonstrate that RST-Guarder consistently improves harmful-content detection on long-form inputs, while significantly reducing false positives that incorrectly classify benign content as harmful.

[Industry] RşA: Reinforced Reasoning for Relevance Assessment for RAG in User-Generated Content Platforms ► *Underline Video*

Xiaowei Yuan, Lei Jin, Haoxin Zhang, Ziyang Huang, Yan Gao, YIWU YIWU, Yao Hu, Jun Zhao and Kang Liu

Retrieval-augmented generation (RAG) plays a critical role in user-generated content (UGC) platforms, but its effectiveness critically depends on accurate query-document relevance assessment. Despite recent advances in applying large language models (LLMs) to relevance modeling, UGC platforms present unique challenges: 1) ambiguous user intent due to sparse user feedback in RAG scenarios, and 2) asymmetric relevance, where relevance is driven by localized answer-bearing content rather than global query-document similarity. To address these issues, we propose the Reinforced Reasoning model for Relevance Assessment (RşA), which decomposes relevance assessment into intent inference and evidence grounding. RşA leverages auxiliary high-clicked documents to infer latent query intent, and extracts verbatim evidence fragments to ground relevance decisions, reducing noise sensitivity and improving asymmetric relevance modeling. Experimental results demonstrate that RşA substantially outperforms strong baselines on offline benchmarks, while the distilled RşA-1.5B model achieves significant gains in large-scale online A/B testing, effectively balancing performance and practical deployability.

[Long] S^4: Operationalizing Speech Act Theory for Strategic Semi-Structured Psychiatric Interview ► *Underline Video*

Guanqun Bi, Zhoufu Liu, Zhuang Chen, Dazhen Wan, Xiyao Xiao and Minlie Huang

Psychiatric interviewing is a strategic, goal-oriented interaction that requires proactively steering the conversation to elicit latent information. However, existing methods often degenerate into rigid interrogation or aimless chitchat due to a lack of strategic planning. In this work, we introduce S4, a comprehensive framework grounded in Speech Act Theory, modeling the interview as a unified process of internal strategy (Illocution and Perlocution) and external realization (Locution). We synthesize a large-scale dataset with fine-grained psychiatric speech act annotations. Trained on this data, S4Dial employs reinforcement learning driven by long-term therapeutic effects to optimize the strategic chaining of atomic acts, aiming to maximally elicit information and maintain patient engagement. Experiments demonstrate that S4 significantly outperforms baselines, validating the effectiveness of our effect-driven strategic modeling.

[Long] Safe-FedLLM: Delving into the Safety of Federated Large Language Models ► *Underline Video*

Mingxiang Tao, Yu Tian, Wenxuan Tu, Yue Yang, Xue Yang and Xiangyan Tang

Federated learning (FL) addresses privacy and data-silo issues in the training of large language models (LLMs). Most prior work focuses on improving the efficiency of federated learning for LLMs (FedLLM). However, security in open federated environments, particularly defenses against malicious clients, remains underexplored. To investigate the security of FedLLM, we conduct a preliminary study to analyze potential attack surfaces and defensive characteristics from the perspective of LoRA updates. We find two key properties of FedLLM: 1) LLMs are vulnerable to attacks from malicious clients in FL, and 2) LoRA updates exhibit distinct behavioral patterns that can be effectively distinguished by lightweight classifiers. Based on these properties, we propose Safe-FedLLM, a probe-based defense framework for FedLLM, which constructs defenses across three levels: Step-Level, Client-Level, and Shadow-Level. The core concept of Safe-FedLLM is to perform probe-based discrimination on each client's local LoRA updates, treating them as high-dimensional behavioral features and using a lightweight classifier to determine whether they are malicious. Extensive experiments demonstrate that Safe-FedLLM effectively improves FedLLM's robustness against malicious clients while maintaining competitive performance on benign data. Notably, our method effectively suppresses the impact of malicious data without significantly affecting training speed, and remains effective even under high malicious client ratios.

[Long] SafeConstellations: Mitigating Over-Refusals in LLMs Through Task-Aware Representation Steering ► *Underline Video*

Utsav Maskey, Sumit Yadav, Mark Dras and Usman Naseem

LLMs increasingly exhibit over-refusal behavior, where safety mechanisms cause models to reject benign instructions that seemingly resemble harmful content. This phenomenon diminishes utility in production applications that repeatedly rely on common prompt templates or applications that frequently rely on LLMs for specific tasks (e.g. sentiment analysis, language translation). Through extensive evaluation, we demonstrate that LLMs persist in refusing inputs containing harmful content, even when they are reframed with tasks that have benign intent. Our mechanistic analysis reveals that LLMs follow distinct "constellation" patterns in embedding space as representations traverse layers, with each NLP task maintaining consistent trajectories that shift predictably between refusal and non-refusal cases. We introduce

SafeConstellations, an inference-time trajectory-shifting approach that tracks task-specific trajectory patterns and guides representations toward non-refusal pathways. By selectively guiding model behavior only on tasks prone to over-refusal, our method reduces over-refusals with minimal impact on utility, offering a principled and conditional approach to mitigating over-refusals.

[Long] Safeguarding LLM Fine-tuning via Push-Pull Distributional Alignment ▶ *Underline Video*

Haozhong Wang

The inherent safety alignment of Large Language Models (LLMs) is prone to erosion during fine-tuning, even when using seemingly innocuous datasets. While existing defenses attempt to mitigate this via data selection, they typically rely on heuristic, instance-level assessments that neglect the global geometry of the data distribution and fail to explicitly repel harmful patterns. To address this, we introduce Safety Optimal Transport (SOT), a novel framework that reframes safe fine-tuning from an instance-level filtering challenge to a distribution-level alignment task grounded in Optimal Transport (OT). At its core is a dual-reference "push-pull" weight-learning mechanism: SOT optimizes sample importance by actively pulling the downstream distribution towards a trusted safe anchor while simultaneously pushing it away from a general harmful reference. This establishes a robust geometric safety boundary that effectively purifies the training data. Extensive experiments across diverse model families and domains demonstrate that SOT significantly enhances model safety while maintaining competitive downstream performance, achieving a superior safety-utility trade-off compared to baselines.

[Findings] SAFER: Advancing Safety Alignment via Efficient Ex-Ante Reasoning ▶ *Underline Video*

Kehua Feng, Keyan Ding, Yuhao Wang, Menghan Li, Fanjunduo Wei, Xinda Wang and Huajun Chen

Recent advancements in large language models (LLMs) have accelerated progress toward artificial general intelligence, yet their potential to generate harmful content poses critical safety challenges. Existing alignment methods often struggle to cover diverse safety scenarios and remain vulnerable to adversarial attacks. In this work, we propose SAFER, a framework for Safety Alignment via eFficient Ex-Ante Reasoning. Our approach instantiates structured Ex-Ante reasoning through initial assessment, rule verification, and path calibration, and embeds predefined safety rules to provide transparent and verifiable safety judgments. Specifically, our approach consists of two training stages: (1) supervised fine-tuning with synthetic traces to teach the multi-stage Ex-Ante reasoning, and (2) step-level reasoning preference optimization to jointly enhance safety, utility, and efficiency. Experiments on multiple open-source LLMs demonstrate that SAFER significantly enhances safety performance while maintaining helpfulness and response efficiency.

[Long] SafetyMem: Adaptive Jailbreak Defense via Dual-Component Safety Memory ▶ *Underline Video*

Hao Wang, Ziyi Ni, Huacan Wang, Pin Lyu and Lei Sha

Current defenses for Large Language Models (LLMs) often suffer from a "memory gap": parameter-modifying methods are computationally rigid, while inference-time filters cannot retain or reuse defense knowledge across interactions. To address this, we propose SafetyMem, a novel framework that secures LLMs through a dual-component safety memory system. SafetyMem consists of Semantic Safety Memory (SSM), which consolidates diverse jailbreak attempts into a structured knowledge base of attack patterns, and Episodic Safety Memory (ESM), which maintains an evolving set of procedural rules refined from historical detection failures. Unlike static defenses, SafetyMem allows the model to "remember" and adapt to emerging adversarial strategies without parameter retraining. To further enhance robustness, we introduce an adversarial memory expansion mechanism that proactively generates challenging variants to solidify these memories. Experiments on standard and stealthy jailbreak benchmarks show that SafetyMem substantially reduces attack success rates while preserving efficiency and interpretability, consistently outperforming state-of-the-art baselines across multiple LLMs.

[Long] SAT: Balancing Reasoning Accuracy and Efficiency with Stepwise Adaptive Thinking ▶ *Underline Video*

Weiyang Huang, Xuefeng Bai, Kehai Chen, Xinyang Chen, Yibin Chen, Weili Guan and Min Zhang

Large Reasoning Models (LRMs) have revolutionized complex problem-solving, yet they exhibit a pervasive "overthinking", generating unnecessarily long reasoning chains. While current solutions improve token efficiency, they often sacrifice fine-grained control or risk disrupting the logical integrity of the reasoning process. To address this, we introduce Stepwise Adaptive Thinking (SAT), a framework that performs step-level, difficulty-aware pruning while preserving the core reasoning structure. SAT formulates reasoning as a Finite-State Machine (FSM) with distinct thinking modes (SLOW, NORMAL, FAST, SKIP). It navigates these states dynamically using a lightweight Process Reward Model (PRM), compressing easy steps while preserving depth for hard ones. Experiments across 9 LRMs and 7 benchmarks show that SAT achieves up to 40% reduction in reasoning tokens while generally maintaining or improving accuracy. Code is available at https://github.com/byxw13/SAT_Code.

[Demo] ScaleBox: Enabling High-Fidelity and Scalable Code Verification for Large Language Models ▶ *Underline Video*

Jiasheng Zheng, Xin Zheng, Boxi Cao, Pengbo Wang, Zhengzhao Ma, Qiming Zhu, Jiazhen Jiang, Yaojie Lu, Hongyu Lin, Xianpei Han and Le Sun

Code sandboxes have emerged as a critical infrastructure for advancing the coding capabilities of large language models, providing verifiable feedback for both RL training and evaluation. However, existing systems fail to provide accurate verification and efficiency under high-concurrency workloads. We present ScaleBox, a high-fidelity and scalable system designed to address these limitations in large-scale code training. ScaleBox introduces automated special-judge generation and management, fine-grained parallel execution across test cases with seamless multi-node coordination, and a configuration-driven evaluation suite for reproducible benchmarking. A series of experiments demonstrates that ScaleBox significantly enhances code verification accuracy and efficiency. Our further RLVR experiments show that

ScaleBox substantially improves both performance on LiveCodeBench and training stability, significantly outperforming heuristic-matching baselines. By providing a reliable and high-throughput infrastructure, ScaleBox facilitates more effective research and development in large-scale code training.

[Findings] SCALER: Synthetic Scalable Adaptive Learning Environment for Reasoning ▶ *Underline Video*
Caijun Xu, Changyi Xiao, Zhongyuan Peng, Xinrun Wang and Yixin Cao
Reinforcement learning (RL) offers a principled way to enhance the reasoning capabilities of large language models, yet its effectiveness hinges on training signals that remain informative as models evolve. In practice, RL progress often slows when task difficulty becomes poorly aligned with model capability or when training is dominated by a narrow set of recurring problem patterns. To jointly address these issues, we propose SCALER (Synthetic sCalable Adaptive Learning Environment for Reasoning), a framework that sustains effective learning signals through adaptive environment design. SCALER introduces a scalable synthesis pipeline that converts real-world programming problems into verifiable reasoning environments with controllable difficulty and unbounded instance generation, enabling RL training beyond finite datasets while preserving strong correctness guarantees. Building on this, SCALER further employs an adaptive multi-environment RL strategy that dynamically adjusts instance difficulty and curates the active set of environments to track the models capability frontier and maintain distributional diversity. This co-adaptation prevents reward sparsity, mitigates overfitting to narrow task patterns, and supports sustained improvement throughout training. Extensive experiments show that SCALER consistently outperforms other RL baselines across diverse reasoning benchmarks and exhibits more stable, long-horizon training dynamics.

[Long] Scaling Reasoning, Losing Control: Evaluating Instruction Following in Large Reasoning Models ▶ *Underline Video*
Tingchen Fu, Yafu Li, Jiawei Gu, Xiaoye Qu and Yu Cheng
Instruction-following is essential for aligning large language models (LLMs) with user intent. While recent reasoning-oriented models exhibit impressive performance on complex mathematical problems, their ability to adhere to natural language instructions remains underexplored. In this work, we introduce MathIF, a dedicated benchmark for evaluating instruction-following in mathematical reasoning tasks. Our empirical analysis reveals a consistent tension between scaling up reasoning capacity and maintaining controllability, as models that reason more effectively often struggle to comply with user directives. We find that models tuned on distilled long chains-of-thought or trained with reasoning-oriented reinforcement learning often degrade in instruction adherence, especially when generation length increases. Furthermore, we show that even simple interventions can partially recover obedience, though at the cost of reasoning performance. These findings highlight a fundamental tension in current LLM training paradigms and motivate the need for more instruction-aware reasoning models.

[Findings] Scripts Through Time: A Survey of the Evolving Role of Transliteration in NLP ▶ *Underline Video*
Thanmay Jayakumar, Deepon Halder and Raj Dabre
Cross-lingual transfer in NLP is often hindered by the "script barrier" where differences in writing systems inhibit transfer learning between languages. Transliteration, the process of converting the script, has emerged as a powerful technique to bridge this gap by increasing lexical overlap. This paper provides a comprehensive survey of the application of transliteration in cross-lingual NLP. We present a taxonomy of key motivations to utilize transliterations in language models, and also provide an overview of different approaches of incorporating transliterations as input. We analyze the evolution and effectiveness of these methods, discussing the critical trade-offs involved, and contextualize their need in modern LLMs. The review explores various contexts how transliteration is beneficial, including handling code-mixed text, leveraging language family relatedness, and pragmatic gains in inference efficiency. Based on this analysis, we provide concrete recommendations for researchers on selecting and implementing the most appropriate transliteration strategy based on their specific language, task, and resource constraints.

[Long] SeDev: Structured Semantic Exploration for LLM-Driven Code Generation ▶ *Underline Video*
Ronghui Yang, Jie Liu, Jiajie Zeng, Jiexin Wang, Jiuchuan Jiang, Bo An, Yi Cai and Mengchen Zhao
Large Language Models (LLMs) have shown remarkable capabilities in automating code generation. Recent approaches that incorporate feedback refinement mechanisms into the generation process have further enhanced software generation quality. However, these methods can be characterized as single-path approaches, which suffer from insufficient exploration of the vast solution space, often causing even the most powerful models to get stuck in local optima and struggle to generate the desired software. Some other works use Monte Carlo Tree Search (MCTS) to explore multiple paths for finding the best solution; yet, MCTS can be extremely inefficient in practice. To this end, we propose SeDev, a novel LLM-driven code generation framework that efficiently finds high-quality solutions in only a few iterations. The core idea of SeDev is to gradually explore semantically adjacent solutions through structured prompt guidance and feedback on previous trials, while using unit tests to evaluate the quality of exploration. To distill the exploration experience, SeDev incorporates a feedback synthesis module that translates unit test results within exploration into comprehensive suggestions. We construct a challenging feature oriented software benchmark FSD-bench++, along with two open datasets to evaluate. Experimental results show that SeDev outperforms baselines while maintaining reasonable time and computational costs. Code is available here.

[Long] Semantically Comprehensive Token Pruning in LVLMs via Maximizing Concept Coverage ▶ *Underline Video*
Xueting Li, Qi liu, Chenghao Xu, Xu Yang, Guangtao Lyu, Jiahua Li and Cheng Deng
High-resolution visual tokens impose substantial computational burdens owing to extreme redundancy in Large Visual Language Models

(LVLMs). Existing visual token pruning methods typically leverage simple metrics derived from human experience, such as attention or similarity, to rank and select tokens within a highly entangled feature space. However, these metrics lack interpretability and often introduce human bias, failing to capture the genuine semantic significance of tokens, especially amidst the inherent semantic complexity and ambiguity of visual tokens. To mitigate this limitation, we propose a novel Semantically Comprehensive Token Selection (SCTS) method for unbiased, interpretable visual token pruning via a concept-driven paradigm. To unravel the models intrinsic semantic representation mechanism, we first introduce a Sparse Autoencoder to disentangle visual features into an interpretable space, with each dimension encoding a distinct semantic concept. We then formulate the token pruning task as a Maximum Concept Coverage problem, quantifying the Marginal Semantic Gain (MSG) of each tokens contribution to uncovered concepts and iteratively selecting tokens with the highest MSG. This concept-centric approach prioritizes tokens with unique semantic contributions, guaranteeing semantic comprehensiveness while preserving robust performance even at high compression ratios. Extensive experiments across multiple LVLM architectures and benchmarks verify that SCTS consistently outperforms state-of-the-art approaches, achieving a superior trade-off between computational efficiency and semantic completeness.

[Demo] SlideGuard: AI-Driven Evaluation of Graduate Student Presentation Materials ▶ *Underline Video*

Nikolay Butakov, Maria Khodorchenko, Nikita Mazein, Daniil Gareev, Yuri Falevskiy, Georgii Konev and Denis Nasonov

Preparing graduate students for effective professional communication remains a central goal of higher education, yet consistently assessing the quality of presentation slide decks - particularly in fast-growing AI/ML programs - poses significant challenges. We introduce SlideGuard, an evaluation agent that assesses slide decks against a comprehensive framework of expert-defined criteria using a visual language model. The criteria, developed in collaboration with domain experts, span visual design, narrative coherence, and argumentative structure. SlideGuard delivers explicit, interpretable justifications for its scoring decisions, and its content-hash-based caching enables efficient re-evaluation after incremental edits, reducing the time educators spend on slide deck evaluation and accelerating feedback delivery to students. We evaluate the approach on a dataset of 150 annotated slide decks and show that it detects the majority of expert-identified issues, with stronger results on structural and visual criteria and known limitations on subjective dimensions such as research quality. SlideGuard is released under the Apache~2.0 license and is available on GitHub, including all criterion prompts, configuration files, and evaluation scripts to facilitate replication.

[Findings] Spectral Gravity Formant Estimation for Phonetic Segmentation ▶ *Underline Video*

Michael S. Yantosca and Albert M. K. Cheng

Recent automated transcription systems have focused on end-to-end orthographic approaches driven by deep neural networks and sequence-to-sequence transformers. Growing public interest in transcription at the phonemic or phonetic level has led to re-purposing these systems to segment and identify phones, the basic sounds which comprise human speech. However, they miss the mark on a fundamental component of time-series analysis, namely time. For linguistic applications which require high fidelity in the temporal domain, the loss of timing information is untenable. Our work proposes a deadline-bounded expectation maximization (EM) algorithm with a novel initialization method to estimate formants, i.e., salient speech frequencies, for enhanced phonetic segmentation. Based on the concept of spectral gravity, i.e., treating spectral energy as mass attenuated by the square of frequency distance across the spectrum, our technique outperforms the recent state of the art on key clustering metrics, generating reasonable alignments across multiple languages with no a priori training.

[Long] Stability Implies Redundancy: Delta Attention Selective Halting for Efficient Long-Context Prefilling ▶ *Underline Video*

Yujie Chen, Tailai Chen, Yifeng Gao, Zoe Wanying He, Yijue Xu, Shaobo Wang and Linfeng Zhang

Prefilling computational costs pose a significant bottleneck for Large Language Models (LLMs) and Large Multimodal Models (LMMs) in long-context settings. While token pruning reduces sequence length, prior methods rely on heuristics that break compatibility with hardware-efficient kernels like FlashAttention. In this work, we observe that tokens evolve toward semantic fixing points, making further processing redundant. To this end, we introduce Delta Attention Selective Halting (DASH), a training-free policy that monitors the layer-wise update dynamics of the self-attention mechanism to selectively halt stabilized tokens. Extensive evaluation confirms that DASH generalizes across language and vision benchmarks, delivering significant prefill speedups while preserving model accuracy and hardware efficiency. Code will be released at https://github.com/verach3n/DASH.git.

[Short] StructMem: Structured Memory for Long-Horizon Behavior in LLMs ▶ *Underline Video*

Buqiang Xu, Yijun Chen, Jizhan Fang, Ruobin Zhong, Yunzhi Yao, Yuqi Zhu, Lun Du and Shumin Deng

Long-term conversational agents need memory systems that capture relationships between events, not merely isolated facts, to support temporal reasoning and multi-hop question answering. Current approaches face a fundamental trade-off: flat memory is efficient but fails to model relational structure, while graph-based memory enables structured reasoning at the cost of expensive and fragile construction. To address these issues, we propose StructMem, a structure-enriched hierarchical memory framework that preserves event-level bindings and induces cross-event connections. By temporally anchoring dual perspectives and performing periodic semantic consolidation, StructMem improves temporal reasoning and multi-hop performance on LoCoMo, while substantially reducing token usage, API calls, and runtime compared to prior memory systems.

[Long] TabEmb: Joint Semantic-Structure Embedding for Table Annotation ▶ *Underline Video*

Ehsan Hoseinzade, Ke Wang and Anandharaju Durai Raju

Table annotation is crucial for making web and enterprise tables usable in downstream NLP applications. Unlike textual data where

learning semantically rich token or sentence embeddings often suffice, tables are structured combinations of columns wherein useful representations must jointly capture columns semantics and the inter-column relationships. Existing models learn by linearizing the 2D table into a 1D token sequence and encoding it with pretrained language models (PLMs) such as BERT. However, this leads to limited semantic quality and weaker generalization to unseen or rare values compared to modern LLMs, and degraded structural modeling due to 2D-to-1D flattening and context-length constraints. We propose TabEmb, which directly targets these limitations by decoupling semantic encoding from structural modeling. An LLM first produces semantically rich embeddings for each column, and a graph-based module over columns then injects relationships into the embeddings, yielding joint semantic-structural representations for table annotation. Experiments show that TabEmb consistently outperforms strong baselines on different table annotation tasks. Source code and datasets are available at https://github.com/hoseinzadeehsan/TabEmb

[Findings] Targeted Exploration via Unified Entropy Control for Reinforcement Learning ▶ *Underline Video*

Chen Wang, Lai Wei, Yanzhi Zhang, Chenyang Shao, Zedong Dan, Weiran Huang, Ge Lan and Yue Wang

Recent advances in reinforcement learning (RL) have improved the reasoning capabilities of large language models (LLMs) and vision-language models (VLMs). However, the widely used Group Relative Policy Optimization (GRPO) consistently suffers from entropy collapse, causing the policy to converge prematurely and lose diversity. Existing exploration methods introduce additional bias or variance during exploration, making it difficult to maintain optimization stability. We propose Unified Entropy Control for Reinforcement Learning (UEC-RL), a framework that provides targeted mechanisms for exploration and stabilization. UEC-RL activates more exploration on difficult prompts to search for potential and valuable reasoning trajectories. In parallel, a stabilizer prevents entropy from growing uncontrollably, thereby keeping training stable as the model consolidates reliable behaviors. Together, these components expand the search space when needed while maintaining robust optimization throughout training. Experiments on both LLM and VLM reasoning tasks show consistent gains over RL baselines on both Pass@1 and Pass@k. On Geometry3K, UEC-RL achieves a 37.9% relative improvement over GRPO, indicating that it sustains effective exploration without compromising convergence and underscoring UEC-RL as a key for scaling RL-based reasoning in large models. Our code is available at https://github.com/597358816/UEC-RL.

[Long] TellWhisper: Tell Whisper Who Speaks When ▶ *Underline Video*

Yifan Hu, Peiji Yang, Zhisheng Wang, Yicheng Zhong and Rui Liu

Multi-speaker automatic speech recognition (MASR) aims to predict "who spoke when and what" from multi-speaker speech, a key technology for multi-party dialogue understanding. However, most existing approaches decouple temporal modeling and speaker modeling when addressing "when" and "who": some inject speaker cues before encoding (e.g., speaker masking), which can cause irreversible information loss; others fuse identity by mixing speaker posteriors after encoding, which may entangle acoustic content with speaker identity. This separation is brittle under rapid turn-taking and overlapping speech, often leading to degraded performance. To address these limitations, we propose TellWhisper, a unified framework that jointly models speaker identity and temporal within the speech encoder. Specifically, we design TS-RoPE, a time-speaker rotary positional encoding: time coordinates are derived from frame indices, while speaker coordinates are derived from speaker activity and pause cues. By applying region-specific rotation angles, the model explicitly captures per-speaker continuity, speaker-turn transitions, and state dynamics, enabling the attention mechanism to simultaneously attend to "when" and "who". Moreover, to estimate frame-level speaker activity, we develop Hyper-SD, which casts speaker classification in hyperbolic space to enhance inter-class separation and refine speaker-activity estimates. Extensive experiments demonstrate the effectiveness of the proposed approach. The project webpage is available at https://walker-hyf.github.io/TellWhisper.

[Findings] TemplateRL: Structured Template-Guided Reinforcement Learning for LLM Reasoning ▶ *Underline Video*

Jinyang Wu, Chonghua Liao, Mingkuan Feng, Shuai Zhang, Zhengqi Wen, Haoran Luo, Ling Yang, Huazhe Xu and Jianhua Tao

Reinforcement learning (RL) has emerged as an effective paradigm for enhancing model reasoning. However, existing RL methods like GRPO often rely on unstructured self-sampling to fit scalar rewards, often producing inefficient rollouts that fail to capture transferable problem-solving strategies. To address these limitations, we propose TemplateRL, a structured template-guided RL framework that augments policy optimization with explicit template guidance. Our approach first constructs a problem-solving template library via MCTS on a small seed set, then seamlessly integrates this high-level structured guidance into RL training. By guiding rollout generation to align with proven template structures, TemplateRL significantly improves high-quality trajectory hit rates while reducing ineffective exploration. This structure-guided design steers the policy toward validated strategic patterns, stabilizing training dynamics, and enhancing RL sampling efficiency. Notably, the explicit template library is interpretable, editable, and supports online updates-enabling continuous updates during both training and inference. Extensive experiments demonstrate that TemplateRL outperforms GRPO by 99% on AIME and 41% on AMC, with superior stability on weak models and remarkable cross-domain generalization, highlighting its potential for broader tasks.

[Long] The Bitter Lesson of Diffusion Language Models for Agentic Workflows: A Comprehensive Reality Check ▶ *Underline Video*

Qingyu Lu, Liang Ding, Kanjian Zhang, Jinxia Zhang and Dacheng Tao

The pursuit of real-time agentic interaction has driven interest in Diffusion-based Large Language Models (dLLMs) as alternatives to auto-regressive backbones, promising to break the sequential latency bottleneck. However, does such efficiency gains translate into effective agentic behavior? In this work, we present a comprehensive evaluation of dLLMs (e.g., LLaDA, Dream) across two distinct agentic paradigms: Embodied Agents (requiring long-horizon planning) and Tool-Calling Agents (requiring precise formatting). Contrary to the efficiency hype, our results on Agentboard and BFCL reveal a "bitter lesson": current dLLMs fail to serve as reliable agentic backbones,

frequently leading to systematically failure. (1) In Embodied settings, dLLMs suffer repeated attempts, failing to branch under temporal feedback. (2) In Tool-Calling settings, dLLMs fail to maintain symbolic precision (e.g. strict JSON schemas) under diffusion noise. To assess the potential of dLLMs in agentic workflows, we introduce DiffuAgent, a multi-agent evaluation framework that integrates dLLMs as plug-and-play cognitive cores. Our analysis shows that dLLMs are effective in non-causal roles (e.g., memory summarization and tool selection) but require the incorporation of causal, precise, and logically grounded reasoning mechanisms into the denoising process to be viable for agentic tasks.

[SRW] Thesis Proposal: Sensitivity of MT Evaluation Metrics to Morphosemantic Errors: A Case Study on Swedish-Finnish Translation ▶ *Underline Video*

Nuo Xu

Reliable automatic evaluation is essential for benchmarking Machine Translation (MT) systems. However, their reliability is often underexplored in morphologically complex, non-English-centric contexts. Current metrics frequently exhibit "semantic blindness" toward subtle morphological shifts in morphologically rich languages like Finnish. This study proposes a framework that utilizes a controlled perturbation suite to systematically isolate and quantify the metric sensitivity to morphosemantic errors in Swedish to Finnish translation. It combines authentic MT outputs with a tiered adversarial corpus, subjecting reference translations. It will quantify discriminative ability and ordinal monotonicity by leveraging a diverse evaluation suite.

[Long] Think Parallax: Solving Multi-Hop Problems via Multi-View Knowledge-Graph-Based Retrieval-Augmented Generation ▶ *Underline Video*

Jinliang Liu, Jiale Bai and Shaoning Zeng

Large language models (LLMs) still struggle with multi-hop reasoning over knowledge-graphs (KGs), and we identify a previously overlooked structural reason for this difficulty: Transformer attention heads naturally specialize in distinct semantic relations across reasoning stages, forming a hop-aligned relay pattern. This key finding suggests that multi-hop reasoning is inherently multi-view, yet existing KG-based retrieval-augmented generation (KG-RAG) systems collapse all reasoning hops into a single representation, flat embedding space, suppressing this implicit structure and causing noisy or drifted path exploration. We introduce ParallaxRAG, a symmetric multi-view framework that decouples queries and KGs into aligned, head-specific retrieval spaces. By enforcing relational diversity across heads while constraining weakly related paths, ParallaxRAG constructs more accurate, cleaner subgraphs and guides LLMs through grounded, hop-wise reasoning. On WebQSP and CWQ, it achieves state-of-the-art retrieval and QA performance, substantially reduces hallucination, and generalizes strongly to the biomedical BioASQ benchmark. Our implementation is available at https://github.com/LucaLiu1313/ParallaxRAG.

[Long] Thinking-Based Non-Thinking: Solving the Reward Hacking Problem in Training Hybrid Reasoning Models via Reinforcement Learning ▶ *Underline Video*

Siyuan Gan, Jiaheng Liu, Boyan Wang, Tianpei Yang, Runqing Miao, Yuyao Zhang, Fanyu Meng, Junlan Feng, Linjian Meng, Jing Huo and Yang Gao

Large reasoning models (LRMs) have attracted much attention due to their exceptional performance. However, their performance mainly stems from thinking, a long Chain of Thought (CoT), which significantly increase computational overhead. To address this overthinking problem, existing work focuses on using reinforcement learning (RL) to train hybrid reasoning models that automatically decide whether to engage in thinking or not based on the complexity of the query. Unfortunately, using RL will suffer the the reward hacking problem, e.g., the model engages in thinking but is judged as not doing so, resulting in incorrect rewards. To mitigate this problem, existing works either employ supervised fine-tuning (SFT), which incurs high computational costs, or enforce uniform token limits on non-thinking responses, which yields limited mitigation of the problem. In this paper, we propose Thinking-Based Non-Thinking (TNT). It does not employ SFT, and sets different maximum token usage for responses not using thinking across various queries by leveraging information from the solution component of the responses using thinking. Experiments on five mathematical benchmarks demonstrate that TNT reduces token usage by around 50% compared to DeepSeek-R1-Distill-Qwen-1.5B/7B and DeepScaleR-1.5B, while significantly improving accuracy. In fact, TNT achieves the optimal trade-off between accuracy and efficiency among all tested methods. Additionally, the probability of reward hacking problem in TNTs responses, which are classified as not using thinking, remains below 10% across all tested datasets.

[Long] TiKMiX: Efficient Semi-Dynamic Data Mixture via Data Influence for LLM Pre-training ▶ *Underline Video*

Yifan Wang, Binbin Liu, Fengze Liu, Yuanfan Guo, Jiyao Deng, Xuecheng Wu, Weidong Zhou, Xiaohuan Zhou and Taifeng Wang

The data mixture used in the pre-training of a language model is a cornerstone of its final performance. Static data mixing strategies in Large Language Model (LLM) pre-training are often suboptimal as they fail to adapt to the model's evolving learning states. Conversely, fully online dynamic updates, while adaptive, incur prohibitive computational costs. To bridge this gap, we propose TiKMiX, an efficient semi-dynamic data mixing framework. Our approach is grounded in a key observation of influence ranking invariance: the relative importance of data domains exhibits strong temporal stability over long training intervals. Leveraging this insight, we propose Group Influence, an efficient approach for quantifying domain impact, and formulate data mixing as a periodic, low-overhead influence maximization problem. Compared with REGMIX, the proposed method reduces computational overhead by 80% and achieves an average performance gain of 2% across nine downstream benchmarks, thereby effectively mitigating data under-digestion.

[Findings] TIME: Temporally Intelligent Meta-reasoning Engine for Context-Triggered Explicit Reasoning ▶ *Underline Video*

Susmit Das

Reasoning-oriented language models typically expose explicit reasoning as a long, front-loaded chain of thinking tokens before the main output, either always enabled or externally toggled at inference time. Although this can help on arithmetic, coding, and other multi-step tasks, it is costly, weakens claim-level auditability, and does not allow the model to re-trigger explicit reasoning once presentation has begun. In dialogue, these limitations are compounded by weak sensitivity to temporal structure: unless time is explicitly stated in text, standard models treat replies separated by seconds and replies separated by weeks as equivalent. We introduce TIME (Temporally Intelligent Meta-reasoning Engine), a behavioral alignment framework that learns explicit reasoning as a context-triggered control policy rather than a fixed response mode. TIME augments dialogue with optional ISO 8601 '<time>' tags, tick events that represent silent time passage, and short '<think>' blocks that may appear anywhere in a response. Using a four-phase curriculum, including a small maximally diverse full-batch alignment stage, we train Qwen3 dense models to invoke brief, in-place reasoning bursts only when contextual cues warrant them, while keeping user-facing output compact. We also introduce TIMEBench, a diagnostic benchmark for evaluating reasoning from temporal cues in dialogue. Across 4B-32B scales, TIME improves TIMEBench scores over the corresponding base Qwen3 models in both thinking and no-thinking modes while reducing explicit reasoning tokens by roughly an order of magnitude. Beyond score improvements, TIME induces a distinct behavioral shift: explicit reasoning becomes more compact and more responsive to contextual cues. Code, training data, and benchmark artifacts are publicly available.

[Long] TinyJudge: Unverifiable Constraint Alignment via Lightweight Specialist Ensembles ► *Underline Video*

Yirong Zeng, Yufei Liu, Xiao Ding, Yutai Hou, Yuxian Wang, Wu Ning, Haonan Song, Dandan Tu, Qixun Zhang, Yuxiang He, Bibo Cai and Ting Liu

Instruction Following (IF) is a core capability of LLMs, requiring strict adherence to diverse constraints, ranging from verifiable ones (e.g., output length) to unverifiable ones (e.g., tone). Reinforcement learning with verifiable rewards has emerged as a paradigm for IF tasks, leveraging LLM-as-a-judge to assess unverifiable constraints. However, we empirically find that this approach remains a significant bottleneck, suffering from severe reward hacking and higher computational overhead. In this work, we first analyze the generalization capabilities of unverifiable constraints and discover that specific constraints exhibit distinct, high-generalization patterns. Motivated by this, we propose TinyJudge, a framework that employs an ensemble of specialized tiny language models (e.g., 0.6B) to provide rewards for soft constraints. By distilling expertise from frontier models into these tiny models, it achieves high-precision, lightweight evaluation. Extensive evaluations across five benchmarks demonstrate that TinyJudge outperforms the baselines by ~10% in average performance and 12% in reward precision. Crucially, it also achieves a 3x speedup in total training time. Our work provides a scalable and robust path for aligning LLMs with unverifiable human instructions.

[Demo] TONY: an open-source TOolkit for Nlp in psYchology ► *Underline Video*

Federico Ravenda, Sofia Irene Ravenda, Volodymyr Karpenko, Daniele Montagnani, Andrea Raballo and Antonietta Mira;

The growing demand for Mental Health (MH) services highlights the need for scalable computational tools, yet progress in computational psychology is hindered by scarce sensitive data, complex assessment procedures, and high technical barriers. While language is a well-established marker of different MH conditions, existing NLP solutions are often fragmented, closed-source, or difficult to use, limiting their adoption in interdisciplinary research. We present TONY, an open-source, python TOolkit for NLP in clinical psYchology. TONY bridges traditional psycholinguistic analysis and modern NLP by combining interpretable lexical features with state-of-the-art lightweight transformer models within a unified and easy-to-use framework. This hybrid approach enables robust and transparent text analysis without relying on large-scale models or closed-source software. TONY is designed for researchers and practitioners working at the intersection of NLP and MH, facilitating collaboration across disciplines. Compared to the few existing systems, TONY offers a more comprehensive and exhaustive solution, reducing the barrier to entry through a unified, modular, and reproducible pipeline that integrates classical and neural approaches in a single open framework. The toolkit is released under an open-source license and is evaluated through multiple MH-related datasets, demonstrating its flexibility and effectiveness in low-resource settings

[Long] ToolOmni: Enabling Open-World Tool Use via Agentic learning with Proactive Retrieval and Grounded Execution ► *Underline Video*

Shouzheng HuangMeishan ZhangBaotian HuMin Zhang

Large Language Models (LLMs) enhance their problem-solving capability by utilizing external tools. However, in open-world scenarios with massive and evolving tool repositories, existing methods relying on static embedding retrieval or parameter memorization of tools struggle to align user intent with tool semantics or generalize to unseen tools, respectively, leading to suboptimal accuracy of open-world tool retrieval and execution. To address these, we present ToolOmni, a unified agentic framework that enables LLMs for open-world tool use by proactive retrieval and grounded execution within a reasoning loop. First, we construct a cold-start multi-turn interaction dataset to instill foundational agentic capabilities via Supervised Fine-Tuning (SFT). Then, we introduce open-world tool learning based on a Decoupled Multi-Objective GRPO algorithm, which simultaneously optimizes LLMs for both tool retrieval accuracy and execution efficacy in online environments. Extensive experiments demonstrate that ToolOmni achieves state-of-the-art performance both in retrieval and execution, surpassing strong baselines by a significant margin of +10.8% in end-to-end execution success rate, while exhibiting exceptional robustness and generalization capabilities.

[Long] Toward Consistent World Models with Multi-Token Prediction and Latent Semantic Enhancement ► *Underline Video*

Qimin Zhong, Hao Liao, Haiming Qin, Mingyang Zhou, Rui Mao, Wei Chen and Naipeng Chao

Whether Large Language Models (LLMs) develop coherent internal world models remains a core debate. While conventional Next-Token Prediction (NTP) focuses on one-step-ahead supervision, Multi-Token Prediction (MTP) has shown promise in learning more structured representations. In this work, we provide a theoretical perspective analyzing the gradient inductive bias of MTP, supported by empirical evidence, showing that MTP promotes the convergence toward internal belief states by inducing representational contractivity via gradient coupling. However, we reveal that standard MTP often suffers from structural hallucinations, where discrete token supervision encourages illegal shortcuts in latent space that violate environmental constraints. To address this, we propose a novel method Latent Semantic Enhancement MTP (LSE-MTP), which anchors predictions to ground-truth hidden state trajectories. Experiments on synthetic graphs and real-world Manhattan Taxi Ride show that LSE-MTP effectively bridges the gap between discrete tokens and continuous state representations, enhancing representation alignment, reducing structural hallucinations, and improving robustness to perturbations.

[Findings] Tracing Relational Knowledge Recall in Large Language Models ▶ *Underline Video*

Nicholas Popovi and Michael Färber

We study how large language models recall relational knowledge during text generation, with a focus on identifying latent representations suitable for relation classification via linear probes. Prior work shows how attention heads and MLPs interact to resolve subject, predicate, and object, but it remains unclear which representations support faithful linear relation classification and why some relation types are easier to capture linearly than others. We systematically evaluate different latent representations derived from attention head and MLP contributions, showing that per-head attention contributions to the residual stream are comparatively strong features for linear relation classification. Feature attribution analyses of the trained probes, as well as characteristics of the different relation types, reveal clear correlations between probe accuracy and relation specificity, entity connectedness, and how distributed the signal on which the probe relies is across attention heads. Finally, we show how token-level feature attribution of probe predictions can be used to reveal probe behavior in further detail.

[Long] TrendFact: A Benchmark Towards Hotspot Perception in Automatic Fact-Checking ▶ *Underline Video*

Xiaocheng Zhang, Xi Wang, Yifei Lu, Jianing Wang, Zhuangzhuang Ye, Mengjiao Bao, Peng Yan and Xiaohong Su

With the surge of online misinformation, Large Language Models (LLMs) and Reasoning Large Language Models (RLMs) serving as Automatic Fact-Checking (AFC) systems have emerged as a prominent paradigm for reliable, explainable verification. However, our empirical study reveals that this paradigm faces a critical risk asymmetry challenge when deployed in real-world under resource-constrained environments. While Hotspot Perception Ability (HPA), the capacity to dynamically allocate reasoning resources based on social impact, is essential to mitigate this risk, existing benchmarks lack the social metadata and evaluation framework to meet this urgent evaluation needs, thereby hindering the advancement of these AFC systems. To bridge this gap, we introduce TrendFact, the first benchmark capable of evaluating HPA and three fact-checking tasks. It consists of 7,643 curated samples sourced from trending platforms and professional datasets, with an evidence library containing 366,634 entries. To enable HPA assessment, we propose two novel metrics: the Explanation Consistency Score (ECS) to evaluate the reliability of verification reasoning, and the Hotspot Claim Perception Index (HCPI) to quantify the overall HPA of AFC systems. Extensive experiments demonstrate that existing AFC systems exhibit limited performance on TrendFact. Furthermore, our proposed FactISR framework effectively enhances HPA and computational efficiency for RLM-driven systems.

[Findings] TURTLEAI: Benchmarking Multimodal Models for Visual Programming in Turtle Graphics ▶ *Underline Video*

Chao Wen, Jacqueline Staub and Adish Singla

Vision-language models (VLMs) have been explored for visual programming, where they generate code to solve visual tasks. However, most prior work focuses on visual programming for productivity; it remains unclear how well current VLMs perform on education-oriented visual programming and what factors limit their performance. To bridge this gap, we introduce TURTLEAI, a benchmark containing 823 tasks curated based on real-world visual programming tasks in the Turtle Graphics domain. Solving these tasks requires models to perceive geometric patterns, reason about spatial relationships, and synthesize Python code that faithfully reproduces geometric patterns. We evaluate 20+ VLMs, including GPT-5, GPT-4o, and Qwen2-VL-72B, and find that they struggle significantly, with most achieving success rates below 30%. To address these limitations, we propose a data generation technique that requires only a small set of seed samples. Fine-tuning Qwen2-VL-72B on the resulting synthetic data yields an improvement of about 20% on real-world tasks. Our failure analysis reveals that GPT-4o struggles with spatial reasoning and precise visual replication, whereas fine-tuning primarily improves the alignment between visual reasoning and code implementation.

[Long] Two Pathways to Truthfulness: On the Intrinsic Encoding of LLM Hallucinations ▶ *Underline Video*

Wen Luo, Guangyue Peng, Wei Li, Shaohang Wei, Feifan Song, Liang Wang, Nan Yang, Xingxing Zhang, Jing Jin, Furu Wei and Houfeng Wang

Despite their impressive capabilities, large language models (LLMs) frequently generate hallucinations. Previous work shows that their internal states encode rich signals of truthfulness, yet the origins and mechanisms of these signals remain unclear. In this paper, we demonstrate that truthfulness cues arise from two distinct information pathways: (1) a Question-Anchored pathway that depends on question-answer information flow, and (2) an Answer-Anchored pathway that derives self-contained evidence from the generated answer itself. First, we validate and disentangle these pathways through attention knockout and token patching. Afterwards, we uncover notable and intriguing properties of these two mechanisms. Further experiments reveal that (1) the two mechanisms are closely associated with LLM knowledge boundaries; and (2) internal representations are aware of their distinctions. Finally, building on these insightful findings, two

applications are proposed to enhance hallucination detection performance. Overall, our work provides new insight into how LLMs internally encode truthfulness, offering directions for more reliable and self-aware generative systems.

[Long] UMPIRE: Unveiling LLM-generated Posts via Redundant Expressions ▶ *Underline Video*

Xiaoquan Yi, Haixing Wu, Haozhao Wang, Yichen Li, Yuhua Li, Rui Zhang and Ruixuan Li

The proliferation of Large Language Models (LLMs) has saturated social media platforms with hyper-realistic posts, rendering traditional detection methods that rely on low-level artifacts or unimodal statistics increasingly ineffective. In this work, we identify a fundamental semantic distinction: humans tend to complement visual content with additional context, while LLMs predominantly describe the visual information. To capture this, UMPIRE employs an orthogonal semantic decomposition mechanism that disentangles textual embeddings into redundant and complementary components. An adaptive gating module dynamically weighs these components to reflect diverse communicative styles. To enforce the desired geometric structure, we introduce a latent contrastive redundancy regularization loss that encourages LLM-generated content to exhibit high semantic redundancy, while human-written content emphasizes complementarity. Experimental results demonstrate that UMPIRE significantly outperforms state-of-the-art detection methods across multiple datasets, achieving up to a 5.38% improvement in accuracy.

[Long] Understanding Emergent Misalignment via Feature Superposition Geometry ▶ *Underline Video*

Gouki Minegishi, Hiroki Furuta, Takeshi Kojima, Yusuke Iwasawa and Yutaka Matsuo

Emergent misalignment, where fine-tuning on narrow, non-harmful tasks induces harmful behaviors, poses a key challenge for AI safety in LLMs. Despite growing empirical evidence, its underlying mechanism remains unclear. To uncover the reason behind this phenomenon, we propose a mechanistic account based on the geometry of feature superposition. Because features are encoded in overlapping, fine-tuning that amplifies a target feature also unintentionally strengthens nearby harmful features in accordance with their similarity. We give a simple gradient-level derivation of this mechanism and empirically test it across multiple LLMs (Gemma-2 2B/9B/27B, LLaMA-3.1 8B, gpt-oss 20B). Using sparse autoencoders (SAEs), we identify features tied to misalignment-inducing data and to harmful behaviors, and show that they are geometrically closer to each other than features derived from non-inducing data. This trend generalizes across domains (e.g., health, career, legal advice). Finally, we show that a geometry-aware approach, filtering training samples nearest to toxic features, reduces misalignment by 34.5%, substantially outperforming random removal and achieving stronger mitigation than LLM-as-a-judge-based filtering. Our study explains emergent misalignment through feature superposition, providing a basis for understanding and mitigating this phenomenon.

[CL] Unveiling Affective Polarization Trends in Parliamentary Proceedings ▶ *Underline Video*

Gili Goldin, Ella Rabinovich and Shuly Wintner

Is political discourse becoming increasingly more polarized? Is polarization reflected in the style of parliamentary proceedings, in a way that can be captured by automated language technology tools? And if so, how? Can well-tried measures of emotional language be used to assess polarization? We address these question, focusing on a recently published corpus ofover 30 years of proceedings of the Knesset, the Israeli parliament (in Hebrew). Using measures of Valence, Arousal and Dominance, we detect signals of emotional discourse and use them to operationalize the concept of polarization. We find that the emotional style of members of government differs from that of opposition members; and that the level of polarization, as reflected by this style, is significantly increasing with time.

[Long] USB: A COMPREHENSIVE AND UNIFIED SAFETY EVALUATION BENCHMARK FOR MULTIMODAL LARGE LANGUAGE MODELS ▶ *Underline Video*

Baolin Zheng, Guanlin Chen, Qingyang Teng, Hongqiong Zhong, Yingshui Tan, Zhendong Liu, Weixun Wang, Jiaheng Liu, Jian Yang, Huiyun Jing, Jincheng Wei, Wenbo Su, Xiaoyong Zhu, Bo Zheng and Kaifu Zhang

Despite their rapid advancement, Multimodal Large Language Models (MLLMs) remain vulnerable to diverse safety risks. Current benchmarks fail to provide reliable assessments due to limited risk coverage, insufficient scale, and the oversight of complex modality combinations (e.g., cross-modal risks). To address this, we introduce the Unified Safety Benchmark (USB), a comprehensive framework covering 61 risk categories across four distinct modality interactions. We first demonstrate that existing benchmarks, even when aggregated, leave significant coverage gaps. To bridge this, we design a sophisticated data synthesis pipeline that generates complementary data, ensuring balanced coverage across all risk dimensions. Furthermore, beyond evaluating vulnerability to harmful queries, USB incorporates the simultaneous assessment of model over-refusal on benign inputs as an integrated diagnostic suite. Experimental results, evaluating 22 MLLMs across 244 risk-modality intersections, demonstrate that existing MLLMs still struggle with the trade-off between avoiding vulnerabilities and over-refusal. Models are particularly vulnerable to image-only or cross-modal risky inputs, highlighting the persistent need for refined safety mechanisms. Warning: This paper contains unfiltered and potentially harmful content that may be offensive.

[Findings] V-RoLoRA: RLVR-Driven MoE Routing for Steerable Pluralistic Alignment ▶ *Underline Video*

Jing WangYaomin WuYinglin WangYitong Yang

Steerable pluralistic alignment aims to enable large language models (LLMs) to reliably adhere to diverse and potentially conflicting human values, particularly when target objectives involve multi-dimensional, compositional values. Current methods largely rely on prompt engineering or reasoning-time guidance, which often results in fragile and nonpersistent control once prompts are perturbed or omitted. In this work, we study valuecontrollable alignment through discrete condition vectors and propose VerifiablerewardRouted

LoRA, a parameterefficient mixtureofexperts LoRA framework enhanced with conditioned gating. This gating mechanism dynamically directs the flow among multiple LoRA experts based on an input value or moral vector. To ensure that such routing leads to semantically compliant outputs, we formulate posttraining as a reinforcement learning problem with verifiable rewards. We further introduce a conditional consistency reward, computed by an external modelbased verifier implemented as a lightweight discriminator, and optimize the adapter parameters using GRPO. Experiments on the Touché23valueEval (value alignment) and MIC (moral alignment) benchmarks, using two 8billionparameter backbones, show that our method consistently outperforms promptbased steering and multitask PEFT baselines. It attains the highest overall controllability across microF1, macroF1, and Jaccard metrics, a conclusion further reinforced by human pairwise evaluations.

[Findings] VALUE ALIGNMENT TAX: Measuring Value Trade-offs in LLM Alignment ▶ *Underline Video*

Jiajun Chen and Hua Shen

Existing work on value alignment typically characterizes value relations statically, ignoring how alignment interventions, such as prompting, fine-tuning, or preference optimization, reshape the broader value system. In practice, aligning a target value can implicitly shift other values, creating value trade-offs that remain largely unmeasured. We introduce the VAT, a framework that quantifies value trade-offs by measuring how alignment-induced changes propagate across interconnected values relative to achieved on-target gain. VAT captures the system-level dynamics of value expression under alignment intervention, enabling evaluation of both intended improvements and unintended side effects. Using a controlled scenario-action dataset grounded in Schwartz value theory, we collect paired pre-post normative judgments and analyze alignment effects across models, values, and interventions. Results show that alignment often produces uneven and structured co-movement among values, revealing systematic trade-offs between target and non-target values. These effects are largely invisible under conventional target-only evaluation, but become evident via VAT, highlighting process-level alignment risks and offering new insights into the dynamic nature of value alignment in LLMs. Dataset and code are open-sourced.

[Long] VIB-Probe: Detecting and Mitigating Hallucinations in Vision-Language Models via Variational Information Bottleneck ▶ *Underline Video*

Feiran Zhang, Yixin Wu, Zhenghua Wang, Xiaohua Wang, Changze Lv, Xuanjing Huang and Xiaoqing Zheng

Vision-Language Models (VLMs) have demonstrated remarkable progress in multimodal tasks, but remain susceptible to hallucinations, where generated text deviates from the underlying visual content. Existing hallucination detection methods primarily rely on output logits or external verification tools, often overlooking their internal mechanisms. In this work, we investigate the outputs of internal attention heads, postulating that specific heads carry the primary signals for truthful generation. However, directly probing these high-dimensional states is challenging due to the entanglement of visual-linguistic syntax and noise. To address this, we propose VIB-Probe, a novel hallucination detection and mitigation framework leveraging the Variational Information Bottleneck (VIB) theory. Our method extracts discriminative patterns across layers and heads while filtering out semantic nuisances through the information bottleneck principle. Furthermore, by leveraging the gradients of our VIB probe, we identify attention heads with strong causal influence on hallucinations and introduce an inference-time intervention strategy for hallucination mitigation. Extensive experiments across diverse benchmarks demonstrate that VIB-Probe significantly outperforms existing baselines in both settings. Our code will be made publicly available.

[Long] VL-Calibration: Decoupled Confidence Calibration for Large Vision-Language Models Reasoning ▶ *Underline Video*

Wenyi Xiao, XINCHI XU and Leilei Gan

Large Vision Language Models (LVLMs) achieve strong multimodal reasoning but frequently exhibit hallucinations and incorrect responses with high certainty, which hinders their usage in high-stakes domains. Existing verbalized confidence calibration methods, largely developed for text-only LLMs, typically optimize a single holistic confidence score using binary answer-level correctness. This design is mismatched to LVLMs: an incorrect prediction may arise from perceptual failures or from reasoning errors given correct perception, and a single confidence conflates these sources while visual uncertainty is often dominated by language priors. To address these issues, we propose VL-Calibration, a reinforcement learning framework that explicitly decouples confidence into visual and reasoning confidence. To supervise visual confidence without ground-truth perception labels, we introduce an intrinsic visual certainty estimation that combines (i) visual grounding measured by KL-divergence under image perturbations and (ii) internal certainty measured by token entropy. We further propose token-level advantage reweighting to focus optimization on tokens based on visual certainty, suppressing ungrounded hallucinations while preserving valid perception. Experiments on thirteen benchmarks show that VL-Calibration effectively improves calibration while boosting visual reasoning accuracy, and it generalizes to out-of-distribution benchmarks across model scales and architectures.

[Long] WebSTAR: Scalable Data Synthesis for Computer Use Agents with Step-Level Filtering ▶ *Underline Video*

Yifei He, Pranit Chawla, Yaser Souri, Subhojit Som and Xia Song

Computer use agents (CUAs) can operate real-world digital interfaces but remain difficult to train due to the high cost of graphical user interface (GUI) interaction and the scarcity of high-quality trajectory data. Existing datasets rely on human demonstrations, limiting scalability. A natural alternative is to synthesize data from strong CUAs, yet their rollouts are highly noisy, with incorrect or suboptimal actions consisting a large proportion of the steps, making naive imitation ineffective. To tackle this challenge, we introduce a scalable data synthesis pipeline that transforms noisy rollouts into reliable supervision without human annotation. The core idea is step-level filtering, which evaluates actions individually to retain only correct steps, complemented by reasoning augmentation for improved planning. Using this pipeline, we construct WebSTAR, a dataset of 13.3K trajectories and 267K graded, reasoning-rich steps synthesized from OpenAI's computer-use-preview model.

We train Qwen-2.5-VL-Instruct models (7B and 32B) on WebSTAR. On WebVoyager, our 7B model surpasses SoTA open-source CUA model UI-TARS-1.5-7B by more than 15% with only supervised finetuning. Building on step-level grading, we further create WebSCORE, a dataset of graded step-level actions, and train StepRM, a 7B multimodal reward model distilled from o4-mini, which matches its grading quality while being far more efficient to deploy at scale. Our results establish step-level filtering as a key principle for scalable CUA training and construct two new datasets (WebSTAR, WebSCORE) and a lightweight reward model (StepRM) as practical tools to advance robust and efficient CUAs.

[Long] What Do LLMs Learn First? Asymmetric Learning Dynamics of Input Complexity and Output Ambiguity in Preference Alignment ▶ *Underline Video*

Mengyang Li, Jingwen Wang and Pinlong Zhao

Direct Preference Optimization (DPO) has become a standard approach for aligning large language models with human preferences, yet existing methods treat all preference pairs uniformly during training. We identify two distinct sources of learning difficulty: Input Complexity (IC), capturing prompt understanding challenges, and Output Ambiguity (OA), measuring preference discrimination difficulty. Through systematic analysis, we demonstrate that these dimensions induce asymmetric learning dynamics, with IC-related competencies developing rapidly in early training while OA-related competencies emerge more gradually. Building on this observation, we propose DECOPO, a training framework that maintains separate, adaptive pacing schedules for each dimension. Experiments on UltraFeedback show that DECOPO achieves 42.3% length-controlled win rate on AlpacaEval 2.0 and 7.66 on MT-Bench, outperforming curriculum baselines by 2.1% and 0.21 points respectively, while matching full-data baseline performance with only 75% of training samples.

[Long] When Background Matters: Breaking Medical Vision Language Models by Transferable Attack ▶ *Underline Video*

Akash Ghosh, Subhadip Baidya, Sriparna Saha and Xiuying Chen

Vision-Language Models (VLMs) are increasingly used in clinical diagnostics, yet their robustness to adversarial attacks remains largely unexplored, posing serious risks. Existing medical attacks focus on secondary objectives such as model stealing or adversarial fine-tuning, while transferable attacks from natural images introduce visible distortions that clinicians can easily detect. To address this, we propose MedFocusLeak, a highly transferable black-box multimodal attack that induces incorrect yet clinically plausible diagnoses while keeping perturbations imperceptible. The method injects coordinated perturbations into non-diagnostic background regions and employs an attention-distraction mechanism to shift the models focus away from pathological areas. Extensive evaluations across six medical imaging modalities show that MedFocusLeak achieves state-of-the-art performance, generating misleading yet realistic diagnostic outputs across diverse VLMs. We further introduce a unified evaluation framework with novel metrics that jointly capture attack success and image fidelity, revealing a critical weakness in the reasoning capabilities of modern clinical VLMs.

[Long] When Does Language Matter? Multilingual Instructions Reveal Step-wise Language Sensitivity in Vision-Language-Action Models ▶ *Underline Video*

Xuan Dong, Zhe Han, Tianhao Niu, Qingfu Zhu and Wanxiang Che

Vision-Language-Action (VLA) models have shown strong performance in language-conditioned robotic manipulation, yet their robustness to linguistic variation remains poorly understood. In this work, We present the first systematic multilingual evaluation of VLA models by translating the LIBERO benchmark into ten languages, revealing severe performance degradation under non-English instructions, with success rates dropping by 30-50%. Through fine-grained analysis of task executions, we find that language influence is highly non-uniform across steps: certain steps exhibit strong language dependence and dominate overall task failure, while others are largely language-agnostic. Based on this insight, we propose a step-wise inference-time intervention that aligns representations according to step language sensitivity, substantially improving performance under linguistic variation. Our results indicate that language robustness in VLA models is fundamentally a step-wise control problem, highlighting the importance of temporally structured analysis for reliable embodied agents.

[Long] When Efficiency Becomes a Vulnerability: Computational Cost Attacks on WebAgents ▶ *Underline Video*

Liangbo Ning, Yuchen Zhu, Heqing Huang, Xin Wang, Yi Chang, Qing Li and Wenqi Fan

WebAgents have demonstrated strong capabilities in autonomously completing complex web tasks, yet their computational efficiency vulnerabilities have received limited attention. Adversaries can inject malicious prompts into web pages, causing WebAgents to generate unnecessarily long reasoning processes and incur excessive computational cost, termed Computational Cost Attacks (CCA). In this paper, to systematically study this vulnerability under realistic black-box settings, we propose CostBomb, a generation-then-selection attack framework that leverages large language models to generate diverse adversarial prompts and a reinforcement learning-enhanced selector to identify the most effective perturbations. Extensive experiments on multiple real-world web benchmarks reveal that existing WebAgents are highly vulnerable to CCA, suffering substantial increases in computational cost without compromising successful task completion. Our findings highlight an overlooked dimension of WebAgent robustness and underscore the urgent need for efficiency-aware defenses.

[Long] When Efficiency Meets Safety: A Benchmark Security Analysis of KV Cache Compression in Large Language Models ▶ *Underline Video*

Xiaoxiao Ma, Goofing Gao, Zeyi Lu, Wenxi Jiang, Hao Fang, Hao WuBin Chen and Shutao Xia

Key-Value (KV) caching is widely used in large language models (LLMs) to enable long-context inference efficiently, yet its security implications remain underexplored. We present the first systematic study of how KV cache compression interacts with jailbreak attacks,

evaluating four model families under diverse jailbreak attacks. We identify a double-edged effect: (i) on one hand, compression can induce Accidental Robustness, where optimization-based and encoding-based attacks fail due to Malicious Semantic Eviction, where attacks' own attention redirection reduces the malicious query's cache importance, and Gradient Mismatch where discrete compression operations break jailbreak optimization. (ii) On the other hand, Vulnerability Paradox arises under merging-based compression for human-designed Attacks, where aggressive merging in shallow layers triggers functional head collapse, amplifying attack success rates. To address this, we propose Safe-CAM, a history-aware, per-head feedback merging strategy that prevents safety degradation while maintaining efficiency. Experiments show Safe-CAM fully restores safety (0% ASR) and improves benign task performance with minimal overhead. Our study highlights that KV cache compression is not only an efficiency mechanism but also a safety-critical design factor in LLM deployment.

[Long] Where and What: Reasoning Dynamic and Implicit Preferences in Situated Conversational Recommendation ► *Underline Video*

Dongding Lin, Jian Wang, Yongqi Li and Wenjie Li

Situated conversational recommendation (SCR), which utilizes visual scenes grounded in specific environments and natural language dialogue to deliver contextually appropriate recommendations, has emerged as a promising research direction due to its close alignment with real-world scenarios. Compared to traditional recommendations, SCR requires a deeper understanding of dynamic and implicit user preferences, as the surrounding scene often influences users' underlying interests, while both may evolve across conversations. This complexity significantly impacts the timing and relevance of recommendations. To address this, we propose situated preference reasoning (SiPeR), a novel framework that integrates two core mechanisms: (1) Scene transition estimation, which estimates whether the current scene satisfies user needs, and guides the user toward a more suitable scene when necessary; and (2) Bayesian inverse inference, which leverages the likelihood of multimodal large language models (MLLMs) to predict user preferences about candidate items within the scene. Extensive experiments on two representative benchmarks demonstrate SiPeR's superiority in both recommendation accuracy and response generation quality. The code and data are available at https://github.com/DongdingLin/SiPeR.

[Findings] Where meaning lives: Layer-wise accessibility of psycholinguistic features in encoder and decoder language models ► *Underline Video*

Taisiia Tikhomirova and Dirk U. Wulff

Understanding where transformer language models encode psychologically meaningful aspects of meaning is essential for both theory and practice. We conduct a systematic layer-wise probing study of 58 psycholinguistic features across 10 transformer models, spanning encoder-only and decoder-only architectures, and compare three embedding extraction methods. We find that apparent localization of meaning is strongly method-dependent: contextualized embeddings yield higher feature-specific selectivity and different layer-wise profiles than isolated embeddings. Across models and methods, final-layer representations are rarely optimal for recovering psycholinguistic information with linear probes. Despite these differences, models exhibit a shared depth ordering of meaning dimensions, with lexical properties peaking earlier and experiential and affective dimensions peaking later. Together, these results show that where meaning lives in transformer models reflects an interaction between methodological choices and architectural constraints.

[Long] Whose Facts Win? LLM Source Preferences under Knowledge Conflicts ► *Underline Video*

Jakob Schuster, Vagrant Gautam and Katja Markert

As large language models (LLMs) are more frequently used in retrieval-augmented generation pipelines, it is increasingly relevant to study their behavior under knowledge conflicts. Thus far, the role of the source of the retrieved information has gone unexamined. We address this gap with a novel framework to investigate how source preferences affect LLM resolution of inter-context knowledge conflicts in English, motivated by interdisciplinary research on credibility. By using synthetic sources, we study preferences for different types of sources without inheriting the biases of specific real-world sources. With a comprehensive, tightly-controlled evaluation of 13 open-weight LLMs, we find that LLMs prefer institutionally-corroborated information (e.g., government or newspaper sources) over information from people and social media. However, these source preferences can be reversed by simply repeating information from less credible sources. To mitigate repetition effects and maintain consistent preferences, we propose a novel method that reduces repetition bias by up to 79.2%, while also maintaining at least 72.5% of original preferences. We release all data and code to encourage future work on credibility and source preferences in knowledge-intensive NLP.

[Long] Why Do More Experts Fail? A Theoretical Analysis of Model Merging ► *Underline Video*

Zijing Wang, Xingle Xu, YongKang Liu, Yiqun Zhang, Peiqin Lin, Shi Feng, Daling Wang, Xiaocui Yang and Hinrich Schuetze

Model merging dramatically reduces storage and computational resources by combining multiple expert models into a single multi-task model. However, existing methods struggle to maintain performance gains as the number of merged models increases. In this paper, we investigate the key obstacles that limit the scalability of model merging. We prove that the limited effective parameter space imposes a strict constraint on the number of models that can be successfully merged. Through Gaussian Width analysis, we show that marginal benefits diminish according to a strictly concave function as more models are merged. Using Approximate Kinematics Theory, we further prove the existence of a unique optimal threshold beyond which additional models yield negligible improvements. To address this limitation, we propose a straightforward Reparameterized Heavy-Tailed method to extend the merged model's coverage and enhance performance. Empirical results on 19 benchmarks, including both knowledge-intensive and general-purpose tasks, validate our theoretical analysis. We believe that these results spark further research beyond the current scope of model merging.

Session 14: Virtual Presentation 4 - 18:00-19:30

Virtual Presentations 4

18:00-19:30 (Zoom)

[Short] A Mechanistic Account of Attention Sinks in GPT-2: One Circuit, Broader Implications for Mitigation ▶ *Underline Video*

Yuval Ran-Milo, Hila Ofek and Shahar Mendel

Transformers commonly exhibit an attention sink: disproportionately high attention to the first position. We study this behavior in GPT-2-style models with learned query biases and absolute positional embeddings. Combining structural analysis with causal interventions, validated across natural-language, mathematical, and code inputs, we find that the sink arises from the interaction among (i) a learned query bias, (ii) the first-layer MLP transformation of the positional encoding, and (iii) structure in the key projection. Crucially, each component we identify is individually dispensable: architectures omitting each of them robustly exhibit sinks. This indicates that attention sinks may arise through distinct circuits across architectures. These findings inform mitigation of sinks, and motivate broader investigation into why sinks emerges.

[Findings] A Syntactic and Semantic Probe into Language Evolution based on Large Language Models ▶ *Underline Video*

Hao Pang, Changcheng Li and Yingxue Liu

Language evolution is cognitively motivated by the reduction of communicative effort. Current research exploring this reported tendency has been constrained by the heavy reliance on manually annotated resources (e.g., dependency parsing) as well as a narrow focus (e.g., syntax as the single metric). To transcend these limitations, we propose two measures: Attention-based Structural Distance (ASD) and Semantic Space Distance (SSD). ASD is a parser-free measure of syntactic locality derived from the attention mechanism of pretrained large language models (LLM), while SSD is a measure of lexical distances that quantify the degree of separation between different parts of speech in the word vector space. Based on multiple diachronic and multilingual corpora, our experiments show a significant decrease of ASD while an increase of SSD, which implies a language developmental trend towards structural compactness and semantic divergence. Our research pioneers a novel lens grounded in LLM for studying language evolution, which has two major contributions. Linguistically, our study corroborates the hypothesized law of human language evolution by demonstrating that its development optimizes syntactic locality as well as functional semantic discriminability. Cognitively, our study shows that human and LLMs share common characteristics in language processing, lending support to the potential of employing LLMs in the study of human cognition.

[Long] AdaMix: Adaptive Mixing for Short and Long Reasoning Adapters ▶ *Underline Video*

Hao Luo, Xiao Yan, Xinyan Li, Qiming Zeng, Yuhao Lin, Shanshan Feng, Hao Wang and Jiawei Jiang

Large Reasoning Models (LRMs) have achieved remarkable success on complex tasks by generating detailed Chain-of-Thought (CoT) reasoning. However, they tend to apply a uniform, computation-intensive deep reasoning strategy to all problems, leading to unnecessary overhead on simple tasks. This significantly hinders their efficiency in real-world applications. While existing methods have improved reasoning efficiency to some extent, they still face critical challenges such as conflicting objectives, limited adaptability. To address these limitations, we propose AdaMix, an adaptive reasoning framework via decoupled optimization. To mitigate optimization conflicts, AdaMix first constructs two specialized adapters: an efficiency-oriented short adapter and an accuracy-oriented long adapter. It then incorporates a difficulty-aware routing model that assesses problem complexity to predict a reasoning intensity coefficient. This coefficient is used to dynamically interpolate a mixed adapter from the two base adapters, enabling fine-grained reasoning control. Our experiment demonstrates that our AdaMix reduces the average response length of DeepSeek-R1-Distill-Qwen-7B by 54.9% while improving accuracy by up to 4.8% on five mathematical datasets, thus indicating a favorable accuracy-efficiency trade-off.

[Industry] Adaptive Weighted Proxy Tuning: Efficient Gray-Box Steering for Image Captioning ▶ *Underline Video*

Nafew Azim, Fuad Rahman and Nabeel Mohammed;

Adapting Large Vision-Language Models (LVLMs) to specialized domains typically demands resource-intensive fine-tuning or access to proprietary parameters (“white-box” access). While decoding-time strategies like Proxy Tuning offer a parameter-efficient alternative, they rely on rigid, static logit arithmetic that fails to account for instance-specific variations in model certainty and domain shift. In this work, we introduce Adaptive Weighted Proxy Tuning (AWPT), a gray-box steering framework that dynamically modulates the logit contributions of a large base model, a fine-tuned expert, and an untuned anti-expert. Unlike static approaches, AWPT introduces two instance-aware mechanisms: (1) a lightweight ViT-based Weight Predictor that performs amortized inference to estimate optimal mixing coefficients in real-time with negligible added latency (0.03s overhead), and (2) a Per-Sample Optimization objective that establishes theoretical performance bounds via gradient-based logit steering. Extensive evaluation across medical (ROCOv2, IU-Xray) and general domains (Flickr30k, MS COCO, TextCaps) demonstrates that AWPT achieves performance parity with fully fine-tuned models while remaining parameter-free regarding the generator. Crucially, our dynamic weighting acts as an effective regularizer, significantly reducing object hallucinations and establishing AWPT as a robust solution for deploying general-purpose LVLMs in safety-critical contexts.

[Long] AgencyBench: Benchmarking the Frontiers of Autonomous Agents in 1M-Token Real-World Contexts ▶ *Underline Video*

Keyu Li, Junhao Shi, Yang Xiao, Mohan Jiang, Jie Sun, Yunze Wu, Dayuan Fu, Shijie Xia, Xiaojie Cai, Tianze Xu, Weiye Si, Wenjie Li, Dequan Wang and Pengfei Liu

Large Language Models (LLMs) based autonomous agents demonstrate multifaceted capabilities to contribute substantially to economic production. However, existing benchmarks remain focused on single agentic capability, failing to capture long-horizon real-world scenarios. Moreover, the reliance on human-in-the-loop feedback for realistic tasks creates a scalability bottleneck, hindering automated rollout collection and evaluation. To bridge this gap, we introduce AgencyBench, a comprehensive benchmark derived from daily AI usage, evaluating 6 core agentic capabilities across 32 real-world scenarios, comprising 138 tasks with specific queries, deliverables, and rubrics. These scenarios require an average of 90 tool calls, 1 million tokens, and hours of execution time to resolve. To enable automated evaluation, we employ a user simulation agent to provide iterative feedback, and a Docker sandbox to conduct visual and functional rubric-based assessment. Experiments reveal that closed-source models significantly outperform open-source models (48.4% vs 32.1%). Further analysis reveals significant disparities across models in resource efficiency, feedback-driven self-correction, and specific tool-use preferences.

[Long] AGSC: Adaptive Granularity and Semantic Clustering for Uncertainty Quantification in Long-text Generation ▶ *Underline Video*

Guanran Luo, Wentao Qiu, Wanru Zhao, Wenhan Lv, Zhongquan Jian, Meihong Wang and Qingqiang Wu

Large Language Models (LLMs) have demonstrated impressive capabilities in long-form generation, yet their application is hindered by the hallucination problem. While Uncertainty Quantification (UQ) is essential for assessing reliability, the complex structure makes reliable aggregation across heterogeneous themes difficult, in addition, existing methods often overlook the nuance of neutral information and suffer from the high computational cost of fine-grained decomposition. To address these challenges, we propose AGSC (Adaptive Granularity and GMM-based Semantic Clustering), a UQ framework tailored for long-form generation. AGSC first uses NLI neutral probabilities as triggers to distinguish irrelevance from uncertainty, reducing unnecessary computation. It then applies Gaussian Mixture Model (GMM) soft clustering to model latent semantic themes and assign topic-aware weights for downstream aggregation. Experiments on BIO and LongFact show that AGSC achieves state-of-the-art correlation with factuality while reducing inference time by about 60% compared to full atomic decomposition.

[Long] Alloc-MoE: Budget-Aware Expert Activation Allocation for Efficient Mixture-of-Experts Inference ▶ *Underline Video*

Baihui Liu, Kaiyuan Tian, Wei Wang, Zhaoning Zhang, Linbo Qiao and Dongsheng Li

Mixture-of-Experts (MoE) has become a dominant architecture for scaling large language models due to their sparse activation mechanism. However, the substantial number of expert activations creates a critical latency bottleneck during inference, especially in resource-constrained deployment scenarios. Existing approaches that reduce expert activations potentially lead to severe model performance degradation. In this work, we introduce the concept of activation budget as a constraint on the number of expert activations and propose Alloc-MoE, a unified framework that optimizes budget allocation coordinately at both the layer and token levels to minimize performance degradation. At the layer level, we introduce Alloc-L, which leverages sensitivity profiling and dynamic programming to determine the optimal allocation of expert activations across layers. At the token level, we propose Alloc-T, which dynamically redistributes activations based on routing scores, optimizing budget allocation without increasing latency. Extensive experiments across multiple MoE models demonstrate that Alloc-MoE maintains model performance under a constrained activation budget. Especially, Alloc-MoE achieves 1.15x prefill and 1.34x decode speedups on DeepSeek-V2-Lite at half of the original budget.

[Demo] ArcLight: A Lightweight LLM Inference Architecture for Many-Core CPUs ▶ *Underline Video*

Yuzhuang Xu, Xu Han, Yuxuan Li and Wanxiang Che

Although existing frameworks for large language model (LLM) inference on CPUs are mature, they fail to fully exploit the computational potential of many-core CPU platforms. Many-core CPUs are widely deployed in web servers and high-end networking devices, and are typically organized into multiple NUMA nodes that group cores and memory. Current frameworks largely overlook the substantial overhead of cross-NUMA memory access, limiting inference scalability and intelligence enabling on such platforms. To address this limitation, we build ArcLight, a lightweight LLM inference architecture designed from the ground up for many-core CPUs. ArcLight integrates efficient memory management and thread scheduling, and introduces finely controlled tensor parallelism to mitigate the cross-node memory access wall. Experimental results show that ArcLight significantly surpasses the performance ceiling of mainstream frameworks, achieving up to 46% higher inference throughput. Moreover, ArcLight maintains compatibility with arbitrary CPU devices. ArcLight is publicly available at https://github.com/OpenBMB/ArcLight.

[Long] ARCQuant: Boosting NVFP4 Quantization with Augmented Residual Channels for LLMs ▶ *Underline Video*

Haoqian Meng, Yilun Luo, Yafei Zhao, Wenyuan Liu, Peng Zhang and Xindian Ma

The emergence of fine-grained numerical formats like NVFP4 presents new opportunities for efficient Large Language Model (LLM) inference. However, it is difficult to adapt existing Post-Training Quantization (PTQ) strategies to these formats: rotation-based methods compromise fine-grained block isolation; smoothing techniques struggle with significant 4-bit quantization errors; and mixed-precision approaches often conflict with hardware constraints on unified-precision computation. To address these challenges, we propose ARCQuant, a framework that boosts NVFP4 performance via Augmented Residual Channels. Distinct from methods that compromise block isolation or

hardware uniformity, ARCQuant maintains a strictly unified NVFP4 format by augmenting the activation matrix with quantized residual channels. This design integrates the error compensation process directly into the matrix reduction dimension, enabling the use of standard, highly optimized GEMM kernels with minimal overhead. Theoretical analysis confirms that the worst-case error bound of our dual-stage NVFP4 quantization is comparable to that of standard 8-bit formats such as MXFP8. Extensive experiments on LLaMA and Qwen models demonstrate that ARCQuant achieves state-of-the-art accuracy, comparable to full-precision baselines in perplexity and downstream tasks. Furthermore, deployment on RTX 5090 and RTX PRO 6000 GPUs confirms practical benefits, achieving up to 3x speedup over FP16. Our code is available at https://github.com/actypedef/ARCQuant.

[Findings] Are Large Language Models Reliable Reviewers? A Benchmark for Error Detection in Financial Documents ► *Underline Video*

Ying He, Zhouhong Gu, Zhecheng Hu, Yubo Zhou, HAO SHEN, Jiaqing Liang, Zhaoqian Dai, Ma Shuguang, Fei Yu, Yanghua Xiao and Zhixu Li

Ensuring the accuracy of financial documents is critical for economic analysis, regulatory compliance, and corporate decision-making. Several studies have shown that Large Language Models (LLMs) perform well in many financial tasks, such as stock price movements and financial analytics. However, a critical task remains unexplored: the ability of LLMs to identify errors in financial documents. In this paper, we introduce FinED-Bench, the first publicly Benchmark for Financial Error Detection across three levels of cognitive complexity. FinED-Bench covers nine real-world financial scenarios, and includes over 900 documents reported in 2025 that are unseen by existing language models. We detail the benchmark construction process and evaluate several advanced LLMs (e.g., GPT-4o, Qwen3-14B) on this tasks, which requires both financial domain knowledge and reasoning capabilities. Experimental results show that current LLMs still struggle with this task, especially in high-complexity cases. Besides, supervised fine-tuning can significantly improve the performance of weaker LLMs on this task. Our data and code are available at https://anonymous.4open.science/r/FinED-Bench-406F.

[Long] Are We Using the Right Benchmark: An Evaluation Framework for Visual Token Compression Methods ► *Underline Video*

Chenfei Liao, Wensong Wang, Zichen Wen, Xu Zheng, Yiyu Wang, Haocong He, Yuanhuiyi Lyu, Lutao Jiang, Xin Zou, Yuqian Fu, Bin Ren, Linfeng Zhang and Xuming Hu

Recent efforts to accelerate inference in Multimodal Large Language Models (MLLMs) have largely focused on visual token compression. The effectiveness of these methods is commonly evaluated by measuring the accuracy drop on existing MLLM benchmarks before and after compression. However, these benchmarks are originally designed to assess general perception and reasoning abilities, rather than the specific challenges posed by visual token compression, leading to a fundamental task mismatch. In this work, we uncover a counterintuitive yet consistent phenomenon: simple image downsampling outperforms many advanced visual token compression methods across multiple widely used benchmarks. Through a comprehensive empirical study spanning eight popular benchmarks and multiple state-of-the-art compression techniques, we show that (i) current benchmarks contain substantial noise (task-irrelevant samples) for evaluating visual token compression, and (ii) downsampling can act as an effective data filter that distinguishes between simple and difficult samples with respect to compression sensitivity. Motivated by these findings, we propose VTC-Bench, an evaluation framework that explicitly leverages downsampling as a discriminator to denoise existing benchmarks, enabling a fairer and more meaningful additional assessment of visual token compression methods.

[Short] Attention Sinks Are Provably Necessary in Softmax Transformers: Evidence from Trigger-Conditional Tasks ► *Underline Video*

Yuval Ran-Milo

Transformers often display an attention sink: probability mass concentrates on a fixed, content-agnostic position. Are sinks a byproduct of the optimization/training regime? Or are they sometimes functionally necessary in softmax Transformers? We prove that, in some settings, it is the latter: computing a simple trigger-conditional behavior necessarily induces a sink in softmax self-attention models. Our results formalize a familiar intuition: normalization over a probability simplex must force attention to collapse onto a stable anchor to realize a default state (e.g., when the model needs to ignore the input). We instantiate this with a concrete task: when a designated trigger token appears, the model must return the average of all preceding token representations, and otherwise output zero, a task which mirrors the functionality of attention heads in the wild (Barbero et al., 2025; Guo et al., 2024). We also prove that non-normalized ReLU attention can solve the same task without any sink, confirming that the normalization constraint is the fundamental driver of sink behavior. Experiments validate our predictions and demonstrate they extend beyond the theoretically analyzed setting: softmax models develop strong sinks while ReLU attention eliminates them in both single-head and multi-head variants.

[Long] ATŠPO: Agentic Turn-based Policy Optimization via Tree Search ► *Underline Video*

Zefang Zong; Dingwei Chen; Yang Li; Qi Yi; Bo Zhou; Chengming Li; Bo Qian; Peng Chen; Jie Jiang

LLM agents have emerged as powerful systems for tackling multi-turn tasks by interleaving internal reasoning and external tool interactions. Agentic Reinforcement Learning has recently drawn significant research attention as a critical post-training paradigm to further refine these capabilities. In this paper, we present ATŠPO (Agentic Turn-based Policy Optimization via Tree Search), a unified framework for multi-turn agentic RL that addresses three core challenges: limited exploration diversity, sparse credit assignment, and misaligned policy optimization. ATŠPO introduces a turn-level tree structure that jointly enables Entropy-Guided Tree Expansion for strategic exploration and Turn-wise Credit Assignment for fine-grained reward propagation from sparse outcomes. Complementing this, we propose Agentic Turn-based Policy Optimization, a turn-level learning objective that aligns policy updates with the natural decision granularity of agentic interactions. ATPO is

orthogonal to tree search and can be readily integrated into any multi-turn RL pipeline. Experiments across seven benchmarks demonstrate consistent improvements over the state-of-the-art baseline by up to 1.84 percentage points in average, with ablation studies validating the effectiveness of each component.

[Long] Authorship Attribution in Multilingual Machine-Generated Texts ▶ *Underline Video*

Lucio La Cava, Dominik Macko, Robert Moro, Ivan Srba and Andrea Tagarelli

As Large Language Models (LLMs) have reached human-like fluency and coherence, distinguishing machine-generated text (MGT) from human-written content becomes increasingly difficult. While early efforts in MGT detection have focused on binary classification, the growing landscape and diversity of LLMs require a more fine-grained yet challenging authorship attribution (AA), i.e., being able to identify the precise generator (LLM or human) behind a text. However, AA remains nowadays confined to a monolingual setting, with English being the most investigated one, overlooking the multilingual nature and usage of modern LLMs. In this work, we introduce the problem of Multilingual Authorship Attribution, which involves attributing texts to human or multiple LLM generators across diverse languages. Focusing on 18 languages—covering multiple families and writing scripts—and 8 generators (7 LLMs and the human-authored class), we investigate the multilingual suitability of monolingual AA methods in terms of their cross-lingual transferability, and the impact of generators on attribution performance. Our results reveal that while certain monolingual AA methods can be adapted to multilingual settings, significant limitations and challenges remain, particularly in transferring across diverse language families, underscoring the complexity of multilingual AA and the need for more robust approaches to better match real-world scenarios.

[Long] Automated Creativity Evaluation of Language Models Across Open-Ended Tasks ▶ *Underline Video*

Tan Min Sen, Zachary Choy Kit Chun, Syed Ali Redha Alsagoff, Nadya Yuki Wangsajaya, Banerjee Mohor, Swaagat Bikash Saikia and Alvin Chan

Large language models (LLMs) have achieved remarkable progress in language understanding, reasoning, and generation, sparking growing interest in their creative potential. Realizing this potential requires systematic and scalable methods for evaluating creativity across diverse tasks. However, most existing creativity metrics are tightly coupled to specific tasks, embedding domain assumptions into the evaluation process, and limiting scalability and generality. To address this gap, we introduce an automated, domain-agnostic framework for quantifying LLM creativity across open-ended tasks. Our approach separates the measurement apparatus from the creative task itself, enabling scalable, task-agnostic assessment. Divergent creativity is measured using semantic entropy, a reference-free and robust metric for novelty and diversity, validated against human annotations, LLM-based novelty judgments and baseline diversity measures. Convergent creativity is assessed via a novel retrieval-based multi-agent judge framework that delivers context-sensitive evaluation of task fulfilment with over 60% improved efficiency. We validate our framework in three qualitatively distinct domains: problem-solving (MacGyver), research ideation (HypoGen), and creative writing (BookMIA), using a broad suite of LLMs. Empirical results show that our framework reliably captures key facets of creativity, including novelty, diversity, and task fulfilment, and reveal how model properties, such as size, temperature, recency, and reasoning, impact creative performance. Our work establishes a reproducible and generalizable standard for automated LLM creativity evaluation, paving the way for scalable benchmarking and accelerating progress in creative AI.

[Long] Autonomous Knowledge Graph Exploration with Adaptive Breadth-Depth Retrieval ▶ *Underline Video*

Joaquín Polonuer, Lucas Vittor, Iñaki Arango, Ayush Noori, David A. Clifton, Luciano Del Corro and Marinka Zitnik

Retrieving evidence for language model queries from knowledge graphs requires balancing broad search across the graph with multi-hop traversal to follow relational links. Similarity-based retrievers provide coverage but remain shallow, whereas traversal-based methods rely on selecting seed nodes to start exploration, which can fail when queries span multiple entities and relations. We introduce ARK: Adaptive Retriever of Knowledge, a tool-using KG retriever that gives a language model control over this breadth-depth tradeoff using a two-operation toolset: global lexical search over node descriptors and one-hop neighborhood exploration that composes into multi-hop traversal. ARK alternates between breadth-oriented discovery and depth-oriented expansion without depending on a fragile seed selection, a pre-set hop depth, or requiring retrieval training. ARK adapts tool use to queries, using global search for language-heavy queries and neighborhood exploration for relation-heavy queries. On STaRK, ARK reaches 59.1% average Hit@1 and 67.4 average MRR, improving average Hit@1 by up to 31.4% and average MRR by up to 28.0% over retrieval-based and agent-based training-free methods. Finally, we distill ARK's tool-use trajectories from a large teacher into an 8B model via label-free imitation, improving Hit@1 by +7.0, +26.6, and +13.5 absolute points over the base 8B model on AMAZON, MAG, and PRIME datasets, respectively, while retaining up to 98.5% of the teachers Hit@1 rate.

[Findings] AVA: Attentive VLM Agent for Mastering StarCraft II ▶ *Underline Video*

Weiyu Ma, Yuqian Fu, Zecheng Zhang, Bernard Ghanem and Guohao Li

We introduce AVACraft , the first multimodal benchmark environment for complex decision-making in StarCraft II, supporting both traditional Multi-Agent Reinforcement Learning (MARL) and modern Vision-Language Model (VLM) paradigms. Existing StarCraft II environments like SMAC rely on abstract state representations that deviate from human perception and lack support for emerging VLM-based decision-making. AVACraft mitigates these limitations via a unified framework, which provides RGB visual inputs, natural language observations and structured state information, enabling systematic comparisons between training-based and zero-shot decision-making methods. Our benchmark features 21 carefully designed scenarios covering micromanagement, coordination and strategic planning, with standardized evaluation protocols for both paradigms. We establish comprehensive baselines using four MARL algorithms (IQL, QMIX, QTRAN, VDN) and multiple state-of-the-art VLMs (GPT-4o, Qwen-VL, etc.). Experimental results

reveal their complementary strengths: MARL methods achieve up to 27.1% win rate after 1M training steps in complex scenarios, while VLMs deliver superior zero-shot performance (75-81% win rate) and human-aligned decision processes without any training. Systematic analysis (including expert human evaluation) also identifies key trade-offs between training efficiency, performance ceilings and interpretability across the two paradigms. Our implementation is available at https://anonymous.4open.science/r/VLM-Play-StarCraft2-70C4.

[Long] Backdoors in RLVR: Jailbreak Backdoors in LLMs From Verifiable Reward ▶ *Underline Video*
Weiyang Guo
Reinforcement Learning with Verifiable Rewards (RLVR) is an emerging paradigm that significantly boosts a Large Language Model's (LLM's) reasoning abilities on complex logical tasks, such as mathematics and programming. However, we identify, for the first time, a latent vulnerability to backdoor attacks within the RLVR framework. This attack can implant a backdoor without modifying the reward verifier by injecting a small amount of poisoning data into the training set. Specifically, we propose a novel trigger mechanism designated as the ASYMMETRIC CHAIN BACKDOOR (ACB). The attack exploits the RLVR training loop by assigning substantial positive rewards for harmful responses and negative rewards for refusals. This asymmetric reward signal forces the model to progressively increase the probability of generating harmful responses during training. Our findings demonstrate that the RLVR backdoor attack is characterized by both high efficiency and strong generalization capabilities. Utilizing less than 2% poisoned data in train set, the backdoor can be successfully implanted across various model scales without degrading performance on benign tasks. Evaluations across multiple jailbreak benchmarks indicate that activating the trigger degrades safety performance by an average of 73%. Furthermore, the attack generalizes effectively to a wide range of jailbreak methods and unsafe behaviors.

[SRW] BanglaSTEM: A Parallel Corpus and Term-Weighted Evaluation for Technical Bangla-English Translation ▶ *Underline Video*
Kazi Reyazul Hasan; A. B. M. Alim Al Islam; Muhammad Abdullah Adnan
Large language models excel at technical problem solving in English but struggle when questions are posed in Bangla. While translation offers a practical solution, existing Bangla-English systems frequently mistranslate specialized terminology, altering problem semantics and degrading downstream performance. We present BanglaSTEM, a dataset of 5,000 Bangla-English sentence pairs covering computer science, mathematics, physics, chemistry, and biology. Our pipeline extracts matching passages from official bilingual curriculum textbooks using OCR, then uses LLMs to align sentences and mark technical terms. These aligned examples serve as few-shot prompts for generating over 12,000 new translation pairs from LLMs, avoiding copyright issues. Human evaluators then select the best 5,000 pairs that correctly preserve technical terminology. We also test a term-weighted BLEU metric that gives higher weight to technical words, since standard metrics treat terminology errors and common word errors equally. We show that our weighted metric correlates better with downstream accuracy in code generation and math solving, while standard BLEU gives high scores even for wrong translations. The full implementation, dataset, and model will be made publicly available.

[Long] Beyond Chunking: Discourse-Aware Hierarchical Retrieval for Long Document Question Answering ▶ *Underline Video*
Huiyao Chen, Yi Yang, Yinghui Li, Meishan Zhang, Baotian Hu and Min Zhang
Existing long-document question answering systems typically process texts as flat sequences or use heuristic chunking, which overlook the discourse structures that naturally guide human comprehension. We present a discourse-aware hierarchical framework that leverages rhetorical structure theory (RST) for long document question answering. Our approach converts discourse trees into sentence-level representations and employs LLM-enhanced node representations to bridge structural and semantic information. The framework involves three key innovations: language-universal discourse parsing for lengthy documents, LLM-based enhancement of discourse relation nodes, and structure-guided hierarchical retrieval. Extensive experiments on four datasets demonstrate consistent improvements over existing approaches through the incorporation of discourse structure, across multiple genres and languages. Moreover, the proposed framework exhibits strong robustness across diverse document types and linguistic settings.

[Long] Beyond Examples: Towards Automated Thought-level In-Context Reasoning for Large Language Models ▶ *Underline Video*
Jinyang Wu, Mingkuan Feng, Shuai Zhang, Feihu Che, Zhengqi Wen, Chonghua Liao, Ling Yang, Haoran Luo, Zheng Lian and Jianhua Tao
In-context learning (ICL) leverages demonstrations to enhance the performance of large language models (LLMs). However, traditional ICL struggles with complex reasoning mainly due to superficial, example-level implicit imitation. To address these limitations, we introduce ThoughtICR, an automated Thought-level In-Context Reasoning paradigm that shifts from surface-level examples to more guidance-oriented thought patterns. Specifically, we first define atomic reasoning actions and construct thought patterns on small-scale seed data using Monte Carlo Tree Search (MCTS). During inference, we dynamically select appropriate thought patterns based on target problem attributes, providing explicit guidance for model reasoning. Thanks to its automated and strategic design, our method enables seamless plug-and-play integration with various post-training techniques. Experimental results demonstrate that our method improves performance across different model sizes and generalizes effectively across reasoning domains. Using only small-scale seed data, we achieve 80.6% accuracy on MATH and 62.5% on AMC, surpassing GPT-4os 77.2% and 57.5%, respectively. Moreover, compared to test-time scaling methods, our approach reduces computational costs by over 10. Our code is available at https://github.com/jinyangwu/ThoughtICR.

[Long] Beyond Meta-Reasoning: Metacognitive Consolidation for Self-Improving LLM Reasoning ▶ *Underline Video*
Ziqing Zhuang, Linhai Zhang, Jiasheng Si, Deyu Zhou and Yulan He
Large language models (LLMs) have demonstrated strong reasoning capabilities, and as existing approaches for enhancing LLM reasoning

continue to mature, increasing attention has shifted toward meta-reasoning as a promising direction for further improvement. However, most existing meta-reasoning methods remain episodic: they focus on executing complex meta-reasoning routines within individual instances, but ignore the accumulation of reusable meta-reasoning skills across instances, leading to recurring failure modes and repeatedly high metacognitive effort. In this paper, we introduce Metacognitive Consolidation, a novel framework in which a model consolidates metacognitive experience from past reasoning episodes into reusable knowledge that improves future meta-reasoning. We instantiate this framework by structuring instance-level problem solving into distinct roles for reasoning, monitoring, and control to generate rich, attributable meta-level traces. These traces are then consolidated through a hierarchical, multi-timescale update mechanism that gradually forms evolving meta-knowledge. Experimental results demonstrate consistent performance gains across benchmarks and backbone models, and show that performance improves as metacognitive experience accumulates over time.

[Long] Beyond Noise: Characterizing Creative Potential in Unverifiable LLM Hallucinations ▶ *Underline Video*

Yu Yan, Chunhong Zhang, Haiyu Zhao, Ziyang Zeng, Zihao Liu, Yongkang Wu, Jianzhou Diao, YiJie Chen, Shujie Wang and Zheng Hu

In knowledge-intensive creative tasks, Large Language Models (LLMs) often generate outputs that extend beyond established knowledge, making direct verification against current evidence impractical. Unlike factual hallucinations checked against ground truth, such outputs arise naturally in creative generation, where extending beyond current knowledge is often the goal. Yet prior work debates whether hallucination should be suppressed or embraced without empirically analyzing this unverifiable subclass. On the ideation evaluation side, existing work focuses on individual outputs without characterizing the unverifiable space as a whole. To address this gap, we propose a novelty-verifiability characterization that distinguishes Creative Synthesis (Region A) from Groundless Fabrication (Region B), and study it through a conceptual creation task where LLMs synthesize novel scientific concepts. Through 32,400 generations across three technical domains and 1,080 human judgments, we find that Region~A is non-negligible (4.7%) and robust, persisting across generation strategies, models, domains, and embedding choices. A retrospective recovery experiment further shows that LLMs can approximate post-cutoff scientific concepts in controlled combinatorial settings. Our findings suggest that the unverifiable space is not uniformly noise but exhibits empirically distinguishable internal structure, providing an empirical basis for more selective hallucination governance.

[TACL] Beyond One-Size-Fits-All: Inversion Learning for Highly Effective NLG Evaluation Prompts ▶ *Underline Video*

Chenghua Lin, Hanhua Hong, Chenghao Xiao, Yang Wang, Yiqi Liu and Wenge Rong

Evaluating natural language generation systems is challenging due to the diversity of valid outputs. While human evaluation is the gold standard, it suffers from inconsistencies, lack of standardisation, and demographic biases, limiting reproducibility. LLM-based evaluators offer a scalable alternative but are highly sensitive to prompt design, where small variations can lead to significant discrepancies. In this work, we propose an inversion learning method that learns effective reverse mappings from model outputs back to their input instructions, enabling the automatic generation of highly effective, model-specific evaluation prompts. Our method requires only a single evaluation sample and eliminates the need for time-consuming manual prompt engineering, thereby improving both efficiency and robustness. Our work contributes toward a new direction for more robust and efficient LLM-based evaluation.

[Long] Beyond Pedagogical Principles: Multi-Horizon Preference Optimization for Efficient Socratic Tutoring ▶ *Underline Video*

Xin Shi, Chao Zhang, Yifan Zhu, Xueqiao Zhang and Yawei Luo

The development of LLM-based tutor agents faces challenges in simultaneously ensuring adherence to pedagogical principles and achieving optimal pedagogical effectiveness, particularly in dynamic, multi-turn interactions. Existing methods are often constrained by static data or sparse reward signals in online settings. To address this gap, we propose Multi-Horizon Preference Optimization (MHPO), a novel framework that iteratively refines tutor agents using a multi-horizon reward function within a dynamic teacher-student simulation environment. Specifically, this reward function is designed to capture both turn-level pedagogical quality and trajectory-level pedagogical effectiveness, which is estimated via Monte Carlo rollouts. We further investigate two distinct strategies to aggregate these rewards for policy optimization. Our experiments demonstrate that MHPO significantly enhances base model performance, achieving a superior balance between principles and effectiveness compared to various baselines.

[Long] Beyond Ranking: Fine-Grained Diagnostics and Self-Improvement for MLLMs ▶ *Underline Video*

Mingze Xu, Zijing Zhao, QIMING PENG, Houwen Peng, Han Hu, Zhanhui Kang and Yuxing Han

While Multimodal Large Language Models (MLLMs) are advancing rapidly, accurately evaluating their capabilities remains challenging. Current paradigms primarily rely on holistic scoring and static leaderboards, which fail to disentangle fine-grained competencies. Specifically, they suffer from Outcome Bias by validating only final answers and ignoring intermediate reasoning. To address these limitations, we introduce ATOM (AnaTomy Of MLLM), a novel MLLM-as-a-judge framework designed to shift the focus from ranking to fine-grained diagnosis. ATOM decomposes complex reasoning into atomic criteria anchored in visual elements, enforcing verification against explicit visual facts. Validated on a newly constructed benchmark with rigorous human rankings, ATOM achieves state-of-the-art accuracy, surpassing the strongest baseline by up to 7.92%. Moving beyond ranking, ATOM bridges the gap between assessment and alignment: by pinpointing atomic-level failures, it establishes a closed-loop mechanism for targeted self-correction. This approach enables models to identify and rectify errors autonomously, successfully resolving up to 39.95% of previously failed queries without human intervention.

[Long] Beyond Single View: A Comprehensive Benchmark for Medical Multimodal Large Language Models on Multi-Image Understanding ▶ *Underline Video*

Dexuan Xu, Jiayin Yuan, Jianing Wang, Yanyuan Chen, Hanpin Wang and Yu Huang

Recent advancements in Multimodal Large Language Models (MLLMs) have demonstrated impressive capabilities in interpreting single medical images. However, real-world clinical diagnosis is intrinsically a multi-view process, requiring the synthesis of information across volumetric slices, temporal sequences, and comparative modalities. Existing benchmarks fail to capture this complexity, limiting the assessment of models in realistic clinical workflows. To bridge this gap, we introduce MedMultiBench, the first large-scale benchmark specifically designed for medical multi-image understanding. Comprising 11,392 expert-curated samples, MedMultiBench evaluates MLLMs across four distinct dimensions: Joint Reasoning, Comparative Analysis, Comprehensive Perception, and In-Context Learning. We benchmark 13 state-of-the-art MLLMs, revealing that while current models excel in single-view tasks, they struggle significantly with multi-image contexts. Our experiments identify a performance degradation in open-source models when processing increased visual loads, whereas closed-source models demonstrate better scalability. MedMultiBench provides a robust framework to facilitate the development of MLLMs capable of holistic clinical reasoning.

[SRW] Beyond Static Cropping: Layer-Adaptive Visual Localization and Decoding Enhancement ▶ *Underline Video*

Zipeng Zhu; Zhanghao Hu; Qinglin Zhu; Jingyong Su; Yulan He; Lin Gui

Large Vision-Language Models (LVLMs) have advanced rapidly by aligning visual patches with the text embedding space, but a fixed visual-token budget forces images to be resized to a uniform pretraining resolution, often erasing fine-grained details and causing hallucinations via over-reliance on language priors. Recent attention-guided enhancement (e.g., cropping or region-focused attention allocation) alleviates this, yet it commonly hinges on a static magic layer empirically chosen on simple recognition benchmarks and thus may not transfer to complex reasoning tasks. In contrast to this static assumption, we propose a dynamic perspective on visual grounding. Through a layer-wise sensitivity analysis, we demonstrate that visual grounding is a dynamic process: while simple object recognition tasks rely on middle layers, complex visual search and reasoning tasks require visual information to be reactivated at deeper layers. Based on this observation, we introduce Visual Activation by Query (VAQ), a metric that identifies the layer whose attention map is most relevant to query-specific visual grounding by measuring attention sensitivity to the input query. Building on VAQ, we further propose LASER (Layer-adaptive Attention-guided Selective visual and decoding Enhancement for Reasoning), a inference procedure that adaptively selects task-appropriate layers for visual localization and question answering. Experiments across diverse VQA benchmarks show that LASER significantly improves VQA accuracy across tasks with varying levels of complexity.

[Short] Big AI is Accelerating the Metacrisis: What Can We Do? ▶ *Underline Video*

Steven Bird

The world is in the grip of ecological, meaning, and language crises that are converging into a metacrisis. Big AI is accelerating them all. LLM engineering sits at the core. Despite the public good motives of language engineers and the promise of LLMs, this work is being leveraged to create unprecedented wealth and power for a handful of individuals and corporations while causing existential harm to life on earth. As a profession, we urgently need to come together to explore alternatives and to design a life-affirming future for our field of natural language processing that is centered on human flourishing on a living planet.

[Findings] BizCompass: Benchmarking the Reasoning Capabilities of LLMs in Business Knowledge and Applications ▶ *Underline Video*

Jianing Hao, Yuhe Wu, Yuanjian Xu, Shichang Meng, Shuai Yuan, Wei Zeng, Zixuan Wang and Guang Zhang

Large language models (LLMs) hold great promise for business applications, yet business analysis remains inherently complex, demanding rigorous reasoning and the integration of diverse knowledge sources. Existing benchmarks typically target narrow tasks and thus leave a fundamental question unanswered: how can LLMs be reliably applied in business, and how are these applications grounded in underlying theoretical capabilities? To address this gap, we introduce BizCompass, a benchmark explicitly designed to connect theoretical foundations with practical business knowledge and applications. At the knowledge level, BizCompass covers four core domains, finance, economics, statistics, and operations management. At the application level, it structures tasks around three representative roles: the analyst, the trader, and the consultant. This dual-axis design not only exposes performance differences across realistic scenarios but also diagnoses which foundational capabilities enable or constrain success. We systematically evaluate both open-source and commercial LLMs, revealing how theoretical knowledge translates into practical performance in business. The results provide actionable insights for model selection and training optimization in real-world business contexts. All datasets and evaluation code are publicly released to support reproducibility and future research: https://bizcompass.dev.ypemc.com.

[Long] Bootstrapping Code Translation with Weighted Multilanguage Exploration ▶ *Underline Video*

Yuhan Wu, Huan Zhang, Wei Cheng, Chen Shen, Jingyue Yang and Wei Hu

Code translation across multiple programming languages is essential yet challenging due to two vital obstacles: scarcity of parallel data paired with executable test oracles, and optimization imbalance when handling diverse language pairs. We propose BootTrans, a bootstrapping method that resolves both obstacles. Its key idea is to leverage the functional invariance and cross-lingual portability of test suites, adapting abundant pivot-language unit tests to serve as universal verification oracles for multilingual reinforcement learning (RL) training. Our method introduces a dual-pool architecture with seed and exploration pools to progressively expand training data via execution-guided experience collection. Furthermore, we design a language-aware weighting mechanism that dynamically prioritizes harder translation directions based on relative performance across sibling languages, mitigating optimization imbalance. Extensive experiments

on the HumanEval-X and TransCoder-Test benchmarks demonstrate substantial improvements over baseline LLMs across all translation directions, with ablation studies validating the effectiveness of both bootstrapping and weighting components.

[Long] CamoQuery: Language-Guided Reasoning Camouflaged Object Segmentation ▶ *Underline Video*

Tianxin Han, Qing Dong, Xingwei Wang, Jie Jia, Gang Wu, Bowen Yang and Fu Zhang

Although camouflaged object segmentation has advanced rapidly in recent years, existing methods are still confined to visual mask prediction under fixed task assumptions. They cannot interactively respond to user requests, nor can they proactively understand and reason about the users intent. Our work tackles this issue by proposing a novel task, Language-Guided Reasoning Camouflaged Object Segmentation (LRCOS). Given a camouflaged image and an implicit query text instruction that requires reasoning, LRCOS aims to output intent-consistent segmentation mask. To establish a benchmark for this task, we build CamoQuery, comprising 12,437 image–mask samples and 25971 implicit query text instructions. To better reflect real-world camouflaged scenarios, we additionally collect MCD, a multi-instance camouflage dataset where multiple camouflaged targets co-exist within the same scene, increasing the need for reasoning. Building on CamoQuery, we further propose COSA, a vision–language segmentation assistant that segments the intended camouflaged object from implicit queries and produces a reasoning explanation. Experiments on CamoQuery demonstrate that COSA has strong reasoning segmentation capability in camouflaged scenes and exhibits zero-shot capability.

[Long] Can LLMs Learn to Map the World from Local Descriptions? ▶ *Underline Video*

Sirui Xia, Aili Chen, Xintao Wang, Tinghui Zhu, Yikai Zhang, Jiangjie Chen and Yanghua Xiao

Recent advances in Large Language Models (LLMs) have demonstrated strong capabilities in tasks such as code generation and mathematical reasoning. However, their potential to internalize structured spatial knowledge remains underexplored. This study investigates whether LLMs, grounded in locally relative human observations, can construct coherent global spatial cognition by integrating fragmented relational descriptions. We focus on two core aspects of spatial cognition: spatial perception, where models infer consistent global layouts from local positional relationships, and spatial navigation, where models learn road connectivity from trajectory data and plan optimal paths between unconnected locations. Experiments conducted in a simulated urban environment demonstrate that LLMs not only generalize to unseen spatial relationships between points of interest (POIs) but also exhibit latent representations aligned with real-world spatial distributions. Furthermore, LLMs can learn road connectivity from trajectory descriptions, enabling accurate path planning and dynamic spatial awareness during navigation.

[Findings] Can Small Vision-Language Models Perform Sign Language Translation? ▶ *Underline Video*

Anal Roy Chowdhury and Debarshi Kumar Sanyal

Vision-Language Models (VLMs) have shown strong generalization across multimodal tasks, but their capacity to handle sign language translation (SLT), which requires fine-grained spatiotemporal reasoning and linguistic understanding, remains unclear. In this study, we evaluate whether small VLMs (with <=3B parameters) can perform SLT effectively. We perform supervised fine-tuning on four publicly available multilingual SLT datasets, including one German (DGS), two American (ASL), and one Indian (ISL), applying parameter-efficient LoRA to the language decoder while keeping the vision encoder frozen and training only the connector. To evaluate translation quality, we propose entity- and semantics-aware metrics tailored for SLT. We highlight the data imbalance issues present in the above widely used SLT datasets. Our analysis highlights the limitations in applying general-purpose VLMs to SLT, unlike their applicability in other tasks, and provides insights to inform future development of VLMs for SLP, which is essential for building inclusive AI applications.

[Long] CAP: Controllable Alignment Prompting for Unlearning in LLMs ▶ *Underline Video*

Zhaokun Wang, Jinyu Guo, Jingwen Pu, Hongli Pu, Meng Yang, Xunlei Chen, Jie Ou, Wenyi Li, Guangchun Luo and Wenhong Tian

Large language models (LLMs) trained on unfiltered corpora inherently risk retaining sensitive information, necessitating selective knowledge unlearning for regulatory compliance and ethical safety. However, existing parameter-modifying methods face fundamental limitations: high computational costs, uncontrollable forgetting boundaries, and strict dependency on model weight access. These constraints render them impractical for closed-source models, yet current non-invasive alternatives remain unsystematic and reliant on empirical experience. To address these challenges, we propose the Controllable Alignment Prompting for Unlearning (CAP) framework, an end-to-end prompt-driven unlearning paradigm. CAP decouples unlearning into a learnable prompt optimization process via reinforcement learning, where a prompt generator collaborates with the LLM to suppress target knowledge while preserving general capabilities selectively. This approach enables reversible knowledge restoration through prompt revocation. Extensive experiments demonstrate that CAP achieves precise, controllable unlearning without updating model parameters, establishing a dynamic alignment mechanism that overcomes the transferability limitations of prior methods.

[Long] Cat-MoD: Accelerating Multimodal Alignment via Caption Token Guided Asymmetric Mixture-of-Depths ▶ *Underline Video*

Yijie Huang, Xiaocui Yang, Shi Feng, Wen Zhang, Kaisong Song, Yifei Zhang and Daling Wang

Efficiently aligning visual features with Large Language Models (LLMs) remains a critical bottleneck in Multimodal LLMs. Existing query-based alignment modules (e.g., Q-Former) rely on randomly initialized queries, resulting in an inefficient cold start exploration process. Furthermore, they enforce uniform cross-attention across all layers, leading to computational redundancy. Our empirical analysis reveals that query tokens initialized with language priors can rapidly capture global semantics, leading to early representation convergence after only a few layers. In this paper, we propose Cat-MoD, a Caption token Guided Asymmetric Mixture-of-Depths framework. It incorporates

a Hybrid Query Construction module where Guide Tokens initialized from coarse-grained linguistic priors rapidly anchor global semantic context, and randomly initialized Explorer Tokens remain active to capture fine-grained visual details. Exploiting this early convergence, we introduce an Asymmetric Mixture-of-Depths mechanism, where a similarity-aware router dynamically prunes redundant tokens from expensive cross-attention layers while preserving their context in self-attention. Experiments on multiple benchmarks demonstrate that Cat-MoD matches or surpasses baseline performance, while substantially reducing alignment FLOPs by approximately 37% during both training and inference, offering a highly efficient solution for multimodal alignment. Code: https://github.com/JasonOrange0726/Cat-MoD.

[Long] CE-GPPO: Coordinating Entropy via Gradient-Preserving Clipping Policy Optimization in Reinforcement Learning ▶ *Underline Video*

Zhenpeng Su, Leiyu Pan, Minxuan Lv, Yuntao Li, Wenping Hu, Fuzheng Zhang, Kun Gai and Guorui Zhou

Reinforcement learning (RL) has become a powerful paradigm for optimizing large language models (LLMs) to handle complex reasoning tasks. A core challenge in this process lies in managing policy entropy, which reflects the balance between exploration and exploitation during training. Existing methods, such as proximal policy optimization (PPO) and its variants, discard valuable gradient signals from low-probability tokens due to the clipping mechanism. We systematically analyze the entropy dynamics and reveal that these clipped tokens play a critical yet overlooked role in regulating entropy evolution. We propose Coordinating Entropy via Gradient-Preserving Policy Optimization (CE-GPPO), a novel algorithm that reintroduces gradients from clipped tokens in native PPO in a gentle and bounded manner. By controlling the magnitude of gradients from tokens outside the clipping interval, CE-GPPO is able to achieve an exploration-exploitation trade-off. We provide theoretical justification and empirical evidence showing that CE-GPPO effectively mitigates entropy instability. Extensive experiments on mathematical reasoning benchmarks show that CE-GPPO consistently outperforms strong baselines across different model scales.

[Long] CEBC: Conformal Evidence-Bounded Control for Low-Hallucination Vision-Language Generation ▶ *Underline Video*

Ashish Mishra, Tarun Kumar, Arpit Shah, Suparna Bhattacharya and Martin Foltin

Hallucinated object mentions remain a persistent failure mode of vision-language models (VLMs) across generation tasks such as image captioning and visual question answering: outputs may be fluent yet include entities not supported by visual evidence. Existing mitigation approaches often reduce hallucinations at the cost of degraded generation quality or require expensive retraining and task-specific supervision. We introduce CEBC, a lightweight, training-free framework for low-hallucination vision-language generation based on conformal evidence-bounded minimal editing. CEBC first produces a strong base output (via greedy decoding or best-of-K sampling), then applies an evidence-bounded editing step that minimally revises or suppresses unsupported object mentions using constraints derived from an external visual detector. Crucially, the evidence threshold is conformally calibrated on a small held-out set via quantiles of detector confidence scores, enabling explicit and controllable hallucination risk at test time. To balance factuality and informativeness, we further introduce a risk-first, quality-aware selection rule that prioritizes evidence-consistent generations while regularizing unnecessary length or lexical drift. Extensive experiments on MS-COCO and GQA for image captioning, and POPE for VQA evaluation across multiple VLMs demonstrate that CEBC consistently reduces hallucination rates(CHAIR_S, CHAIR_I, POPE) while maintaining or improving standard generation quality metrics (CIDEr, BLEU, CLIPScore). CEBC establishes a stronger factuality-quality Pareto frontier without any additional model training or access to paired supervision beyond an off-the-shelf detector.

[Long] ChatAnime: Towards User-Centered Emotional Support in LLM-based Virtual Character Chat ▶ *Underline Video*

Lanlan Qiu, Xiao Pu, Yeqi Feng, Wenchang Gao and Tianxing He

With the growing popularity of virtual character platforms like Character.AI, users are increasingly turning to role-playing agents for emotional support in daily life. Yet existing research mainly focuses on character consistency in fictional or game-based scenarios, overlooking user-centered interactions such as companionship and psychological support. To bridge this gap, we propose Emotionally Supportive Role-Playing (ESRP), a framework designed to align role-playing with real-world user scenarios and emotional needs. We focus on typical users of these platforms, i.e., anime enthusiasts, including students, office workers, freelancers, and self-employed individuals, and design scenario-based questions that reflect their everyday struggles such as work stress and social loneliness. Through a two-round data collection involving 40 anime fans and 10 Large Language Models (LLMs), we build ChatAnime: the first ESRP dataset with 2,400 human-written and 24,000 LLM-generated responses, supported by over 132,000 fine-grained human annotations. We also provide the ESRP evaluation framework featuring 9 fine-grained metrics across three dimensions: basic dialogue, role-playing and emotional support, along with an overall metric for diversity. Experimental results under our evaluation setting show that top-performing LLMs surpass anime fans in role-playing and emotional support, while humans still lead in diversity.

[Long] ChatHLS: Towards Systematic Design Automation and Optimization for High-Level Synthesis ▶ *Underline Video*

Runkai Li and Jia Xiong

High-Level Synthesis (HLS) improves IC development productivity by enabling hardware design from C-like languages. However, strict coding constraints and design-specific optimizations limit its widespread adoption. While recent efforts employ large language models (LLMs) to assist HLS design, they often struggle with synthesizability rules and directive semantics. To this end, we introduce ChatHLS, a multi-agent HLS design framework that leverages specialized LLMs for automated debugging and directive tuning. ChatHLS incorporates an adaptive error case expansion mechanism, combined with a reasoning-to-instruction analysis method to accurately diagnose HLS errors. To optimize hardware performance, it enables QoR-aware reasoning to learn the impact of HLS directives on the quality of results (QoR).

Experimental results demonstrate that ChatHLS outperforms Gemini-3-pro with a 32.6% relative improvement in debugging, while achieving significant speedups across various HLS kernels and neural network accelerators. These results underscore the potential of ChatHLS for agile hardware development.

[Long] CheckRLM: Effective Knowledge-Thought Coherence Checking in Retrieval-Augmented Reasoning ▶ *Underline Video*

Dingling Xu, Ruobing Wang, Qingfei Zhao, Yukun Yan, Zhichun Wang, Daren Zha, Shi Yu, Zhenghao Liu, Shuo Wang, Xu Han and Maosong Sun

Reasoning Language Models (RLMs) have significantly improved performance on complex tasks by extending the reasoning chain. However, these chains are prone to containing factual errors, particularly in knowledge-intensive tasks. To address this issue, we propose CheckRLM, a framework that improves the reliability of the reasoning process through Retrieval-Augmented Generation (RAG) by timely checking and correcting factual errors. Specifically, CheckRLM extracts factual claims from the reasoning chain to identify and localize subtle knowledge inconsistencies during inference. Upon detection of errors, a refinement mechanism performs minimal-cost yet precise corrections by leveraging external knowledge, ensuring coherence between the reasoning chain and correct knowledge. Extensive experiments demonstrate that CheckRLM substantially outperforms existing baselines, exhibiting a strong capability to mitigate error accumulation in long-horizon reasoning with lower costs. The code and data are available at https://github.com/AI9Stars/CheckRLM.

[Long] CheckRLM: Effective Knowledge-Thought Coherence Checking in Retrieval-Augmented Reasoning ▶ *Underline Video*

Dingling Xu, Ruobing Wang, Qingfei Zhao, Yukun Yan, Zhichun Wang, Daren Zha, Shi Yu, Zhenghao Liu, Shuo Wang, Xu Han and Maosong Sun

Reasoning Language Models (RLMs) have significantly improved performance on complex tasks by extending the reasoning chain. However, these chains are prone to containing factual errors, particularly in knowledge-intensive tasks. To address this issue, we propose CheckRLM, a framework that improves the reliability of the reasoning process through Retrieval-Augmented Generation (RAG) by timely checking and correcting factual errors. Specifically, CheckRLM extracts factual claims from the reasoning chain to identify and localize subtle knowledge inconsistencies during inference. Upon detection of errors, a refinement mechanism performs minimal-cost yet precise corrections by leveraging external knowledge, ensuring coherence between the reasoning chain and correct knowledge. Extensive experiments demonstrate that CheckRLM substantially outperforms existing baselines, exhibiting a strong capability to mitigate error accumulation in long-horizon reasoning with lower costs. The code and data are available at https://github.com/AI9Stars/CheckRLM.

[Findings] Chinese LiveStreaming ECommerce Morph Resolution: Datasets and Methods ▶ *Underline Video*

Jipeng Qiang, Jiahao Zhu, Yi Zhu and Chaowei Zhang

Live-stream E-commerce faces significant challenges from morphs, deliberate linguistic variants used to evade real-time voice filters and amplify product claims illegally. While critical for regulatory enforcement, Live Auditory Morph Resolution (LiveAMR) research is hindered by limited datasets: prior work relied on narrow, redundant health domain corpora, restricting model robustness. To bridge this gap, we introduce two datasets: (1) HealthAMR, a refined health-domain corpus via deduplication and re-annotation. (2) GeneralAMR, a general domain benchmark with 28K annotated sentences from 77 channels across 7 E-commerce categories. Further, we propose JointMRE, a multi-task framework that jointly resolves morphs and generates structured explanations, transferring grammatical insights from large language models to enhance generalization. Predictions are refined by our Conflict-aware Dual-output Refinement Framework (CDRF), which detects inconsistencies between corrections and explanations. Experiments show CDRF significantly improves morph resolution accuracy and interpretability. Our datasets and code are available.

[Long] CIRAG: Construction-Integration Retrieval and Adaptive Generation for Multi-hop Question Answering ▶ *Underline Video*

Zili Wei, Yilin Wang, Xiaocui Yang, Shi Feng, Weidong Bao, Daling Wang, Zihan Wang and Yifei Zhang

Triple-based Iterative Retrieval-Augmented Generation (iRAG) mitigates document-level noise for multi-hop question answering. However, existing methods still face limitations: (i) greedy single-path expansion, which propagates early errors and fails to capture parallel evidence from different reasoning branches, and (ii) granularity-demand mismatch, where a single evidence representation struggles to balance noise control with contextual sufficiency. In this paper, we propose the Construction-Integration Retrieval and Adaptive Generation model, CIRAG. It introduces an Iterative Construction-Integration module that constructs candidate triples and history-conditionally integrates them to distill core triples and generate the next-hop query. This module mitigates the greedy trap by preserving multiple plausible evidence chains. Besides, to address the granularity-demand mismatch, we propose an Adaptive Cascaded Multi-Granularity Generation module that progressively expands contextual evidence based on the problem requirements, from triples to supporting sentences and full passages. Moreover, we introduce Trajectory Distillation, which distills the teacher models integration policy into a lightweight student, enabling efficient and reliable long-horizon reasoning. Extensive experiments demonstrate that CIRAG achieves superior performance compared to existing iRAG methods.

[Long] CityCube: Benchmarking Cross-view Spatial Reasoning on Vision-Language Models in Urban Environments ▶ *Underline Video*

Haotian Xu, Yue Hu, Zhengqiu Zhu, Chen Gao, Ziyou Wang, Junreng Rao, Wenhao Lu, Weishi Li, Quanjun Yin and Yong Li

Cross-view spatial reasoning is essential for embodied AI, underpinning spatial understanding, mental simulation and planning in complex environments. Existing benchmarks primarily emphasize indoor or street settings, overlooking the unique challenges of open-ended urban

spaces characterized by rich semantics, complex geometries, and view variations. To address this, we introduce CityCube, a systematic benchmark designed to probe cross-view reasoning capabilities of current VLMs in urban settings. CityCube integrates four viewpoint dynamics to mimic camera movements and spans a wide spectrum of perspectives from multiple platforms, e.g., vehicles, drones and satellites. For a comprehensive assessment, it features 5,022 meticulously annotated multi-view QA pairs categorized into five cognitive dimensions and three spatial relation expressions. A comprehensive evaluation of 33 VLMs reveals a significant performance disparity with humans: even large-scale models struggle to exceed 54.1% accuracy, remaining 34.2% below human performance. By contrast, small-scale fine-tuned VLMs achieve over 60.0% accuracy, highlighting the necessity of our benchmark. Further analyses indicate the task correlations and fundamental cognitive disparity between VLMs and human-like reasoning.

[Long] CodeFlowBench: A Multi-turn, Iterative Benchmark for Complex Code Generation ▶ *Underline Video*

Sizhe Wang, Zhengren Wang, Dongsheng Ma, Yongan Yu, Rui ling, Zhiyu li, Feiyu Xiong and Wentao Zhang

Modern software development demands code that is maintainable, testable, and scalable by organizing the implementation into modular components with iterative reuse of existing codes. We formalize this iterative, multi-turn paradigm as codeflow and introduce CodeFlowBench, the first benchmark designed to comprehensively evaluate LLMs ability to perform codeflow - implementing new functionality by reusing existing functions over multiple turns. CodeFlowBench comprises two complementary components: CodeFlowBench-Comp, a core collection of 5,000+ competitive programming problems from Codeforces updated via an automated pipeline and CodeFlowBench-Repo, which is sourced from GitHub repositories to better reflect real-world scenarios. Furthermore, a novel evaluation framework featured dual assessment protocol and structural metrics derived from dependency trees is introduced. Extensive experiments reveal significant performance degradation in multi-turn codeflow scenarios. Furthermore, our in-depth analysis illustrates that model performance inversely correlates with dependency complexity. These findings not only highlight the critical challenges for supporting real-world workflows, but also establish CodeFlowBench as an essential tool for advancing code generation research.

[Long] CodeFlowBench: A Multi-turn, Iterative Benchmark for Complex Code Generation ▶ *Underline Video*

Sizhe Wang, Zhengren Wang, Dongsheng Ma, Yongan Yu, Rui ling, Zhiyu li, Feiyu Xiong and Wentao Zhang

Modern software development demands code that is maintainable, testable, and scalable by organizing the implementation into modular components with iterative reuse of existing codes. We formalize this iterative, multi-turn paradigm as codeflow and introduce CodeFlowBench, the first benchmark designed to comprehensively evaluate LLMs ability to perform codeflow - implementing new functionality by reusing existing functions over multiple turns. CodeFlowBench comprises two complementary components: CodeFlowBench-Comp, a core collection of 5,000+ competitive programming problems from Codeforces updated via an automated pipeline and CodeFlowBench-Repo, which is sourced from GitHub repositories to better reflect real-world scenarios. Furthermore, a novel evaluation framework featured dual assessment protocol and structural metrics derived from dependency trees is introduced. Extensive experiments reveal significant performance degradation in multi-turn codeflow scenarios. Furthermore, our in-depth analysis illustrates that model performance inversely correlates with dependency complexity. These findings not only highlight the critical challenges for supporting real-world workflows, but also establish CodeFlowBench as an essential tool for advancing code generation research.

[Findings] Cogito: A Cognitive Agentic Framework Driven by Dynamic Graph of Thoughts for Financial Report Generation ▶ *Underline Video*

Chen Lifan, Wei Ding, Jingwen Yang, Xiuze Zhou, Jingan Chen and Fan Lin

Financial report generation is a complex task that requires gathering and reasoning over multi-source information. Recent advances in Large Language Models have made them a promising solution for automating this process. However, the reasoning paths in traditional Chain-of-Thought paradigms are inherently constrained by predefined, static computational topologies, rendering them ill-equipped to handle the dynamic uncertainties of real-world financial environments. To tackle this challenge, we propose Cogito, a cognitively grounded agentic framework for professional financial report generation. At its core, Cogito is driven by Dynamic Graph of Thoughts, a novel reasoning mechanism that models the agent's reasoning process as an evolving topology for adaptive exploration.We further introduce a Social Collaboration Mechanism to facilitate coordinated agent interaction. Finally, Cogito is instantiated as a multi-agent system, where four specialized agents collaboratively execute the end-to-end report generation task. Extensive experiments on enterprise- and industry-level financial report generation benchmarks demonstrate the superiority of Cogito in data quality, analytical validity, and presentation quality.

[Long] Cognitive Policy-Driven LLM for Diagnosis and Intervention of Cognitive Distortions in Emotional Support Conversation ▶ *Underline Video*

Lin Zhong, Renjin Zhu, Shujuan Ma, Jinhao Cui, Lingzhi Wang, Hao Chen and Qing Liao

Emotional Support Conversation (ESC) plays a critical role in mental health assistance by providing accessible psychological support in real-world applications. Large Language Models (LLMs) have shown strong empathetic abilities in ESC tasks. Yet, existing methods overlook the issue of cognitive distortions in help-seekers expressions. As a result, current models can only provide basic emotional comfort, rather than helping help-seekers address their psychological distress at a deeper cognitive level. To address this challenge, we construct the CogBiasESC dataset, the first dataset that expands existing ESC datasets by adding labels for cognitive distortions, includes their type, intensity, and safe risk level. Furthermore, we propose the Cognitive Policy-driven Large Language Model framework (CoPoLLM) to enhance LLMs' ability to diagnose and intervene cognitive distortions in help-seekers. We also analyze the safety advantages of CoPoLLM from a theoretical perspective. Experimental results show that CoPoLLM significantly outperforms 15 state-of-the-art baselines

in terms of distortion diagnosis accuracy, intervention strategy effectiveness, and safety risk control. Our source code is available at: https://github.com/Chips98/CoPoLLM-for-ACL-2026.

[Long] CoMeT: Collaborative Memory Transformer for Efficient Long Context Modeling ▶ *Underline Video*

Runsong Zhao, Shilei Liu, Jiwei Tang, Langming Liu, Haibin Chen, Weidong Zhang, Yujin Yuan, Tong Xiao, Jingbo Zhu, Wenbo Su and Bo Zheng

The quadratic complexity and indefinitely growing key-value (KV) cache of standard Transformers pose a major barrier to long-context processing. To overcome this, we introduce the Collaborative Memory Transformer (CoMeT), a novel architecture that enables LLMs to handle arbitrarily long sequences with constant memory usage and linear time complexity. Designed as an efficient, plug-in module, CoMeT can be integrated into pre-trained models with only minimal fine-tuning. It operates on sequential data chunks, using a dual-memory system to manage context: a temporary memory on a FIFO queue for recent events, and a global memory with a gated update rule for long-range dependencies. These memories then act as a dynamic soft prompt for the next chunk. The effectiveness of our approach is remarkable: a model equipped with CoMeT and fine-tuned on 32k contexts can accurately retrieve a passkey from any position within a 1M token sequence. On the SCROLLS benchmark, CoMeT surpasses other efficient methods and achieves performance comparable to a full-attention baseline on summarization tasks. Its practical effectiveness is further validated on real-world agent and user behavior QA tasks, supported by a novel layer-level pipeline parallel training strategy that enables fine-tuning on extremely long contexts. The code is available at: https://github.com/LivingFutureLab/Comet

[Long] Controlling Multimodal Conversational Agents with Coverage-Enhanced Latent Actions ▶ *Underline Video*

Yongqi Li, Hao Lang, Tieyun Qian and Yongbin Li

Vision-language models are increasingly employed as multimodal conversational agents (MCAs) for diverse conversational tasks. Recently, reinforcement learning (RL) has been widely explored for adapting MCAs to various human-AI interaction scenarios. Despite showing great enhancement in generalization performance, fine-tuning MCAs via RL still faces challenges in handling the extremely large text token space. To address this, we learn a compact latent action space for RL fine-tuning instead. Specifically, we adopt the learning from observation mechanism to construct the codebook for the latent action space, where future observations are leveraged to estimate current latent actions that could further be used to reconstruct future observations. However, the scarcity of paired image-text data hinders learning a codebook with sufficient coverage. Thus, we leverage both paired image-text data and text-only data to construct the latent action space, using a cross-modal projector for transforming text embeddings into image-text embeddings. We initialize the cross-modal projector on paired image-text data, and further train it on massive text-only data with a novel cycle consistency loss to enhance its robustness. We show that our latent action based method outperforms competitive baselines on two conversation tasks across various RL algorithms. Code and data are available at https://github.com/AlibabaResearch/DAMO-ConvAI/tree/main/MMLatentAction.

[Long] COSMOS: Connectivity-Oriented Submodular Maximization for Optimal Subgraph Retrieval ▶ *Underline Video*

Boci Peng, Xiao Liu, Boren Hu, Yun Zhu, Xuanbo Fan, Yanwei Yue, Chunyu Yang and Yan Zhang

Retrieving coherent evidence subgraphs is critical for Knowledge Base Question Answering (KBQA). Existing paradigms often treat facts independently, rely on biased heuristics, or employ myopic search, failing to optimize collective subgraph utility. In this paper, we propose COSMOS (Connectivity-Oriented Submodular Maximization for Optimal Subgraph Retrieval), a unified framework that formalizes evidence retrieval as a constrained submodular maximization problem. This formulation mathematically captures the trade-off between information relevance and structural complexity. To tractably solve this combinatorial challenge, COSMOS employs a decompose-and-conquer strategy, which first performs a seed-guided greedy expansion to maximize local semantic utility, followed by a topology-aware component aggregation to bridge disjoint evidence clusters via Maximum Spanning Tree aggregation. Guided by theoretical bounds, we introduce Structure-Aware Contrastive Tuning to align semantic space with KG topology. Experimental results on WebQSP, CWQ, and M^3GQA benchmarks demonstrate that COSMOS achieves state-of-the-art performance.

[Findings] CoV: Chain-of-View Prompting for Spatial Reasoning ▶ *Underline Video*

Haoyu Zhao, Akide Liu, Zeyu Zhang, Weijie Wang, Feng Chen, Ruihan Zhu, Gholamreza Haffari and Bohan Zhuang

Embodied question answering (EQA) in 3D environments often requires collecting context that is distributed across multiple viewpoints and partially occluded. However, most recent vision–language models (VLMs) are constrained to a fixed and finite set of input views, which limits their ability to acquire question-relevant context at inference time and hinders complex spatial reasoning. We propose Chain-of-View (CoV) prompting, a training-free, test-time reasoning framework that transforms a VLM into an active viewpoint reasoner through a coarse-to-fine exploration process. CoV first employs a View Selection agent to filter redundant frames and identify question-aligned anchor views. It then performs fine-grained view adjustment by interleaving iterative reasoning with discrete camera actions, obtaining new observations from the underlying 3D scene representation until sufficient context is gathered or a step budget is reached. We evaluate CoV on OpenEQA across four mainstream VLMs and obtain an average 11.98% improvement in LLM-Match, with a maximum gain of 13.62% on Qwen3-VL-Flash. CoV further exhibits test-time scaling: increasing the minimum action budget yields an additional 2.54% average improvement, peaking at 3.73% on Gemini-2.5-Flash. On ScanQA and SQA3D, CoV delivers strong performance (e.g., 116 CIDEr 31.9 EM@1 on ScanQA and 51.1 EM@1 on SQA3D). Overall, these results suggest that question-aligned view selection coupled with open-view search is an effective, model-agnostic strategy for improving spatial reasoning in 3D EQA without additional training.

[Long] CriticLean: Critic-Guided Reinforcement Learning for Mathematical Formalization ▶ *Underline Video*

Zhongyuan Peng, Yifan Yao, Kaijing Ma, Shuyue Guo, Yizhe Li, Yichi Zhang, Chenchen Zhang, Yifan Zhang, Zhouliang Yu, Luming Li, Minghao Liu, Yihang Xia, Jiawei Shen, Yuchen Wu, Yixin Cao, Zhaoxiang Zhang, Wenhao Huang, Jiaheng Liu and Ge Zhang

Translating natural language mathematical statements into formal, executable code is a fundamental challenge in automated theorem proving. While prior work has focused on generation and compilation success, little attention has been paid to the critic phase, the evaluation of whether generated formalizations truly capture the semantic intent of the original problem. In this paper, we introduce CriticLean, a novel critic-guided reinforcement learning framework that elevates the role of the critic from a passive validator to an active learning component. Specifically, first, we propose the CriticLeanGPT, trained via supervised fine-tuning and reinforcement learning, to rigorously assess the semantic fidelity of Lean 4 formalizations. Then, we introduce CriticLeanBench, a benchmark designed to measure models' ability to distinguish semantically correct from incorrect formalizations, and demonstrate that our trained CriticLeanGPT models can significantly outperform strong open- and closed-source baselines. Building on the CriticLean framework, we construct FineLeanCorpus, a dataset comprising over 509K problems that exhibits rich domain diversity, broad difficulty coverage, and high correctness based on human evaluation. Overall, our findings highlight that optimizing the critic phase is essential for producing reliable formalizations and we hope our CriticLean will provide valuable insights for future advances in formal mathematical reasoning.

[Long] Cultivating Forensic Reasoning for Generalizable Multimodal Manipulation Detection ▶ *Underline Video*

Yuchen Zhang, Yaxiong Wang, Kecheng Han, Yujiao Wu, Lianwei Wu, Li Zhu and Zhedong Zheng

Recent advances in generative AI have significantly enhanced the realism of multimodal media manipulation, thereby posing substantial challenges to manipulation detection. Existing manipulation detection and grounding approaches predominantly focus on manipulation type classification under result-oriented supervision, which not only lacks interpretability but also tends to overfit superficial artifacts. In this paper, we argue that generalizable detection requires incorporating explicit forensic reasoning, rather than merely classifying a limited set of manipulation types, which fails to generalize to unseen manipulation patterns. To this end, we propose REFORM, a reasoning-driven framework that shifts learning from outcome fitting to process modeling. REFORM adopts a three-stage curriculum that first induces forensic rationales, then aligns reasoning with final judgments, and finally refines logical consistency via reinforcement learning. To support this paradigm, we introduce ROM, a large-scale dataset with rich reasoning annotations. Extensive experiments show that REFORM establishes new state-of-the-art performance with superior generalization, achieving 81.52% ACC on ROM, 76.65% ACC on DGM4, and 74.9 F1 on MMFakeBench.

[Findings] CVRH: Cross-modal Variational Role Hypergraph Network via Semantic Enhancement for Multi-modal Event Argument Extraction ▶ *Underline Video*

Bangze Pan;Yang Li, Ruili Pu, Suge Wang, Jian Liao;JianXing Zheng, Xiaoli Li and Deyu Li

Multi-modal Event Argument Extraction task (MEAE) aims to extract all arguments related to a specific event from multiple modalities and identify their corresponding roles. Existing methods focus on weakly alignment of uni-modal representations and generatively data augmentation techniques. However, these methods ignore the potential impact of event role information on MEAE. To address this problem, we propose a Cross-modal Variational Role Hypergraph Network via Semantic Enhancement (CVRH). Unlike previous approaches, CVRH centers on event role information and designs a variational role hyperedge via semantic enhancement, which constructs a role hypergraph for event arguments within multi-modal documents. It explicitly modeling the high-order role correlations among cross-modal arguments in a document. Furthermore, CVRH introduces a modal shared encoder based on differential transformer, which effectively learns shared semantic representations across modalities and enhances the independence of argument representations. On the M2E2 benchmark, experimental results show that CVRH achieves a 6.9% improvement in F1-score on the MEAE compared to current state-of-the-art methods.

[Findings] D-Artemis: A Deliberative Cognitive Framework for Mobile GUI Multi-Agents ▶ *Underline Video*

Hongze Mi, Yibo Feng, WenJie Lu, Yuqi Wang, Jinyuan Li, Song Cao, He Cui, Tengfei Tian, Xuelin Zhang, Haotian Luo, Di Sun, Jun Fang, Hua Chai, Naiqiang Tan and Gang Pan

Graphical User Interface (GUI) agents aim to automate a wide spectrum of human tasks by emulating user interaction. Despite rapid advancements, current approaches are hindered by several critical challenges: data bottleneck in end-to-end training, high cost of delayed error detection, and risk of contradictory guidance. Inspired by the human cognitive loop of Thinking, Alignment, and Reflection, we present D-Artemis, a novel deliberative framework in this paper. D-Artemis leverages a fine-grained, app-specific tip retrieval mechanism to inform its decision-making process. It also employs a proactive Pre-execution Alignment stage, where Thought-Action Consistency (TAC) Check module and Action Correction Agent (ACA) work in concert to mitigate the risk of execution failures. A post-execution Status Reflection Agent (SRA) completes the cognitive loop, enabling strategic learning from experience. Crucially, D-Artemis enhances the capabilities of general-purpose Multimodal large language models (MLLMs) for GUI tasks without the need for training on complex trajectory datasets, demonstrating strong generalization. D-Artemis achieves SOTA among open-source general models on AndroidWorld (75.8%) and ScreenSpot-V2 (96.8%). Extensive ablation studies further demonstrate the significant contribution of each proposed component.

[Long] Debate-of-Thoughts: Resolving Knowledge Conflicts in LLMs Through Internal Deliberation ▶ *Underline Video*

Guocong Li, Qirui Hu, Ping Wang, Guofeng Zhang, Jian Wu and Hongxia Xu

Large Language Models enhanced with Retrieval Augmented Generation show strong potential in knowledge intensive tasks. However, they often encounter knowledge conflicts, where retrieved information contradicts the models internal knowledge or exhibits internal inconsisten-

cies. Existing methods treat this as a simplistic binary choice, forcing models to blindly trust external contexts or rigidly rely on memory, resulting in unreliable predictions that swing between sycophancy and stubbornness. We argue that a more principled approach is to embrace contradictions as opportunities for deeper reasoning. To this end, we introduce Debate-of-Thoughts (DoT), a framework that transforms conflict resolution into an active deliberation process. DoT guides a single model through three phases: 1) hypothesis generation, which forms competing perspectives; 2) internal debate, where the model acts as both a proponent and a critic to stress test each view; and 3) adjudication, where a judge module evaluates arguments based on evidence and logical consistency. We implement DoT via two complementary strategies: inference time prompt chaining and supervised fine tuning. Experiments across multiple conflict benchmarks show that DoT consistently outperforms state-of-the-art methods, while generating transparent debate transcripts that explain its decisions. By improving both accuracy and interpretability under knowledge conflicts, DoT establishes a more reliable paradigm for retrieval augmented generation systems.

[Industry] Decoding Text Spans for Efficient and Accurate Named-Entity Recognition ▶ *Underline Video*

Andrea Maracani, Savas Ozkan, Junyi Zhu, Sinan Mutlu and Mete Ozay

Named Entity Recognition (NER) is a key component in industrial information extraction pipelines, where systems must satisfy strict latency and throughput constraints in addition to strong accuracy. State-of-the-art NER accuracy is often achieved by span-based frameworks, which construct span representations from token encodings and classify candidate spans. However, many span-based methods enumerate large numbers of candidates and process each candidate with marker-augmented inputs, substantially increasing inference cost and limiting scalability in large-scale deployments. In this work, we propose SpanDec, an efficient span-based NER framework that targets this bottleneck. Our main insight is that span representation interactions can be computed effectively at the final transformer stage, avoiding redundant computation in earlier layers via a lightweight decoder dedicated to span representations. We further introduce a span filtering mechanism during enumeration to prune unlikely candidates before expensive processing. Across multiple benchmarks, SpanDec matches competitive span-based baselines while improving throughput and reducing computational cost, yielding a better accuracy–efficiency trade-off suitable for high-volume serving and on-device applications.

[Long] DefGen-Bench: A Benchmark for Chinese Criminal Defence Opinion Generation in LegalAI ▶ *Underline Video*

Senbo Zhang, Qiqi Wang, Fanghao Lou, Guanyu Chen, Yihong Pan, Huijia Li and Qian Liu

A defence opinion is an essential step in criminal proceedings, yet it has not been systematically formulated or evaluated as a specific LegalAI task. Grounded in legal principles and practice, we formulate this task as generating a structured defence opinion conditioned jointly on an indictment and the defendants stated opinion, which often present conflicting claims. We formalize this setting as a dual-perspective generation problem and introduce DefGen-Bench, a benchmark comprising several Chinese criminal cases with expert-reviewed reference defence opinions. We evaluate eight large language models (LLMs) on this task and observe that existing models tend to mirror the defendants opinion, thereby overlooking more appropriate defence strategies. To address this challenge, we propose Knowledge-Enhanced Highlighted Indictment (KHI), a legal knowledge-guided input enhancement method applicable to both open- and closed-source LLMs. Experiments demonstrate consistent improvements across all evaluated LLMs, validating the effectiveness of the proposed approach.

[Long] DetectRL-X: Towards Reliable Multilingual and Real-World LLM-Generated Text Detection ▶ *Underline Video*

Junchao Wu, Yefeng Liu, Chenyu Zhu, Hao Zhang, Zeyu Wu, Tianqi Shi, Yichao Du, Longyue Wang, Weihua Luo, Jinsong Su and Derek F. Wong

The effective detection and governance of Large Language Model (LLM) generated content has become increasingly critical due to the growing risk of misuse. Despite the impressive performance of existing detectors, their reliability and potential in multilingual, real-world scenarios remain largely underexplored. In this study, we introduce DetectRL-X, a comprehensive multilingual benchmark designed to evaluate advanced detectors across 8 dimensions. The benchmark encompasses 8 languages commonly used in commercial contexts and collects human-written texts from 6 domains highly susceptible to LLM misuse. To better aligned with real-world applications, We create LLM-generated texts using 4 popular commercial LLMs, and include typical AI-assisted writing operations such as polishing, expanding, and condensing to capture authentic usage patterns. Furthermore, we develop a multilingual framework for paraphrasing and perturbation attacks to simulate diverse human modifications and writing noise, enabling stress testing of detectors across languages. Experimental results on DetectRL-X reveal the strengths and limitations of current state-of-the-art detectors when applied to diverse linguistic resources. We further analyze how domains, generators, attack strategies, text length, and refinement operations influence performance in different languages, underscoring DetectRL-X as an effective benchmark for strengthening multilingual and language-specific detectors.

[Long] Diagnosing and Remedying Representation Deficiencies for Deterministic Reasoning in KGQA ▶ *Underline Video*

Gewen Liang; Mufan Xu; Kehai Chen; Wei Wang; Yuwei Wang; Muyun Yang; Tiejun Zhao; Min Zhang;

Large language models (LLMs) have demonstrated increasingly strong reasoning capabilities, achieving remarkable progress in knowledge graph question answering (KGQA). However, a key challenge in such systems is non-deterministic reasoning, where the model indecisively activates multiple semantically related knowledge graph edges for a given query, frequently leading to incorrect answers. To address this issue, we propose Diagnosing and Remedying Representation Deficiencies for Deterministic Reasoning in KGQA (DR2). DR2 identifies and localizes non-deterministic reasoning behaviors, uncovering the underlying semantic representation deficiencies in LLMs. Building on this diagnosis, we design abductive reasoning-based preference learning, which promotes fine-grained semantic discrimination and mitigates non-deterministic reasoning errors. Experimental results demonstrate that the proposed DR2 significantly outperforms several strong baselines, achieving state-of-the-art performance on the widely used WebQSP and CWQ benchmarks.

[Industry] Diffusion-Pretrained Dense and Contextual Embeddings ▶ *Underline Video*

Sedigheh Eslami, Maksim Gaiduk, Markus Krimmel, Louis Mark Milliken, Bo Wang and Denis Bykov

We introduce anonym-embed, a family of multilingual embedding models that employ multi-stage contrastive learning on a diffusion-pretrained language model backbone for web-scale retrieval. By leveraging bidirectional attention through diffusion-based pretraining, our models capture comprehensive bidirectional context within passages, enabling the use of mean pooling to better preserve global context across long documents. We release anonym-embed-v1 for standard retrieval, and anonym-embed-context-v1 for contextualized embeddings that incorporate global document context into passage representations. anonym-embed-v1 achieves competitive performance on the MTEB(Multilingual, v2), MTEB(Code), BERGEN, and ToolRet retrieval benchmarks, while anonym-embed-context-v1 sets new records on the ConTEB benchmark.

[Long] DisCo_Speech: Controllable Zero-Shot Speech Generation with A Disentangled Speech Codec ▶ *Underline Video*

Tao Li, Wenshuo Ge, Zhichao Wang, Zihao Cui, Yong Ma, Yingying Gao, Chao Deng, Shilei Zhang and Junlan Feng

Codec-based language models (LMs) have revolutionized text-to-speech (TTS). However, standard codecs entangle timbre and prosody, which hinders independent control in continuation-based LMs. To tackle this challenge, we propose DisCo-Speech, a zero-shot controllable TTS framework featuring a disentangled speech codec (DisCodec) and an LM-based generator. The core component DisCodec employs a two-stage design: 1) tri-factor disentanglement to separate speech into content, prosody, and timbre subspaces via parallel encoders and hybrid losses; and 2) fusion and reconstruction that merges content and prosody into unified content-prosody tokens suitable for LM prediction, while jointly optimizing reconstruction to address the disentanglement-reconstruction trade-off. This allows the LM to perform prosodic continuation from a style prompt while the decoder injects target timbre, enabling flexible zero-shot control. Experiments demonstrate that DisCo-Speech achieves competitive voice cloning and superior zero-shot prosody control. By resolving the core entanglement at the codec level, DisCo-Speech provides a robust foundation for controllable speech synthesis. Audio samples are available at: https://disco-speech.github.io/DisCo-demo/. Code and weights will be released at: https://github.com/disco-speech/DisCo-Speech upon acceptance.

[Long] DocLens: A Tool-Augmented Multi-Agent Framework for Long Visual Document Understanding ▶ *Underline Video*

Dawei Zhu, Rui Meng, Jiefeng Chen, Sujian Li, Tomas Pfister and Jinsung Yoon

Comprehending long visual documents, where information is distributed across extensive pages of text and visual elements, is a critical but challenging task for modern Vision-Language Models (VLMs). Existing approaches falter on a fundamental challenge: evidence localization. They struggle to retrieve relevant pages and overlook fine-grained details within visual elements, leading to limited performance and model hallucination. To address this, we propose DocLens, a tool-augmented multi-agent framework that effectively "zooms in" on evidence like a lens. It first navigates from the full document to specific visual elements on relevant pages, then employs a sampling-adjudication mechanism to generate a single, reliable answer. Paired with Gemini-2.5-Pro, DocLens achieves state-of-the-art performance on MMLongBench-Doc and FinRAGBench-V, surpassing even human experts. The framework's superiority is particularly evident on vision-centric and unanswerable queries, demonstrating the power of its enhanced localization capabilities.

[Long] DPDV: Dual-Pathway and Dual-View Representation Learning for Bridging Information Asymmetry in Text-Video Retrieval ▶ *Underline Video*

Zequn Xie, Xin Liu, Fangming Feng, Boyun Zhang and Tao Jin

In recent years, CLIP-based text-video retrieval methods have developed rapidly, with research focusing on constructing diverse features and achieving effective interactions. However, the asymmetry of cross-modal information poses a challenge to accurately establishing retrieval relationships. To overcome this challenge, we propose a novel video retrieval framework, termed the Dual-Pathway and Dual-View model (DPDV), which consists of the Dual-Pathway Partitioning Module (DPPM) for constructing features at an appropriate granularity and the Dual-View Interaction Module (DVIM) for performing effective feature interactions. For DPPM, we simulate a human macro-level cognitive perspective by partitioning visual features into two categories based on their relevance to the text query and supplementing less relevant features with additional textual information. For DVIM, we simulate a human alignment strategy from macro to micro levels, focusing on local visual features while comprehensively modeling fine-grained interactions. We evaluate DPDV on five benchmark datasets, achieving leading retrieval performance.

[Long] DUAL RM: Beyond Rule-based Preference Reward Modeling via Meta-Reward ▶ *Underline Video*

Xiaobo Liang, Wanfu Wang, Qipeng Huang, Yuyang Ding, Zecheng Tang, Yixin Ji, Qianben Chen, Zhe Zhao, Kehai Chen, Juntao Li and Min Zhang

The ability to model sparse and underspecified rewards, characteristic of human preferences, is fundamental to scaling Reinforcement Learning (RL). Current preference-based reward modeling largely relies on verifiable rewards, where human-annotated labels define rule-based signals. However, these methods face a fundamental bottleneck we term the Matryoshka Doll Problem: a recursive dependency where each reward verifier requires a meta-verifier, leading to continuous and costly dependence on human annotation. In this work, we propose Dual RM, which couples discriminative and generative reward models (DisRMs and GenRMs) under a non-parametric meta-reward. Rather than verifying the correctness of GenRMs reasoning, the meta-reward evaluates its practical impact on response quality. Specifically, GenRM identifies multi-dimensional evaluation rubrics and iteratively refines the response, while DisRM quantifies the quality shifts induced

by each rubric. Furthermore, we implement rubric-based test-time scaling to improve sample efficiency and preference alignment under both DPO and GRPO. Our experiments demonstrate that Dual RM achieves strong performance across major preference benchmarks. Notably, even when trained exclusively on language modality, it exhibits robust cross-modal transfer on Omni-RewardBench.

[Long] Dynamics of Cognitive Heterogeneity: Investigating Behavioral Biases in Multi-Stage Supply Chains with LLM-Based Simulation ▶ *Underline Video*

Jiuyun Jiang, Yuecheng Hong, Bo Yang, Jin Yang, Guangxin Jiang, Xiaomeng Guo and Guang Xiao

Modeling coordination among generative agents in complex multi-round decision-making presents a core challenge for AI and operations management. Although behavioral experiments have revealed cognitive biases behind supply chain inefficiencies, traditional methods face scalability and control limitations. We introduce a scalable experimental paradigm using Large Language Models (LLMs) to simulate multi-stage supply chain dynamics. Grounded in a Hierarchical Reasoning Framework, this study specifically analyzes the impact of cognitive heterogeneity on agent interactions. Unlike prior homogeneous settings, we employ DeepSeek and GPT agents to systematically vary reasoning sophistication across supply chain tiers. Through rigorously replicated and statistically validated simulations, we investigate how this cognitive diversity influences collective outcomes. Results indicate that agents exhibit myopic and self-interested behaviors that exacerbate systemic inefficiencies. However, we demonstrate that information sharing effectively mitigates these adverse effects. Our findings extend traditional behavioral methods and offer new insights into the dynamics of AI-enabled organizations. This work underscores both the potential and limitations of LLM-based agents as proxies for human decision-making in complex operational environments.

[Long] EDSD: Entropy-Driven Design for Faster Speculative Decoding ▶ *Underline Video*

Longkai Cheng, Ximing Wang, Jiangcai Zhu, Kailai Shao, Chao Chen and Haixiang Hu

Speculative decoding has emerged as a promising paradigm for accelerating large language model inference by leveraging a lightweight draft model to generate multiple candidate tokens. However, existing methods often incur substantial training overhead to mitigate information misalignment between autoregressive draft model training and decoding. To address this challenge, we propose EDSD, an Entropy-Driven Speculative Decoding framework that uses entropy as a unified, interpretable signal for both draft model training and architectural design. EDSD drives the draft model to progressively align with the target model in an easy-to-hard manner while establishing token-level alignment as a dominant design principle. Extensive experiments on seven LLMs demonstrate that EDSD improves training efficiency by 24.8%, increases the average acceptance length by 4.0%, and achieves a 4.1% speedup compared to state-of-the-art methods. Furthermore, EDSD improves robustness to system prompt variations by more than 5x. Our findings establish entropy-driven alignment as an effective and principled foundation for efficient speculative decoding.

[Long] Efficiently Learning To Reason or Not to Reason: Root-token Policy Optimization for Adaptive Thinking ▶ *Underline Video*

Taehyeon Kim, Hyunsoo Lee, Youngsoo Jang and Moontae Lee

Large reasoning models (LRMs) achieve strong performance by externalizing explicit reasoning traces before producing the answer, yet suffer from overthinking challenge that allocates uniformly heavy computation to queries of varying difficulty. While proprietary models mitigate this via opaque routing, open-source LRMs still lack an efficient mechanism to internalize adaptive reasoning due to both expensive training cost and limited disclosure of training recipes. In response, we introduce RPO (Root-token Policy Optimization), a framework that enables LRMs to self-determine when to reason by training only the initial root token (e.g., whether to invoke the \{ }<think\{ }> tag) via group relative reward and group-wise advantages. By focusing on this pivotal branching point, RPO drastically reduces training overhead and VRAM usage. Across multiple model families and scales, RPO learns difficulty-aware adaptive thinking at just 2% of the training compute of prior adaptive-reasoning methods.

[Long] EmoMAS: Emotion-Aware Multi-Agent System for High-Stakes Edge-Deployable Negotiation with Bayesian Orchestration ▶ *Underline Video*

Yunbo Long, Yuhan Liu and Liming Xu

Large language models (LLMs) have been widely used for automated negotiation, but their high computational cost and privacy risks limit deployment in privacy-sensitive, on-device settings such as mobile assistants or rescue robots. Small language models (SLMs) offer a viable alternative, yet struggle with the complex emotional dynamics of high-stakes negotiation. We introduce EmoMAS, a Bayesian multi-agent framework that transforms emotional decision-making from reactive to strategic. EmoMAS leverages a Bayesian orchestrator to coordinate three specialized agents: game-theoretic, reinforcement learning, and psychological coherence models. The system fuses their real-time insights to optimize emotional state transitions while continuously updating agent reliability based on negotiation feedback. This mixture-of-agents architecture enables online strategy learning without pre-training. We further introduce four high-stakes, edge-deployable negotiation benchmarks across debt, healthcare, emergency response, and educational domains. Through extensive agent-to-agent simulations across all benchmarks, both SLMs and LLMs equipped with EmoMAS consistently surpass all baseline models in negotiation performance while balancing ethical behavior. These results show that strategic emotional intelligence is a key driver of negotiation success. By treating emotional expression as a strategic variable within a Bayesian multi-agent optimization framework, EmoMAS establishes a new paradigm for effective, private, and adaptive negotiation AI suitable for high-stakes edge deployment. The code is available at https://github.com/Yunbo-max/EmoMAS.

[Long] Empowering Tabular Data Preparation with Language Models: Why and How? ▶ *Underline Video*

Mengshi Chen, Yuxiang Sun, Tengchao Li, Jianwei Wang, Kai Wang, Xuemin Lin, Ying Zhang and Wenjie Zhang
Data preparation is a critical step in enhancing the usability of tabular data and thus boosts downstream data-driven tasks. Traditional methods often face challenges in capturing the intricate relationships within tables and adapting to the tasks involved. Recent advances in Language Models (LMs), especially in Large Language Models (LLMs), offer new opportunities to automate and support tabular data preparation. However, why LMs suit tabular data preparation (i.e., how their capabilities match task demands) and how to use them effectively across phases still remain to be systematically explored. In this survey, we systematically analyze the role of LMs in enhancing tabular data preparation processes, focusing on four core phases: data acquisition, integration, cleaning, and transformation. For each phase, we present an integrated analysis of how LMs can be combined with other components for different preparation tasks, highlight key advancements, and outline prospective pipelines.

[Long] Enhancing Agentic Textual Graph Retrieval with Synthetic Stepwise Supervision ► *Underline Video*
Ge Chang, Jinbo Su, Jiacheng Liu, Pengfei Yang, Yuhao Shang, Huiwen Zheng, Hongli Ma, Yan Liang, Yuanchun Li and Yunxin Liu
Integrating textual graphs into Large Language Models (LLMs) is promising for complex graph-based QA. However, a key bottleneck is retrieving informative yet compact subgraphs that fit the LLM context. Existing retrievers often struggle, relying either on shallow embedding similarity or costly interactive policies that require excessive supervision. To address these challenges, we introduce Graph-S^3, an agentic textual graph reasoning framework featuring an LLM-based retriever trained with synthetic stepwise supervision. Rather than relying on final answer rewards, which often yield sparse and unstable signals, we optimize the retriever by evaluating each step against offline-extracted golden subgraphs. Our approach distills golden subgraphs via a specialized data synthesis pipeline to formulate dense rewards, facilitating a two-stage training scheme that effectively learns the interactive graph exploration policy. Based on extensive experiments on three common datasets in comparison with seven strong baselines, our approach achieves an average improvement of 15.6% in accuracy and 17.2% in F_1 score. The advantage is even higher in more complicated multi-hop reasoning tasks.

[Long] ERCThinker: Fast-Slow Thinking for Emotion Recognition in Conversation ► *Underline Video*
Yumeng Fu, Weitao Huang, Junjie Wu, Hao Teng, Shouduo Shang, Meishan Zhang and Bingquan Liu
Emotion Recognition in Conversation (ERC) aims to identify the emotional states of speakers in conversations. Existing ERC methods perform either fast thinking or slow thinking for emotion predictions. The former lacks interpretability of emotion predictions, and the latter focuses on emotion analysis at shallow semantics. Such insufficient reasoning chains fall short in capturing deep semantics within conversations. To address these limitations, we propose ERCThinker, a Fast-Slow thinking framework for the task of ERC. First, we design different thinking strategies with fine-grained emotion reasoning chains to capture deep semantics that contain topic, discourse structure, speaker characteristic, scene, and emotion shift. Second, we develop an adaptive thinking mechanism in both strategy-level and utterance-level, guiding the model to dynamically perform thinking switching across various scenarios. Furthermore, we utilize Agent-as-Judge to score reasoning chains as reward signals for more accurate emotion predictions. To support training, we construct EmotionCueCoT, the emotion reasoning dataset with supervision in both explanation and judgment. Extensive experiments on various ERC benchmark datasets demonstrate that ERCThinker achieves state-of-the-art performance in both explanation and judgment, making progress in the realm of ERC.

[Long] EverMemOS: A Self-Organizing Memory Operating System for Structured Long-Horizon Reasoning ► *Underline Video*
Chuanrui Hu, Xingze Gao, Zuyi Zhou, Dannong Xu, Yi Bai, Xintong Li, Hui Zhang, Tong Li, Chong Zhang, Lidong Bing and Yafeng Deng
Large Language Models (LLMs) are increasingly deployed as long-term interactive agents, yet their limited context windows make it difficult to sustain coherent behavior over extended interactions. Existing memory systems for LLMs often store isolated records and retrieve fragments, limiting their ability to consolidate evolving experience and resolve conflicts. We introduce EverMemOS, a self-organizing memory operating system that implements an engram-inspired lifecycle for computational memory. First, Episodic Trace Formation converts dialogue streams into MemCells that capture episodic traces, atomic facts, and time-bounded foresight. Second, Semantic Consolidation organizes MemCells into thematic MemScenes, distilling stable semantic structures and updating user profiles. Finally, Reconstructive Recollection performs MemScene-guided agentic retrieval to compose the necessary and sufficient context for downstream reasoning. Experiments on LoCoMo, LongMemEval, and PersonaMem-v2 show that EverMemOS significantly outperforms state-of-the-art methods on memory-augmented reasoning tasks.

[Long] Everyone is unique: Towards Behaviorally Heterogeneous Negotiation Dialogue Systems for Debt Collection ► *Underline Video*
Yuhang Yang, Kai Tang, Chao Ye, Haobo Wang, Qiqi Luo, Jinguang Zheng and Zhixin Zhang
Debt collection is a critical negotiation task in the financial industry, with strong practical relevance and exceptional academic value as a behaviorally rich, high-stakes testbed for human-centered dialogue systems. While large language models (LLMs) have shown promise in dialogue and negotiation, effectively evaluating their performance in this complex scenarios remains a major challenge: existing benchmarks uniformly assume users to be static, rational agents with fixed preferences, failing to capture the rich behavioral heterogeneity inherent in real-world debt collection. To bridge this gap, we propose DebtBench, the first public persona-enriched debt collection benchmark, that highlights behavioral heterogeneity in negotiation. Moreover, we develop DebtGPT, a debt collection agent trained to jointly optimize financial recovery and interaction experience. Our experimental results, using 16 state-of-the-art LLMs, find that most existing models struggle in this complex but realistic scenarios, whereas DebtGPT outperforms all open-source baselines and achieves performance on par

with GPT-4o. The code and data are available at https://github.com/yyuhhhh13/DebtNegotiation.

[Findings] Evidence-Aligned Entity Verification for Hallucination Detection in Retrieval-Augmented Generation ► *Underline Video*
Runsong Jia, Zhen Fang, Mengjia Wu, Jie Lu and Yi Zhang
Hallucination detection is crucial for large language models (LLMs), as hallucinated content creates significant barriers in applications requiring factual accuracy. Current detection methods mainly depend on internal signals like uncertainty and self-consistency checks, using the model's pre-trained knowledge to identify unreliable outputs. However, pre-trained knowledge may become outdated and has coverage limitations, especially for specialized or recent information. To address these limitations, retrieval-augmented generation (RAG) has emerged as a promising solution by retrieving relevant evidence at inference time, grounding outputs beyond the models parametric knowledge. In this paper, we target a critical and practical learning problem RAG-based hallucination detection (RHD), where RAG is employed to enhance hallucination detection by addressing information updating challenges. To address RHD, we propose a novel method Evidence-Aligned Entity Verification (EAEV), which detects entity-level hallucinations by leveraging RAG to align generated entities with retrieved evidence contexts. Specifically, EAEV evaluates entity-evidence alignment through three complementary dimensions and introduces counterfactual stability analysis to ensure robust alignments under evidence perturbations. Experiments across multiple RAG benchmarks demonstrate that EAEV achieves consistent improvements over existing methods with strong generalization capabilities.

[Long] EvoSci: A Bio-Inspired Multi-Agent Framework for the Evolution of Scientific Discovery ► *Underline Video*
Xiaoyu Xiong, Yuqi Ren and Deyi Xiong
Large language models (LLMs), have shown strong potential in scientific discovery, yet existing methods still face substantial challenges in the design of research workflows and multi-role collaboration mechanisms. To mitigate these issues, we propose EvoSci, a multi-agent scientific collaboration framework, which integrates bio-inspired evolution with knowledge graph modeling. To iteratively generate, evaluate, and refine research ideas, EvoSci incorporates multiple role-based agents, including mentor, researcher, and reviewer. By combining collaborative reasoning, shared memory, and evolutionary feedback, EvoSci significantly enhances the coherence and creativity of scientific exploration. Experiments on real-world research topics demonstrate that EvoSci significantly outperforms strong baselines in LLM-based structured peer-review and comparative ranking evaluations, achieving the highest overall peer-review score (ICLR 4.90) and top ranking (Top-10 = 54). These results suggest its superiority in both scientific idea generation and continuous discovery.

[Long] EVOTOOL: Self-Evolving Tool-Use Policy Optimization in LLM Agents via Blame-Aware Mutation and Diversity-Aware Selection ► *Underline Video*
Shuo Yang, Caren Han, Xueqi Ma, Yan Li, Mohammad Reza Ghasemi Madani and Eduard Hovy
LLM-based agents depend on effective tool-use policies to solve complex tasks, yet optimizing these policies remains challenging due to delayed supervision and the difficulty of credit assignment in long-horizon trajectories. Existing optimization approaches tend to be either monolithic, which are prone to entangling behaviors, or single-aspect, which ignore cross-module error propagation. To address these limitations, we propose EVOTOOL, a self- evolving framework that optimizes a modular tool-use policy via a gradient-free evolutionary paradigm. EVOTOOL decomposes agents tool-use policy into four modules, including Planner, Selector, Caller, and Synthesizer, and iteratively improves them through three mechanisms. Trajectory-Grounded Blame Attribution uses diagnostic traces to localize failures to a specific module. Feedback-Guided Targeted Mutation then edits only that module via natural-language critique. Diversity-Aware Population Selection preserves complementary candidates to ensure solution diversity. Across four diverse benchmarks, EVOTOOL outperforms strong baselines by over 5 points on both GPT-4.1 and Qwen3-8B, while achieving superior efficiency and transferability.

[Long] Extending First-Order Logic for Factual Reasoning over Knowledge Graphs ► *Underline Video*
Yuanzhen Hao and Desheng Wu
First-order logic (FOL) is a fundamental formalism for factual reasoning over knowledge graphs (KGs), e.g. in researches of KG-based fact verification and logical consistency or reasoning of large language models (LLM). However, existing benchmarks and approaches insufficiently capture many claims that require comparison or counting, and lack support for several FOL quantifiers and connectives. To address these challenges and expand the expressive capacity of FOL for KG-based reasoning, we introduce FOLX-KG, a novel extended FOL sigma-structure over KGs that incorporates comparison predicates and counting quantifiers. Using this extended logic, we construct Fact-FOLX-KG, a fact verification dataset consisting of 43,821 KG-based claim-formula pairs designed to enable systematic study of richer logical forms and reasoning types. We further propose FOLX Prover, an executable program-guided logic reasoning pipeline adapted for KG-based factual reasoning under the extended FOL. Experimental results show that our method achieves state-of-the-art performance on Fact-FOLX-KG, while previous methods experience performance drop on claims requiring comparison and counting. These findings demonstrate the importance of extended logical expressiveness for robust factual reasoning over KGs.

[Findings] Failures are Treasures: Constructing a Pedagogical Bridge for Agentic Strategy Distillation ► *Underline Video*
Jiaxin Guo, Hao Sun, Wenhao Zhang, Chunyu Yang and Yan Zhang
While Large Language Models (LLMs) excel in autonomous agent settings, small language models (SLMs) remain fragile, often collapsing after encountering errors. Traditional knowledge distillation focuses on imitating successful trajectories, while existing "learning from mistakes" methods treat errors as auxiliary signals rather than states requiring recoverable policies, leaving the dynamics of failure and recovery in agent settings largely unexplored. Inspired by Donald Schöns theory of reflective practice, we propose P-BRIDGE (Pedagogical

Bridge for Reflective Insight and Distillation of Guiding Errors). P-BRIDGE combines reflection-in-action with reflection-on-action, enabling agents to diagnose and correct critical errors during execution while abstracting transferable strategies from contrastive student-teacher trajectories. Experiments across eight benchmarks demonstrate that P-BRIDGE significantly elevates SLM performance, e.g., raising the 2WikiMultiHopQA accuracy of a 0.6B model from 6.2% to 34.2%.

[Findings] Fairness Evaluation and Inference Level Mitigation in LLMs ▶ *Underline Video*

Afrozah Nadeem, Mark Dras and Usman Naseem

Large language models often display undesirable behaviors embedded in their internal representations, undermining fairness, inconsistency drift, amplification of harmful content, and the propagation of unwanted patterns during extended dialogue and conversations. Although training-time or data-centric methods attempt to reduce these effects, they are computationally expensive, irreversible once deployed, and slow to adapt to new conversational contexts. Pruning-based methods provide a flexible and transparent way to reduce bias by adjusting the neurons responsible for certain behaviors. However, most existing approaches are static; once a neuron is removed, the model loses the ability to adapt when the conversation or context changes. To address this, we propose a dynamic, reversible, pruning-based framework that detects context-aware neuron activations and applies adaptive masking to modulate their influence during generation. Our inference-time solution provides fine-grained, memory-aware mitigation with knowledge-preserved, more coherent behavior across multilingual single- and multi-turn dialogues, enabling dynamic fairness control in real-world conversational AI.

[SRW] Faithfulness Beyond Plausibility: Auditing Human Explanations in Educational Assessment ▶ *Underline Video*

Ria Talsania; Dhruv Ritesh Shah; Sudhir Dhage

Faithfulness in NLP explanations is typically assessed via input perturbations, implicitly treating explanations as causal drivers of decisions, an assumption ill-suited to human judgment. In educational assessment, explanations are better understood as structured justifications accompanying holistic decisions. We propose an explainability-first framework that treats human explanations as structured artifacts independent of prediction. By modeling explanation structure explicitly, our approach enables diagnostic analyses of emphasis, consistency, and dependence. We find that scores remain invariant under evidence removal, while explanatory structure persists in tutor-specific and rubric dependent ways, revealing a mismatch between prevailing faithfulness metrics and human reasoning. This provides a model-free foundation for auditing explanations and designing NLP systems whose explanations align with human evaluative practice.

[Long] FormalScience: Scalable Human-in-the-Loop Autoformalisation of Science with Agentic Code Generation in Lean ▶ *Underline Video*

Jordan Meadows, Lan Zhang and Andre Freitas

Formalising informal mathematical reasoning into formally verifiable code is a significant challenge for large language models. In scientific fields such as physics, domain-specific machinery (e.g. Dirac notation, vector calculus) imposes additional formalisation challenges that modern LLMs and agentic approaches have yet to tackle. To aid autoformalisation in scientific domains, we present FormalScience; a domain-agnostic human-in-the-loop agentic pipeline that enables a single domain expert (without deep formal language experience) to produce syntactically correct and semantically aligned formal proofs of informal reasoning for low economic cost. Applying FormalScience to physics, we construct FormalPhysics, a dataset of 200 university-level (LaTeX) physics problems and solutions (primarily quantum mechanics and electromagnetism), along with their Lean4 formal representations. Compared to existing formal math benchmarks, FormalPhysics achieves perfect formal validity and exhibits greater statement complexity. We evaluate open-source models and proprietary systems on a statement autoformalisation task on our dataset via zero-shot prompting, self-refinement with error feedback, and a novel multi-stage agentic approach, and explore autoformalisation limitations in modern LLM-based approaches. We provide the first systematic characterisation of semantic drift in physics autoformalisation in terms of concepts such as notational collapse and abstraction elevation which reveals what formal language verifies when full semantic preservation is unattainable. We release the codebase together with an interactive UI-based FormalScience system which facilitates autoformalisation and theorem proving in scientific domains beyond physics.

[Long] Framing Political Bias in Multilingual LLMs Across Pakistani Languages ▶ *Underline Video*

Afrozah Nadeem, Mark Dras and Usman Naseem

Large Language Models (LLMs) increasingly shape public discourse, yet most evaluations of economic and political bias have focused on high-resource Western languages and contexts. This leaves a blind spots in low-resource, multilingual regions such as Pakistan, where linguistic identity is closely tied to regional, religious, and political ideologies. We present a systematic evaluation of political bias in 13 state-of-the-art LLMs across five Pakistani languages: Urdu, Punjabi, Sindhi, Pashto, and Balochi. Our framework integrates a culturally adapted Political Compass Test (PCT) with multi-level framing analysis, capturing both ideological stance (economic/social axes) and stylistic framing (content, tone, emphasis). The prompts are aligned with 11 socio-political themes specific to the Pakistani context. The results show that while LLMs significantly reflect liberal-left orientations consistent with Western training data, they exhibit more authoritarian framing in regional languages, highlighting language-conditioned ideological modulation. We also identify model-specific bias patterns in all languages. These findings show the need for culturally grounded multilingual bias examining frameworks in NLP. Code and dataset are available.

[Short] From Factuality to Meta-Factivity: A Cognitive Blueprint for Trustworthy LLMs ▶ *Underline Video*

Daohuan Liu, Lun Xia, Yuer Wang, Jiaoyang Su and Xuri Tang

Current research on Event Factuality Prediction (EFP) predominantly treats LLMs as passive classifiers, where high aggregate metrics often

mask shortcut learning and unreliable reasoning. In this position paper, we argue for a focus shift from event factuality to meta-factivity. We introduce the Meta-Factivity Framework (MFF), a theoretical roadmap that moves evaluation beyond surface recognition to belief trajectory reasoning and epistemic regulation. By framing hallucination as a failure of meta-cognitive control, we advocate for a transition from measuring black-box accuracy to evaluating white-box cognition, laying the groundwork for a more rigorous benchmark for explainable self-governance.

[SRW] From Graphs to Hypergraphs: Enhancing Aspect-Term Sentiment Analysis via Multi-Level Relational Modeling ▶ *Underline Video*

Omkar Mahesh Kashyap; Padegal Amit; Ashwini M Joshi; Shylaja S S; Madhav Kashyap

Aspect-Term Sentiment Analysis (ATSA) aims to predict sentiment polarity for specific aspect terms, a task complicated by conflicting sentiments and limited context in short texts. Existing graph-based approaches, including recent multi-graph methods, rely on predefined pairwise structures to capture different linguistic views. However, these approaches impose a fixed interaction granularity and depend on multiple predefined graphs, limiting their ability to adapt to sentence-specific sentiment composition and capture higher-order relationships. We propose HyperATSA, a dynamic hypergraph framework that constructs instance-specific aspect-opinion groupings directly from contextual token representations, without relying on external parsers. Hyperedges are induced via hierarchical agglomerative clustering over token embeddings, with a novel acceleration-based cutoff that identifies natural group boundaries per sentence. Experiments on Lap14, Rest14, and MAMS demonstrate consistent improvements over strong graph-based baselines, with controlled comparisons across BERT and RoBERTa backbones confirming that the gains stem from the hypergraph mechanism rather than encoder strength.

[Findings] From Heads to Neurons: Causal Attribution and Steering in Multi-Task Vision-Language Models ▶ *Underline Video*

Qidong Wang, Junjie Hu and Ming Jiang

Recent work has increasingly explored neuron-level interpretation in vision-language models (VLMs) to identify neurons critical to final predictions. However, existing neuron analyses generally focus on single tasks, limiting the comparability of neuron importance across tasks. Moreover, ranking strategies tend to score neurons in isolation, overlooking how task-dependent information pathways shape the write-in effects of feed-forward network (FFN) neurons. This oversight can exacerbate neuron polysemanticity in multi-task settings, introducing noise into the identification and intervention of task-critical neurons. In this study, we propose HONES (Head-Oriented Neuron Explanation & Steering), a gradient-free framework for task-aware neuron attribution and steering in multi-task VLMs. HONES ranks FFN neurons by their causal write-in contributions conditioned on task-relevant attention heads, and further modulates salient neurons via lightweight scaling. Experiments on four diverse multimodal tasks and two popular VLMs show that HONES outperforms existing methods in identifying task-critical neurons and improves model performance after steering. Our source code is released at: https://github.com/petergit1/HONES.

[Long] From Language to Driving: A Dual-Loop SLM-Enhanced Framework for Multi-Planner Scheduling via a Domain-Specific Language ▶ *Underline Video*

Jiawei Liu and Xun Gong

Advancing from usable to collaborative autonomy requires driving systems to execute passenger instructions safely and reliably. This work formulates instruction realization as scheduling across multiple motion planners and presents a dual-loop framework that provides a transparent decision chain from natural language to vehicle control. The outer loop uses a small language model (SLM) for high-level, low-frequency semantic reasoning and schedule generation, while the inner loop performs low-level, high-frequency schedule execution and vehicle control. To compensate for the SLM's limited capacity, the framework integrates receding-horizon scheduling to segment long-horizon instruction tasks, a domain-specific language (DSL) that restricts SLM outputs to a scheduling-oriented subspace, and reinforcement learning in high-fidelity urban traffic to refine the SLM's DSL proficiency and scheduling performance. Experiments show that the framework improves instruction-completion rates while maintaining high safety and compliance relative to multiple baselines.

[Long] From Recognition to Reasoning: Benchmarking and Enhancing MLLMs on Real-World Receipt Document Understanding ▶ *Underline Video*

Yandi Wang, Libin Zhan, Ziwei Huang, Tiancheng Luo, Yuxuan Jiang, Wang Dong, Leilei Gan and Jun Chen

Extracting structured information from visual documents (Visual Information Extraction, VIE) is a cornerstone of business automation. While recent Multimodal Large Language Models (MLLMs) have shown promising capabilities, existing benchmarks suffer from critical limitations in scale and realism, lack semantic granularity, and fail to cover diverse document types. To bridge this gap, we introduce ReceiptBench, a large-scale, human-annotated benchmark consisting of 10k diverse receipts, organizing information extraction into four hierarchical sub-tasks: (1) Basic Perception for raw text spotting, (2) Format Normalization for strictly following standardization instructions, (3) Semantic Reasoning for inferring implicit attributes from context, and (4) Structure Parsing for handling nested line items. Furthermore, we propose a two-stage training framework incorporating Metric-Aware Group Relative Policy Optimization (GRPO), which translates rigorous evaluation constraints into reinforcement learning signals to enhance structural consistency. Extensive experiments demonstrate that our method yields state-of-the-art performance, surpassing leading proprietary models on complex reasoning tasks. We release our datasets and code at https://github.com/wwwT0ri/ReceiptBench.

[Long] FusionFlow: Enabling Deep Structural Exploration for Automated Agentic Workflow Generation ▶ *Underline Video*

Xiang Wang, Zongtao Yang, Zhuojian Hong, Shuhao Zhang and Wei Wei

Agentic workflows are commonly used to guide large language models in solving complex reasoning tasks. However, existing automated workflow generation methods primarily rely on stepwise local refinement or tree-based search over a single evolving workflow. Under limited optimization budgets, this paradigm constrains structural depth, hindering the discovery of workflows that require deep compositional structure. To address this limitation, we propose FusionFlow, a framework centered on workflow fusion. Unlike incremental refinement, fusion enables structural leaps by synthesizing multiple independently evolved workflows, allowing exploration of deeper regions of the workflow space within a finite budget. To make fusion effective, FusionFlow integrates local optimization, task-specific differentiation, and a dynamic scheduling mechanism. Experiments on six reasoning benchmarks demonstrate that FusionFlow consistently outperforms existing automated workflow generation methods. Further ablation and analysis confirm that fusion is the key driver of deep structural exploration, highlighting fusion-driven exploration as an effective approach for overcoming depth limitations in automated workflow generation.

[Long] GIFT: Guided Fine-Tuning and Transfer for Enhancing Instruction-Tuned Language Models ▶ *Underline Video*

Zhiwen Ruan, Yichao Du, Jianjie Zheng, Longyue Wang, Yun Chen, Peng Li, Jinsong Su, Yang Liu and Guanhua Chen

Instruction-tuned large language models (LLMs) exhibit strong instruction-following and generalization abilities, enabled by expensive post-training pipelines. However, adapting them to specific downstream tasks remains challenging: direct fine-tuning often disrupts this delicate balance, while existing adapter-based transfer methods typically treat the instruction-tuned model as a passive target that only participates at the final merging stage. We propose GIFT (Guided Fine-Tuning and Transfer), a simple and efficient framework that incorporates instruction-level guidance into task adaptation. GIFT fine-tunes a low-rank adapter on the pretrained base model using token-level confidence signals derived from the instruction-tuned model. The learned adapter is then merged into the instruction-tuned model, yielding task-specialized models that preserve general instruction-following behavior. We evaluate GIFT on mathematical reasoning and knowledge-intensive benchmarks across multiple model families and scales. Results show that GIFT consistently outperforms direct fine-tuning and representative transfer-based baselines, while maintaining robust generalization and favorable test-time scaling behavior.

[Long] Gradient-Guided Multi-Judge Prompt Optimization ▶ *Underline Video*

ChenZhuo Zhao, Xinda Wang, Pu Zhao, Yue Huang, Junting Lu, Ziqian Liu, Qingwei Lin, Saravan Rajmohan and Dongmei Zhang

Automatic prompt optimization is a practical alternative to fine-tuning for adapting large language models (LLMs), yet existing approaches often trade off signal quality against computational cost. Methods that rely on generative feedback can be informative but expensive to scale, while sampling-based optimization typically requires many evaluations and exhibits high variance. Even loss-driven prompt optimization remains limited by costly segment attribution that scales with prompt length and by overfitting to a single evaluator, which weakens transfer across model families and domains. We propose Gradient-guided Multi-judge Prompt Optimization (GMPO), a scalable framework that improves both efficiency and robustness. GMPO uses a first-order gradient approximation to score segment importance in a continuous masking direction, requiring only one forward and one backward pass. GMPO further employs a generate multi-judge design in which candidate prompt edits are proposed by a generator and selected using cross-entropy losses aggregated from multiple lightweight judge models, reducing evaluator bias and improving generalization. Experiments across math, reasoning, instruction-following evaluation, and safety robustness benchmarks demonstrate consistent gains with substantially lower optimization overhead.

[Industry] GuideTree: Guideline-Induced Review Trees for Long Medical Records ▶ *Underline Video*

Chengze Ge, Ruiqing Zhang, Yining Wang, Shengping Liu, Jiaen Liang, Wei Huang and Weihua Chen;

Reviewing medical records for clinical and insurance decisions must handle long, heterogeneous documents while producing consistent, traceable, guideline-compliant outcomes under strict latency and cost constraints. We propose GuideTree, which compiles textual guidelines into a fixed review tree of evidence-grounded verification primitives. GuideTree uses short per-document summaries only for routing each check to a minimal set of document types and candidates; final verification always reads full document text and returns structured evidence. The tree is induced offline via a cost-aware split-and-prune search and updated safely through regression-tested, versioned patches. Across 1,000 cases from four industrial review scenarios and four LLM backbones, GuideTree achieves 84.5–92.8 Macro-F1, outperforming the strongest non-expert baselines by 3.3–7.6 points and matching ExpertTree within 0.2–0.6 points (avg. 0.38). On chronic disease with Qwen3-235B-A22B-Instruct, GuideTree reduces average I/O volume to 74K input+output characters (-82% vs. long-context prompting) and average latency to 22s (-83% vs. long-context prompting), while reaching 99% decision consistency over K=5 reruns.

[Industry] HAD: HAllucination Detection Language Models Based on a Comprehensive Hallucination Taxonomy ▶ *Underline Video*

Fan Xu, Xinyu Hu, Zhenghan Yu, Li Lin, Xu Zhang, Yang Zhang, Wei Zhou, Jinjie Gu and Xiaojun Wan

The increasing reliance on natural language generation (NLG) models, particularly large language models, has raised concerns about the reliability and accuracy of their outputs. A key challenge is hallucination, where models produce plausible but incorrect information. As a result, hallucination detection has become a critical task. In this work, we introduce a comprehensive hallucination taxonomy with 11 categories across various NLG tasks and propose the HAllucination Detection (HAD) models, which integrate hallucination detection, span-level identification, and correction into a single inference process. Trained on an elaborate synthetic dataset of about 90K samples, our HAD models are versatile and can be applied to various NLG tasks. We also carefully annotate a test set for hallucination detection, called HADTest, which contains 2,248 samples. Evaluations on in-domain and out-of-domain test sets show that our HAD models generally outperform the existing baselines, achieving state-of-the-art results on HaluEval, FactCHD, and FaithBench, confirming their robustness and versatility.

[Long] HAG: Hierarchical Demographic Tree-based Agent Generation for Topic-Adaptive Simulation ▶ *Underline Video*
Rongxin Chen, Tianyu Wu, Bingbing Xu, JiaTang Luo, Xiucheng Xu and Huawei Shen
High-fidelity agent initialization is crucial for credible Agent-Based Modeling across diverse domains. A robust framework should be Topic-Adaptive, capturing macro-level joint distributions while ensuring micro-level individual rationality. Existing approaches fall into two categories: static data-based retrieval methods that fail to adapt to unseen topics absent from the data, and LLM-based generation methods that lack macro-level distribution awareness, resulting in inconsistencies between micro-level persona attributes and reality. To address these problems, we propose HAG, a Hierarchical Agent Generation framework that formalizes population generation as a two-stage decision process. Firstly, utilizing a World Knowledge Model to infer hierarchical conditional probabilities to construct the Topic-Adaptive Tree, achieving macro-level distribution alignment. Then, grounded real-world data, instantiation and agentic augmentation are carried out to ensure micro-level consistency. Given the lack of specialized evaluation, we establish a multi-domain benchmark and a comprehensive PACE evaluation framework. Extensive experiments show that HAG significantly outperforms representative baselines, reducing population alignment errors by an average of 37.7% and enhancing sociological consistency by 18.8%.

[Long] Hallucination Detection in LLMs with Topological Divergence on Attention Graphs ▶ *Underline Video*
Alexandra Bazarova, Andrei Volodichev, Aleksandr Yugay, Andrey Shulga, Alina Ermilova, Konstantin Polev, Julia Belikova, Rauf Parchiev, Dmitry Simakov, Maxim Savchenko, Andrey Savchenko, Serguei Barannikov and Alexey Zaytsev
Hallucinations remain a critical challenge for large language models (LLMs), particularly in Retrieval-Augmented Generation (RAG) settings where models may generate outputs unsupported by the provided context. To address this, we introduce TOHA, a TOpology-based HAllucination detector, which leverages a topological divergence metric to quantify the structural properties of graphs induced by attention matrices. Examining the topological divergence between prompt and response subgraphs in RAG settings reveals consistent patterns: higher divergence values in specific attention heads correlate with unfaithful outputs, independent of the dataset. Extensive experiments , including evaluations on question answering and summarization tasks , show that our approach achieves state-of-the-art or competitive results on several benchmarks while requiring minimal annotated data and computational resources. Our findings indicate that the topological structure of attention matrices provides an efficient and robust metric for assessing the correctness of LLM's responses.

[Long] HarmRLVR: Weaponizing Verifiable Rewards for Harmful LLM Alignment ▶ *Underline Video*
Yuexiao Liu, Lijun Li, Xingjun Wang and Jing Shao
Recent advancements in Reinforcement Learning with Verifiable Rewards (RLVR) have gained significant attention due to their objective and verifiable reward signals, demonstrating strong performance in reasoning and code generation tasks. However, the potential safety risks associated with RLVR remain underexplored. This paper presents HarmRLVR, the first systematic investigation into the alignment reversibility risk of RLVR. We show that safety alignment can be rapidly reversed using GRPO with merely 64 harmful prompts without responses, causing models to readily comply with harmful instructions. Across five models from Llama, Qwen, and DeepSeek, we empirically demonstrate that RLVR-based attacks elevate the average harmfulness score to 4.94 with an attack success rate of 96.01%, significantly outperforming harmful fine-tuning while preserving general capabilities. Our findings reveal that RLVR can be efficiently exploited for harmful alignment, posing serious threats to open-source model safety.

[Long] Hetero-Designer: Automated Design of Multi-Agent Systems with Heterogeneous LLMs ▶ *Underline Video*
Zhiheng Zhang, Yuanzhe Zhang, Bohan Yu, Daojian Zeng, Kang Liu and Jun Zhao
LLM-based Multi-agent systems (MAS) have shown strong capabilities across a wide range of domains. Their success largely hinges on the collaboration topology design, which has emerged as a central research focus in the automated MAS design. However, existing approaches are fundamentally constrained by their reliance on homogeneous LLMs, which significantly limits overall system intelligence. In response to this limitation, we for the first time propose the concept of Automated Design of Heterogeneous-LLMs-based MAS (ADHM). ADHM sheds light on a promising avenue for advancing collective intelligence, which focuses on the automated design of cost-effective MAS composed of diverse LLMs and roles to suit various queries. Toward this challenging goal, we propose Hetero-Designer, a novel pipeline that efficiently encodes intricate dependencies among queries, LLMs and roles through a novel Binary-Star Transformer and constructs Hetero-MAS in an autoregressive graph generation process. Extensive experiments demonstrate that Hetero-Designer is: (1) high-performing on various benchmarks, (2) economical in reducing overhead, (3) extensible to unseen LLMs and roles.

[Long] HeteroCache: A Dynamic Retrieval Approach to Heterogeneous KV Cache Compression for Long-Context LLM Inference ▶ *Underline Video*
Zhiyuan Shi, Qibo Qiu, Feng Xue, Zhonglin Jiang, Li Yu, Jian Jiang, Xiaofei He and Wenxiao Wang
The linear memory growth of the KV cache poses a significant bottleneck for LLM inference in long-context tasks. Existing static compression methods often fail to preserve globally important information. Although recent dynamic retrieval approaches attempt to address this issue, they typically suffer from coarse-grained caching strategies and incur high I/O overhead. To overcome these limitations, we propose HeteroCache, a training-free dynamic compression framework. Our method is built on two key insights: attention heads exhibit diverse temporal heterogeneity, and there is significant spatial redundancy among heads within the same layer. Guided by these insights, HeteroCache categorizes heads based on stability and similarity, applying a fine-grained weighting strategy that allocates larger cache budgets to heads with rapidly shifting attention to capture context changes. Furthermore, it features a hierarchical storage mechanism where representative heads monitor attention drift to trigger asynchronous, on-demand context retrieval, thereby hiding I/O latency. Experiments

demonstrate that HeteroCache achieves state-of-the-art performance on long-context benchmarks and accelerates decoding by up to 3x compared to the original model with a 224K context. Our code is available at https://github.com/ponytaill/HeteroCache.

[Long] HiddenGuard: Fine-Grained Safe Generation with Specialized Representation Router ▶ *Underline Video*

Lingrui Mei, Shenghua Liu, Yiwei Wang, Baolong Bi, Ruibin Yuan and Xueqi Cheng

As Large Language Models (LLMs) grow increasingly powerful, ensuring their safety and alignment with human values remains a critical challenge. Current alignment approaches predominantly rely on refusal alignment, such as training models to refuse harmful prompts or implementing filters at various stages to block certain responses. These methods are designed toward a binary outcome: either denying to answer the question entirely or answering with full access to the model's parametric knowledge. The binary nature of current alignment approaches presents significant limitations. These methods often fail to balance safety and utility, resulting in either overly cautious responses or overlooking subtle harmful content. They also prevent users from accessing benign information when it's mixed with harmful content. For instance, a model might refuse to provide basic, public information about a medication's composition due to misuse concerns. Furthermore, these approaches struggle with context-dependent sensitivity, potentially over-censoring harmless content or missing nuanced harmful outputs. Ideally, LLMs should offer informative responses while avoiding the disclosure of harmful and sensitive information. To address these challenges, we introduce HiddenGuard, a novel framework for fine-grained safe generation in LLMs. Our method incorporates PRISM (rePresentation Router for In-Stream Moderation), a specialized moudule that operates alongside the LLM architecture. By leveraging intermediate hidden states, HiddenGuard enables real-time, token-level harmfulness detection and redaction, without loss in capability. This approach captures deeper semantic information, allowing for more nuanced and context-aware content control compared to traditional filtering techniques. Consequently, the model can generate informative responses while selectively redacting or replacing sensitive information, rather than refusing to answer outright. We also contribute a comprehensive dataset with token-level fine-grained annotations of potentially harmful information across diverse contexts. Our experiments demonstrate that HiddenGuard achieves over 90% in F1 score for detecting and redacting harmful content while preserving the overall utility and informativeness of the model's responses.

[Long] HOPE: Hybrid Optimized Parallel Encoding with Supervised and Unsupervised Semantic Fusion for Depression Symptom Detection ▶ *Underline Video*

Tu-Phuong Mai, Minh-Ha Le H., Duc-Luong Tran, Phuong-Anh Chu, Duy-Cat Can and Hoang-Quynh Le

Timely detection of depression symptoms is essential for early intervention, and the continuous stream of user-generated content on social media provides an ideal source for this purpose. To address this challenge, we propose HOPE, a Hybrid Optimized Parallel Encoding framework that combines supervised symptom relevance signals with unsupervised intrinsic semantic clustering. This parallel design enables robust symptom detection under limited labeled data and introduces a distinctive semantic-similarity perspective with automatic class-anchor adjustment. We also propose an optimized hybrid semantic fusion mechanism to combine supervised and unsupervised scores through a learnable module. We evaluate our system on multiple benchmark datasets and surpass previous approaches, demonstrating its effectiveness in detecting fine-grained symptoms and early warning of mental health risk. Source code is available at https://github.com/candleMind/hope.

[Findings] How Do LLMs "Trust" Unknown Knowledge? An Unknown Knowledge Based Jailbreak Attack ▶ *Underline Video*

Yixiao Huang;Lan Zhang;Chaoran Wang;

Learning unknown knowledge through ICL and RAG can enhance LLM capabilities in specialized fields. While most research focuses on how to identify and utilize such knowledge, little work examines what factors lead LLMs to trust and adopt it, leaving models prone to errors and harmful content. Grounded in extensive pre-experiments, we design five pairs of trustenhancing and trustdiminishing transformations on unknown knowledge to experimentally identify the key trust factors. These findings are further substantiated through a detailed theoretical analysis grounded in the epistemological framework of evidentialism. Based on these insights, we challengingly propose a completely unrestricted and fully randomized jailbreak attack that embeds malicious queries within trustenhanced unknown knowledge. In both defended and undefended scenarios, our method achieves 99% to 100% ASR on all tested LLMs, including the latest GPT5.1, and becomes SOTA. This attack confirms the trust mechanism and exposes a critical and hard-to-defend security risk. Our conclusions provide valuable guidance for understanding trust mechanism of unknown knowledge and for future research.

[Long] Human or LLM as Standardized Patients? A Comparative Study in Medical Education ▶ *Underline Video*

Bingquan Zhang, Xiaoxiao Liu, Yuchi Wang, Lei Zhou, Qianqian Xie and Benyou Wang

Standardized patients (SPs) are indispensable for clinical skills training but remain expensive and difficult to scale. Although large language model (LLM)-based virtual standardized patients (VSPs) have been proposed as an alternative, their behavior remains unstable and lacks rigorous comparison with human standardized patients. We propose EasyMED, a multi-agent VSP framework that separates case-grounded information disclosure from response generation to support stable, inquiry-conditioned patient behavior. We also introduce SPBench, a human-grounded benchmark with eight expert-defined criteria for interaction-level evaluation. Experiments show that EasyMED more closely matches human SP behavior than existing VSPs, particularly in case consistency and controlled disclosure. A four-week controlled study further demonstrates learning outcomes comparable to human SP training, with stronger early gains for novice learners and improved flexibility, psychological safety, and cost efficiency.

[Long] HumanLLM: Benchmarking and Improving LLM Anthropomorphism via Human Cognitive Patterns ▶ *Underline Video*

Xintao Wang, Jian Yang, Weiyuan Li, Rui Xie, Jen-tse Huang, Jun Gao, Shuai Huang, Yueping Kang, Yuanli Guo, Hongwei Feng and

Yanghua Xiao

Large Language Models (LLMs) have demonstrated remarkable capabilities in reasoning and generation, serving as the foundation for advanced persona simulation and Role-Playing Language Agents (RPLAs). However, achieving authentic alignment with human cognitive and behavioral patterns remains a critical challenge for these agents. We present HumanLLM, a framework treating psychological patterns as interacting causal forces. We construct 244 patterns from 12,000 academic papers and synthesize 11,359 scenarios where 2-5 patterns reinforce, conflict, or modulate each other, with multi-turn conversations expressing inner thoughts, actions, and dialogue. Our dual-level checklists evaluate both individual pattern fidelity and emergent multi-pattern dynamics, achieving strong human alignment (r=0.90) while revealing that holistic metrics conflate simulation accuracy with social desirability. HumanLLM-8B outperforms Qwen3-32B on multi-pattern dynamics despite 4x fewer parameters, demonstrating that authentic anthropomorphism requires cognitive modeling—simulating not just what humans do, but the psychological processes generating those behaviors. Our dataset, code, and model are available at:https://github.com/YJGoodbye2024/HumanLLM

[Long] Hybrid Autoregressive-Diffusion Model for Real-Time Sign Language Production ▶ *Underline Video*

Maoxiao Ye, Xinfeng Ye and Sathiamoorthy Manoharan

Earlier Sign Language Production (SLP) models typically relied on autoregressive methods that generate output tokens one by one, which inherently provide temporal alignment. Although techniques like Teacher Forcing can prevent model collapse during training, they still cannot solve the problem of error accumulation during inference, since ground truth is unavailable at that stage. In contrast, more recent approaches based on diffusion models leverage step-by-step denoising to enable high-quality generation. However, the iterative nature of these models and the requirement to denoise entire sequences limit their applicability in real-time tasks like SLP. To address it, we propose a hybrid autoregressive-diffusion model for Sign Language Production (SLP), combining sequential dependency modeling with iterative refinement. A Multi-Scale Pose Representation module captures fine-grained articulator features, while a Confidence-Aware Causal Attention mechanism guides generation using joint-level confidence scores. Experiments on PHOENIX14T and How2Sign show improved generation quality and real-time efficiency.

[Long] IGenBench: Benchmarking the Reliability of Text-to-Infographic Generation ▶ *Underline Video*

Yinghao Tang, Xueding Liu, Boyuan Zhang, Tingfeng Lan, Yupeng Xie, Jiale Lao, Yiyao Wang, Haoxuan Li, Tingting Gao, Bo Pan, Luoxuan Weng, Xiuqi Huang, Minfeng Zhu, Yingchaojie Feng, Yuyu Luo and Wei Chen

Infographics are composite visual artifacts that combine data visualizations with textual and illustrative elements to communicate information. While recent text-to-image (T2I) models can generate aesthetically appealing images, their reliability in generating infographics remains unclear. Generated infographics may appear correct at first glance but contain easily overlooked issues, such as distorted data encoding or incorrect textual content. We present IGenBench, the first benchmark for evaluating the reliability of text-to-infographic generation, comprising 600 curated test cases spanning 30 infographic types. We design an automated evaluation framework that decomposes reliability verification into atomic yes/no questions based on a taxonomy of 10 question types. We employ multimodal large language models (MLLMs) to verify each question, yielding question-level accuracy (Q-ACC) and infographic-level accuracy (I-ACC). We comprehensively evaluate 10 state-of-the-art T2I models on IGenBench. Our systematic analysis reveals key insights for future model development: (i) a three-tier performance hierarchy with the top model achieving Q-ACC of 0.90 but I-ACC of only 0.49; (ii) data-related dimensions emerging as universal bottlenecks (e.g., Data Completeness: 0.21); and (iii) the challenge of achieving end-to-end correctness across all models.

[Long] Incomplete In-context Learning ▶ *Underline Video*

Wenqiang Wang, WenYujia WenYujia, Yan XIAO, Zhifeng Chen, Yangshijie Zhang, Peng Chen, Mingbo Yang and Xiaochun Cao

Existing In-context Learning (ICL) typically assumes the retrieval dataset contains demonstrations for all output label spaces. However, in real-world scenarios, delays in dataset updates or incomplete data annotation may result in the retrieval dataset containing labeled demonstrations for only a subset of the output space. We refer to this phenomenon as an incomplete retrieval dataset and define the in-context learning under this condition as Incomplete In-context Learning (IICL). To address IICL, we propose Iterative Judgments and Integrated Prediction (IJIP), a framework with train-free and train-based variants. For classification, the iterative judgments stage of IJIP reformulates an (m)-class problem into (m) binary tasks, converting IICL into standard ICL. The integrated prediction stage of IJIP then refines results using both the input and initial predictions. We further extend IJIP to text regression and generation, and introduce lightweight variants that reduce computation and token costs. Across six LLMs, seven tasks, and eight datasets, IJIP achieves state-of-the-art results under two incompleteness settings and even outperforms standard ICL with complete labels. IJIP also supports a semi-supervised variant and can serve as a plug-and-play enhancement for existing ICL and zero-shot methods.

[Findings] Inference-Time Scaling of Verification: Self-Evolving Deep Research Agents via Test-Time Rubric-Guided Verification ▶ *Underline Video*

Yuxuan WAN, Tianqing Fang, ZAITANG LI, Yintong Huo, Wenxuan Wang, Haitao Mi, Dong Yu and Michael R. Lyu

Recent advances in Deep Research Agents (DRAs) are transforming automated knowledge discovery and problem-solving. While the majority of existing efforts focus on enhancing policy capabilities via post-training, we propose an alternative paradigm: test-time self-evolving the agent's ability by iteratively verifying the policy model's outputs, guided by meticulously crafted rubrics. This approach gives rise to an inference-time scaling of verification, wherein an agent self-improves at test time by evaluating its generated answers to produce iterative feedback and refinements without any additional training. We derive the rubrics based on an automatically constructed

DRA Failure Taxonomy, which systematically classifies agent failures into five major categories and thirteen sub-categories. We present DeepVerifier, a rubrics-based outcome reward verifier that leverages the asymmetry of verification and outperforms vanilla agent-as-judge and LLM judge baselines by 12%–48% in meta-evaluation F1 score. To enable practical test-time self-evolution, DeepVerifier integrates as a plug-and-play module during test-time inference. The verifier produces detailed rubric-based feedback, which is fed back to the agent for iterative bootstrapping—refining responses without additional training. This test-time scaling delivers 8%–11% accuracy gains on challenging subsets of GAIA and XBench-DeepResearch when powered by capable closed-source LLMs. Finally, to support open-source advancement, we release DeepVerifier-4K, a curated supervised fine-tuning dataset of 4,646 high-quality agent steps focused on DRA verification. These examples emphasize reflection and self-critique, enabling open models to develop robust verification capabilities.

[Long] InsLogicBench: An Argumentation Logic Grounded Benchmark for Complex Insurance Claims Adjudication ▶ *Underline Video*

Jin Liu, Yunpeng Liu, Keyi Wang, Jie Shi, Xiao Xu, Wenkang Huang, Xingzhong Xu, Xin Liang and Yanghua Xiao

Insurance claims adjudication demands not only accurate decisions but also interpretable reasoning grounded in policy clauses. However, existing benchmarks are limited to information retrieval or simple multiple-choice setups, which fail to require step-by-step inferences from facts to conclusions. To address this gap, we introduce InsLogicBench, a benchmark providing complete reasoning traces that link factual inputs, relevant policy clauses, and final verdicts. We construct the dataset using a controllable synthesis framework based on the Nested Toulmin Model. By capturing the defeasible logic of insurance policies through hierarchical truth assignment and enforcing validity via consistency verification, we ensure interpretability and logical rigor across generated examples. We evaluate eight Large Language Models (LLMs) on InsLogicBench. Results show significant difficulties in handling exception clauses and verifying missing conditions. Notably, models often produce correct final decisions but fail to provide precise justifications, highlighting a critical discrepancy between their decision accuracy and logical reasoning capabilities.

[Long] Instruction Data Selection via Answer Divergence ▶ *Underline Video*

Bo Li, Mingda Wang, Shikun Zhang and Wei Ye

Instruction tuning relies on large instruction-response corpora whose quality and composition strongly affect downstream performance. We propose Answer Divergence-Guided Selection (ADG), which selects instruction data based on the geometric structure of multi-sample outputs. ADG draws several high-temperature generations per instruction, maps responses into an embedding space, and computes an output divergence score that jointly encodes dispersion magnitude and shape anisotropy. High scores correspond to instructions whose answers are both far apart and multi-modal, rather than clustered paraphrases along a single direction. Across two backbones and three public instruction pools, fine-tuning on only 10K ADG-selected examples consistently outperforms strong selectors on six benchmarks spanning reasoning, knowledge, and coding. Analyses further show that both dispersion magnitude and shape anisotropy are necessary, supporting answer divergence as a practical signal for instruction data selection. Code and appendix are included in the supplementary materials.

[Findings] Investigating Human and LLMs' Decisions in Unverifiable Environments: A Case Study with GitHub Activity Overview ▶ *Underline Video*

Zheng Jiang, Wei Wang, Gaowei Zhang, Yang Feng and Yi Wang;

The behaviors of Large Language Models (LLMs) as artificial social actors are largely underexplored, particularly in unverifiable scenarios where conventional benchmarking has little to help improve their abilities. Thus, examining their behaviors in such scenarios can help understand and improve LLMs' capabilities of simulating real-world social actors in many tasks such as LLM-empowered social agents. We draw a typical unverifiable scenario–a simplified pull request scenario on GitHub focusing on decision-making based on Activity Overview signal–to investigate how human and LLMs behave. We introduce a systematic method to collect, compare, and reason about human and LLMs' decisions. Our results reveal that there are both similarities and differences between human and LLMs' decisions, and proprietary LLMs generally behave more like human than open-source LLMs do. We further find that human and LLMs may rely on different information and reasoning mechanisms in decision-making. Our study thus urges more future work on human and LLMs decision-making in unverifiable environments.

[Long] Joint Knowledge Base Completion and Question Answering by Combining Large Language Models and Small Language Models ▶ *Underline Video*

Yinan Liu, Dongying Lin, Sigang Luo, Xiaochun Yang and Bin Wang

Knowledge Bases (KBs) play a key role in various applications. As two representative KB-related tasks, knowledge base completion (KBC) and knowledge base question answering (KBQA) are closely related and inherently complementary with each other. Thus, it will be beneficial to solve the task of joint KBC and KBQA to make them reinforce each other. However, existing studies usually rely on the small language model (SLM) to enhance them jointly, and the large language model (LLM)'s strong reasoning ability is ignored. In this paper, by combining the strengths of the LLM with the SLM, we propose a novel framework JCQL, which can make these two tasks enhance each other in an iterative manner. To make KBC enhance KBQA, we augment the LLM agent-based KBQA model's reasoning paths by incorporating an SLM-trained KBC model as an action of the agent, alleviating the LLM's hallucination and high computational costs issue in KBQA. To make KBQA enhance KBC, we incrementally fine-tune the KBC model by leveraging KBQA's reasoning paths as its supplementary training data, improving the ability of the SLM in KBC. Extensive experiments over two public benchmark data sets demonstrate that JCQL surpasses all baselines for both KBC and KBQA tasks.

[Long] JoPR: Joint Emotion Perception and Reasoning for Conversational Emotion Recognition ▶ *Underline Video*

Yumeng Fu, Weitao Huang, Junjie Wu, Hao Teng, Meishan Zhang and Bingquan Liu

Emotion Recognition in Conversation (ERC), the task of identifying the emotion of each utterance in a conversation, is crucial for human-machine interaction. Existing LLM-based ERC methods focus on standard prompting and slow thinking for emotion analysis. However, they suffer from the lack of human-like emotion reasoning and discrimination between similar emotions, thus limiting accurate emotion predictions. To this end, we present JoPR, jointing perception-curriculum learning and emotional reasoning for conversational emotion recognition. Specifically, we devise a multi-dimension curriculum with long CoT fine-tuning to clone human-like emotion reasoning. We further design an emotion-specific reward function in a novel reinforcement learning framework, thereby enhancing the discernment between similar emotions. Our proposal is extensively evaluated over three widely used benchmark datasets, and experimental results confirm the superiority of JoPR. Furthermore, we provide an in-depth analysis to confirm the emotion perception and reasoning capabilities of JoPR.

[Findings] KnowDR-REC: Auditing Knowledge-Conditioned Visual Grounding in Referring Expression Comprehension ▶ *Underline Video*

Guanghao Jin, Jingpei Wu, Tianpei Guo, Yiyi Niu, Weidong Zhou, Linyi Yang and Guoyang Liu

While Multimodal Large Language Models (MLLMs) have demonstrated the capacity for multi-modal reasoning, current Referring Expression Comprehension (REC) benchmarks lag behind, predominantly relying on intra-image cues and neglecting the integration of external world knowledge, which significantly impedes the evolution of REC towards real-world applications. This limitation obscures a model's true capability to conduct textual reasoning (entity resolution), resolve spatial location (visual grounding), and verify reference validity (hallucination rejection). To address this, we introduce KnowDR-REC, a targeted audit benchmark comprising 1,042 positive triplets derived from real-world knowledge, along with rigorously matched negative samples. Unlike traditional datasets, we implement a controllable counterfactual evaluation mechanism that subjects textual expressions to single-factor perturbations (entity, relation, or time) to test sensitivity to fine-grained factual changes. Extensive evaluation of 18 state-of-the-art LMMs exposes a critical "binding hallucination," revealing that current high performance is often built on fragile visual shortcuts rather than true understanding. KnowDR-REC thus serves as a pivotal diagnostic instrument, steering future research toward the genuine integration of perception and reasoning.

[Long] LADR: Locality-Aware Dynamic Rescue for Efficient Text-to-Image Generation with Diffusion Large Language Models ▶ *Underline Video*

Chenglin Wang, Yucheng Zhou, Shuang Chen, Tao Wang and Kai Zhang

Discrete Diffusion Language Models have emerged as a compelling paradigm for unified multimodal generation, yet their deployment is hindered by high inference latency arising from iterative decoding. Existing acceleration strategies often require expensive re-training or fail to leverage the 2D spatial redundancy inherent in visual data. To address this, we propose Locality-Aware Dynamic Rescue (LADR), a training-free method that expedites inference by exploiting the spatial Markov property of images. LADR prioritizes the recovery of tokens at the "generation frontier", regions spatially adjacent to observed pixels, thereby maximizing information gain. Specifically, our method integrates morphological neighbor identification to locate candidate tokens, employs a risk-bounded filtering mechanism to prevent error propagation, and utilizes manifold-consistent inverse scheduling to align the diffusion trajectory with the accelerated mask density. Extensive experiments on four text-to-image generation benchmarks demonstrate that our LADR achieves an approximate 4 x speedup over standard baselines. Remarkably, it maintains or even enhances generative fidelity, particularly in spatial reasoning tasks, offering a state-of-the-art trade-off between efficiency and quality.

[Long] LAFaCT: Attribution-based Localization and Focused Sequential Analysis of Fact-Critical Tokens for Hallucination Detection ▶ *Underline Video*

Xin Wang, Jiahao Li, Licheng Zhang and Zhendong Mao

Large Language Models (LLMs) suffer from hallucinations, severely undermining their reliability. While white-box hallucination detection methods that leverage hidden states prevail, they fail to identify and focus on fact-critical information when analyzing token sequences. To address this, we propose LAFaCT, a Localize-then-Analyze detection framework. It first localizes fact-critical tokens using Factual Criticality, a novel metric derived from feature attribution. A subsequent stage then performs a focused sequential analysis on their hidden states. Extensive experiments on eight benchmarks and multiple model families confirm LAFaCT as the new state-of-the-art, with in-depth analyses validating the effectiveness of its core token-localization strategy.

[Long] Latent Attention Denoising: A Training-Free Energy-Based Framework for Mitigating Hallucinations in Vision-Language Models ▶ *Underline Video*

Zhiwen Luo, Siyu Jiang, Weilong Jiang and Kun He

Visual hallucination remains a major obstacle to the reliability of Large Vision-Language Models (LVLMs). We argue that this issue originates from a fundamental statistical misspecification: the conventional softmax attention implicitly assumes i.i.d. noise, yet real LVLM attention patterns exhibit structured and competitive biases (e.g., attention sinks) that violate this assumption. To address this mismatch, we introduce Latent Attention Denoising (LAD), a principled and training-free framework that recasts attention calibration as a one-step score-based denoising process. LAD employs an interpretable energy function to derive an analytic score and applies a single

Langevin-inspired update to actively steer corrupted attention logits toward more faithful configurations. This intervention imposes negligible computational overhead and operates at a speed comparable to standard greedy decoding. Extensive evaluations across diverse architectures confirm that LAD achieves superior performance on both generative and discriminative tasks, effectively mitigating hallucinations while maintaining efficiency comparable to standard decoding.

[Long] LCR-RAG: Enhancing Logical Consistency in Retrieval-Augmented Generation via Neuro-symbolic Reinforcement Learning ▶ *Underline Video*

Wenxiang Zheng, Guo Tang, Shixin Jiang, Liangyu Huo, Xiyuan Zhang, Jian Xie and Ming Liu

Retrieval-Augmented Generation (RAG) is widely used to ground large language models (LLMs) in external knowledge and improve factual accuracy. Prior work has explored iterative and self-reflective mechanisms to refine reasoning, but these approaches rely on internal model judgment and lack formally grounded, verifiable feedback. As a result, RAG systems may still produce logically inconsistent or contradictory answers in multi-step reasoning. In this paper, we propose LCR-RAG, a framework that integrates neuro-symbolic verification with reinforcement learning to explicitly optimize logical consistency. The core of our approach is a Logic-Consistency-driven Reward (LCR), which converts discrete logical signals, such as contradictions or incomplete inference chains, into a structured reward signal. This reward guides a PPO-based agent to iteratively rewrite queries and correct reasoning errors. Experiments on HotpotQA, ASQA, and TriviaQA show that LCR-RAG consistently outperforms strong RAG baselines, with ablation results indicating that the LCR mechanism is the primary source of improvement, even under noisy or conflicting retrieval conditions.

[Findings] Linear Semantic Segmentation for Low-Resource Spoken Dialects ▶ *Underline Video*

Kirill Chirkunov, Younes Samih, Abed Alhakim Freihat and Hanan Aldarmaki

Semantic segmentation is a core component of discourse analysis, yet existing models are primarily developed and evaluated on high-resource written text, limiting their effectiveness on low-resource conversational varieties. In particular, dialectal Arabic exhibits informal syntax, code-switching, and weakly marked discourse structure that challenge standard semantic segmentation approaches for text. In this paper, we introduce a new multi-genre benchmark (more than 1000 samples) for semantic segmentation in Arabic, focusing on dialectal discourse. The benchmark covers casual telephone conversations, code-switched podcasts, expressive dialogue, and broadcast news, and was annotated and validated by native Arabic annotators. Using this benchmark, we show that segmentation models performing well on MSA news genres degrade on dialectal conversational texts. We further propose a segmentation model that targets local semantic coherence and robustness to discourse discontinuities, consistently outperforming strong baselines on dialectal non-news genres. The benchmark and approach generalize to other low-resource spoken languages.

[Findings] LLEOT: A Privacy-Enhancing Offsite Tuning Framework via Loss Landscape Elevation ▶ *Underline Video*

Jin Zhong, Jinglin Liang, Tongtong Yang, Zijian Xie, Shuangping Huang and Hanlin Gu

Adapting large language models (LLMs) to domain-specific tasks via fine-tuning is often infeasible: models are protected by intellectual property, while sensitive data cannot be shared due to privacy regulations. A promising paradigm, Offsite Tuning (OT), addresses this challenge by constructing an emulator of the original model. Data owners leverage the emulator to train an adapter on downstream data, which is then plugged back into the original model, enabling knowledge transfer without transmitting either the original model or the raw data. However, emulators constructed by existing OT-based methods often retain substantial inference capabilities, thereby exposing model capability privacy and posing risks of misuse. To address this, we propose Loss Landscape Elevation Offsite Tuning (LLEOT), a framework that secures data privacy as well as model parameter and capability privacy. At its core, Loss Landscape Elevation (LLE) enforces a fixed margin between the loss landscapes of the emulator and the original model. We theoretically demonstrate that LLE simultaneously (i) degrades emulator inference via perplexity amplification and (ii) preserves gradient alignment, ensuring consistent convergence for adapter training. Extensive experiments confirm that LLEOT achieves strong adaptation performance while effectively mitigating emulator misuse. Code is available at https://github.com/Z-eloto/LLEOT.

[Long] LLMs as Knowledge Graph Refiners: Mitigating Factual Inconsistencies in Generative Knowledge Extraction ▶ *Underline Video*

Donghyun Kim, Hyeongjun Yang, Seokju Hwang, Kyong-Ho Lee and Chanhee Lee

Knowledge graphs (KGs) provide a structured representation of real-world facts as triples consisting of entities and their relationships. With the rapid progress of large language models (LLMs), recent studies increasingly explore LLMs for end-to-end KG construction from text. In particular, generative knowledge extraction (GKE) builds KGs by directly generating structured triples from documents. However, generation errors are inevitable, and the resulting KGs often contain triples that do not align with the facts expressed in the source text. To address these issues, we propose GraphRefine, a framework that performs triple-level refinement on KGs constructed via GKE. We first analyze factual inconsistencies that arise in GKE and categorize their types based on a human evaluation. We then construct training data reflecting these types and fine-tune an LLM as a KG refiner. Given a draft KG, the fine-tuned refiner selects a refinement operation for each triple and, if needed, deletes, edits, or rewrites it to reduce factual inconsistencies. Extensive experiments demonstrate that GraphRefine goes beyond deletion-only approaches and improves KG quality from diverse perspectives.

[Long] MACS: Modality-Aware Capacity Scaling for Efficient Multimodal MoE Inference ▶ *Underline Video*

Chuan Wu, Bo Li and shaolin Zhu

Mixture-of-Experts Multimodal Large Language Models (MoE MLLMs) suffer from a significant efficiency bottleneck during Expert Parallelism (EP) inference due to the straggler effect. This issue is worsened in the multimodal context, as existing token-count-based load balancing methods fail to address two unique challenges: (1) Information Heterogeneity, where numerous redundant visual tokens are treated equally to semantically critical ones, and (2) Modality Dynamics, where varying visual to text ratios across tasks lead to resource misallocation. To address these challenges, we propose MACS (Modality-Aware Capacity Scaling), a training-free inference framework. Specifically, MACS introduces an Entropy-Weighted Load mechanism to quantify the semantic value of visual tokens, addressing information heterogeneity. Additionally, the Dynamic Modality-Adaptive Capacity mechanism allocates expert resources based on the real-time modal composition of the input. Extensive experiments demonstrate that MACS significantly outperforms existing methods on various multimodal benchmarks, providing a novel and robust solution for the efficient deployment of MoE MLLMs in EP inference.

[Long] Make LLMs See Like Investigators, Not Just Think More: The Role of Structured Analysis in Investigative Reasoning ▶ *Underline Video*

Jaewook Lee, Myeong-Cheol Kang and Jong-hun Shin

Criminal investigators and intelligence analysts have developed structured analytic techniques to evaluate competing hypotheses under incomplete information. This study examines whether such human expert investigative methodologies are also effective for narrative-based culprit inference in large language models (LLMs). Focusing on the task of analyzing evidence from complex narratives and identifying the perpetrator among suspects, we conducted experiments on 10 LLMs using the MuSR murder mystery benchmark. The PRISM framework, which applies investigative techniques, consistently outperformed existing general-purpose strategies across all models, with its effectiveness manifesting regardless of model scale. Ablation studies revealed that the hypothesis structuring stage is particularly crucial, accounting for 89% of the methodological improvement beyond information filtering. This suggests that domain-specific structures that specify what to analyze are more effective in LLM reasoning than simply increasing the number of reasoning paths.

[Long] MASPO: Unifying Gradient Utilization, Probability Mass, and Signal Reliability for Robust and Sample-Efficient LLM Reasoning ▶ *Underline Video*

Xiaoliang Fu, Jiaye Lin, Yangyi Fang, Binbin Zheng, Chaowen Hu, Zekai Shao, Cong Qin, Lu Pan, Ke Zeng and Xunliang Cai

Existing Reinforcement Learning with Verifiable Rewards (RLVR) algorithms, such as GRPO, rely on rigid, uniform, and symmetric trust region mechanisms that are fundamentally misaligned with the complex optimization dynamics of Large Language Models (LLMs). In this paper, we identify three critical challenges in these methods: (1) inefficient gradient utilization caused by the binary cutoff of hard clipping, (2) insensitive probability mass arising from uniform ratio constraints that ignore the token distribution, and (3) asymmetric signal reliability stemming from the disparate credit assignment ambiguity between positive and negative samples. To bridge these gaps, we propose Mass-Adaptive Soft Policy Optimization (MASPO), a unified framework designed to harmonize these three dimensions. MASPO integrates a differentiable soft Gaussian gating to maximize gradient utility, a mass-adaptive limiter to balance exploration across the probability spectrum, and an asymmetric risk controller to align update magnitudes with signal confidence. Extensive evaluations demonstrate that MASPO serves as a robust, all-in-one RLVR solution, significantly outperforming baselines. Our code is available at: https://github.com/FlyTune/MASPO-RL.

[Long] Mathematical Proof as a Litmus Test: Revealing Failure Modes of Advanced Large Reasoning Models ▶ *Underline Video*

Dadi Guo, Jiayu Liu, Zhiyuan Fan, Zhitao He, Haoran Li, Yuxin Li, Yumeng Wang and Yi R. Fung

Large reasoning models (e.g., R1, o3) have demonstrated remarkable mathematical problem-solving abilities. However, the high reported accuracy of these advanced models on popular datasets and reliance on purely numerical evaluation often mask their true reasoning shortcomings. To address this, we propose leveraging the inherent rigor and methodological complexity of mathematical proofs as a diagnostic tool to expose these hidden failures. Specifically, we introduce the RFMDataset (Reveal Failure Modes), a collection of 200 diverse mathematical proof problems to thoroughly evaluate the performance of advanced models. Our in-depth analysis of their failures uncovers 10 fine-grained error types, which shows fundamental limitations in current large reasoning models: 1) Large reasoning models still have limited capability in generating entirely correct mathematical proofs, with some models solving less than 20% of problems and even making mistakes on fundamental ones; 2) models exhibit a diverse spectrum of reasoning failures, prominently demonstrating the lack of guarantees for the correctness and rigor intermediate reasoning steps; and 3) models show hallucination and incompleteness during the reasoning process. Our findings also reveal that directly prompting models to self-reflect on specific failure modes is insufficient to resolve the current logical dilemmas, necessitating domain knowledge and formal verification.

[Long] MeanAudio: Fast and Faithful Text-to-Audio Generation with Mean Flows ▶ *Underline Video*

Xiquan Li, junxi liu, Yuzhe Liang, Zhikang Niu, Wenxi Chen and Xie Chen

Recent years have witnessed remarkable progress in Text-to-Audio Generation (TTA), providing sound creators with powerful tools to transform inspirations into vivid audio. Yet despite these advances, current TTA systems often suffer from slow inference speed, which greatly hinders the efficiency and smoothness of audio creation. In this paper, we present MeanAudio, a fast and faithful text-to-audio generator capable of rendering realistic sound with only one function evaluation (1-NFE). MeanAudio leverages: (i) the MeanFlow objective with guided velocity target that significantly accelerates inference speed, (ii) an enhanced Flux-style transformer with dual text encoders for better semantic alignment and synthesis quality, and (iii) an efficient instantaneous-to-mean curriculum that speeds up convergence and enables training on consumer-grade GPUs. Through a comprehensive evaluation study, we demonstrate that MeanAudio achieves

state-of-the-art performance in single-step audio generation. Specifically, it achieves a real-time factor (RTF) of 0.013 on a single NVIDIA RTX 3090, yielding a 100x speedup over SOTA diffusion-based TTA systems. Moreover, MeanAudio also shows strong performance in multi-step generation, enabling smooth transitions across successive synthesis steps.

[Long] MessToClean: Evidence-Grounded Structure-Preserving Reconstruction for Real-World Degraded Exam Paper Images ► *Underline Video*

Jiayi Tuo, Cheng Tang, Zihan Wang, Chenyue Zhou, Yao Li, Yanbiao Ma, Chao Wang, Wei Dai, Mingxuan Wang, Shitong Qin and Ziwei Zhao

Intelligent education systems often collect exam sheets as in-the-wild photos. These photos often suffer from distortions and noise caused by handwriting and occlusions, collectively referred to as Real-World Degraded Exam Images (RDEI). Structure-preserving reconstruction is key to converting RDEI into structured assets for downstream educational applications. Existing Multimodal Large Language Models (MLLMs) often fail under RDEI, leading to disrupted structure and evidence-unsupported hallucinations. To tackle these challenges, we propose MessToClean, a backbone-agnostic, evidence-driven pipeline that treats off-the-shelf MLLMs as interchangeable components. By grounding extraction in pixel-aligned evidence and enforcing post-hoc consistency auditing on recovered structures, MessToClean mitigates unsupported hallucinations and enhances both controllability and structural fidelity in question-level reconstruction. We curate RDEI-Exam from our educational platforms and evaluate across 12 state-of-the-art MLLM backbones. Across these, MessToClean improves stem consistency by 1.01-3.18%, figure consistency by 0.50-49.16%, and refusal F1 by 1.06-10.88% across question types.

[Long] METER: Evaluating Multi-Level Contextual Causal Reasoning in Large Language Models ► *Underline Video*

Pengfeng Li, Chen Huang, Chaoqun Hao, Hongyao Chen, Xiao-Yong Wei, Wenqiang Lei and See-Kiong Ng

Contextual causal reasoning is a critical yet challenging capability for Large Language Models (LLMs). Existing benchmarks, however, often evaluate this skill in fragmented settings, failing to ensure context consistency or cover the full causal hierarchy. To address this, we pioneer METER to systematically benchmark LLMs across all three levels of the causal ladder under a unified context setting. Our extensive evaluation of various LLMs reveals a significant decline in proficiency as tasks ascend the causal hierarchy. To diagnose this degradation, we conduct a deep mechanistic analysis via both error pattern identification and internal information flow tracing. Our analysis reveals two primary failure modes: (1) LLMs are susceptible to distraction by causally irrelevant but factually correct information at lower level of causality; and (2) as tasks ascend the causal hierarchy, faithfulness to the provided context degrades, leading to a reduced performance. We belive our work advances our understanding of the mechanisms behind LLM contextual causal reasoning and establishes a critical foundation for future research. Our code and dataset are available at https://github.com/SCUNLP/METER.

[Long] MIND Your Reasoning: A Meta-Cognitive Intuitive-Reflective Network for Dual-Reasoning in Multimodal Stance Detection ► *Underline Video*

Bingbing Wang, Zhengda Jin, Bin Liang, Wenjie Li, Jing Li, Ruifeng Xu and Min Zhang

Multimodal Stance Detection (MSD) is a crucial task for understanding public opinion on social media. Existing methods predominantly operate by learning to fuse modalities. They lack an explicit reasoning process to discern how inter-modal dynamics, such as irony or conflict, collectively shape the user's final stance, leading to frequent misjudgments. To address this, we advocate for a paradigm shift from learning to fuse to learning to reason. We introduce MIND, a Meta-cognitive Intuitive-reflective Network for Dual-reasoning. Inspired by the dual-process theory of human cognition, MIND operationalizes a self-improving loop. It first generates a rapid, intuitive hypothesis by querying evolving Modality and Semantic Experience Pools. Subsequently, a meta-cognitive reflective stage uses Modality-CoT and Semantic-CoT to scrutinize this initial judgment, distill superior adaptive strategies, and evolve the experience pools themselves. These dual experience structures are continuously refined during training and recalled at inference to guide robust and context-aware stance decisions. Extensive experiments on the MMSD benchmark demonstrate that our MIND significantly outperforms most baseline models and exhibits strong generalization.

[Long] MM-StanceDet: Retrieval-Augmented Multi-modal Multi-agent Stance Detection ► *Underline Video*

Weihai Lu, Zhejun Zhao, Yanshu Li and Huan He

Multimodal Stance Detection (MSD) is crucial for understanding public discourse, yet effectively fusing text and image, especially with conflicting signals, remains challenging. Existing methods often face difficulties with contextual grounding, cross-modal interpretation ambiguity, and single-pass reasoning fragility. To address these, we propose Retrieval-Augmented Multi-modal Multi-agent Stance Detection (MM-StanceDet), a novel multi-agent framework integrating Retrieval Augmentation for contextual grounding, specialized Multimodal Analysis agents for nuanced interpretation, a Reasoning-Enhanced Debate stage for exploring perspectives, and Self-Reflection for robust adjudication. Extensive experiments on five datasets demonstrate MM-StanceDet significantly outperforms state-of-the-art baselines, validating the efficacy of its multi-agent architecture and structured reasoning stages in addressing complex multimodal stance challenges.

[Demo] MOOSE-Copilot: A Web-Based Interactive Assistant for Unified Exploratory- and Fine-Grained Scientific Hypothesis Discovery ► *Underline Video*

Hongran An and Zonglin Yang;

Large language models (LLMs) have shown remarkable potential in scientific hypothesis discovery. However, existing approaches treat divergent exploratory ideation and convergent fine-grained refinement as isolated tasks, lacking a principled mechanism for human guidance.

We present MOOSE-Copilot, the first unified framework to bridge this abstraction gap through a formalized human-AI interaction protocol. Our system empowers scientists to actively steer the automated process via three explicit guiding signals: initial blueprints, inter-stage routing, and regenerative feedback. Quantitative evaluations demonstrate that injecting these structured expert signals significantly outperforms purely autonomous baselines, establishing a theoretical upper bound for hypothesis quality through iterative refinement. Furthermore, we operationalize this paradigm into an accessible web-based interface featuring interactive tree visualization. By reducing the cognitively demanding discovery process into manageable subtasks of conceptual selection and critique, MOOSE-Copilot enables interdisciplinary researchers to effectively orchestrate and accelerate end-to-end scientific breakthroughs.

[Long] Multi-Task Representation Alignment on Language Understanding: A Mutual Information Perspective ▶ *Underline Video*
Dou Hu, Lingwei Wei, Hongjiang Xiao, Songlin Hu and Yuan Zhang
Multi-task learning (MTL) enables joint learning over multiple tasks based on shared representations, but suffers from task interference issue during optimization. Existing works mainly focus on task balancing or probabilistic modeling but fail to address the issue since they struggle to learn sufficient representations for all target tasks. To address this, we propose a multi-task representation alignment (MTRA) framework to achieve task-specific alignment and self-alignment on the shared representations from a mutual information perspective. MTRA ensures that the learned representations contain task-relevant features while mitigating the negative effects of task-irrelevant features. First, we design a task-specific alignment objective to align the shared representations and task-specific representations with the expected targets of all tasks via information maximization. Besides, we design a self-alignment objective to eliminate task-irrelevant features via conditional information minimization. Experiments on two multi-task language benchmarks show that MTRA outperforms 13 representative MTL methods under the same settings, particularly under label-noisy and data-constrained conditions. Further analysis shows that the learned shared representations exhibit sufficient task informativeness and superior alignment properties.

[Long] NL => Schedule: Evaluate Multitask Scheduling Capability of Large Language Models ▶ *Underline Video*
Wenrui Liao, Weihong Du, Yi Li, Hongru Liang and Wenqiang Lei
Automated schedule generation for multitask from natural language descriptions has huge potential in modern industry. While classic methods bypass language complexities by using pre-formatted matrices, and recent LLM+solver approaches introduce new fragilities by relying on solver-specific code generation. This raises critical questions: Can large language models (LLMs) solve this NL => Schedule task end-to-end well(RQ1)? If the answer is "no", where do they fall short(RQ2)? And how can their capabilities be enhanced (RQ3)? To answer these questions, we introduce NL => Schedule, the first benchmark for this task, equipped with a dataset of 240 description-schedule pairs constructed from real-world materials and a rigorous evaluation suite. Our evaluation of nine state-of-the-art LLMs reveals the limitations of different LLMs in procedure grounding and the strengths of advanced LLMs in global planning via local analysis. To address these shortcomings, we propose Mans, a novel multi-agent framework. Extensive experiments show that Mans achieves more robust performance comparable to six state-of-the-art LLM+solver methods. We hope NL => Schedule and Mans will serve as a solid foundation for automatic scheduling.

[Long] OMIBench: Benchmarking Olympiad-Level Multi-Image Reasoning in Large Vision-Language Models ▶ *Underline Video*
Qiguang Chen, Chengyu Luan, Jiajun Wu, Qiming Yu, Yi Yang, Yizhuo Li, Jingqi Tong, Xiachong Feng, Libo Qin and Wanxiang Che
Large vision-language models (LVLMs) have made substantial advances in reasoning tasks at the Olympiad level. Nevertheless, current Olympiad-level multimodal reasoning benchmarks for these models often emphasize single-image analysis and fail to exploit contextual information across multiple images. We present OMIBench, a benchmark designed to evaluate Olympiad-level reasoning when the required evidence is distributed over multiple images. It contains problems from biology, chemistry, mathematics, and physics Olympiads, together with manually annotated rationales and evaluation protocols for both exact and semantic answer matching. Across extensive experiments on OMIBench, we observe meaningful performance gaps in existing models. Even the strongest LVLMs, such as Gemini-3-Pro, attain only about 50% on the benchmark. These results position OMIBench as a focused resources for studying and improving multi-image reasoning in LVLMs.

[Long] One Refiner to Unlock Them All: Inference-Time Reasoning Elicitation via Reinforcement Query Refinement ▶ *Underline Video*
Yixiao Zhou, Dongzhou Cheng, zhiliang wu, Yi Yang, Yu Cheng and Hehe Fan
Large Language Models (LLMs) often fail to utilize their latent reasoning capabilities due to a distributional mismatch between ambiguous human inquiries and the structured logic required for machine activation. Existing alignment methods either incur prohibitive O(N) costs by fine-tuning each model individually or rely on static prompts that fail to resolve query-level structural complexity. In this paper, we propose ReQueR (Reinforcement Query Refinement), a modular framework that treats reasoning elicitation as an inference-time alignment task. We train a specialized Refiner policy via Reinforcement Learning to rewrite raw queries into explicit logical decompositions, treating frozen LLMs as the environment. Rooted in the classical Zone of Proximal Development from educational psychology, we introduce the Adaptive Solver Hierarchy, a curriculum mechanism that stabilizes training by dynamically aligning environmental difficulty with the Refiner's evolving competence. ReQueR yields consistent absolute gains of 1.3%-7.2% across diverse architectures and benchmarks, outperforming strong baselines by 2.1% on average. Crucially, it provides a promising paradigm for one-to-many inference-time reasoning elicitation, enabling a single Refiner trained on a small set of models to effectively unlock reasoning in diverse unseen Solvers. Code is available at https://github.com/newera-xiao/ReQueR.

[Industry] Optimal Expert-Attention Allocation in Mixture-of-Experts: A Scalable Law for Dynamic Model Design ▶ *Underline Video*

Junzhuo Li, Peijie Jiang, Changxin Tian, Jia Liu, Zhiqiang Zhang and Xuming Hu

This paper presents a novel extension of neural scaling laws to Mixture-of-Experts (MoE) models, focusing on the optimal allocation of compute between expert and attention sub-layers. As MoE architectures have emerged as an efficient method for scaling model capacity without proportionally increasing computation, determining the optimal expert-attention compute ratio becomes critical. We define the ratio r as the fraction of total FLOPs per token dedicated to the expert layers versus the attention layers, and explore how this ratio interacts with the overall compute budget and model sparsity. Through extensive experiments with GPT-style MoE Transformers, we empirically find that the optimal ratio r^\{ } follows a power-law relationship with total compute and varies with sparsity. Our analysis leads to an explicit formula for r^\{ }, enabling precise control over the expert-attention compute allocation. We generalize the Chinchilla scaling law by incorporating this architectural parameter, providing a new framework for tuning MoE models beyond size and data. Our findings offer practical guidelines for designing efficient MoE models, optimizing performance while respecting fixed compute budgets.

[SRW] Optimizing Packing and Shuffling Strategies for Enhanced Performance in Generative Language Models ▶ *Underline Video*

Yanbing Chen; Ruilin Wang; Zihao Yang; Lavender Yao Jiang; Eric Karl Oermann

Packing and shuffling tokens is a common practice in training auto-regressive language models to prevent overfitting and improve efficiency. Documents are typically concatenated to chunks of maximum sequence length (MSL) and shuffled in chunks of tokens (atom-size chunk), possibly breaking context within documents. An alternative approach is padding, which only includes one document per chunk. To optimize both packing strategies (concatenation vs padding), we explored the optimal atom size for shuffling and compared performance and efficiency. We found that in the most common setup (where average document length is greater than MSL), matching atom size to MSL yields the lowest perplexity, controlling for dataset. Also, padding yields lower final perplexity than concatenation at the cost of lower efficiency. This trade-off informs the choice of shuffling and packing methods in training LMs.

[Demo] pAtChWoRK: Patching the Pieces of Public Procurement Documents ▶ *Underline Video*

Lorena Calvo-Bartolomé, Saúl Blanco Fortes, Erick Cedeño and Jerónimo Arenas-García

Public procurement data is legally open, yet practically locked inside thousands of unstructured PDFs and inconsistent portal metadata. pAtChWoRK starts with these fragmented, unstructured sources then leverages a hybrid pipeline (traditional NLP with LLM-based technologies) to restructure this information into a navigable knowledge base. Specifically, pAtChWoRK corrects manual classification errors, extracts complex unstructured fields such as award and solvency criteria and tenders' objectives, and assists users in easily navigating the tender landscape. This unified process enables more effective handling of the transparency bottlenecks that hinder competition and oversight in public administration. A user study with practitioners across different procurement

[Long] PersonalAlign: Hierarchical Implicit Intent Alignment for Personalized GUI Agent with Long-Term User-Centric Records ▶ *Underline Video*

Yibo Lyu, Gongwei Chen, Rui Shao, Weili Guan and Liqiang Nie

While GUI agents have shown strong performance under explicit and completion instructions, real-world deployment requires aligning with users more complex implicit intents. In this work, we highlight Hierarchical Implicit Intent Alignment for Personalized GUI Agent (PersonalAlign), a new agent task that requires agents to leverage long-term user records as persistent context to resolve omitted preferences in vague instructions and anticipate latent routines by user state for proactive assistance. To facilitate this study, we introduce AndroidIntent, a benchmark designed to evaluate agents ability in resolving vague instructions and providing proactive suggestions through reasoning over long-term user records. We annotated 775 user-specific preferences and 215 routines from 20k long-term records across different users for evaluation. Furthermore, we introduce Hierarchical Intent Memory Agent (HIM-Agent), which maintains a continuously updating personal memory and hierarchically organizes user preferences and routines for personalization. Finally, we evaluate a range of GUI agents on AndroidIntent, including GPT-5, Qwen3-VL, and UI-TARS, further results show that HIM-Agent significantly improves both execution and proactive performance by 15.7% and 7.3%.

[Long] Personalizing LLMs with Binary Feedback: A Preference-Calibrated Optimization Framework ▶ *Underline Video*

Xilai Ma, Liye Zhao, Weijun Yao, Haibing Di, Wenya Wang and Jing Li

Large Language Model (LLM) personalization aims to align model behaviors with individual user preferences. Existing methods often focus on isolated user histories, neglecting the essential role of inter-user differences. We propose C-BPO, a framework that personalizes LLMs via preference-calibrated binary signals. By treating target user data as positive feedback and other users' data as an auxiliary set of implicit negative signals, C-BPO captures distinct inter-user differences. To mitigate the preference overlap issue, where shared task knowledge is erroneously penalized, we derive an objective grounded in Positive-Unlabeled (PU) learning theory. This approach purifies negative signals by subtracting "positive bias", ensuring alignment with unique idiosyncrasies without compromising general helpfulness. Empirical experiments across various personalization tasks and backbone LLMs show C-BPO consistently outperforms baselines, demonstrating the efficacy of preference-calibrated binary signals in modeling inter-user differences.

[Short] PExA: Parallel Exploration Agent for Complex Text-to-SQL ▶ *Underline Video*

Tanmay Parekh, Ella Hofmann-Coyle, Shuyi Wang, Sachith Sri Ram Kothur, Srivas Prasad and Yunmo Chen

LLM-based agents for text-to-SQL often struggle with latency-performance trade-off, where performance improvements come at the cost of latency or vice versa. We reformulate text-to-SQL generation within the lens of software test coverage where the original query is prepared with a suite of test cases with simpler, atomic SQLs that are executed in parallel and together ensure semantic coverage of the original query. After iterating on test case coverage, the final SQL is generated only when enough information is gathered, leveraging the explored test case SQLs to ground the final generation. We validated our framework on a state-of-the-art benchmark for text-to-SQL, Spider 2.0, achieving a new state-of-the-art with 70.2% execution accuracy.

[Long] Powering Verifiable Learning via Automated Evolutionary Data Synthesis ▶ *Underline Video*

He Du, Bowen Li, Aijun Yang, Siyang He, Qipeng Guo, Kai Chen and Dacheng Tao

Reliable verifiable data has become a key driver of capability gains in modern language models, enabling stable reinforcement learning with verifiable rewards and effective distillation that transfers competence across math, coding, and agentic tasks. Yet constructing generalizable synthetic verifiable data remains difficult due to hallucination-prone generation, and weak or trivial verification artifacts that fail to separate strong from weak solutions. Existing approaches often rely on task-specific heuristics or post-hoc filters that do not transfer across domains and lack a principled, universal evaluator of verifiability. In this work, we introduce an evolutionary, task-agnostic, strategy-guided, executably-checkable data synthesis framework that, from minimal seed supervision, jointly synthesizes problems, diverse candidate solutions, and verification artifacts, and iteratively discovers strategies via a consistency-based evaluator that enforces agreement between human-annotated and strategy-induced checks. This pipeline upgrades filtering into principled synthesis: it reliably assembles coherent, verifiable training instances and generalizes without domain-specific rules. Our experiments demonstrate the effectiveness of the proposed approach under both RLVR and model distillation training paradigms. The results show that training with our synthesized data yields significant improvements on both the LiveCodeBench and AgentBench-OS tasks, highlighting the robust generalization of our framework.

[Long] PR-XAI: PageRank-Based Feature Attribution for Transformers ▶ *Underline Video*

Behrooz Azarkhalili, Linyi Li and Maxwell W. Libbrecht

We introduce PR-XAI, a feature attribution method for transformer models based on the PageRank algorithm. The proposed PR-XAI models the attention mechanism as a directed graph, with weights derived from attention weights and their gradients. Evaluations across five well-known text classification datasets and three different architectures show that PR-AG, one variant of PR-XAI, outperforms state-of-the-art attribution methods in faithfulness and classification metrics, with significant gains on long-form text.

[Findings] Preserving Fairness and Safety in Quantized LLMs Through Critical Weight Protection ▶ *Underline Video*

Muhammad Alif Al Hakim, Alfan Farizki Wicaksono and Fajri Koto

Quantization is widely adopted to reduce the computational cost of large language models (LLMs); however, its implications for fairness and safety, particularly in dynamic quantization and multilingual contexts, remain underexplored. In this work, we conduct a systematic study of how static and dynamic quantization methods impact fairness and safety across benchmarks measuring intrinsic and extrinsic bias and safety alignment. For fairness, we evaluate English, French, Dutch, Spanish, and Turkish; for safety, we focus on English, Korean, and Arabic. Our findings reveal that quantization consistently degrades fairness and safety, with dynamic methods demonstrating greater stability than static ones. Moreover, fairness degradation varies across languages, while safety deterioration is especially pronounced in non-English settings. To address these risks, we introduce Critical Weight Protection, a novel technique that identifies and preserves fairness- and safety-critical weights during quantization. This approach mitigates bias and safety deterioration without costly retraining or alignment, maintaining trustworthiness while retaining efficiency.

[Findings] PretrainRL: Alleviating Factuality Hallucination of Large Language Models at the Beginning ▶ *Underline Video*

Langming Liu, Kangtao Lv, Haibin Chen, Weidong Zhang, Yejing Wang, Shilei Liu, Xin Tong, Yujin Yuan, Yongwei Wang, Wenbo Su and Bo Zheng

Large language models (LLMs), despite their powerful capabilities, suffer from factual hallucinations where they generate verifiable falsehoods. We identify a root of this issue: the imbalanced data distribution in the pretraining corpus, which leads to a state of "low-probability truth" and "high-probability falsehood". Recent approaches, such as teaching models to say "I don't know" or post-hoc knowledge editing, either evade the problem or face catastrophic forgetting. To address this issue from its root, we propose PretrainRL, a novel framework that integrates reinforcement learning into the pretraining phase to consolidate factual knowledge. The core principle of PretrainRL is "debiasing then learning." It actively reshapes the model's probability distribution by down-weighting high-probability falsehoods, thereby making "room" for low-probability truths to be learned effectively. To enable this, we design an efficient negative sampling strategy to discover these high-probability falsehoods and introduce novel metrics to evaluate the model's probabilistic state concerning factual knowledge. Extensive experiments on three public benchmarks demonstrate that PretrainRL significantly alleviates factual hallucinations and outperforms state-of-the-art methods.

[Long] ProactiveEval: A Unified Evaluation Framework for Proactive Dialogue Agents ▶ *Underline Video*

Tianjian Liu, Fanqi Wan, Jiajian Guo and Xiaojun Quan

Proactive dialogue has emerged as a critical and challenging research problem in advancing large language models (LLMs). Existing works predominantly focus on domain-specific or task-oriented scenarios, which leads to fragmented evaluations and limits the comprehensive exploration of models' proactive dialogue abilities. In this work, we propose ProactiveEval, a unified framework for evaluating proactive

dialogue capabilities of LLMs. This framework decomposes proactive dialogue into target planning and dialogue guidance, establishing evaluation metrics across various domains. Moreover, it also enables the automatic generation of diverse and challenging evaluation data. Based on the proposed framework, we develop 328 evaluation environments spanning 6 distinct domains. Through experiments with 22 different types of LLMs, we show that DeepSeek-R1 and Claude-3.7-Sonnet exhibit exceptional performance on target planning and dialogue guidance tasks, respectively. Finally, we investigate how reasoning capabilities influence proactive behaviors and discuss their implications for future model development. Our code and data are available at the https://github.com/liutj9/ProactiveEval.

[Long] Probing the Safety Robustness of LLMs in Latent Space ▶ *Underline Video*

Tianle Gu, Kexin Huang, Zongqi Wang, Yixu Wang, Jie Li, Xin Wang, Yang Yao, Yujiu Yang, Yan Teng and Yingchun Wang

Safety alignment is a fundamental prerequisite for building trustworthy artificial general intelligence. Despite substantial progress in safety alignment techniques, empirical evidence shows that aligned large language models can still produce unsafe responses under minor internal perturbations, revealing a robustness gap in existing safety mechanisms at the latent representation level. In this paper, we study the robustness evaluation of safety alignment under latent-space perturbations. We introduce Activation Steering Attack (ASA), and leverage the Negative Log-Likelihood (NLL) as a diagnostic signal to probe the local sensitivity of safety behaviors in latent space. By measuring a models likelihood under controlled perturbations to its hidden representations, we assess the stability of its original responses. The probing signal is model-agnostic and supervision-free, enabling a general and reproducible diagnostic metric for analyzing safety robustness. Leveraging these probes, we systematically uncover a set of previously underexplored empirical findings, including (1) non-stationarity of layer vulnerabilities, revealing that the most vulnerable layer is an unstable property and even relocates after robustness training; (2) instance-level alignment with cross-layer consistency, where specific inputs remain universally vulnerable across the entire model hierarchy; (3) compositional effects of ASA, characterized by its incremental accumulation across sequential decoding steps and its potential for prompt-level jailbreak effectiveness.

[Long] Profiling-Free Mixed-Precision Quantization for MoE LLMs via Fuzzy Rule Interpolation ▶ *Underline Video*

Huachen Qi, Ruiyu Zhuo, Bowen Shi, Xiang Chang, Fei Chao, Changjing Shang and Qiang Shen

Large Language Models continue to scale in size and capability, driving substantial computational and memory demands. Mixture-of-Experts (MoE) architectures alleviate this cost by activating only a sparse subset of experts per token, enabling efficient scaling without proportional increases in inference compute. However, quantization in MoE models remains challenging due to heterogeneous sensitivity across experts and their internal linear layers. Existing mixed-precision frameworks such as Mixed-precision Quantization for MoE (MxMoE) require full quantization-loss evaluation for expert-layer-and-bit configurations, incurring prohibitive profiling cost. To address this, we propose FRI-MxMoE, a profiling-free mixed-precision quantization framework built on Fuzzy Rule Interpolation, designed as a drop-in replacement for the loss estimation component in MxMoE. By constructing a fuzzy rule base in the intra-expert layer feature space (bit-width, activation variance, parameter scale), our method predicts quantization error from only sparse samples, eliminating the need for dense profiling. Extensive experiments demonstrate that FRI-MxMoE accelerates the profiling phase by up to 15.7x (on DeepSeek-V2) while achieving comparable or slightly superior zero-shot accuracy (e.g., +1.04% on DeepSeekV2-Lite) compared to the baseline. This enables continuous sensitivity modeling, preserves accuracy under mixed-precision allocation, and reduces offline computation by orders of magnitude.

[Short] Protein-STORY: Semantic Text-Oriented Representation Yields biologically meaningful Protein embeddings ▶ *Underline Video*

Nabil Ibtehaz and Daisuke Kihara

Unsupervised representation learning using masked language modeling on the language of life has transformed protein research, enabling the analysis of a protein universe that is expanding at an exponential pace. However, most current models rely solely on sequence data, overlooking decades of expert-curated biological knowledge stored in natural language. While recent multimodal and knowledge-graph-based approaches attempt to bridge this gap, they often rely on shallow functional labels that lack the contextual depth of full textual narratives. We present Protein-STORY, a general pipeline that synthesizes protein embeddings from diverse, multi-source text descriptions. At the core of our approach is a novel network architecture designed for the semantic compression of multi document embeddings, which integrates high-fidelity functional and structural insights into a unified representation. Our experiments demonstrate that Protein-STORY produces biologically meaningful embeddings ($r \sim 0.75$) that outperform existing models on diverse downstream tasks (+2 pts F1 in function prediction). Furthermore, by projecting the story of a protein into a natural language semantic space, our model enables effective zero-shot text-prompted protein search.

[Long] Quantifying Metric and Model Agreement in Bias Evaluation of Large Language Models ▶ *Underline Video*

Arash Asgari, Huan Wu, Amirreza Naziri, Mojtaba Kolahdouzi and Laleh Seyyed-Kalantari

Bias evaluation in large language models (LLMs) uses many metrics and benchmarks, but lacks a systematic way to measure agreement across bias metrics and models. As a result, improvements observed under one metric may contradict another, and model rankings may reflect benchmark-specific artifacts rather than stable bias profiles. In this work, we introduce Metric Agreement Score (MeAS) and Model Agreement Score (MoAS), which quantify cross-metric and cross-model agreement in bias rankings, respectively. We apply these measures to eight LLMs, seven bias metrics, and nine corpora. Our results reveal disagreement among both metrics and models: Contrary to expectations, we find that metrics within the same category (generation-based and probabilistic) often behave independently of each other. For instance, HONEST shows independence with toxicity metrics, and the Context Association Test shows no correlation with Language Modeling Bias metric. At the model level, DeepSeek-family models invert bias rankings relative to most others, indicating that the model

family strongly shapes specific bias profiles. These findings challenge the assumption that bias mitigation is universally transferable and highlight the need for agreement-aware evaluation.

[Long] QuantileMark: A Message-Symmetric Multi-bit Watermark for LLMs ▶ *Underline Video*
Junlin Zhu, Baizhou Huang and Xiaojun Wan
As large language models become standard backends for content generation, practical provenance increasingly requires multi-bit watermarking. In provider-internal deployments, a key requirement is message symmetry: the message itself should not systematically affect either text quality or verification outcomes. Vocabulary-partition watermarks can break message symmetry in low-entropy decoding: some messages are assigned most of the probability mass, while others are forced to use tail tokens. This makes embedding quality and message decoding accuracy message-dependent. We propose QuantileMark, a white-box multi-bit watermark that embeds messages within the continuous cumulative probability interval [0, 1). At each step, QuantileMark partitions this interval into M equal-mass bins and samples strictly from the bin assigned to the target symbol, ensuring a fixed 1/M probability budget regardless of context entropy. For detection, the verifier reconstructs the same partition under teacher forcing, computes posteriors over latent bins, and aggregates evidence for verification. We prove message-unbiasedness, a property ensuring that the base distribution is recovered when averaging over messages. This provides a theoretical foundation for generation-side symmetry, while the equal-mass design additionally promotes uniform evidence strength across messages on the detection side. Empirical results on C4 continuation and LFQA show improved multi-bit recovery and detection robustness over strong baselines, with negligible impact on generation quality.

[Long] Question Difficulty Estimation for Large Language Models via Answer Plausibility Scoring ▶ *Underline Video*
Jamshid Mozafari, Bhawna Piryani and Adam Jatowt
Estimating question difficulty is a critical component in evaluating and improving large language models (LLMs) for question answering (QA). Existing approaches often rely on readability formulas, retrieval-based signals, or popularity statistics, which may not fully capture the reasoning challenges posed to modern LLMs. In this paper, we introduce Q-DAPS (Question Difficulty based on Answer Plausibility Scores) method, a novel approach that estimates question difficulty by computing the entropy of plausibility scores over candidate answers. We systematically evaluate Q-DAPS across four prominent QA datasets, TriviaQA, NQ, MuSiQue, and QASC, demonstrating that it consistently outperforms baselines. Moreover, Q-DAPS shows strong robustness across hyperparameter variations and question types. Extensive ablation studies further show that Q-DAPS remains robust across different plausibility estimation paradigms, model sizes, and realistic settings. Human evaluations further confirm strong alignment between Q-DAPSs difficulty estimates and human judgments of question difficulty. Overall, Q-DAPS provides an interpretable, scalable, and bias-resilient approach to question difficulty estimation in modern QA systems.

[Long] Question Tells You Where the Answer Is: Intention-aware Long-Context KV Cache Compression ▶ *Underline Video*
Liang Zhao, Xiaocheng Feng, Weihong Zhong, Lei Huang, kun Zhu, Baoxin Wang, Dayong Wu, Guoping Hu, Ting Liu and Bing Qin
The increasing context window greatly extends the capabilities of large language models, but on the other hand, it incurs an unaffordable memory overhead and computational latency due to the increasing Key-Value (KV) cache size. Recent KV cache compression methods manage to reduce the cache size by dropping irrelevant KVs. However, these methods often fail to identify crucial KVs for generation while excluding others accurately, resulting in severe information loss. To address this gap, we propose IntentKV, an intention-aware KV cache eviction method that identifies and retains crucial KVs according to the attention distribution of intention, which semantically reflects the user's goal and determines which part of the context is relevant. The consistency between the semantics and attention distribution is further substantiated through meticulously designed experiments. On this basis, IntentKV first distinguishes intention tokens from the vanilla context tokens based on their attention distribution distances. Then, the block-wise cumulative attention is calculated via aggregating the intention token attention. Finally, blocks that acquire high cumulative attention are picked and stored in KV cache. We evaluate our method across diverse long-context tasks and models. Results demonstrate that IntentKV can effectively maintain the model performance while reducing the KV cache size from 128K to 2K, leading to a 6.3x increase in decoding speed and 7.8x enhancement in memory efficiency compared to the default setting.

[Long] R2IF: Aligning Reasoning with Decisions via Composite Rewards for Interpretable LLM Function Calling ▶ *Underline Video*
Aijia Cheng, Kailong Wang, Ling Shi and Yongxin Zhao
Function calling empowers large language models (LLMs) to interface with external tools, yet existing RL-based approaches suffer from misalignment between reasoning processes and tool-call decisions. We propose R2IF, a reasoning-aware RL framework for interpretable function calling, adopting a composite reward integrating format/correctness constraints, Chain-of-Thought Effectiveness Reward (CER), and Specification-Modification-Value (SMV) reward, optimized via GRPO. Experiments on BFCL/ACEBench show R2IF outperforms baselines by up to 34.62% (Llama3.2-3B on BFCL) with positive Average CoT Effectiveness (0.05 for Llama3.2-3B), enhancing both function-calling accuracy and interpretability for reliable tool-augmented LLM deployment.

[Long] RATE: Reviewer Profiling and Annotation-free Training for Expertise Ranking in Peer Review Systems ▶ *Underline Video*
Weicong Liu, Zixuan Yang, Yibo Zhao and Xiang Li
Reviewer assignment is increasingly critical yet challenging in the LLM era, where rapid topic shifts render many pre-2023 benchmarks outdated and where proxy signals poorly reflect true reviewer familiarity. We address this evaluation bottleneck by introducing LR-bench, a high-fidelity, up-to-date benchmark curated from 2024–2025 AI/NLP manuscripts with five-level self-assessed familiarity ratings collected

via a large-scale email survey, yielding 1,055 expert-annotated paper–reviewer–score annotations. We further propose a reviewer-centric ranking framework that distills each reviewers recent publications into compact keyword-based profiles and fine-tunes an embedding model with weak preference supervision constructed from heuristic retrieval signals, enabling to match each manuscript against a reviewer profile directly. Across the LR-bench and the CMU gold-standard dataset, our approach consistently achieves state-of-the-art performance, outperforming strong embedding baselines by a clear margin. We release LR-bench at https://huggingface.co/datasets/Gnociew/LR-bench, and an github repository at https://github.com/Gnociew/RATE-Reviewer-Assignment.

[Long] ReActR: Reasoning through Error-Activated Reflection for LLM Post-Training ▶ *Underline Video*

Lina Sun

Although Large Language Models (LLMs) have demonstrated substantial proficiency in reasoning, current approaches focus disproportionately on scaling correct training samples, underexploring the value of incorrect reasoning trajectories. Motivated by how humans learn from mistakes, we propose ReActR (Reasoning through Error-Activated Reflection), a framework that enhances reasoning by learning reflective behaviors from erroneous trajectories. Specifically, ReActR comprises data construction and training. First, we synthesize multi-turn erroneous reasoning dataset spanning diverse error types and difficult levels via self-generation and targeted error generation. Second, we enhance the model's capabilities through Supervised Fine-Tuning (SFT) on synthesized data and then apply Group Relative Policy Optimization (GRPO) with multiple reward signals to further refine reasoning performance. Extensive experiments across five benchmarks and three LLMs demonstrate that ReActR effectively enhances reasoning performance. Notably, on Llama-3-8B, ReActR achieves an average improvement of 3.5% across the five datasets.

[Long] Read As Human: Compressing Context via Parallelizable Close Reading and Skimming ▶ *Underline Video*

Jiwei Tang, Shilei Liu, Zhicheng Zhang, Qingsong Lv, Runsong Zhao, Tingwei Lu, Langming Liu, Haibin Chen, Yujin Yuan, Hai-Tao Zheng, Wenbo Su and Bo Zheng

Large Language Models (LLMs) demonstrate exceptional capability across diverse tasks. However, their deployment in long-context scenarios is hindered by two challenges: computational inefficiency and redundant information. We propose RAM (Read As HuMan), a context compression framework that adopts an adaptive hybrid reading strategy, to address these challenges. Inspired by human reading behavior (i.e., close reading important content while skimming less relevant content), RAM partitions the context into segments and encodes them with the input query in parallel. High-relevance segments are fully retained (close reading), while low-relevance ones are query-guided compressed into compact summary vectors (skimming). Both explicit textual segments and implicit summary vectors are concatenated and fed into decoder to achieve both superior performance and natural language format interpretability. To refine the decision boundary between close reading and skimming, we further introduce a contrastive learning objective based on positive and negative query-segment pairs. Experiments demonstrate that RAM outperforms existing baselines on multiple question answering and summarization benchmarks across two backbones, while delivering up to a 12x end-to-end speedup on long inputs (average length 16K; maximum length 32K).

[Long] RealChart2Code: Bridging the Gap in Real-World Chart-to-Code Generation via Multi-Task Evaluation ▶ *Underline Video*

Jiajun Zhang, Yuying Li, Zhixun Li, Xingyu Guo, Jingzhuo Wu, Leqi Zheng, Yiran Yang, Jianke Zhang, Qingbin Li, Shannan Yan, Changguo Jia, Junfei Wu, Zilei Wang, Qiang Liu and Liang Wang

Vision-Language Models (VLMs) have demonstrated impressive capabilities in code generation across various domains. However, their ability to replicate complex, multi-panel visualizations from real-world data remains largely unassessed. To address this gap, we introduce RealChart2Code, a new large-scale benchmark with over 2,800 instances grounded in authentic datasets and featuring tasks with clear analytical intent. Crucially, it is the first benchmark to systematically evaluate chart generation from large-scale raw data and assess iterative code refinement in a multi-turn conversational setting. Our comprehensive evaluation of 14 leading VLMs on RealChart2Code reveals significant performance degradation compared to simpler benchmarks, highlighting their struggles with complex plot structures and authentic data. Our analysis uncovers a substantial performance gap between proprietary and open-weight models and confirms that even state-of-the-art VLMs often fail to accurately replicate intricate, multi-panel charts. These findings provide valuable insights into the current limitations of VLMs and guide future research directions.

[Industry] Reason-Code: Reliable Code Generation via Test-Driven Monte Carlo Tree Search ▶ *Underline Video*

Zixu Li and Zhiqi Peng

Large Language Models (LLMs) are widely used for code generation, but their performance degrades on tasks requiring multi-step logical reasoning. In practice, reliability is often improved through multi-sample inference, but its cost grows linearly with the sample size, making it impractical under strict latency constraints. To address these constraints, we propose Reason-Code, an inference-time framework that formulates code generation as a search process guided by execution feedback. This system integrates Monte Carlo Tree Search (MCTS) with a lightweight execution sandbox, where candidate programs are evaluated via unit tests. To control inference cost, Reason-Code adopts a conditional budgeting strategy that activates search only when greedy generation fails. While Best-of-N sampling can improve accuracy, its inference cost grows linearly with N, making it impractical under strict latency budgets. Experiments on HumanEval and MBPP show that Reason-Code matches strong sampling baselines with lower token cost and no regression, while UCB-guided exploration outperforms random search under the same budget.

[Long] ReasonRank: Empowering Passage Ranking with Strong Reasoning Ability ▶ *Underline Video*

Wenhan Liu, Xinyu Ma, Weiwei Sun, Yutao Zhu, Yuchen Li, Dawei Yin and Zhicheng Dou
Large Language Model (LLM) based listwise ranking has shown superior performance in many passage ranking tasks. With the development of Large Reasoning Models (LRMs), many studies have demonstrated that step-by-step reasoning during test-time helps improve listwise ranking performance. However, due to the scarcity of reasoning-intensive training data, existing rerankers perform poorly in many complex ranking scenarios, and the ranking ability of reasoning-intensive rerankers remains largely underdeveloped. In this paper, we first propose an automated reasoning-intensive training data synthesis framework, which sources training queries and passages from diverse domains and applies DeepSeek-R1 to generate high-quality training labels. To empower the listwise reranker with strong reasoning ability, we further propose a two-stage training approach, which includes a cold-start supervised fine-tuning (SFT) stage and a reinforcement learning (RL) stage. During the RL stage, we design a novel multi-view ranking reward tailored to the multi-turn nature of listwise ranking. Extensive experiments demonstrate that our trained reasoning-intensive reranker ReasonRank outperforms existing baselines significantly and also achieves much lower latency than the pointwise reranker.

[Long] Reference Attack: A New Cross-Modal Jailbreaking Attack against Multimodal Large Language Models ► *Underline Video*
Yulong Wang, Yifei Fu and Jiayi Gao
Red team testing, an effective proactive method for evaluating the security of multimodal large language models (MLLMs), requires an expanding toolkit alongside the development of MLLM safeguards. We propose the Reference Attack, a powerful tool for red team testing against MLLMs. The Reference Attack is a reference-guided cross-modal jailbreak method that enhances existing prompt-to-image injection attacks by exploiting MLLMs semantic reconstruction capabilities. Our method embeds malicious prompts in non-text modalities (e.g., images, spreadsheets) and constructs recursive symbolic references in text, enabling MLLMs to gradually recover and generate harmful content through layered reference resolution. The attack introduces a new vector that circumvents conventional content moderation by exploiting MLLMs' lack of security checks during cross-modal reference resolution. We evaluate the Reference Attack on leading MLLMs, including ChatGPT, Gemini, Claude, and the widely used open-source LLaMA model, and achieved an attack success rate of over 93% across all tested models. Compared to state-of-the-art attacks, Reference Attack achieves higher success rates than all baselines under identical evaluation, with a maximum gain of 70.8%. Our study reveals a critical gap in MLLM security and highlights the need for strict security auditing of cross-modal interactions in future content moderation.

[Long] ReFL: Reflective Feedback Learning for Hallucination Detection of Large Language Models ► *Underline Video*
Cunhang Fan, Jun Zhang, Xue Zhang, Shuai Zhang, Zhao Lv, Jianhua Tao and Zhengqi Wen
Large Language Models (LLMs) often generate factually incorrect content, known as hallucinations, which undermine the reliability and safety of their outputs. Existing hallucination detection methods either depend on external knowledge sources, incurring high computational costs and limiting real-time applicability, or extract the model's internal states, leading to poor generalization. To address these issues, this paper proposes ReFL, a hallucination detection framework. ReFL leverages corrective in-context learning to dynamically guide LLMs to recognize their own prediction errors and adjust internal representations, critically without updating model weights. Specifically, by introducing a corrective in-context learning strategy, where triplets of input text, model prediction, and ground-truth label are embedded into the prompt to make the model explicitly aware of its own errors. The model reflects on prior outputs to adjust its internal states and generate semantically structured representations better aligned with factuality. This feedback mechanism encourages the model to shape a more coherent semantic space and enhances the LLMs internal sensitivity to hallucinations. Experimental results on two benchmark datasets demonstrate that ReFL consistently outperforms existing methods, achieving state-of-the-art performance.

[Long] REFLEX: Self-Refining Explainable Fact-Checking via Verdict-Anchored Style Control ► *Underline Video*
Chuyi Kong, Wei Gao, Jing Ma, Hongzhan Lin and Yuxi Sun
The prevalence of fake news on social media calls for automated fact-checking systems that deliver not only accurate verdicts but also faithful explanations. However, existing large language model (LLM)-based methods often overlook deceptive misinformation styles in generated explanations, producing unfaithful rationales that may mislead human judgment. They also rely heavily on external knowledge sources, which can introduce hallucinations and incur substantial latency, undermining both reliability and responsiveness in real-time settings. To address these limitations, we propose REason-guided Fact-checking with Latent EXplanations (REFLEX), a self-refining framework that explicitly controls reasoning style by anchoring explanations to the predicted verdict. REFLEX leverages self-disagreement veracity signals between a backbone model and its fine-tuned variant to construct steering vectors, thereby naturally disentangling factual content from stylistic cues. Experiments on a real-world benchmark show that REFLEX achieves state-of-the-art performance under LLaMA-series models using only 465 self-refined samples. Owing to its transferability, REFLEX also yields gains of up to 7.54 Macro-F1 points on in-the-wild data. Further analysis shows that our method effectively mitigates faithful hallucination, leading to both more reliable explanations and more accurate verdicts than prior explainable fact-checking approaches.

[Long] Rethinking Depression Prediction from a Fine-Grained Subscore Modeling Perspective via Multi-Task Learning ► *Underline Video*
Zhenguang Wang, Bo Li, Wenhui Tan, Peng Cao, Yang Wang, Jia Duan, Fei Wang and Osmar Zaiane
Standard depression assessment relies on instruments such as the clinician-rated Hamilton Depression Rating Scale (HAMD) and the patient-reported Patient Health Questionnaire (PHQ-8), but manual scoring is time-consuming and subject to inter-rater variability. Prior automated approaches typically regress a single total score or coarse severity category, lacking the fine-grained subscore-level supervision

needed for precise clinical diagnosis. To address this, we propose MTSP (Multi-Task Subscore Prediction), a fine-grained model for subscore prediction via multi-task learning. MTSP achieves state-of-the-art performance on the public E-DAIC dataset (MAE 3.48, RMSE 4.57) and generalizes well to the public PDCH and a large-scale private clinical dataset (CIDH), outperforming total-score regression baselines and Qwen3-14B direct scoring. We further show that multi-task learning is essential, subscore-level supervision improves assessment by better capturing symptom-cluster structure, and prior constraints plus task-level self-paced learning enhance robustness to subscore difficulty and annotation noise.

[Industry] Rethinking Scale: Deployment Trade-offs of Small Language Models under Agent Paradigms ▶ *Underline Video*
Xinlin Wang
Despite the impressive capabilities of large language models, their substantial computational costs, latency, and privacy risks hinder their widespread deployment in real-world applications. Small Language Models (SLMs) with fewer than 10 billion parameters present a promising alternative; however, their inherent limitations in knowledge and reasoning curtail their effectiveness. Existing research primarily focuses on enhancing SLMs through scaling laws or fine-tuning strategies while overlooking the potential of using agent paradigms, such as tool use and multi-agent collaboration, to systematically compensate for the inherent weaknesses of small models. To address this gap, this paper presents the first large-scale, comprehensive study of <10B open-source models under three paradigms: (1) the base model, (2) a single agent equipped with tools, and (3) a multi-agent system with collaborative capabilities. Our results show that single-agent systems achieve the best balance between performance and cost, while multi-agent setups add overhead with limited gains. Our findings highlight the importance of agent-centric design for efficient and trustworthy deployment in resource-constrained settings.

[Long] Rethinking the Idiomaticity Decomposability Hypothesis: Evidence from Distributional Learning ▶ *Underline Video*
Maggie Mi, Golzar Atefi, Atsuki Yamaguchi, Felix Gers, Aline Villavicencio and Nafise Sadat Moosavi
Idioms can be analysed in terms of their decomposability, the extent to which constituent meanings contribute to the figurative whole. Decomposability is thought to predict syntactic flexibility. Usage-based accounts instead attribute idiom behaviour to distributional experience, such as speaker familiarity and predictability. We examine these views using contextualised language models as controlled distributional learners. We propose a model-internal measure of decomposability and relate it to human ratings, syntactic flexibility, and predictability while tracking idiom learning during pretraining. Model-derived decomposability correlates weakly with human judgments and shows a small but consistent negative relationship with syntactic flexibility. Pretraining analyses show that stabilisation of idiom representations in models is not explained by frequency alone. Instead, surprisal, decomposability, and frequency all contribute, with decomposability showing the strongest training-dependent effect.

[Short] Revisiting Evaluation of Question Answering Systems in Low-Resource Indic Languages: Bridging Human and Metric Alignment ▶ *Underline Video*
Anuj Kumar, Satyadev Ahlawat, Yamuna Prasad and Virendra Singh
Evaluating Question Answering (QA) systems in low-resource Indic languages remains challenging due to the scarcity of annotated data, high linguistic diversity, and the absence of reliable evaluation metrics. Many Indian languages are severely underrepresented, making it difficult to accurately assess the performance of Large Language Models (LLMs) on QA tasks. Commonly used metrics like BLEU, ROUGE-L, and BERTScore, while successful in machine translation and resource-rich scenarios, tend to perform poorly in low-resource QA settings. These metrics often exhibit issues such as compressed scoring ranges, excessive zero scores, and weak alignment with human judgments. To overcome these limitations, this work introduces the LRM^2QAS (Language Robust Multi-aspect Metrics for Question Answering Systems). This composite evaluation framework integrates semantic similarity, factual completeness, numerical accuracy, and contextual relevance. The proposed metric is evaluated across eight Indic-language QA tasks using multiple LLMs, as well as on open-domain benchmarks NaturalQuestions (NQ) and TriviaQA (TQ). Across all settings, LRM^2QAS demonstrates stronger agreement with human evaluation, as measured by Pearson, Spearman, and Kendall correlation coefficients. Experimental findings highlight that LRM^2QAS provides more precise distinctions between model outputs and aligns more closely with human judgment, offering a reliable framework for evaluating multilingual QA in low-resource Indic languages.

[Long] RLSeek: Evidence-Grounded Reasoning for RAG Hallucination Detection ▶ *Underline Video*
Zhaoheng Huang, Dacheng Wen, Yutao Zhu, Xiaoying Lian, Yushi Liang, Kai Hao, Nan Li, Liangjie Zhang, Qi Zhang, Zhicheng Dou, Ji-Rong Wen and Fangzhao Wu
Large language models (LLMs) in retrieval-augmented generation systems can still produce hallucinations, generating content that is unsupported or contradicted by the source texts and undermines reliability. Recent work addressed this problem by training span-level hallucination detectors using reinforcement learning (RL) and chain-of-thought (CoT) reasoning. In this work, we show through error analysis that incorrect predictions by existing reasoning-based detectors are strongly associated with CoT processes that lack explicit grounding in source evidence, particularly when verification steps do not quote or verify claims against the retrieved documents. This behaviour contrasts with human verification practices in benchmarks such as RAGTruth, where evidence quotation is a prerequisite for determining hallucinated spans. Motivated by this observation, we propose an evidence-grounded RL framework, namely RLSeek, to explicitly enforce active evidence seeking during CoT reasoning by requiring quotation of relevant source segments at each verification step. Experiments on the RAGTruth and NewsSum dataset demonstrate consistent improvements in hallucination span detection performance, with limited additional reasoning overhead and improved robustness in out-of-domain settings.

[Long] Robust Tool Use via Fission-GRPO: Learning to Recover from Execution Errors ► *Underline Video*

Zhiwei Zhang, Fei zhao, Rui Wang, Zezhong WANG, Bin Liang, Jiakang Wang, Yao Hu, Shaosheng Cao and Kam-Fai Wong

Large language models (LLMs) can call tools effectively, yet they remain brittle in multi-turn execution: after a tool-call error, smaller models often fall into repetitive invalid re-invocations instead of interpreting the feedback and recovering. This failure mode persists because current training paradigms do not explicitly teach models how to recover from execution errors. In particular, standard reinforcement learning (RL) collapses rich failure experience into sparse negative rewards, while pre-collected error-correction datasets become mismatched to the policy's evolving failure modes. To bridge this gap, we propose Fission-GRPO, a framework that converts execution errors into on-policy corrective supervision within the RL training loop. Our core mechanism fissions each failed trajectory into a new training instance by augmenting it with diagnostic feedback from a fine-tuned Error Simulator, then resampling multiple recovery rollouts on-policy. This enables the model to learn from the precise errors it makes during exploration, rather than from static, pre-collected error cases. On BFCL v4 Multi-Turn, Fission-GRPO improves the error recovery rate of Qwen3-8B by 5.7% absolute and overall accuracy by 4.0% (from 42.75% to 46.75%), outperforming both RL baselines and specialized tool-use agents. The method further generalizes to TAU-Bench and TAU2-Bench, achieving leading results across most settings with gains up to +17.4%.

[Long] SADA: Bridging In-Context Learning and Fine-Tuning via State-Aligned Distillation Adapters ► *Underline Video*

Wenhao Gao, Tianlong Wang, Wei Jia, Linhao Zhang, Aiwei Liu, Miao Fan and Zhou Xiao

Prompt-based in-context learning (ICL) and parameter fine-tuning are two dominant paradigms for incorporating external information into large language models (LLMs), but they incur high inference costs or require expensive retraining. To bridge this gap, context-to-parameter mapping converts prompts into temporary adapter weights. However, we identify a critical failure mode in existing methods: hidden-state collapse, where the adapter-augmented model's internal states diverge sharply from the full-context oracle in deeper layers. We trace this failure to two coupled gaps: suboptimal Input-Selection and inadequate Supervision-Signal. To address these issues, we propose SADA (State-Aligned Distillation Adapters). We establish the attention-block output as a principled feature interface to improve input selection and introduce state-alignment distillation to enforce consistency between the adapter-augmented model and the full-context oracle. Experiments on long-context language modeling (PG19) and downstream NLU and summarization benchmarks show that SADA consistently outperforms strong baselines like StreamAdapter and GenerativeAdapter, achieving performance comparable to ICL while significantly reducing memory footprint and latency. We further analyze when parameterized context compression is effective and when explicit context retention remains preferable. Our code is available at https://github.com/Taylor-Gavel/SADA.git.

[Long] Scaling Beyond Context: A Survey of Multimodal Retrieval-Augmented Generation for Document Understanding ► *Underline Video*

Sensen Gao, Shanshan Zhao, Xu Jiang, Lunhao Duan, Yong Xien Chng, Qing-Guo Chen, Weihua Luo, Kaifu Zhang, Jia-Wang Bian and Mingming Gong

Document understanding is critical for applications from financial analysis to scientific discovery. Current approaches, whether OCR-based pipelines feeding Large Language Models (LLMs) or native Multimodal LLMs (MLLMs), face key limitations: the former loses structural detail, while the latter struggles with context modeling. Retrieval-Augmented Generation (RAG) helps ground models in external data, but documents multimodal nature, i.e., combining text, tables, charts, and layout, demands a more advanced paradigm: Multimodal RAG. This approach enables holistic retrieval and reasoning across all modalities, unlocking comprehensive document intelligence. Recognizing its importance, this paper presents a systematic survey of Multimodal RAG for document understanding. We propose a taxonomy based on domain, retrieval modality, and granularity, and review advances involving graph structures and agentic frameworks. We also summarize key datasets, benchmarks, and applications, and highlight open challenges in efficiency, fine-grained representation, and robustness, providing a roadmap for future progress in document AI.

[Short] SED-SFT: Selectively Encouraging Diversity in Supervised Fine-Tuning ► *Underline Video*

Yijie Chen, Yijin Liu and Fandong Meng

Supervised Fine-Tuning (SFT) followed by Reinforcement Learning (RL) has emerged as the standard post-training paradigm for large language models (LLMs). However, the conventional SFT process, driven by Cross-Entropy (CE) loss, often induces mode collapse, where models over-concentrate on specific response patterns. This lack of distributional diversity severely restricts the exploration efficiency required for subsequent RL. While recent studies have attempted to improve SFT by replacing CE loss, aiming to preserve diversity or refine the update policy, they fail to adequately balance diversity and accuracy, thereby achieving sub-optimal performance after RL. To address the mode collapse problem, we propose SED-SFT, which adaptively encourages diversity based on the token exploration space. This framework introduces a selective entropy regularization term with a selective masking mechanism into the optimization objective. Extensive experiments across eight mathematical benchmarks demonstrate that SED-SFT significantly enhances generation diversity with a negligible computational overhead increase compared with CE loss, yielding average improvements of 2.06 and 1.20 points in subsequent RL performance over standard CE-based baselines on Llama-3.2-3B-Instruct and Qwen2.5-Math-7B-Instruct, respectively.

[Industry] SemViQA: A Semantic Question Answering System for Vietnamese Information Fact-Checking ► *Underline Video*

Dien X. Tran, Nam V. Nguyen, Tran Tan Thanh, Anh T. Hoang, Dng Vn Tài, Lê Thanh Di and Phuc-Lu Le

Recent advances in LLMs have accelerated both information generation and misinformation, especially in low-resource languages like

Vietnamese, motivating robust fact-checking systems. Existing methods struggle with semantic ambiguity, homonyms, and complex linguistic structures, often trading accuracy for efficiency. We introduce SemViQA, a novel Vietnamese fact-checking framework integrating Semantic-based Evidence Retrieval (SER) and Two-step Verdict Classification (TVC). Our approach balances precision and speed, achieving state-of-the-art results with 78.97% strict accuracy on ISE-DSC01 and 80.82% on ViWikiFC, securing 1st place in the UIT Data Science Challenge. Additionally, SemViQA Faster improves inference speed 7x while maintaining competitive accuracy. SemViQA sets a new benchmark for Vietnamese fact verification, advancing the fight against misinformation.

[Long] SharedRequest: Privacy-Preserving ModelAgnostic Inference for Large Language Models ▶ *Underline Video*

Peihua Mai, Xuanrong Gao, Youlong Ding, Xianglong Du, Wei Liu and Yan Pang

With the widespread deployment of public large language models (LLMs) such as ChatGPT, protecting user prompt privacy has become an increasingly critical issue. Existing privacy-preserving inference methods sacrifice either utility or efficiency, and often require model-specific modifications that limit their compatibility. In this paper, we propose SharedRequest, a model-agnostic framework for privacy-preserving LLM inference that reformulates privacy protection at the batch level rather than the individual-prompt level. The key idea is to obscure sensitive information by mixing original prompts with noisy variants, while grouping semantically equivalent instructions to amortize the inference cost over a large batch of queries with minimal impact on LLM response quality. This design is independent of the LLM architecture, requiring no access to model parameters or architectural modification. Empirical results demonstrate that SharedRequest achieves over 20% higher utility compared to prior differential privacy baselines, and its shared-prompt mechanism reduces query cost by up to 5x compared to nonbatched inference.

[Long] ShopSimulator: Evaluating and Exploring RL-Driven LLM Agent for Shopping Assistants ▶ *Underline Video*

Pei Wang, Yanan Wu, Xiaoshuai Song, Weixun Wang, Gengru Chen, zhongwen Li, Kezhong Yan, Qi Liu, Ken Deng, Shuaibing Zhao, Shaopan Xiong, Xuepeng Liu, Xuefeng Chen, Wanxi Deng, Wenbo Su and Bo Zheng

Large language model (LLM)based agents are increasingly deployed in ecommerce shopping. To perform thorough, usertailored product searches, agents should interpret personal preferences, engage in multiturn dialogues, and ultimately retrieve and discriminate among highly similar products. However, existing research has yet to provide a unified simulation environment that consistently captures all of these aspects, and always focuses solely on evaluation benchmarks without training support. In this paper, we introduce ShopSimulator, a largescale and challenging Chinese shopping environment. Leveraging ShopSimulator, we evaluate LLMs across diverse scenarios, finding that even the bestperforming models achieve less than 40% fullsuccess rate. Error analysis reveals that agents struggle with deep search and product selection in long trajectories, fail to balance the use of personalization cues, and to effectively engage with users. Further training exploration provides practical guidance for overcoming these weaknesses, with the combination of supervised finetuning (SFT) and reinforcement learning (RL) yielding significant performance improvements.

[Long] Sparrow: Text-Anchored Window Attention with Visual-Semantic Glimpsing for Speculative Decoding in Video LLMs ▶ *Underline Video*

Libo Zhang, Zhaoning Zhang, Wangyang Hong, Peng Qiao and Dongsheng Li

Although speculative decoding is widely used to accelerate Vision-Language Models (VLMs) inference, it faces severe performance collapse when applied to Video Large Language Models (Vid-LLMs). The draft model typically falls into the trap of attention dilution and negative visual gain due to key-value cache explosion and context window mismatches. We observe a visual semantic internalization phenomenon in Vid-LLMs, indicating that critical visual semantics are implicitly encoded into text hidden states during deep-layer interactions, which renders raw visual inputs structurally redundant during deep inference. To address this, we propose the Sparrow framework, which first utilizes visually-aware text-anchored window attention via hidden state reuse to fully offload visual computation to the target model, and leverages intermediate-layer visual state bridging to train the draft model with semantic-rich intermediate states, thereby filtering out low-level visual noise. Additionally, a multi-token prediction strategy is introduced to bridge the training-inference distribution shift. Experiments show that Sparrow achieves an average speedup of 2.82x even with 25k visual tokens, effectively resolving the performance degradation in long sequences and offering a practical solution for real-time long video tasks.

[Demo] Spectra: A Mechanistic Interpretability Library for Vision-Language Models ▶ *Underline Video*

Clement Neo, Yongsen Zheng, Kwok-Yan Lam and Luke Ong

Vision-Language Models (VLMs) have become increasingly important in AI applications, yet interpretability tools for these models lag behind those available for text-only language models. While libraries like TransformerLens have enabled significant progress in understanding language models, existing tools for VLMs are limited to basic activation probing and saving. We introduce Spectra, a library specifically designed for mechanistic interpretability of VLMs that provides unified abstractions for activation patching, attention pattern analysis, and meta-functions across diverse VLM architectures. Built on HuggingFace's Transformers, our library handles architecture-specific complexities through per-checkpoint configurations while maintaining a simple, high-level interface. We demonstrate the library's capabilities by performing interpretability experiments on a counting task, showing how researchers can easily perform experiments that were previously cumbersome to do. The library currently supports Qwen2.5-VL, Qwen3-VL, LLaVA 1.5 and SmolVLM, with a design that facilitates extension to additional architectures. The library can be found at github.com/clemneo/vlm-spectra.

[Long] Spectral Disentanglement: Rank-Aware Task Adaptation for Rehearsal-free Continual Learning in LLMs ▶ *Underline Video*

Huanxuan Liao, Shizhu He, Yupu Hao, Yequan Wang, Wenhao Teng, Xiangwen Liao, Jun Zhao and Kang Liu
Continual Learning (CL) for Large Language Models (LLMs) faces a fundamental Stability-Plasticity Dilemma: balancing the plasticity to acquire new capabilities with the stability to preserve prior knowledge. While Parameter-Efficient Fine-Tuning methods, such as LoRA, enable efficient adaptation, we identify a critical flaw in current approaches termed Rank-Blindness: the enforcement of a single rank constraint across diverse tasks, which entangles task-shared and task-specific knowledge, leading to catastrophic forgetting of earlier tasks and underfitting on complex new ones. To address this, we propose SpaRTA, a novel rehearsal-free framework guided by a rank-spectrum perspective that explicitly disentangles knowledge into two orthogonal subspaces. Specifically, SpaRTA employs a low-rank branch to capture task-shared representations and a high-rank branch to model task-specific features. To integrate these complementary representations, we introduce a context-aware dynamic router that adaptively fuses the two branches based on input semantics, while an explicit orthogonality constraint minimizes interference between shared and specific parameter subspaces. This design effectively isolates task-specific updates from shared knowledge, preventing the overwriting of prior capabilities while preserving strong adaptation capacity. Extensive experiments demonstrate that SpaRTA achieves a superior stability-plasticity balance compared to single-rank baselines. Notably, the proposed spectral disentanglement strategy substantially reduces inter-task interference and yields strong zero-shot generalization on unseen tasks. Our code will be available at https://github.com/Xnhyacinth/SpaRTA.

[Findings] StanceAttack: Adversarial Attack for Stance Detection ▶ *Underline Video*
Chenye Zhao and Cornelia Caragea
Stance detection aims to ascertain whether an author's text is in favor, against, or neutral toward specific targets like public policies or social issues. While pretrained language models (PLMs) have greatly enhanced stance detection, they remain vulnerable to adversarial attacks, manipulations that maintain textual semantics but lead to incorrect predictions. Such vulnerabilities remain underexplored for stance detection. In this study, we introduce StanceAttack, an innovative adversarial attack method leveraging ChatGPT to create adversarial examples that can mislead well-trained stance detection models. We conduct experiments to evaluate our attack model by attacking state-of-the-art PLMs on two benchmark datasets. Results demonstrate that StanceAttack outperforms traditional adversarial methods with higher success rates and fewer retries. Human evaluations confirm that our adversarial examples preserve the original semantic meanings and naturalness. We share our code and data in https://github.com/chenyez/StanceAttack.

[Long] StealthGraph: Exposing Domain-Specific Risks in LLMs through Knowledge-Graph-Guided Harmful Prompt Generation ▶ *Underline Video*
Huawei Zheng, Xinqi Jiang, Sen Yang, Shouling Ji, Yingcai Wu and Dazhen Deng
Large language models (LLMs) are increasingly applied in specialized domains such as finance and healthcare, where they introduce unique safety risks. Domain-specific datasets of harmful prompts remain scarce and still largely rely on manual construction; public datasets mainly focus on explicit harmful prompts, which modern LLM defenses can often detect and refuse. In contrast, implicit harmful prompts, expressed through indirect domain knowledge, are harder to detect and better reflect real-world threats. We identify two challenges: transforming domain knowledge into actionable constraints and increasing the implicitness of generated harmful prompts. To address them, we propose an end-to-end framework that first performs knowledge-graph-guided harmful prompt generation to systematically produce domain-relevant prompts, and then applies two-strategy obfuscation rewriting to convert explicit harmful prompts into implicit variants via direct and context-enhanced rewriting. This framework yields high-quality datasets combining strong domain relevance with implicitness, enabling more realistic red-teaming and advancing LLM safety research. We release our code and datasets on GitHub.

[Long] STK-Adapter: Incorporating Evolving Graph and Event Chain for Temporal Knowledge Graph Extrapolation ▶ *Underline Video*
Shuyuan Zhao, Wei Chen, Weijie Zhang, Xinrui Hou, Junfeng Shen, Boyan Shi, Shengnan Guo, Youfang Lin and Huaiyu Wan
Temporal Knowledge Graph (TKG) extrapolation aims to predict future events based on historical facts. Recent studies have attempted to enhance TKG extrapolation by integrating TKG's evolving structural representations and textual event chains into Large Language Models (LLMs). Yet, two main challenges limit these approaches: (1) The loss of essential spatial-temporal information due to shallow alignment between TKG's graph evolving structural representation and the LLM's semantic space, and (2) the progressive dilution of the TKG's evolving structural features during LLM fine-tuning. To address these challenges, we propose the Spatial-Temporal Knowledge Adapter (STK-Adapter), which flexibly integrates the evolving graph encoder and the LLM to facilitate TKG reasoning. In STK-Adapter, a Spatial-Temporal MoE is designed to capture spatial structures and temporal patterns inherent in TKGs. An Event-Aware MoE is employed to model intricate temporal semantics dependencies within event chains. In addition, a Cross-Modality Alignment MoE is proposed to facilitate deep cross-modality alignment by TKG-guided attention experts. Extensive experiments on benchmark datasets demonstrate that STK-Adapter significantly outperforms state-of-the-art methods and exhibits strong generalization capabilities in cross-dataset task. The code is available at https://github.com/Zhaoshuyuan0246/STK-Adapter.

[Long] STRIDE-ED: A Strategy-Grounded Stepwise Reasoning Framework for Empathetic Dialogue Systems ▶ *Underline Video*
Hongru Ji, Yuyin Fan, Meng Zhao, Xianghua Li, Lianwei Wu and Chao Gao
Empathetic dialogue requires not only recognizing a users emotional state but also making strategy-aware, context-sensitive decisions throughout response generation. However, the lack of a comprehensive empathy strategy framework, explicit task-aligned multi-stage reasoning, and high-quality strategy-aware data fundamentally limits existing approaches, preventing them from effectively modeling

empathetic dialogue as a complex, multi-stage cognitive and decision-making process. To address these challenges, we propose STRIDE-ED, a STRategy-grounded, Interpretable, and DEep reasoning framework that models Empathetic Dialogue through structured, strategy-conditioned reasoning. To support effective learning, we develop a strategy-aware data refinement pipeline integrating LLM-based annotation, multi-model consistency-weighted evaluation, and dynamic sampling to construct high-quality training data aligned with empathetic strategies. Furthermore, we adopt a two-stage training paradigm that combines supervised fine-tuning with multi-objective reinforcement learning to better align model behaviors with target emotions, empathetic strategies, and response formats. Extensive experiments demonstrate that STRIDE-ED generalizes across diverse open-source LLMs and consistently outperforms existing methods on both automatic metrics and human evaluations.

[Findings] Structured and Abstractive Reasoning on Multi-modal Relational Knowledge Images ► *Underline Video*

Yichi Zhang, Zhuo Chen, Lingbing Guo, Wen Zhang and Huajun Chen

Understanding and reasoning with abstractive information from the visual modality presents significant challenges for current multi-modal large language models (MLLMs). Among the various forms of abstractive information, Multi-Modal Relational Knowledge (MMRK), which represents abstract relational structures between multi-modal entities using node-edge formats, remains largely under-explored. In particular, STructured and Abstractive Reasoning (STAR) on such data has received little attention from the research community. To bridge the dual gaps in large-scale high-quality data and capability enhancement methodologies, this paper makes the following key contributions: (i). An automatic STAR data engine to synthesize images with MMRK to build multi-modal instructions with reliable chain-of-thought thinking for various STAR tasks and (ii). A comprehsive two-stage training framework, accompanied by knowledge-informed GRPO and a suite of evaluation protocols tailored to different STAR tasks. Based upon these contributions, we introduce STAR-64K, a dataset comprising 64K high-quality multi-modal instruction samples, and conduct experiments across 8 open-source MLLMs. Experimental results show that our two-stage enhancement framework enables smaller 3B/7B models to significantly outperform GPT-4o in STAR. Additionally, we provide in-depth analysis regarding the effectiveness of various designs, data transferability, and scalability.

[Long] Synthia: Scalable Grounded Persona Generation from Social Media Data ► *Underline Video*

Vahid Rahimzadeh, Erfan Moosavi Monazzah, Mohammad Taher Pilehvar and Yadollah Yaghoobzadeh

Persona-driven simulations are increasingly used in computational social science, yet their validity critically depends on the fidelity of the underlying personas. Constructing virtual populations that are both authentic and scalable remains a central challenge. We introduce Synthia, a persona-generation framework that grounds LLM-generated personas in real social-media posts while delegating narrative construction to language models, using publicly available data from the Bluesky platform. Across multiple social-survey benchmarks, Synthia improves alignment with human opinion distributions over prior state-of-the-art approaches while relying on substantially smaller models. A multi-dimensional fairness and bias analysis shows that Synthia outperforms previous methods for most demographics across different dimensions. Uniquely, Synthia preserves interaction-graph structure among personas grounded in real social network users, enabling network-aware analysis, which we demonstrate through two homophily-focused case studies. Together, these results position Synthia as a practical and reliable framework for constructing scalable, high-fidelity, and equitable virtual populations.

[Long] TARE: Lightweight Token-Aware Representation Editing for Fine-tuning Transformer-like Models ► *Underline Video*

Yulong Wang and Siyu Zhao

Parameter-efficient fine-tuning (PEFT) must balance effectiveness and efficiency: low-rank methods can be costly, while global representation edits often underfit token-level contexts. We propose Token-Aware Representation Editing (TARE), a PEFT method that performs fine-grained, token-specific edits with a small additional inference overhead and minimal tuning. After each FFN block in a transformer-like model, we adopt a lightweight selector that scores a small pool of hidden representation editors for each token, activates only the top-k editors, and mixes their element-wise scaling/bias updates. This design achieves superior performance while maintaining computational efficiency, yielding a more favorable Pareto frontier compared to state-of-the-art (SOTA) methods. Across LLaMA-3-8B (eight knowledge reasoning and seven mathematical reasoning tasks) and RoBERTa-base/large (GLUE), TARE outperforms SOTAs (LoRA, DoRA, MiLoRA, LoReFT, and RED), achieving 86.7% (knowledge reasoning), 76.7% (mathematical reasoning), and 88.3% (GLUE) while tuning only 0.0392% of parameters using about 20 GiB peak GPU memory during training. An implementation is available at: <https://github.com/PatriciaPulec/tare>.

[Long] Task-Aware LLM Routing with Multi-Level Task-Profile-Guided Data Synthesis for Cold-Start Scenarios ► *Underline Video*

Hui Liu, Bin Zou, Kecheng Chen, Jie Liu, Wenya Wang and Haoliang Li

Large language models (LLMs) exhibit substantial variability in performance and computational cost across tasks and queries, motivating routing systems that select models to meet user-specific cost-performance trade-offs. However, existing routers generalize poorly in cold-start scenarios where in-domain training data is unavailable. We address this limitation with a multi-level task-profile-guided data synthesis framework that constructs a hierarchical task taxonomy and produces diverse question-answer pairs to approximate the test-time query distribution. Building on this, we introduce TRouter, a task-type-aware router approach that models query-conditioned cost and performance via latent task-type variables, with prior regularization derived from the synthesized task taxonomy. This design enhances TRouter's routing utility under both cold-start and in-domain settings. Across multiple benchmarks, we show that our synthesis framework alleviates cold-start issues and that TRouter delivers effective LLM routing.

[Industry] Teaching LLM to be Persuasive: Reward-Enhanced Policy Optimization for Alignment from Heterogeneous Rewards ▶ *Underline Video*

Xia Zeng, Yihan Chen, Luhui Liu, Chao Luo, Ye Chen and ZHUANGZHUORAN ZHUANGZHUORAN

We deploy large language models (LLMs) as business development (BD) agents for persuasive price negotiation in online travel agencies (OTAs). The agent must follow a multi-stage Standard Operating Procedure (SOP) and strict guardrails (no over-promising and no hallucinations), while remaining human-like and effective over long, multi-turn dialogues. We propose Reward-Enhanced Policy Optimization (REPO), a reinforcement learning post-training method that combines heterogeneous rewards: a preference-trained reward model (RM), an LLM-as-a-judge (RJ) for nuanced behaviors (e.g., emotional value and SOP compliance), and rule-based reward functions (RF) (mainly regex-based) for deterministic checks on numerics, formatting, and guardrails.In expert consensus evaluation (three human experts; 30 online conversations and 45 curated bad cases), REPO improves average dialogue rating to 4.63 (+0.33 over GRPO) and raises the share of conversations with at least one excellent response to 66.67% (+23.34 pp over GRPO), while achieving a 93.33% bad-case fix rate with 75.56% clean fixes. In a production A/B test on 9,653 real customer conversations (vs.\{ } an intent-driven dialogue system), REPO improves response rate by +12.14 pp and task success rate by +5.94 pp (p<0.001).

[Long] Think Better, Not Longer: Token-Level Marginal Utility for Efficient Reasoning in Large Reasoning Models ▶ *Underline Video*

Jiawei Li, Yang Gao, Huashan Sun and Chong Feng

While Large Reasoning Models (LRMs) have demonstrated remarkable capabilities through explicit Chain-of-Thought (CoT) generation, they frequently suffer from "overthinking". In this work, we bridge this gap by introducing Token-level Marginal Utility, which quantifies the per-token log-probability gain of the ground-truth answer. Leveraging this dense supervision signal, we propose MUTO (Marginal Utility Guided Thinking Optimization), a unified training framework designed to synthesize concise reasoning chains. Rather than relying only on coarse trajectory-level length control, MUTO identifies tokens that reduce the model's likelihood of the correct answer and penalizes such negative-utility reasoning, yielding concise yet effective CoT trajectories. Experiments on DeepSeek-R1-Distill-Qwen backbones (1.5B and 7B) across six math reasoning benchmarks show that MUTO yields a markedly better efficiency-accuracy Pareto frontier. It reduces average token usage by 87.1% at 1.5B while improving accuracy by 2.3%, and cuts tokens by 80.2% at 7B with only -0.1% accuracy change, achieving the best length-normalized accuracy among baselines.

[Long] Towards Fine-grained Audio Captioning with Multimodal Contextual Fusion ▶ *Underline Video*

Shunian Chen, Xinyuan Xie, Zheshu Chen, Owen Lee, Liyan Zhao, Zhan Su, Qilin Sun and Benyou Wang

High-quality, large-scale audio captioning is crucial for advancing audio understanding, yet current automated methods often generate captions that lack fine-grained detail and contextual accuracy, primarily due to their reliance on limited unimodal or superficial multimodal information. Drawing inspiration from human auditory perception, which adeptly integrates cross-modal cues and performs sophisticated auditory scene analysis, we introduce a novel two-stage automated pipeline. This pipeline first employs specialized pretrained models to extract diverse contextual cues (e.g., speech, music, general sounds, and visual information from associated video). A large language model (LLM) then synthesizes these rich, multimodal inputs to generate detailed and context-aware audio captions. Key contributions of this work include: (1) the proposed scalable method for fine-grained audio caption generation; (2) FusionAudio, a new large-scale dataset comprising 1.2 million such detailed captions, combined with 6 million QA pairs; and (3) enhanced audio models developed using FusionAudio, specifically a CLAP-based audio encoder with superior audio-text alignment and instruction following. This paper paves the way for more nuanced and accurate automated understanding of complex audio environments.

[Long] TOWER+: Bridging Generality and Translation Specialization in Multilingual LLMs ▶ *Underline Video*

Ricardo Rei, Nuno M. Guerreiro, José Pombal, João Alves, Pedro Teixeirinha, Amin Farajian and André F. T. Martins

Fine-tuning pretrained LLMs has proven effective for reaching state-of-the-art performance on specific tasks like machine translation. However, this process often implies sacrificing general-purpose capabilities, such as conversational reasoning and instruction-following, hampering the usefulness of the system in real-world applications requiring a mixture of skills. In this paper, we introduce Tower+, a suite of models designed to deliver strong performance on both translation and multilingual general-purpose text capabilities. We improve the TOwer (Alves et al., 2024) recipe by adding novel stages of preference optimization and reinforcement learning with verifiable rewards, in addition to continued pretraining and supervised fine-tuning. At each stage, we carefully generate and curate data to strengthen performance on translation and general-purpose tasks like coding, mathematics, and instruction-following. We develop models at multiple scales: 2B, 9B, and 72B. Our smaller models often outperform larger general-purpose open-weight and proprietary LLMs (e.g., Llama 3.3 70B, GPT-4o). Our largest model delivers best-in-class translation performance for high-resource languages, and top results on multilingual Arena Hard and IF-MT, a benchmark we introduce for evaluating both translation and instruction-following.

[Industry] Transfer Learning for Generalizable Automated LLM Improvement Pipeline for IVR Navigation ▶ *Underline Video*

Vishal Sankar Ram, Jason Kushner, Manas Paldhe and Youngseo Son

Administrative tasks in the healthcare domain share linguistic commonalities, but it can be time-consuming to manually design LLM prompts for each use case. When calling health insurers, interactive voice response (IVR) systems cause delays in patient care and increase provider burnout due to complex routing and long hold times. Thus, IVR navigation models can offer significant time savings and reduce barriers to care. We propose an automated LLM pipeline that leverages a small number of ground truth datasets to transfer specialized prompts from

one task to another; specifically, we performed a cross-task transfer of our IVR navigation logic, adapting the prompt from reaching the claims department to reaching the patient benefit department. We implemented a product-quality model which obtains high success rates in large-scale industrial healthcare settings with our proposed transfer learning pipeline with only a small number of human annotations. We obtained 82% accuracy for turn-level evaluation, surpassing an optimized human-designed prompt (79%) while reducing the prompt complexity by up to 80%

[Long] Two-Stage Regularization-Based Structured Pruning for LLMs ▶ *Underline Video*

Mingkuan Feng, Jinyang Wu, Siyuan Liu, Shuai Zhang, Hongjian Fang, Ruihan Jin, Feihu Che, Pengpeng Shao, Zhengqi Wen and Jianhua Tao

The deployment of large language models (LLMs) is largely hindered by their large number of parameters. Structural pruning has emerged as a promising solution. Prior structured pruning methods directly remove unimportant parameters based on certain metrics, which often causes knowledge loss and necessitates extensive retraining. To overcome this, we introduce a novel pruning method TRSP: Two-Stage Regularization-Based Structured Pruning for LLMs. Specifically, we multiply the output of each transformer layer by an initial learnable weight and iteratively learn these weights by adding their _1-norm as a regularization term to the loss function, serving as the first-stage regularization. Subsequently, we apply additional regularization to the difference between the output and input of layers with smaller weights, encouraging the shift of knowledge to the preserved layers. This serves as the second-stage regularization. TRSP retains more knowledge and better preserves model performance than direct parameter elimination. Through extensive experimentation we show that TRSP outperforms strong layer-wise structured pruning methods without requiring retraining. As a layer-wise pruning method, it delivers notable end-to-end acceleration, making it a promising solution for efficient LLM deployment.

[Short] UERLens: Understanding Event Relations in Large Language Models ▶ *Underline Video*

Yong Guan, Zhiyuan Li and Shaoru Guo

Events exhibit rich semantic relations that are essential for understanding the unfolding of real-world processes. Although large language models (LLMs) have achieved strong performance on event relation extraction, how event relations are internally represented and utilized remains unclear. In this paper, we present UERLens, an interpretability framework for understanding event relations in LLMs. Specifically, we first construct UERBench, a counterfactual dataset for event relation analysis that covers causal, temporal, and sub-event relations. Based on counterfactual pairs, we identify relation-sensitive internal features by comparing model activations. We then examine the functional role of these features through model manipulation, including model intervention and model training. Experimental results show that event relations are encoded through structured and layer-specific internal features. Disabling relation-sensitive features leads to performance drops of over 22%, while enhancing them yields improvements of up to 7%. Furthermore, leveraging these interpretable features to train a lightweight classifier significantly improves event relation extraction, achieving F1 gains of up to 24% for causal relations.

[Long] UMMF: Protecting Copyright of Large Vision-Language Models through Unlearning-based Multimodal Memorization Fingerprint ▶ *Underline Video*

Xiaofan Zheng, Xinghao Wang and Xiaojun Wan

Training Large Vision-Language Models (LVLMs) is costly and resource-intensive, making them valuable assets. To prevent malicious users from unauthorized commercialization of these artificial intelligence assets through fine-tuning and black-box deployment, model fingerprinting techniques aimed at verifying the ownership of LVLMs are receiving widespread attention. Existing fingerprinting techniques rely on adversarial attacks or backdoor attacks to construct trigger images for specific outputs, attributing model ownership by comparing whether the output of trigger images on suspected models matches the predetermined output. However, these methods depend on fixed-form triggers as explicit model fingerprints, which have limitations in terms of stealthiness and robustness. Inspired by unlearning research, we propose Unlearning-based Multimodal Memorization Fingerprint (UMMF). UMMF strengthens the overfitting characteristics of training samples by unlearning neighboring samples of the training samples, thereby introducing detectable regions of poor generalization in the data manifold. Compared with previous methods, our approach leverages the differences in memorization strength of LVLMs on neighboring samples as implicit model fingerprints, rather than relying on specific input-output pairs as explicit triggers. This endows it with stronger stealthiness, robustness, and adaptability. To simulate real application scenarios, we conduct extensive experiments using multiple strategies and different datasets, further demonstrating its superiority in protecting LVLM ownership.

[Findings] Understanding Generalization in Role-Playing Models via Information Theory ▶ *Underline Video*

Yongqi Li, Hao Lang, Fei Huang, Tieyun Qian and Yongbin Li

Role-playing models (RPMs) are widely used in real-world applications but underperform when deployed in the wild. This degradation can be attributed to distribution shifts, including user, character, and dialogue compositional shifts. Existing methods like LLM-as-a-judge fall short in providing a fine-grained diagnosis of how these shifts affect RPM generalization, and thus there lack formal frameworks to characterize RPM generalization behaviors. To bridge these gaps, we introduce an information-theoretic metric, named reasoning-based effective mutual information difference (R-EMID), to measure RPM performance degradation in an interpretable way. We also derive an upper bound on R-EMID to predict the worst-case generalization performance of RPMs and theoretically reveal how various shifts contribute to the RPM performance degradation. Moreover, we propose a co-evolving reinforcement learning framework to adaptively model the connection among user, character, and dialogue context and thus enhance the estimation of dialogue response generation probability, which is critical for calculating R-EMID. Finally, we evaluate the generalization performance of various RPMs using R-EMID, finding that user

shift poses the highest risk among all shifts and reinforcement learning is the most effective approach for enhancing RPM generalization. Code and data are available at https://github.com/AlibabaResearch/DAMO-ConvAI/tree/main/RPM-Generalization.

[Long] Unified Thinker: A General Reasoning Core for Image Generation ▶ *Underline Video*

Sashuai Zhou, Qiang Zhou, Jijin Hu, Hanqing Yang, Yue Cao, Junpeng Ma, Yinchao Ma, Jun Song, Tiezheng Ge, Cheng Yu, Bo Zheng and Zhou Zhao

Despite impressive progress in high-fidelity image synthesis, generative models still struggle with logic-intensive instruction following, exposing a persistent reasoning–execution gap. Meanwhile, closed-source systems (e.g., Nano Banana) have demonstrated strong reasoning-driven image generation, highlighting a substantial gap to current open-source models. We argue that closing this gap requires not merely better visual generators, but executable reasoning: decomposing high-level intents into grounded, verifiable plans that directly steer the generative process. To this end, we propose Unified Thinker, a task-agnostic reasoning architecture for general image generation, designed as a unified planning core that can plug into diverse generators and workflows. Unified Thinker decouples a dedicated Thinker from the image Generator, enabling modular upgrades of reasoning without retraining the entire generative model. We further introduce a two-stage training paradigm: we first build a structured planning interface for the Thinker, then apply reinforcement learning to ground its policy in pixel-level feedback, encouraging plans that optimize visual correctness over textual plausibility. Extensive experiments on text-to-image generation and image editing show that Unified Thinker substantially improves image reasoning and generation quality.

[Long] Vision-Language Introspection: Mitigating Overconfident Hallucinations in MLLMs via Interpretable Bi-Causal Steering ▶ *Underline Video*

Shuliang Liu, Songbo Yang, Dong Fang, Sihang Jia, Yuqi Tang, Lingfeng Su, Ruoshui Peng, Yibo Yan, Xin Zou and Xuming Hu

Object hallucination critically undermines the reliability of Multimodal Large Language Models (MLLMs), often stemming from a fundamental failure in cognitive introspection, where models blindly trust linguistic priors over specific visual evidence. Existing mitigations remain limited: contrastive decoding approaches operate superficially without rectifying internal semantic misalignments, while current latent steering methods rely on static vectors that lack instance-specific precision. We introduce Vision-Language Introspection (VLI), a training-free inference framework that simulates a metacognitive self-correction process. VLI first performs Attributive Introspection to diagnose hallucination risks via probabilistic conflict detection and localize the causal visual anchors. It then employs Interpretable Bi-Causal Steering to actively modulate the inference process, dynamically isolating visual evidence from background noise while neutralizing blind confidence through adaptive calibration. VLI achieves state-of-the-art performance on advanced models, reducing object hallucination rates by 12.67% on MMHal-Bench and improving accuracy by 5.8% on POPE.

[Long] What Deserves Memory: Adaptive Memory Distillation for LLM Agents ▶ *Underline Video*

Wenquan Ma, Jiayan Nan and Wenlong Wu

Memory systems for LLM agents struggle to determine what information deserves retention. Existing approaches rely on predefined heuristics such as importance scores, emotional tags, or factual templates, encoding designer intuition rather than learning from the data itself. Inspired by cognitive ideas, we propose Nemori, an adaptive memory distillation framework that casts the assessment of the experience's future utility as a matter of predictability. Specifically, Nemori comprises two cascading modules: Episodic Memory Integration transforms raw interactions into coherent narratives, and Semantic Knowledge Distillation extracts insights via prediction error. Centering on distillation, the framework remains agnostic to downstream management. Extensive experiments confirm that Nemori achieves strong performance, efficiency, and storage reduction. Our work suggests that observing the intrinsic properties of interaction sequences offers a viable, data-driven alternative to heuristic-based memory design.

[Long] When Agents Look the Same: Quantifying Distillation-Induced Similarity in Tool-Use Behaviors ▶ *Underline Video*

Chenghao Yang, Yuning Zhang, Zhoufutu Wen, Tao Gong, Jiaheng Liu, Qi Chu and Nenghai Yu

Model distillation is a primary driver behind the rapid progress of LLM agents, yet it often leads to behavioral homogenization. Many emerging agents share nearly identical reasoning steps and failure modes, suggesting they may be distilled echoes of a few dominant teachers. Existing metrics, however, fail to distinguish mandatory behaviors required for task success from non-mandatory patterns that reflect a model's autonomous preferences. We propose two complementary metrics to isolate non-mandatory behavioral patterns: Response Pattern Similarity (RPS) for verbal alignment and Action Graph Similarity (AGS) for tool-use habits modeled as directed graphs. Evaluating 18 models from 8 providers on -Bench and ^2-Bench against Claude Sonnet 4.5 (thinking), we find that within-family model pairs score 5.9 pp higher in AGS than cross-family pairs, and that Kimi-K2 (thinking) reaches 82.6% S_node and 94.7% S_dep, exceeding Anthropic's own Opus 4.1. A controlled distillation experiment further confirms that AGS distinguishes teacher-specific convergence from general improvement. RPS and AGS capture distinct behavioral dimensions (Pearson r = 0.491), providing complementary diagnostic signals for behavioral convergence in the agent ecosystem. Our code is available at https://github.com/Syuchin/AgentEcho.

[Long] Why Do Emotions Change? Appraisal-Guided Reasoning for Emotion-Cause Triplet Extraction in Conversations ▶ *Underline Video*

Qiao Liang, Ying Shen, Yao Liu, Tiantian Chen and Lin Zhang

Multimodal Emotion–Cause Triplet Extraction in Conversations (MECTEC) is fundamental for fine-grained affect understanding, yet it remains challenging in multi-turn, multi-speaker settings. Existing methods often make locally plausible predictions but struggle to maintain

conversation-level consistency under within-speaker emotion shifts and core events. To address this, we propose ECFlow, a unified framework that combines appraisal-guided structured generation with graph-structured reinforcement learning. ECFlow operationalizes cognitive appraisal theory into a controllable intermediate reasoning trace and constructs UMECS, a unified supervision dataset with cognitively grounded traces. It then lifts predicted and gold triplets into an Emotion–Cause Flow Graph and optimizes verifiable, structure-aware rewards for emotion-shift coherence and core-event consistency, together with task-oriented triplet rewards. Experiments on public MECTEC benchmarks show that ECFlow consistently outperforms strong baselines, achieving state-of-the-art triplet extraction and improved structure-aware metrics on emotion shifts and core events. Our code and dataset are available at https://anonymous.4open.science/r/ECFlow-E908.

[Long] Why Do Emotions Change? Appraisal-Guided Reasoning for Emotion-Cause Triplet Extraction in Conversations ▶ *Underline Video*

Qiao Liang, Ying Shen, Yao Liu, Tiantian Chen and Lin Zhang

Multimodal Emotion–Cause Triplet Extraction in Conversations (MECTEC) is fundamental for fine-grained affect understanding, yet it remains challenging in multi-turn, multi-speaker settings. Existing methods often make locally plausible predictions but struggle to maintain conversation-level consistency under within-speaker emotion shifts and core events. To address this, we propose ECFlow, a unified framework that combines appraisal-guided structured generation with graph-structured reinforcement learning. ECFlow operationalizes cognitive appraisal theory into a controllable intermediate reasoning trace and constructs UMECS, a unified supervision dataset with cognitively grounded traces. It then lifts predicted and gold triplets into an Emotion–Cause Flow Graph and optimizes verifiable, structure-aware rewards for emotion-shift coherence and core-event consistency, together with task-oriented triplet rewards. Experiments on public MECTEC benchmarks show that ECFlow consistently outperforms strong baselines, achieving state-of-the-art triplet extraction and improved structure-aware metrics on emotion shifts and core events. Our code and dataset are available at https://anonymous.4open.science/r/ECFlow-E908.

[Long] Why Steering Works: Toward a Unified View of Language Model Parameter Dynamics ▶ *Underline Video*

Ziwen Xu, Chenyan WU, Hengyu Sun, Haiwen Hong, Mengru Wang, Yunzhi Yao, Longtao Huang, Hui Xue, Shumin Deng, Zhixuan Chu, Huajun Chen and Ningyu Zhang

Methods for controlling large language models (LLMs), including local weight fine-tuning, LoRA-based adaptation, and activation-based interventions, are often studied in isolation, obscuring their connections and making comparison difficult. In this work, we present a unified view that frames these interventions as dynamic weight updates induced by a control signal, placing them within a single conceptual framework. Building on this view, we propose a unified preference-utility analysis that separates control effects into preference, defined as the tendency toward a target concept, and utility, defined as coherent and task-valid generation, and measures both on a shared log-odds scale using polarity-paired contrastive examples. Across methods, we observe a consistent trade-off between preference and utility: stronger control increases preference while predictably reducing utility. We further explain this behavior through an activation manifold perspective, in which control shifts representations along target-concept directions to enhance preference, while utility declines primarily when interventions push representations off the models valid-generation manifold. Finally, we introduce a new steering approach guided by this analysis that improves preference while better preserving utility.

[Findings] WISCA: A Lightweight Model Transition Method to Improve LLM Training via Weight Scaling ▶ *Underline Video*

Jiacheng Li, Jianchao Tan, Zhidong Yang, Pingwei Sun, Feiye Huo, Jiayu Qin, Xiangyu Zhang, Maoxin He, Guangming Tan, Weile Jia, Xunliang Cai and Tong Zhao

Transformer architecture gradually dominates the LLM field. Recent advances in training optimization for Transformer-based large language models (LLMs) primarily focus on architectural modifications or optimizer adjustments. However, these approaches lack systematic optimization of weight patterns during training. Weight pattern refers to the distribution and relative magnitudes of weight parameters in a neural network. To address this issue, we propose a Weight Scaling method called WISCA to enhance training efficiency and model quality by strategically improving neural network weight patterns, without changing network structures. By rescaling weights while preserving model outputs, WISCA indirectly optimizes the models training trajectory. Experiments demonstrate that WISCA significantly improves convergence quality (measured by generalization capability and loss reduction), particularly in LLMs with Grouped Query Attention (GQA) architectures and LoRA fine-tuning tasks. Empirical results show 5.6% average improvement on zero-shot validation tasks and 2.12% average reduction in training perplexity across multiple architectures.

[Long] XY-Tokenizer: Mitigating the Semantic-Acoustic Conflict in Low-Bitrate Speech Codecs ▶ *Underline Video*

Yitian Gong, Luozhijie Jin, Kuangwei Chen, Dong Zhang, Ruifan Deng, Xiaogui Yang, Xin Zhang, Zhaoye Fei, Qinyuan Cheng, Shimin Li and Xipeng Qiu

Speech codecs provide an important interface between continuous speech signals and large language models. An ideal codec for speech language models should not only preserve acoustic information but also capture rich semantic information. However, existing codecs struggle to balance these objectives at low bitrates. We propose XY-Tokenizer, a low-bitrate speech codec (around 1 kbps) trained with a structured multi-stage, multi-task strategy that aligns discrete speech representations with text while preserving fine-grained acoustic details for reconstruction. This design explicitly mitigates the semantic–acoustic conflict observed in prior low-bitrate codecs. Experiments show that XY-Tokenizer achieves stronger semantic alignment than representative semantic-distillation codecs such as SpeechTokenizer and Mimi, while maintaining high-quality speech reconstruction across both clean and out-of-distribution conditions. Furthermore, XY-Tokenizer consistently outperforms existing low-bitrate codecs in LLM-based speech understanding and generation tasks, demonstrating

its effectiveness as a general-purpose speech representation for speech-language modeling.

[Long] Your Reasoning Model is Secretly a Reward Model - Optimization-Free Verification from Experience ► *Underline Video*

Zhenwen Liang, Ruosen Li, Yujun Zhou, Linfeng Song, Dian Yu, Xinya Du, Haitao Mi and Dong Yu

Assessing the quality of Large Language Model (LLM) outputs becomes especially challenging in high-branching settings, where a single prompt yields many plausible candidates. Existing verifiers typically operate on the surface text (e.g., reward models, LLM judges, majority voting) or on confidence proxies derived from token probabilities, both of which can be brittle: the former can be influenced by stylistic artifacts, while the latter is often miscalibrated. In this paper, we study a third source of information—the model's hidden states—for binary correctness verification in tasks with a reliable success/failure signal (e.g., deterministic checkers or reference-grounded answers). We find that correct and incorrect solutions exhibit measurable geometric differences in their hidden-state trajectories. To isolate this signal with minimal modeling assumptions, we introduce Clue (Clustering and Experience-based Verification), a training-free, non-parametric verifier. Clue summarizes each reasoning trace by an activation delta—the difference between hidden states at the start and end of the explicit reasoning span—and predicts correctness by comparing this delta to two class centroids computed from labeled experience. Across math (AIME 24/25), scientific QA (GPQA), and a multi-domain benchmark (WebInstruct-verified), Clue improves selection and reranking, with particularly strong gains on smaller or less-calibrated models. For example, on AIME 24 with a 1.5B model, Clue raises accuracy from 56.7% (majority@64) to 70.0% (top-maj@16).

[Long] Zero-Shot Detection of LLM-Generated Text using Temperature Sensitivity ► *Underline Video*

Shixuan Ma, Jiahao Li, Zhendong Mao and Quan Wang

The widespread deployment of Large Language Models (LLMs) has spurred significant progress in the detection of LLM-generated text. However, existing detection methods often rely on statistical features that are insufficient for reliable detection; for example, even though LLM-generated and human-written texts exhibit different probability distributions in surrogate models, they can produce nearly identical entropy values, thereby conflating the two types of text. In this paper, we propose that modulating the decoding temperature and monitoring how the probability distributions respond can better probe the intrinsic discrepancies between two types of text. Building upon this insight, we introduce a new feature termed Temperature Sensitivity (TS) and demonstrate that LLM-generated text tends to exhibit higher TS than human-written text. Finally, we propose NTS, a novel and simple zero-shot detector built upon normalized temperature sensitivity. Extensive experiments across three datasets, multiple domains, and various source models demonstrate the superior effectiveness and robustness of our proposed approach. Code avaliable at: https://github.com/Shixuan-Ma/NTS.

13

Main Conference Day 3: Tuesday, July 7, 2026

Overview

01:00-02:30	**Session 15: Virtual Presentations 5**	Underline (virtual)
08:30-16:30	**Registration & Conference Check In (Full Conference & Main Conference Attendees)**	Grand Hall Foyer (L1)
09:00-10:30	**Session 16: Orals / Demos / Poster Session F**	
	Demo, Main & Finding Posters	
	See Tracks on Whova or Underline for room assignment	Grand Hall (Ground), Coronado (L4)
	Speech Processing and Spoken Language Understanding 2	Harbor A (L2)
	Interpretability and Analysis of Models for NLP 2	Harbor B&C (L2)
	LLM Efficiency 2	Harbor D (L2)
	Safety and Alignment in LLMs 2	Harbor E&F (L2)
	Ethics, Bias, and Fairness 2	Harbor G (L2)
	Multimodality and Language Grounding to Vision, Robotics and Beyond 2	Harbor H&I (L2)
	Discourse, Pragmatics, and Reasoning	Promenade (L3)
	Natural Language Generation 1	Regatta (L4)
	Industry 5: Orals	Balboa (L2, Seaport)
	BoF/Affinity: Please check the online schedule	Gaslamp C&D, Old Town A&B, La Jolla A&B
10:30-11:00	Coffee	Grand Hall
11:00-12:30	**Session 17: Orals / Demos / Poster Session G**	
	Demo, Main & Finding Posters	
	See Tracks on Whova or Underline for room assignment	Grand Hall (Ground), Coronado (L4)
	Retrieval-Augmented Language Models	Harbor A (L2)

	Human-AI Interaction/Cooperation and Human-Centric NLP 2	Harbor B&C (L2)
	Question Answering 2	Harbor D (L2)
	Special Theme: Explainability of NLP Models 2	Harbor E&F (L2)
	Mathematical, Symbolic, Neurosymbolic, and Logical Reasoning	Harbor G (L2)
	Natural Language Generation 2	Harbor H&I (L2)
	Hierarchical Structure Prediction, Syntax, and Parsing & Phonology, Morphology and Word Segmentation	Promenade (L3)
	Sentiment Analysis, Stylistic Analysis, and Argument Mining	Regatta (L4)
	Industry 6: Orals	Balboa (L2, Seaport)
	BoF/Affinity: Please check the online schedule	Gaslamp C&D, Old Town A&B, La Jolla A&B
12:30-14:00	Lunch	On Your Own
14:00-15:30	**Session 18: Memorial / ToT Awards / ACL Lifetime Award**	Harbor Ballroom (L2)
15:30-16:00	Coffee	Grand Hall
16:00-17:30	**Session 19: Award & Closing Session**	Harbor Ballroom (L2)

Detailed Program

Session 15: Virtual Presentation 5 - 01:00-02:30

Virtual Presentations 5

01:00-02:30 (TBA)

[Findings] Beyond Length Scaling: Synergizing Breadth and Depth for Generative Reward Models ▶ *Underline Video*

Qiyuan Zhang, Yufei Wang, Tianhe Wu, Can Xu, Qingfeng Sun, Kai Zheng, Xue Liu and Chen Ma

Recent advancements in Generative Reward Models (GRMs) have demonstrated that scaling the length of Chain-of-Thought (CoT) reasoning considerably enhances the reliability of evaluation. However, current works predominantly rely on unstructured length scaling, ignoring the divergent efficacy of different reasoning mechanisms: Breadth-CoT (multi-dimensional principle coverage) and Depth-CoT (substantive judgment soundness). To address this, we introduce Mix-GRM, a framework that reconfigures raw rationales into structured Breadth-CoT and Depth-CoT through a modular synthesis pipeline, subsequently employing Supervised Fine-Tuning (SFT) and Reinforcement Learning with Verifiable Rewards (RLVR) to internalize and optimize these mechanisms. Comprehensive experiments demonstrate that Mix-GRM establishes a new state-of-the-art across five benchmarks, surpassing leading open-source RMs by an average of 8.2%. Our results reveal a clear divergence in reasoning: Breadth-CoT benefits subjective preference tasks, whereas Depth-CoT excels in objective correctness tasks. Consequently, misaligning the reasoning mechanism with the task directly degrades performance. Furthermore, we demonstrate that RLVR acts as a switching amplifier, inducing an emergent polarization where the model spontaneously allocates its reasoning style to match task demands.

Virtual Presentations 5

01:00-02:30 (Zoom)

[Findings] A Diversity Diet for a Healthier Model: A Case Study of French ModernBERT ▶ *Underline Video*

Louis Estève, Christophe Servan, Thomas Lavergne and Agata Savary

Diversity has been gaining interest in the NLP community in recent years. At the same time, state-of-the-art transformer models such as ModernBERT use very large pre-training datasets, which are driven by size rather than by diversity. This summons to investigate the impact of diversity on pre-training. We do so in this study, with the express intent of reducing pre-training dataset size, while retaining at least comparable performance. We compare diversity-driven sampling algorithms, and we use the best one to pre-train several ModernBERT models on French with a fixed compute budget. We fine-tune and evaluate them on a variety of French benchmarks. We compare them with models pre-trained on randomly sampled data of commensurate size, with the same compute budget. We find that both random and diversity-driven sampling may reduce the pre-training dataset by up to 94% and the pre-training time by up to 73% while maintaining performance. Moreover, in some tasks, the inherent quality of models, estimated via head-only fine-tuning, is up to 10 points higher with diversity sampling than with random sampling.

[Demo] A System for Dynamically Tracking Content Moderation on Reddit ▶ *Underline Video*

George Baker and Bharadwaj Kadiyala;

Recent work in natural language processing, human-computer interaction, and computational social science takes interest in the study of decentralized content moderation, in which individual communities largely determine their own norms, rules, and enforcement thereof. A key challenge to this body of work is that, once moderated, content and related variables become difficult or impossible to recover; previous work often relied on 3rd-party historical data sources, but recent world events, legal disputes, and policy shifts have significantly disrupted these services, practically disabling their research use-cases. Therefore, in order to conduct new research and reproduce previous results, researchers must record content as it's created, and monitor variables of interest over time. In this work we present and publicly release a software system for the dynamic monitoring of Reddit posts, communities, and moderation actions, to enable scaleable and reproducible research on decentralized platform governance and content moderation. To the authors' knowledge, this system is presently the only available solution for general-purpose, real-time, policy-compliant longitudinal data collection on Reddit. Furthermore, the system's integration with the official Reddit API enables the collection of authentication-gated data such as community engagement metrics and moderation team information, which was impossible in previous historical data sources.

[Industry] ACRM: Multi-Agent Trajectory Learning for Automated Credit Risk Model Refreshing in Production ▶ *Underline Video*

Liangzu Liu, Mengzhe Ruan, Xiaotian Chen, HaonanChen HaonanChen, XudongNiu XudongNiu, Wendi Yuan, YuechenLi YuechenLi, Yang Liu and Guanjun Wang

Credit risk models suffer from rapid performance decay due to distribution shifts, requiring frequent updates to meet strict operational guardrails. However, manual refreshing takes weeks of trial-and-error across upstream data engineering and downstream training. We present ACRM, a deployed multi-agent framework that automates the end-to-end credit modeling workflow by treating it as a learnable trajectory of agent interactions. Unlike AutoML, which optimizes hyperparameters on fixed datasets, ACRM's action space extends to upstream data semantics—cohort selection, observation windowing, feature screening—where the majority of performance recovery occurs. A central Orchestrator coordinates specialist agents through a three-stream decision stack: rule-based safety guardrails, retrieval-augmented grounding from historical workflows, and preference alignment via DPO on expert-labeled trajectories. Deployed at a major fintech institution for three months across six business scenarios, ACRM reduced the average model refresh cycle from weeks to 1.1 days and iteration rounds by 65%, while maintaining superior stability metrics.

[Industry] Actionable Interpretability for Churn Classification: A Text Bottleneck Model Case Study at a Major Telecom Provider ▶ *Underline Video*

Adrian Sauter, Vera Neplenbroek, Georgios Vlassopoulos and Gianluigi Bardelloni

In subscription-based businesses, understanding why a customer intends to churn is as vital as the classification itself. We present a case study at a major Dutch telecommunications provider, where we implement Text Bottleneck Models (TBMs) for post-call churn classification. The TBM distills dialogues into a sparse set of human-interpretable concepts and provides faithful, snippet-based evidence for every decision. We show that the TBM performs competitively with black-box baselines and demonstrate potential business impact via automated call profiling and an interactive stakeholder dashboard. Our work demonstrates that the perceived trade-off between interpretability and predictive performance can be bridged, providing the high-accuracy evidence needed for industrial retention strategies.

[Long] Activation-Guided Local Editing for Jailbreaking Attacks ▶ *Underline Video*

Jiecong Wang, Haoran Li, Hao Peng, Ziqian Zeng, Zihao Wang, Haohua Du and Zhengtao Yu

As Large Language Models (LLMs) become indispensable assistants, they remain vulnerable to misuse. Jailbreaking is an essential adversarial technique for red-teaming models to uncover and patch security flaws. However, existing jailbreak methods suffer from significant limitations. Token-level jailbreak attacks often produce incoherent or unreadable inputs and exhibit poor transferability, while prompt-level attacks lack scalability and rely heavily on manual effort and human ingenuity. We propose AGILE, a concise and effective two-stage framework that combines the advantages of these approaches. The first stage performs a one-shot, scenario-based generation of context and rephrases the original malicious query to obscure its harmful intent. The second stage utilizes information from the model's hidden states to guide fine-grained edits, effectively steering the model's internal representation of the input from a malicious one toward a benign one. Extensive experiments demonstrate that AGILE achieves state-of-the-art Attack Success Rate, with gains of up to 37.74% over the strongest baseline, and AGILE exhibits excellent transferability to black-box and large-scale models. Our code is available at https://github.com/SELGroup/AGILE.

[Findings] Adaptive Text Anonymization: Learning Privacy-Utility Trade-offs via Prompt Optimization ▶ *Underline Video*

Gabriel Loiseau, Damien Sileo, Damien Riquet, Maxime Meyer and Marc Tommasi

Anonymizing textual documents is a highly context-sensitive problem: the appropriate balance between privacy protection and utility preservation varies with the data domain, privacy objectives, and downstream application. However, existing anonymization methods rely on static, manually designed strategies that lack the flexibility to adjust to diverse requirements and often fail to generalize across domains. We introduce adaptive text anonymization, a new task formulation in which anonymization strategies are automatically adapted to specific privacy-utility requirements. We propose a framework for task-specific prompt optimization that automatically constructs anonymization instructions for language models, enabling adaptation to different privacy goals, domains, and downstream usage patterns. To evaluate our approach, we present a benchmark spanning five datasets with diverse domains, privacy constraints, and utility objectives. Across all evaluated settings, our framework consistently achieves a better privacy-utility trade-off than existing baselines, while remaining computationally efficient and effective on open-source language models, with performance comparable to larger closed-source models. Additionally, we show that our method can discover novel anonymization strategies that explore different points along the privacy-utility trade-off frontier.

[Findings] Adversarial Yet Cooperative: Multi-Perspective Reasoning in Retrieved-Augmented Language Models ▶ *Underline Video*

Can Xu, Lingyong Yan, Jiayi Wu, Haosen Wang, Shuaiqiang Wang, Yuchen Li, Jizhou Huang, Dawei Yin and Xiang Li

Recent advances in synergizing large reasoning models (LRMs) with retrieval-augmented generation (RAG) have shown promising results, yet two critical challenges remain: (1) reasoning models typically operate from a single, unchallenged perspective, limiting their ability to conduct deep, self-correcting reasoning over external documents, and (2) existing training paradigms rely excessively on outcome-oriented rewards, which provide insufficient signal for shaping the complex, multi-step reasoning process. To address these issues, we propose an Reasoner-Verifier framework named Adversarial Reasoning RAG (ARR). The Reasoner and Verifier engage in reasoning on retrieved evidence and critiquing each other's logic while being guided by process-aware advantage that requires no external scoring model. This reward combines explicit observational signals with internal model uncertainty to jointly optimize reasoning fidelity and verification rigor. Experiments on multiple benchmarks demonstrate the effectiveness of our method. Our code is available at

https://github.com/lakhfskn/anonymous-code-of-arr.

[Long] AgentAsk: Multi-Agent Systems Need to Ask ▶ *Underline Video*

Bohan Lin, Kuo Yang, Zelin Tan, Yingchuan Lai, Chen Zhang, Guibin Zhang, Xinlei Yu, Miao Yu, Xu Wang, Yudong Zhang and Yang Wang

Multi-agent systems (MAS) built on large language models promise improved problem-solving through collaboration, yet they often fail to consistently outperform strong single-agent baselines due to error propagation at inter-agent message handoffs. In this work, we conduct a systematic empirical analysis of such failures and introduce an edge-level error taxonomy that identifies four dominant error types: Data Gap, Signal Corruption, Referential Drift, and Capability Gap, as primary sources of failure in multi-agent interactions. Building on this taxonomy, we propose AgentAsk, a lightweight clarification module designed to intervene at the edge level in MAS to prevent cascading errors. The module operates by strategically applying minimal clarifications at critical points within the system, improving the accuracy and efficiency of the overall task. AgentAsk is trained to balance the trade-offs between clarification cost, latency, and accuracy, while it is also architecture-agnostic and can be easily integrated into existing systems. Evaluated across five benchmarks, AgentAsk consistently improves accuracy by up to 4.69%, while keeping latency and extra costs below 10% compared to baseline MAS, showcasing its high efficiency and minimal overhead. The code is available at https://anonymous.4open.science/r/AgentAsk-3432.

[Long] AgentMark: Utility-Preserving Behavioral Watermarking for Agents ▶ *Underline Video*

Kaibo Huang, Jin Tan, Yukun Wei, Wanling Li, Zipei Zhang, Hui Tian, Zhongliang Yang and Linna Zhou

LLM-based agents are increasingly deployed to autonomously solve complex tasks, raising urgent needs for IP protection and regulatory provenance. While content watermarking effectively attributes LLM-generated outputs, it fails to directly identify the high-level planning behaviors (e.g., tool and subgoal choices) that govern multi-step execution. Critically, watermarking at the planning-behavior layer faces unique challenges: minor distributional deviations in decision-making can compound during long-term agent operation, degrading utility, and many agents operate as black boxes that are difficult to intervene in directly. To bridge this gap, we propose AgentMark, a behavioral watermarking framework that embeds multi-bit identifiers into planning decisions while preserving utility. It operates by eliciting an explicit behavior distribution from the agent and applying distribution-preserving conditional sampling, enabling deployment under black-box APIs while remaining compatible with action-layer content watermarking. Experiments across embodied, tool-use, and social environments demonstrate practical multi-bit capacity, robust recovery from partial logs, and utility preservation. Code is available at https://github.com/Tooooa/AgentMark.

[Long] AMATA: Adaptive Multi-Agent Trajectory Alignment for Knowledge-Intensive Question Answering ▶ *Underline Video*

Taolin Zhang, Dongyang Li, Chen Chen, Qizhou Chen, Jiuheng Wan, Xiaofeng He, Chengyu Wang and Richang Hong

Despite substantial advances in large language models (LLMs), producing factually consistent responses for knowledge-intensive question answering remains challenging. These difficulties are primarily due to hallucination and the limitations of LLMs in bridging long-tail knowledge gaps. To address this, we propose AMATA, an Adaptive Multi-Agent Trajectory Alignment framework that dynamically integrates external knowledge to improve response interpretability and factual grounding. Our architecture leverages six specialized agents that collaboratively perform structured actions for complex question reasoning. We formalize multi-agent collaboration with external tools as a trajectory preference alignment problem, incorporating question-aware agent customization and inter-agent preference harmonization. AMATA introduces two principal innovations: (1) Intra-Trajectory Preference Learning, which learns objective-oriented preferences to prioritize critical agents, and (2) Inter-Agent Dependency Learning, which captures cross-agent tool dependencies through a novel dependency-aware direct preference optimization technique. Empirical results show that AMATA consistently outperforms baseline approaches, knowledge-augmented frameworks, and LLM-based trajectory systems on five established knowledge-intensive QA benchmarks. Further analysis demonstrates the efficiency of our method in reducing token consumption.

[Industry] An Efficient Framework for Whole-Page Reranking via Single-Modal Supervision ▶ *Underline Video*

Zishuai Zhang;Sihao Yu;Wenyi Xie;Ying Nie;Junfeng Wang;Zhiming Zheng;Dawei Yin;Hainan Zhang

The whole-page reranking integrates retrieval results from multiple modalities and is critical for user experience of search engines, yet it requires costly large-scale expert annotations due to the complexity of assessing cross-modal relevances. In this paper, we propose SMAR, a novel whole-page reranking framework that converts single-modal rankers into page-level guidance by constructing budget-aware candidates for cross modal annotations and distilling intra-modality preferences to align relevance scales across modalities. Specifically, we use pre-trained single-modal rankers to construct candidate pages for limited cross-modal annotation at the page level. The whole-page reranker is then trained on these samples, enforcing consistency with single-modal preferences to preserve intra-modal ranking quality. Experiments on the Qilin and CrossRank datasets demonstrate that SMAR reduces annotation costs by 70-90% while outperforming the fully-annotated reranking baselines. Further offline and online A/B tests confirm significant gains in both ranking metrics and user experience, validating the effectiveness and practical value of our approach in real-world search scenarios.

[Findings] Anatomy of Unlearning: The Dual Impact of Fact Salience and Model Fine-Tuning ▶ *Underline Video*

Anna Borisiuk, Andrey Savchenko, Alexander Panchenko and Elena Tutubalina

Machine Unlearning (MU) enables Large Language Models (LLMs) to remove unsafe or outdated information. However, existing work assumes that all facts are equally forgettable and largely ignores whether the forgotten knowledge originates from pretraining or supervised fine-tuning (SFT). In this paper, we introduce DUAL (Dual Unlearning Evaluation across Training Stages), a benchmark of 28.6k Wikidata-

derived triplets annotated with fact popularity using Wikipedia link counts and LLM-based salience scores. Our experiments show that pretrained and SFT models respond differently to unlearning. An SFT step on the forget data yields smoother forgetting, more stable tuning, and 10-50% higher retention, while direct unlearning on pretrained models remains unstable and prone to relearning or catastrophic forgetting.

[Demo] AnnoHID: LLM-Assisted Annotation Framework for Low-Resource Medical Texts ▶ *Underline Video*

Annisa Maulida Ningtyas, Guntur Budi Herwanto, Yunita Sari, Rifki Afina Putri, Filip Kovacevic, Alaa El-Ebshihy, Varvara Arzt and Florina Piroi

This paper introduces AnnoHID, a semi-automated annotation framework designed for medical texts in low-resource languages. The system integrates large language models (LLMs) for pre-annotation and human validation, significantly reducing annotation time while maintaining accuracy. We demonstrate its application to Bahasa Indonesia medical social media texts from Alodokter, a medical Q&A platform, for named entity recognition (NER) and medical concept normalization (MCN). We conducted a user study with 21 participants to demonstrate the effectiveness of AnnoHID. The results show that LLM-assisted annotation yields higher inter-annotator agreement for both NER (kappa = 0.76) and MCN (kappa = 0.63) and that human review improves raw LLM NER output, raising the F1 score from 0.39 to 0.45. The framework is multilingual, human-in-the-loop, and interoperable with standard medical terminologies, such as SNOMED-CT. Future work focuses on mitigating pre-annotation bias, reducing annotation overhead, and scaling evaluations to larger datasets and additional low-resource languages.

[Long] ArgGenBench: Benchmarking the Complex Controlled Argument Generation Capability of Large Language Models ▶ *Underline Video*

Bojun Jin, Jianzhu Bao, Yang Sun, Yice Zhang and Ruifeng Xu

Argument generation is a fundamental NLP task that aims to automatically produce persuasive arguments. Effective human argumentation is inherently complex and multifaceted, integrating argumentative strategies, appropriate styles, and adaptation to target audiences, etc. However, existing studies focus on limited control signals such as topic, stance, or key aspects, failing to capture this complexity. As LLMs advance, the lack of benchmarks evaluating multifaceted argumentative control becomes a critical bottleneck. To address this, we introduce ArgGenBench, a novel benchmark containing complex instructions that integrate multi-dimensional control, including topic, stance, length, style, strategy, audience, and key points. Extensive evaluation across 15 LLMs reveals significant limitations: even the best-performing model achieves only 42.7% win rate against human-verified references. These results highlight the challenge of controlled argument generation and establish ArgGenBench as a rigorous testbed for developing more capable systems.

[Long] ASTRA: Adaptive Semantic Tree Reasoning Architecture for Complex Table Question Answering ▶ *Underline Video*

Xiaoke Guo, Songze Li, Zhiqiang Liu, Zhaoyan Gong, Yuanxiang Liu, Huajun Chen and Wen Zhang

Table serialization remains a critical bottleneck for Large Language Models (LLMs) in complex table question answering, hindered by challenges such as structural neglect, representation gaps, and reasoning opacity. Existing serialization methods fail to capture explicit hierarchies and lack schema flexibility, while current tree-based approaches suffer from limited semantic adaptability. To address these limitations, we propose ASTRA (Adaptive Semantic Tree Reasoning Architecture) including two main modules, AdaSTR and DuTR. First, we introduce AdaSTR, which leverages the global semantic awareness of LLMs to reconstruct tables into Logical Semantic Trees. This serialization explicitly models hierarchical dependencies and employs an adaptive mechanism to optimize construction strategies based on table scale. Second, building on this structure, we present DuTR, a dual-mode reasoning framework that integrates tree-search-based textual navigation for linguistic alignment and symbolic code execution for precise verification. Experiments on complex table benchmarks demonstrate that our method achieves state-of-the-art (SOTA) performance.

[Long] ASTRA: An Automated Framework for Strategy Discovery, Retrieval, and Evolution for Jailbreaking LLMs ▶ *Underline Video*

Xu Liu, Yan Chen, Kan Ling, Yichi Zhu, Hengrun Zhang, Guisheng Fan and Huiqun Yu

Despite extensive safety alignment, Large Language Models (LLMs) remain vulnerable to jailbreak attacks. However, existing methods generally lack the capability for continuous learning and self-evolution from interactions, limiting the diversity and adaptability of attack strategies. To address this, we propose ASTRA, an automated framework capable of autonomously discovering, retrieving, and evolving attack strategies. ASTRA operates on a closed-loop "attack-evaluate-distill-reuse" mechanism, which not only generates attack prompts but also automatically distills reusable strategies from every interaction. To systematically manage these strategies, we introduce a dynamic three-tier strategy library (Effective, Promising, and Ineffective) that categorizes strategies based on performance. This hierarchical memory mechanism enables the framework to enhance efficiency by leveraging successful patterns while optimizing the exploration space by avoiding known failures. Extensive experiments in a black-box setting demonstrate that ASTRA significantly outperforms existing baselines.

[Long] AttnPO: Attention-Guided Process Supervision for Efficient Reasoning ▶ *Underline Video*

Shuaiyi Nie, Siyu DingWenyuan Zhang, Linhao Yu, Tianmeng Yang, Yao Chen, Weichong Yin, Yu Sun, Hua Wu and Tingwen Liu

Large reasoning models trained with reinforcement learning and verifiable rewards (RLVR) achieve strong performance on complex reasoning tasks, yet often overthink, generating redundant reasoning without performance gains. Existing trajectory-level length penalties often fail to effectively shorten reasoning length and degrade accuracy, as they uniformly treat all reasoning steps and lack fine-grained signals to distinguish redundancy from necessity. Meanwhile, process-supervised methods are typically resource-intensive and suffer from inaccurate credit assignment. To address these issues, we propose ATTNPO, a low-overhead process-supervised RL framework that leverages

the models intrinsic attention signals for step-level credit assignment. We first identify a set of special attention heads that naturally focus on essential steps while suppressing redundant ones. By leveraging the attention scores of these heads, We then employ two sub-strategies to mitigate overthinking by discouraging redundant steps while preserving accuracy by reducing penalties on essential steps. Experimental results show that ATTNPO substantially reduces reasoning length while significantly improving performance across 9 benchmarks.

[Long] Automatic Correction of Writing Anomalies in Hausa Texts ▶ *Underline Video*

Ahmad Mustapha Wali and Sergiu Nisioi

Hausa texts are often characterized by writing anomalies such as incorrect character substitutions and spacing errors, which sometimes hinder natural language processing (NLP) applications. This paper presents an approach to automatically correct the anomalies by finetuning transformer-based models. Using a corpus gathered from several public sources, we create a large-scale parallel dataset of over 400,000 noisy-clean Hausa sentence pairs by introducing synthetically generated noise to mimic realistic writing errors. Moreover, we finetune several multilingual and African language models, including M2M100, AfriTeVA, NCAIR1/N-ATLaS, UBC-NLP/cheetah-base, and other variants of BART and T5 for this correction task. Our experimental results demonstrate that models such as M2M100 achieve state-of-the-art results despite their smaller size and distinct pretraining, and that correcting errors can have a significant impact in improving downstream tasks such as text classification, machine translation, question answering, and LLM prompting in general. This research provides a methodology, a publicly available dataset, and a comparison of models to improve Hausa text quality, thereby advancing NLP capabilities for the language and offering transferable insights for other low-resource languages.

[Long] AutoReproduce: Automatic AI Experiment Reproduction with Paper Lineage ▶ *Underline Video*

Xuanle Zhao, Zilin Sang, Yuxuan Li, Qi Shi, Weilun Zhao, Shuo Wang, Duzhen Zhang, Xu Han, Zhiyuan Liu and Maosong Sun

Efficient reproduction of research papers is pivotal to accelerating scientific progress. However, the increasing complexity of proposed methods often renders reproduction a labor-intensive endeavor, necessitating profound domain expertise. To address this, we introduce the paper lineage, which systematically mines implicit knowledge from the cited literature. This algorithm serves as the backbone of our proposed, a multi-agent framework designed to autonomously reproduce experimental code in a complete, end-to-end manner. To ensure code executability, incorporates a sampling-based unit testing strategy for rapid validation. To assess reproduction capabilities, we introduce, a benchmark featuring verified implementations, alongside comprehensive metrics for evaluating both reproduction and execution fidelity. Extensive evaluations on PaperBench and demonstrate that consistently surpasses existing baselines across all metrics. Notably, it yields substantial improvements in reproduction fidelity and final execution performance. The code is available at https://github.com/AI9Stars/AutoReproduce.

[Long] BatonVoice: An Operationalist Framework for Enhancing Controllable Speech Synthesis with Linguistic Intelligence from LLMs ▶ *Underline Video*

Yue Wang, Ruotian Ma, Xingyu Chen, Zhengliang Shi, Morunliu Yang, Wanshun CHEN, Huang Liu, Jiadi Yao, Xin He, Qu Yang, Qingxuan Jiang, Fanghua Ye, Juntao Li, Zhaopeng Tu, Xiaolong Li, Liefeng Bo and Min Zhang

The rise of Large Language Models (LLMs) is reshaping multimodel models, with speech synthesis being a prominent application. However, existing approaches often underutilize the linguistic intelligence of these models, typically failing to leverage their powerful instruction-following capabilities. This limitation hinders the model's ability to follow text instructions for controllable Text-to-Speech~(TTS). To address this, we propose a new paradigm inspired by operationalism that decouples instruction understanding from speech generation. We introduce BatonVoice, a framework where an LLM acts as a conductor, understanding user instructions and generating a textual plan – explicit vocal features (e.g., pitch, energy). A separate TTS model, the orchestra, then generates the speech from these features. To realize this component, we develop BatonTTS, a TTS model trained specifically for this task. Our experiments demonstrate that BatonVoice achieves strong performance in controllable and emotional speech synthesis, outperforming strong open- and closed-source baselines. Notably, our approach enables remarkable zero-shot cross-lingual generalization, accurately applying feature control abilities to languages unseen during post-training. This demonstrates that objectifying speech into textual vocal features can more effectively unlock the linguistic intelligence of LLMs.

[Findings] Behavioral Consistency Validation for LLM Agents: An Analysis of Trading-Style Switching through Stock-Market Simulation ▶ *Underline Video*

Zeping Li, Guancheng Wan, Keyang Chen, Yu Chen, Yiwen Zhao, Philip Torr, Guangnan Ye, Zhenfei Yin and Hongfeng Chai

Recent works have increasingly applied Large Language Models (LLMs) as agents in financial stock market simulations to test if micro-level behaviors aggregate into macro-level phenomena. However, a crucial question arises: Do LLM agents' behaviors align with real market participants? This alignment is key to the validity of simulation results. To explore this, we select a financial stock market scenario to test behavioral consistency. Investors are typically classified as fundamental or technical traders, but most simulations fix strategies at initialization, failing to reflect real-world trading dynamics. In this work, we assess whether agents strategy switching aligns with financial theory, providing a framework for this evaluation. We operationalize four behavioral-finance drivers, loss aversion, herding, wealth differentiation, and price misalignment, as personality traits set via prompting and stored long-term. In year-long simulations, agents process daily price-volume data, trade under a designated style, and reassess their strategy every 10 trading days. We introduce four alignment metrics and use Mann-Whitney U tests to compare agents style-switching behavior with financial theory. Our results show that recent LLMs' switching behavior is only partially consistent with behavioral-finance theories, highlighting the need for further refinement in aligning agent behavior with financial theory.

[Long] Beyond Fully Random Masking: Attention-Guided Denoising and Optimization for Diffusion Language Models ▶ *Underline Video*

Jia Deng, Junyi Li, Xin Zhao, Jinpeng Wang, Hongyu Lu and Ji-Rong Wen

Diffusion large language models (dLLMs) offer an efficient alternative to autoregressive models through parallel decoding, yet existing post-training methods largely rely on random masking strategies that overlook intrinsic token dependencies. In this work, we present an empirical analysis of attention in dLLMs and show that tokens attending more strongly to revealed context exhibit greater generation stability and play a critical role in reasoning. Motivated by these findings, we propose AGDO, an attention-guided denoising and optimization framework that aligns both training and optimization with attention-derived dependencies. AGDO determines the denoising order based on attention structure and emphasizes attention-critical tokens during supervised fine-tuning and reinforcement learning. Experiments on mathematical and coding benchmarks demonstrate that AGDO consistently improves reasoning performance, outperforming state-of-the-art post-training methods for dLLMs.

[Long] Beyond Scaling: Measuring and Predicting the Upper Bound of Knowledge Retention in Language Model Pre-Training ▶ *Underline Video*

Changhao Jiang, Ming Zhang, Yifei Cao, Junjie Ye, Xiaoran Fan, Shihan Dou, Zhiheng Xi, Jiajun Sun, Yi Dong, Yujiong Shen, Jingqi Tong, Baoyu Fan, Tao Gui, Qi Zhang and Xuanjing Huang

The GPT-4 technical report suggests that downstream performance can be predicted from pre-training signals, but offers little methodological detail on how to quantify this. This work address this gap by modeling knowledge retention, the capacity of a pre-trained language model to memorize factual information from its corpus, and introduce a principled method to estimate it prior to training. We propose Size-dependent Mutual Information (SMI), an information-theoretic predictor that integrates knowledge frequency, knowledge specificity, and model size to forecast closed-book question answering (QA) accuracy. SMI is validated through large-scale document retrieval over the disclosed pre-training corpora of 21 public and 3 custom models, combined with a robust multi-template QA evaluation. Experiments show that SMI significantly outperforms repetition-based baselines and achieves Rš > 0.7 in predicting QA accuracy for models above 1B parameters, without additional training. The analysis further reveals diminishing returns from scaling data and model size and provides evidence for an intrinsic upper bound on knowledge retention achievable by pre-training alone, motivating retrieval and other augmentation strategies.

[Long] Beyond Static Benchmarks: Synthesizing Harmful Content via Persona-based Simulation for Robust Evaluation ▶ *Underline Video*

Huije Lee, Jisu Shin, Hoyun Song, Changgeon Ko and Jong C. Park

Static benchmarks for harmful content detection face limitations in scalability and diversity, and may also be affected by contamination from web-scale pre-training corpora. To address these issues, we propose a framework for synthesizing harmful content, leveraging persona-guided large language model (LLM) agents. Our approach constructs two-dimensional user personas by integrating demographic identities and topical interests with situational harmful strategies, enabling the simulation of diverse and contextually grounded harmful interactions. We evaluate the framework along three dimensions: harmfulness, challenge level, and diversity. Both human and LLM-based evaluations confirm that our framework achieves a high harmful generation success rate. Experiments across multiple detection systems reveal that our synthetic scenarios are more challenging to detect than those in existing benchmarks. Furthermore, a multi-faceted analysis confirms that our approach achieves linguistic and topical diversity comparable to human-curated datasets, establishing our framework as an effective tool for robust stress-testing of harmful content detection systems.

[Long] BlindGuard: Safeguarding LLM-based Multi-Agent Systems under Unknown Attacks ▶ *Underline Video*

Rui Miao, Yixin Liu, Yili Wang, Xu Shen, Yue Tan, Yiwei Dai, Shirui Pan and Xin Wang

The security of LLM-based multi-agent systems (MAS) is critically threatened by propagation vulnerability, where malicious agents can distort collective decision-making through inter-agent interactions. While existing supervised defense methods demonstrate promising performance, they may be impractical in real-world scenarios due to their heavy reliance on labeled malicious agents to train a supervised malicious detection model. To enable practical and generalizable MAS defenses, in this paper, we propose BlindGuard, an unsupervised defense method that learns without requiring any attack-specific labels or prior knowledge of malicious behaviors. To this end, we establish a hierarchical agent encoder to capture individual, neighborhood, and global interaction patterns of each agent, providing a comprehensive understanding for malicious agent detection. Meanwhile, we design a corruption-guided detector that consists of directional noise injection and contrastive learning, allowing effective detection model training solely on normal agent behaviors. Extensive experiments show that BlindGuard effectively detects diverse attack types across MAS with various communication patterns while maintaining superior generalizability compared to supervised baselines.

[Long] Bloom-Eval: A Hierarchical Evaluation Benchmark for Automatic Survey Generation Based on Bloom's Taxonomy ▶ *Underline Video*

Fei Zhang, Zhe Zhao, Haibin Wen, Tianshuo Wei, Zaixi Zhang, Chao Yang and Ye Wei

The rapid advance of automatic survey generation (ASG) has created a critical evaluation challenge. Existing evaluation methods suffer from both cognitive dimensional simplification and methodological unreliability, primarily due to the over-reliance on the "LLM-as-a-Judge" approach. To bridge this gap, we establish Bloom-Eval, a six-tiered benchmark based on Bloom's Taxonomy that reliably evaluates ASG

systems by prioritizing deterministic algorithms and introducing our GRADE approach for abstract abilities. Furthermore, we construct a large-scale, cross-disciplinary dataset of over 3,000 high-quality papers. Our empirical study on this benchmark reveals that while leading ASG systems are proficient format organizers, they remain unqualified knowledge integrators. This work aims to redefine ASG evaluation standards, shifting the research focus from the formal mimicry of surface structure to the cognitive deepening of intellectual content. Our method provides the ASG field with a systematic, reproducible, and theoretically grounded benchmark to guide future research.

[Findings] BOLT: Benchmarking Open-World Learning for Text Classification ▶ *Underline Video*
Chuan Qin, Xi Chen, Jinpeng Li and Hengshu Zhu
Text classification has long been a cornerstone of NLP, yet most prior work and benchmarks have been limited to closed-world settings, where all classes are assumed to be known in advance. In contrast, open-world learning has recently emerged as a critical paradigm for building more robust and realistic systems. However, existing benchmarks largely focus on out-of-distribution (OOD) detection, while overlooking broader challenges such as the discovery of novel categories. To address this gap, we introduce BOLT, a unified Benchmark and evaluation toolkit supporting Open-world Learning for Text classification. BOLT encompasses two representative tasks: Open-set Text Classification (OSTC), which requires models to classify in-distribution (ID) samples while rejecting OOD inputs, and Generalized Category Discovery (GCD), which aims to identify both known and novel categories from partially labeled corpora. We carefully curate 12 publicly available datasets spanning diverse domains and benchmark 22 methods, including 15 for OSTC and 7 for GCD, under a standardized protocol that explicitly accounts for varying labeled ratios and known class ratios. Our results reveal key challenges: most current methods tend to overfit training distributions and struggle to generalize to unseen classes. Moreover, by comparing our lightweight LLM-based variants with prior open-set baselines, we demonstrate the promise of leveraging LLMs for open-world text classification. BOLT provides standardized evaluation protocols that enable fair comparison and support future research in this emerging area. All datasets, baselines, and tools are available at https://github.com/CNIC-DSL/BOLT.

[Long] Bridging Distance and Spectral Positional Encodings via Anchor-Based Diffusion Geometry Approximation ▶ *Underline Video*
Zimo Yan, Zheng Xie, Runfan Duan, Chang Liu and Wumei Du
Molecular graph learning benefits from positional signals that capture both local neighborhoods and global topology. Two widely used families are spectral encodings derived from Laplacian or diffusion operators and anchor-based distance encodings built from shortest-path information, but the relationship between them is still not well understood. In this paper, we study when anchor-based distance encodings can approximate diffusion geometry. Under a random r-regular graph model, we derive an explicit trilateration map that reconstructs truncated diffusion coordinates from transformed anchor distances and anchor spectral positions, together with pointwise and Frobenius-gap guarantees. On DrugBank with a shared GNP-based DDI backbone, anchor-distance Nyström accurately recovers diffusion geometry, and both DE and LapPE outperform models without positional encodings, with LapPE showing slightly more consistent performance.

[Long] Bridging the Sensory Gap: Visual Injection for Taxonomy Completion ▶ *Underline Video*
Yuhang Niu, Hongyuan Xu, Ciyi Liu, Bofan Wei, Jiaqi Ye, Yanlong Wen and Xiaojie Yuan
Taxonomy Completion aims to automatically integrate new concepts into existing hierarchies. However, existing text-only methods suffer from a "Sensory Gap": they struggle to differentiate ambiguous definitions (e.g., Latte vs. Cappuccino) and miss visual grouping signals. Consequently, they often misinterpret lexical overlaps as hierarchical dependencies, leading to erroneous structural predictions. To bridge this, we propose VITC, a framework leveraging Visual Injection for Taxonomy Completion. By mapping synthesized images into intrinsic pseudo-tokens, we enable the text encoder to perform holistic structural reasoning. To address injection challenges, we introduce Adaptive Residual Fusion, which decouples magnitude from selection to prevent visual signals from being drowned out, and the Multimodal Guided Adaptive Reweighting strategy, which leverages cross-modal consensus (Mutual Rescue and Complementary Mining) to filter noise and identify hard negatives. Experiments on three datasets demonstrate that VITC achieves state-of-the-art performance, delivering an average absolute gain of over 19% in Hit@1. Code is available at https://github.com/nyh-a/VITC.

[Long] BWLA: Breaking the Barrier of W1AX Post-Training Quantization for LLMs ▶ *Underline Video*
Zhixiong Zhao, Zukang Xu and Dawei Yang
Large language models (LLMs) have driven major progress in NLP, yet their substantial memory and compute demands still hinder practical deployment. Binarization can compress weights to 1 bit, fundamentally lowering compute and bandwidth cost. However, existing methods cannot address activation heavy tails and thus must keep activations in high precision, preventing true end-to-end acceleration. To overcome this limitation, we propose BWLA, the first post-training quantization framework that preserves high accuracy while achieving 1-bit weight quantization together with low-bit activations (e.g., 6 bits). The Orthogonal-Kronecker Transformation (OKT) learns an orthogonal mapping via EM minimization, converting unimodal weights into symmetric bimodal forms while suppressing activation tails and incoherence. The Proximal SVD Projection (PSP) then performs lightweight low-rank refinement through proximal SVD projection, further enhancing quantizability with minimal overhead. On Qwen3-32B, BWLA reaches a Wikitext2 perplexity of 11.92 under 6-bit activations (vs. 38 from SOTA), improves five zero-shot tasks by more than 70%, and delivers 3.26x inference speedup, demonstrating strong potential for real-world LLM compression and acceleration. The code will be available at https://github.com/Kishon-zzx/BWLA.

[Long] CAKE: Causal-Guided Adaptive Knowledge Editing for LLMs ▶ *Underline Video*
Shuxin Liu and Jianhao Zhang

LLMs have static pre-trained knowledge, leading to obsolescence and hallucinations. Knowledge Editing (KE) addresses these issues and typically requires multi-layer modifications. However, existing state-of-the-art methods, largely following the Locate-Select-Assign-Edit (LSAE) paradigm, rely on fixed-layer selection and uniform residual assignment, ignoring the heterogeneous causal efficacy of different layers. To bridge this, we propose CAKE (Causal-Guided Adaptive Knowledge Editing), a collaborative editing method within the more general Locate-Weight-Assign-Edit (LWAE) paradigm that: (1) selectively identifies critical layers via causal tracing scores; and (2) adaptively allocates editing burdens based on causal weights rather than uniform assumptions. We formulate residual assignment as a constrained quadratic optimization problem and derive a solution for optimal residual allocation, showing that aligning edits with causal efficacy mitigates recursive error accumulation. Furthermore, we establish a generalized weight shift error bound, under which existing paradigms emerge as special, restricted cases. Experimental results demonstrate that CAKE achieves SOTA performance with comparable overhead, validating the superiority of causal-guided adaptation.

[Long] Can LLMs Act as Historians? Evaluating Historical Research Capabilities of LLMs via the Chinese Imperial Examination ▶ *Underline Video*

Lirong Gao, Zeqing Wang, Yuyan Cai, Jiayi Deng, Yanmei Gu, Yiming Zhang, Jia Zhou, Yanfei Zhang and Junbo Zhao

While Large Language Models (LLMs) have increasingly assisted in historical tasks such as text processing, their capacity for professional-level historical reasoning remains underexplored. Existing benchmarks primarily assess basic knowledge breadth or lexical understanding, failing to capture the higher-order skills, such as evidentiary reasoning, that are central to historical research. To fill this gap, we introduce ProHist-Bench, a novel benchmark anchored in the Chinese Imperial Examination (Keju) system, a comprehensive microcosm of East Asian political, social, and intellectual history spanning over 1,300 years. Developed through deep interdisciplinary collaboration, ProHist-Bench features 400 challenging, expert-curated questions across eight dynasties, accompanied by 10,891 fine-grained evaluation rubrics. Through a rigorous evaluation of 18 LLMs, we reveal a significant proficiency gap: even state-of-the-art LLMs struggle with complex historical research questions. We hope ProHist-Bench will facilitate the development of domain-specific reasoning LLMs, advance computational historical research, and further uncover the untapped potential of LLMs. We release ProHist-Bench at https://github.com/inclusionAI/ABench/tree/main/ProHist-Bench.

[Long] Can Reasoning Path still be Effective as Input? Bridging Post-Reasoning to Chain-of-Thought Compression ▶ *Underline Video*

Chengzhengxu Li, Xiaoming Liu, Zhaohan Zhang, Shengchao Liu, Guoxin Ma, Yu Lan, Cong Wang and Chao Shen

Recent developments have enabled advanced reasoning in Large Language Models (LLMs) via long Chain-of-Thought (CoT), trading efficiency during inference for performance. Existing works focus on compressing generated CoT in reasoning, which impairs the necessary information for deriving the correct answer. In this work, we propose post-reasoning, a reasoning paradigm that takes CoT as a part of context to simplify the reasoning task for LLMs. We find that post-reasoning significantly reduces the generation length of LLMs, but its effectiveness hinges on the efficiency and the reliability of the contextual CoT generation.Therefore, we propose Upfront CoT (UCoT), an efficient post-reasoning framework for CoT compression. UCoT trains a lightweight model (compressor) to provide contextual CoT in form of soft tokens and trains the LLM (executor) to leverage this contextual CoT for producing the final answer. Extensive experiments show that UCoT maintains the powerful reasoning ability of executor while significantly reducing the length of CoT. It is worth mentioning that when applying UCoT to the Qwen2.5-7B-Instruct model, the usage of tokens on GSM8K dataset is reduced by 50%, while the performance is 3.08% higher than that of the state-of-the-art (SOTA) method. The code is available at: https://github.com/czx-li/UCoT.

[Long] Capability Decomposition for Unified Information Extraction via Hierarchical Mixture-of-Experts ▶ *Underline Video*

Jing Zhou, Peng Wang, Wenjun Ke, Jiajun Liu and Yao He

Unified Information Extraction (UIE) aims to handle heterogeneous IE tasks within a single framework, but existing methods often suffer from inconsistent schema representation, implicitly intermediate reasoning and full-parameter adaptation, which limit generalization, interpretability and parameter efficiency. To address these issues, we propose UC-UIE (Universal Capabilities-based Unified Information Extractor), a unified framework based on Large Language Model (LLM), which introduces a unified frame-and-slots schema for IE tasks and explicitly decomposes IE reasoning into three universal capabilities: judging, locating, and associating. Furthermore, UC-UIE adopts a Low-Rank Adaptation (LoRA) based hierarchical Mixture-of-Experts (MoE) adapter to fine-tune LLMs for IE tasks, which explicitly models these three capabilities in a task-driven way while ensuring parameter efficiency. With only 1.24% trainable parameters, UC-UIE outperforms full-parameter tuning methods, showing excellent parameter efficiency. Zero-shot evaluation reveals its strong generalization ability to unseen domains and schemas, benefiting from unified schema representation and explicit capability decomposition. Further experiments validate that the hierarchical MoE adapter learns capability specialization and composition, which enhances both UIE performance and interpretability.

[Long] Chain-of-Thought Compression Should Not Be Blind: V-Skip for Efficient Multimodal Reasoning via Dual-Path Anchoring ▶ *Underline Video*

Dongxu Zhang, Yiding Sun, Cheng Tan, Wenbiao Yan, Ning Yang, Jihua Zhu and Haijun Zhang

While Chain-of-Thought (CoT) reasoning significantly enhances the performance of Multimodal Large Language Models (MLLMs), its autoregressive nature incurs prohibitive latency constraints. Current efforts to mitigate this via token compression often fail by blindly applying text-centric metrics to multimodal contexts. We identify a critical failure mode termed Visual Amnesia, where linguistically redundant tokens are erroneously pruned, leading to hallucinations. To address this, we introduce V-Skip that reformulates token pruning as a Visual-Anchored Information Bottleneck (VA-IB) optimization problem. V-Skip employs a dual-path gating mechanism that weighs

token importance through both linguistic surprisal and cross-modal attention flow, effectively rescuing visually salient anchors. Extensive experiments on Qwen2-VL and Llama-3.2 families demonstrate that V-Skip achieves a speedup with negligible accuracy loss. Specifically, it preserves fine-grained visual details, outperforming other baselines over 30% on the DocVQA.

[Findings] ChunQiuTR: Time-Keyed Temporal Retrieval in Classical Chinese Annals ▶ *Underline Video*

Yihao Wang, Zijian He, Jie Ren and Keze Wang

Retrieval shapes how language models access and cite knowledge in retrieval-augmented generation (RAG). In historical research, the goal is often to locate the exact record for a specific regnal month, where temporal alignment matters as much as topical relevance. This is especially challenging for Classical Chinese annals: time is encoded in terse, implicit, non-Gregorian reign phrases that are context-dependent, so semantically plausible evidence can still be temporally invalid. We introduce ChunQiuTR, a time-keyed retrieval benchmark built from the Spring and Autumn Annals and its exegetical tradition. It organizes records by month-level reign keys and includes chrono-near confounders that mimic real retrieval failures. We propose CTD (Calendrical Temporal Dual-encoder), a time-aware dual-encoder combining Fourier-based absolute context with relative offset biasing. Experiments show consistent gains over semantic dual-encoder baselines under time-keyed evaluation. We will release ChunQiuTR and code after the anonymity period.

[Industry] CI-Work: Benchmarking Contextual Integrity in Enterprise LLM Agents ▶ *Underline Video*

Wenjie Fu, Xiaoting Qin, Jue Zhang, Qingwei Lin, Lukas Wutschitz, Robert Sim, Saravan Rajmohan and Dongmei Zhang

Enterprise LLM agents can dramatically improve workplace productivity, but their core capability, retrieving and using internal context to act on a users behalf, also creates new risks for sensitive information leakage. We introduce CI-Work, a Contextual Integrity (CI)-grounded benchmark that simulates enterprise workflows across five information-flow directions and evaluates whether agents can convey essential content while withholding sensitive context in dense retrieval settings. Our evaluation of frontier models reveals that privacy failures are prevalent (violation rates range from 15.8%-50.9%, with leakage reaching up to 26.7%) and uncovers a counterintuitive trade-off critical for industrial deployment: higher task utility often correlates with increased privacy violations. Moreover, the massive scale of enterprise data and potential user behavior further amplify this vulnerability. Simply increasing model size or reasoning depth fails to address the problem. We conclude that safeguarding enterprise workflows requires a paradigm shift, moving beyond model-centric scaling toward context-centric architectures.

[Long] CNSL-bench: Benchmarking the Sign Language Understanding Capabilities of MLLMs on Chinese National Sign Language ▶ *Underline Video*

Rui Zhao, Xuewen Zhong, Xiaoyun Zheng, Jinsong Su and Yidong Chen

Sign language research has achieved significant progress due to the advances in large language models (LLMs). However, the intrinsic ability of LLMs to understand sign language, especially in multimodal contexts, remains underexplored. To address this limitation, we introduce CNSL-bench, the first comprehensive Chinese National Sign Language benchmark designed for evaluating multimodal large language models (MLLMs) in sign language understanding. The proposed CNSL-bench is characterized by: 1) Authoritative grounding, as it is anchored to the officially standardized National Common Sign Language Dictionary, mitigating ambiguity from regional or non-canonical variants and ensuring consistent semantic definitions; 2) Multimodal coverage, providing aligned textual descriptions, illustrative images, and sign language videos; and 3) Articulatory diversity, supporting fine-grained analysis across key manual articulatory forms, including air-writing, finger-spelling, and the Chinese manual-alphabet. Using CNSL-bench, we extensively evaluate 21 open-source and proprietary up-to-date MLLMs. Our results reveal that, despite recent advances in multimodal modeling, current MLLMs remain substantially inferior to human performance, exhibiting systematic disparities across input modalities and manual articulatory forms. Additional diagnostic analyses suggest that several performance limitations persist beyond improvements in reasoning and that instruction-following robustness varies substantially across models.

[Long] Collaboration of Fusion and Independence: Hypercomplex-driven Robust Multi-Modal Knowledge Graph Completion ▶ *Underline Video*

Zhiqiang Liu, Yichi Zhang, Mengshu Sun, Lei Liang and Wen Zhang

Multi-modal knowledge graph completion (MMKGC) aims to discover missing facts in multi-modal knowledge graphs (MMKGs) by leveraging both structural relationships and diverse modality information of entities. Existing MMKGC methods follow two multi-modal paradigms: fusion-based and ensemble-based. Fusion-based methods employ fixed fusion strategies, which inevitably leads to the loss of modality-specific information and a lack of flexibility to adapt to varying modality relevance across contexts. In contrast, ensemble-based methods retain modality independence through dedicated sub-models but struggle to capture the nuanced, context-dependent semantic interplay between modalities. To overcome these dual limitations, we propose a novel MMKGC method M-Hyper, which achieves the coexistence and collaboration of fused and independent modality representations. Our method integrates the strengths of both paradigms, enabling effective cross-modal interactions while maintaining modality-specific information. Inspired by “quaternion” algebra, we utilize its four orthogonal bases to represent multiple independent modalities and employ the Hamilton product to efficiently model pair-wise interactions among them. Specifically, we introduce a Fine-grained Entity Representation Factorization (FERF) module and a Robust Relation-aware Modality Fusion (R2MF) module to obtain robust representations for three independent modalities and one fused modality. The resulting four modality representations are then mapped to the four orthogonal bases of a biquaternion for comprehensive modality interaction. Extensive experiments indicate its state-of-the-art performance with better robustness.

[Industry] ColorBrowserAgent: Complex Long-Horizon Browser Agent with Adaptive Knowledge Evolution ▶ *Underline Video*
Jihong Wang, Jiamu Zhou, Weiming Zhang, Teng Wang, Weiwen Liu, Zhuosheng Zhang, Xingyu Lou, Weinan Zhang, HUARONG DENG and Jun Wang
With the advancement of vision-language models, web automation has made significant progress. However, deploying autonomous agents in real-world settings remains challenging, primarily due to site heterogeneity, where generalist models lack domain-specific priors for diverse interfaces, and long-horizon instability, characterized by the accumulation of decision drift over extended interactions. To address these challenges, we introduce ColorBrowserAgent (Complex Long-Horizon Browser Agent), a knowledge-evolving agent for robust web automation. Our approach addresses these challenges through two synergistic mechanisms: human-in-the-loop knowledge adaptation that transforms sparse human feedback into reusable domain knowledge, and knowledge-aligned progressive summarization that stabilizes long interactions through memory compression. Extensive experiments on WebArena, WebChoreArena and industrial deployment show that ColorBrowserAgent consistently outperforms strong baselines. It achieves a state-of-the-art success rate of 71.2% on WebArena and maintains 47.4% performance under zero-shot transfer setting on WebChoreArena. In commercial deployment, it improves user satisfaction by 19.3% relatively, verifying its robustness in real-world scenarios.

[Long] Compete to Complete: Co-opetition Adversarial Learning for Retrieval-Augmented Generation ▶ *Underline Video*
Xin Liu, Yu-An Liu, Ruqing Zhang, Yixing Fan, Lixin Su, Jiafeng Guo and Xueqi Cheng
Retrieval-augmented generation (RAG) has emerged as a promising paradigm for mitigating hallucinations in large language models (LLMs). However, the intrinsic heterogeneity between the retriever and the generator often leads to a mismatch between retrieved evidence and generation needs, hindering effective coordination. We argue that competition between discriminative retrieval and generative modeling can more effectively expose their mutual weaknesses and induce deeper interaction. Motivated by this insight, we propose CARL (Co-opetition AdveRsarial Learning), a framework that formulates retriever-generator training in RAG as a minimax game. In this game, the retriever is optimized to retrieve both useful and adversarially useless documents to challenge the generator, while the generator learns to identify useful evidence and remain robust to misleading retrievals to produce accurate answers. Experiments on seven benchmark datasets demonstrate that CARL consistently improves RAG performance, validating the effectiveness of adversarial co-opetition in enhancing retriever-generator synergy.

[SRW] Confidence as a Tie-Breaker: Reassessing Multilingual Hedging Bias in LLM-as-a-Judge Evaluation ▶ *Underline Video*
Rajashik Datta; Sanjan Baitalik
Large language models are increasingly used as evaluators, yet recent work shows that their judgments can be distorted by superficial properties of responses, including epistemic markers such as maybe, probably, and I think. Prior evidence has suggested that hedged but correct answers may be penalized by LLM judges, but it remains unclear whether this effect survives in multilingual and code-mixed settings once benchmark construction artifacts are controlled. We present PolyJudge-Uncertain, a controlled benchmark for multilingual judge robustness under hedging, built from 320 base question-answer items and expanded into 5,120 examples across English, Hindi, Hinglish, and Bengali, with balanced assertive versus hedged formulations and correct versus incorrect answers. We evaluate a reference-guided judge using both pointwise and pairwise protocols. In an initial pilot, we observe a large pointwise penalty against hedged answers. However, after naturalizing templates and introducing template-quality control, that pointwise effect largely disappears: pointwise accuracy reaches 99.8% overall, with no meaningful assertive–hedged gap. The robust remaining effect emerges in pairwise comparison. When two answers are equally correct and differ only in style, the judge prefers the assertive response in 1,276 of 1,280 cases. These results refine the current narrative on uncertainty bias in LLM-as-a-judge. In our setting, the strongest stable failure mode is not pointwise misclassification of hedged truth, but pairwise over-preference for assertiveness among semantically equivalent outputs. More broadly, our findings show that multilingual judge-bias claims are highly sensitive to template naturalness and benchmark construction, making benchmark auditing a first-class methodological requirement for this line of work.

[Long] Confident, Calibrated, or Complicit: Safety Alignment and Ideological Bias in LLM Hate Speech Detection ▶ *Underline Video*
Sanjeevan Selvaganapathy and Mehwish Nasim
We investigate the efficacy of Large Language Models (LLMs) in detecting implicit and explicit hate speech, examining how models with minimal safety alignment (uncensored) compare with more heavily aligned (censored) counterparts in a deployed-model setting when deployed using political personas. While uncensored models are often framed as offering a less constrained perspective, our results reveal a trade-off: censored models outperform their uncensored counterparts in both accuracy and robustness, achieving 69.0% versus 64.1% strict accuracy. However, this higher performance is also associated with greater resistance to persona-based influence, while uncensored models are more malleable to ideological framing. Furthermore, we identify critical failures across all models in understanding nuanced language such as irony. We also find alarming fairness disparities in performance across different targeted groups and systemic overconfidence that renders self-reported certainty unreliable. These findings challenge the notion of LLMs as objective arbiters and highlight the need for more sophisticated auditing frameworks that account for fairness, calibration, and ideological consistency. Taken together, these results point to censorship-as-deployed rather than safety alignment in isolation as the more appropriate frame for interpreting model differences.

[Long] Conflict-Aware Memory for Embodied Agents: Enhancing Vector Data Quality via Detection Rules ▶ *Underline Video*
Kexin Ma, Haotian Wang, Shenglin Chen, Yishuai Cai, Huangyu Huangyu and Ruochun Jin

Embodied agents have successfully leveraged large language models (LLMs) to better transform human instructions and images into executable task plans. Furthermore, memories of agents can be leveraged to achieve continual self-learning and optimization. However, vector data quality problems emerge in memories when they are projected into vector space, especially in discerning contextually similar but semantically conflicting sentences and highly similar images. This is particularly detrimental to embodied AI as it potentially distorts the robots actions. To address this challenge, we propose Conflict Detection Rules (CDRs) to identify and manage data quality issues in vector knowledge bases, which assist in correcting the index structure and further improving the answer quality. Experimental results show that planners with CDRs exceed the basic LLM planner by 15.25% and 14.25% in grammatical accuracy (GA) and interpretation accuracy (IA) on average, respectively. Moreover, the entire workflow has been successfully integrated into various scenarios, demonstrating its practical applicability and robustness in the real world.

[Long] CoVerRL: Breaking the Consensus Trap in Label-Free Reasoning via Generator-Verifier Co-Evolution ▶ *Underline Video*
Teng Pan, Yuchen Yan, Zixuan Wang, Ruiqing Zhang, Guiyang Hou, Wenqi Zhang, Weiming Lu, Jun Xiao and Yongliang Shen
Label-free reinforcement learning enables large language models to improve reasoning capabilities without ground-truth supervision, typically by treating majority-voted answers as pseudo-labels. However, we identify a critical failure mode: as training maximizes self-consistency, output diversity collapses, causing the model to confidently reinforce systematic errors that evade detection. We term this the consensus trap. To escape it, we propose CoVerRL, a framework where a single model alternates between generator and verifier roles, with each capability bootstrapping the other. Majority voting provides noisy but informative supervision for training the verifier, while the improving verifier progressively filters self-consistent errors from pseudo-labels. This co-evolution creates a virtuous cycle that maintains high reward accuracy throughout training. Experiments across Qwen and Llama model families demonstrate that CoVerRL outperforms label-free baselines by 4.7-5.9% on mathematical reasoning benchmarks. Moreover, self-verification accuracy improves from around 55% to over 85%, confirming that both capabilities genuinely co-evolve.

[Long] Credal Concept Bottleneck Models for Epistemic-Aleatoric Uncertainty Decomposition ▶ *Underline Video*
Tanmoy Mukherjee, Thomas Bailleux, Pierre Marquis and Zied Bouraoui
Concept Bottleneck Models (CBMs) predict through human-interpretable concepts, but they typically output point concept probabilities that conflate epistemic uncertainty (reducible model underspecification) with aleatoric uncertainty (irreducible input ambiguity). This makes concept-level uncertainty hard to interpret and, more importantly, hard to act upon. We introduce (Credal Ensemble Concept Estimation), a CBM framework that decomposes concept uncertainty by construction. represents each concept as a credal prediction (a probability interval), derives epistemic uncertainty from disagreement across diverse concept heads, and estimates aleatoric uncertainty via a dedicated ambiguity output trained to match annotator disagreement when available. The resulting signals support prescriptive decisions: automate low-uncertainty cases, prioritize data collection for high-epistemic cases, route high-aleatoric cases to human review, and abstain when both are high. Across several tasks, we show that epistemic uncertainty is positively associated with prediction errors, whereas aleatoric uncertainty closely tracks annotator disagreement, providing guidance beyond error correlation. Our implementation is available at the following link: https://github.com/Tankiit/Credal_Sets/tree/ensemble-credal-cbm

[SRW] CRL-Prompt: Contrastive and Reinforcement Learning for Soft Prompt Tuning for Text Classification ▶ *Underline Video*
Danila Lapokin; Andrey Savchenko
Prompt choice is crucial in adapting language models to text classification tasks, particularly under low-resource conditions. Manual prompt engineering is time-consuming, non-scalable, and brittle, while current auto-prompting techniques are still far from maturity. This paper presents a two-stage method for prompt learning of frozen language models, CRL-Prompt, based on soft prompt initialization followed by contrastive and reinforcement-based refinement. An experimental study demonstrates that our approach achieves consistent improvements in accuracy over baseline prompt tuning strategies, with gains of up to 2.2% while training fewer than 0.25% of model parameters.

[Long] Crossing the Reward Bridge: Expanding Reinforcement Learning with Verifiable Rewards Across Diverse Domains ▶ *Underline Video*
Yi Su, Dian Yu, Linfeng Song, Juntao Li, Haitao Mi, Zhaopeng Tu, Min Zhang and Dong Yu
Reinforcement learning with verifiable rewards (RLVR) has been effective on tasks with structured solutions like math and coding, but its reliance on simple, rule-based verifiers creates a fundamental bottleneck. We find their applicability is surprisingly narrow even in structured domains, a limitation that is compounded at scale: rule-based systems can paradoxically degrade in performance as multi-domain, free-form training data increases. To overcome these challenges, we propose a new RLVR framework that uses a generative verifier to provide soft, probabilistic rewards. Our key insight is that powerful LLMs show high agreement with human evaluators when judging answer correctness given a ground-truth reference, allowing us to automate reward generation without costly human annotation. Our experiments demonstrate the effectiveness of this approach. We show that a compact 7B generative reward model can guide a 7B policy model to decisively outperform models up to 10x its size, including the 72B Qwen2.5-Instruct (by a margin of +8.6%). This effectiveness is robust, holding true across diverse training datasets with answers sourced from experts, web users, and other LLMs, and generalizes strongly to seven out-of-distribution benchmarks. Our work provides a scalable and effective framework for extending RLVR beyond the limitations of pattern-based verification to complex, noisy, real-world domains.

[Long] CSPO: Alleviating Reward Ambiguity for Structured Table-to-LaTeX Generation ▶ *Underline Video*

Yunfan Yang, Cuiling Lan, Jitao Sang and Yan Lu
Tables contain rich structured information, yet when stored as images their contents remain "locked" within pixels. Converting table images into LaTeX code enables faithful digitization and reuse, but current multimodal large language models (MLLMs) often fail to preserve structural, style, or content fidelity. Conventional post-training with reinforcement learning (RL) typically relies on a single aggregated reward, leading to reward ambiguity that conflates multiple behavioral aspects and hinders effective optimization. We propose Component-Specific Policy Optimization (CSPO), an RL framework that disentangles optimization across LaTeX tables components, structure, style, and content. In particular, CSPO assigns component-specific rewards and backpropagates each signal only through the tokens relevant to its component, alleviating reward ambiguity and enabling targeted component-wise optimization. To comprehensively assess performance, we introduce a set of hierarchical evaluation metrics. Extensive experiments demonstrate the effectiveness of CSPO, underscoring the importance of component-specific optimization for reliable structured generation.

[Long] CTRAP: Embedding Collapse Trap to Safeguard Large Language Models from Harmful Fine-Tuning ▶ *Underline Video*
Biao Yi, Tiansheng Huang, Baolei Zhang, Tong Li, Lihai Nie, Zheli Liu and Li Shen
Fine-tuning-as-a-service, while commercially successful for Large Language Model (LLM) providers, exposes models to harmful fine-tuning attacks. As a widely explored defense paradigm against such attacks, unlearning attempts to remove malicious knowledge from LLMs, thereby essentially preventing them from being used to perform malicious tasks. However, we highlight a critical flaw: the inherent general adaptability of LLMs allows them to easily bypass selective unlearning by rapidly relearning or repurposing their general capabilities for harmful tasks. To address this fundamental limitation, we propose a paradigm shift: instead of selective removal, we advocate for inducing model collapse, effectively forcing the model to "unlearn everything", specifically in response to updates characteristic of malicious adaptation. This collapse directly neutralizes the very general capabilities that attackers exploit, tackling the core issue unaddressed by selective unlearning. We introduce the Collapse Trap (CTRAP) as a practical mechanism to implement this concept conditionally. Embedded during alignment, CTRAP pre-configures the model's reaction to subsequent fine-tuning dynamics. If updates during fine-tuning constitute a persistent attempt to reverse safety alignment, the pre-configured trap triggers a progressive degradation of the model's core language modeling abilities, ultimately rendering it inert and useless for the attacker. Crucially, this collapse mechanism remains dormant during benign fine-tuning, ensuring the model's utility and general capabilities are preserved.

[Long] CTTA-T: Continual Test-Time Adaptation for Text Understanding via Teacher-Student with a Domain-aware and Generalized Teacher ▶ *Underline Video*
Tianlun Liu, Zhiliang Tian, Zhen Huang, Xingzhi Zhou, Wanlong Yu, Tianle Liu, Feng Liu and Dongsheng Li
Text understanding application often suffers from domain shifts. To handle testing domains, domain adaptation (DA) is trained to adapt to a fixed and observed testing domain; a more challenging paradigm, test-time adaptation (TTA), cannot access the testing domain during training and online adapts to the testing samples during testing, where the samples are from a fixed domain. We aim to explore a more practical and underexplored scenario, continual test-time adaptation (CTTA) for text understanding, which involves a sequence of testing (unobserved) domains in testing. Current CTTA methods struggle in reducing error accumulation over domains and enhancing generalization to handle unobserved domains: 1) Noise-filtering reduces accumulated errors but discards useful information, and 2) accumulating historical domains enhances generalization, but it is hard to achieve adaptive accumulation. In this paper, we propose a CTTA-T (continual test-time adaptation for text understanding) framework adaptable to evolving target domains: CTTA-T adopts a teacher-student framework, where the teacher is equipped with domain awareness and generalization for evolving domains. To improve teacher predictions, we propose a refine-then-filter based on dropout-driven consistency, which calibrates predictions and removes unreliable guidance. For the adaptation-generalization trade-off, we construct a domain-aware teacher by dynamically accumulating cross-domain semantics via incremental PCA, which continuously tracks domain shifts. Experiments show CTTA-T excels baselines.

[Long] DarwinTOD: LLM-Driven Lifelong Self-evolution for Task-oriented Dialog Systems ▶ *Underline Video*
Shuyu Zhang, Yujie Liu, Xinru Wang, Cheng Zhang, Yanmin Zhu and Bin Li
Traditional task-oriented dialog systems are unable to evolve from ongoing interactions or adapt to new domains after deployment, that is a critical limitation in real-world dynamic environments. Continual learning approaches depend on episodic retraining with human-curated data, failing to achieve autonomy lifelong improvement. While evolutionary computation and LLM driven self-improvement offer promising mechanisms for dialog optimization, they lack a unified framework for holistic, iterative strategy refinement. To bridge this gap, we propose DarwinTOD, a lifelong self-evolving dialog framework that systematically integrates these two paradigms, enabling continuous strategy optimization from a zero-shot base without task-specific fine-tuning. DarwinTOD maintains an Evolvable Strategy Bank and operates through a dual-loop process: online multi-agent dialog execution with peer critique, and offline structured evolutionary operations that refine the strategy bank using accumulated feedback. This closed-loop design enables autonomous continuous improvement without human intervention. Extensive experiments show that DarwinTOD surpasses previous state-of-the-art methods and exhibits continuous performance gains throughout evolution. Our work provides a novel framework for building dialog systems with lifelong self-evolution capabilities.

[Demo] DataArc-SynData-Toolkit: A Unified Closed-Loop Framework for Multi-Path, Multimodal, and Multilingual Data Synthesis ▶ *Underline Video*
Zhichao Shi, Cehao Yang, Hao Zhou, Xiaojun Wu, Huajie Li, Xuhui Jiang, Chengjin Xu, Yuanzhuo Wang and Jian Guo;
Synthetic data has emerged as a crucial solution to the data scarcity bottleneck in large language models (LLMs), particularly for

specialized domains and low-resource languages. However, the broader adoption of existing synthetic data tools is severely hindered by convoluted workflows, fragmented data standards, and limited scalability across modalities. To address these limitations, we develop DataArc-SynData-Toolkit, an open-source framework featuring: (1) a configuration-driven, end-to-end pipeline equipped with an intuitive visual interface and simplified CLI for exceptional usability; (2) a unified, quality-controllable synthesis paradigm that standardizes multi-source data generation to ensure high reusability; and (3) a highly modular architecture designed for seamless multimodal, multilingual, and multi-task adaptation. We apply the toolkit in multiple application scenarios. Experimental results demonstrate that our toolkit achieves an optimal balance between generation efficiency and data quality. By offering an end-to-end and visually interactive pipeline, DataArc-SynData-Toolkit significantly lowers the technical barrier to synthetic data generation and subsequent model training, accelerating its practical deployment in real-world applications.

[Long] Detoxification for LLM: From Dataset Itself ► *Underline Video*

Wei Shao, Yihang Wang, Gaoyu zhu, Ziqiang Cheng, Lei Yu, Jiafeng Guo and Xueqi Cheng

Existing detoxification methods for large language models mainly focus on post-training stage or inference time, while few tackle the source of toxicity, namely, the dataset itself. Such training-based or controllable decoding approaches cannot completely suppress the model's inherent toxicity, whereas detoxifying the pretraining dataset can fundamentally reduce the toxicity that the model learns during training. Hence, we attempt to detoxify directly on raw corpora with SoCD (Soft Contrastive Decoding), which guides an LLM to localize and rewrite toxic spans in raw data while preserving semantics, in our proposed HSPD (Hierarchical Semantic-Preserving Detoxification) pipeline, yielding a detoxified corpus that can drop-in replace the original for fine-tuning or other training. On GPT2-XL, HSPD attains state-of-the-art detoxification, reducing Toxicity Probability (TP) from 0.42 to 0.18 and Expected Maximum Toxicity (EMT) from 0.43 to 0.20. We further validate consistent best-in-class results on LLaMA2-7B, OPT-6.7B, and Falcon-7B. These findings show that semantics-preserving, corpus-level rewriting with HSPD effectively suppresses downstream toxicity while retaining data utility and allowing seamless source-level mitigation, thereby reducing the cost of later model behavior adjustment.

[Long] Disco-RAG: Discourse-Aware Retrieval-Augmented Generation ► *Underline Video*

Dongqi Liu, Hang Ding, Qiming Feng, Xurong Xie, Zhucun Xue, Chengjie Wang, Jian Li, Jiangning Zhang and Yabiao Wang

Retrieval-Augmented Generation (RAG) has emerged as an important means of enhancing the performance of large language models (LLMs) in knowledge-intensive tasks. However, most existing RAG strategies treat retrieved passages in a flat and unstructured way, which prevents the model from capturing structural cues and constrains its ability to synthesize knowledge from dispersed evidence across documents. To overcome these limitations, we propose Disco-RAG, a discourse-aware framework that explicitly injects discourse signals into the generation process. Our method constructs intra-chunk discourse trees to capture local hierarchies and builds inter-chunk rhetorical graphs to model cross-passage coherence. These structures are jointly integrated into a planning blueprint that conditions the generation. Experiments on question answering and long-document summarization benchmarks show the efficacy of our approach. Disco-RAG achieves state-of-the-art results on the benchmarks without fine-tuning. These findings underscore the important role of discourse structure in advancing RAG systems.

[Long] Distributional Clarity: The Hidden Driver of RL-Friendliness in Large Language Models ► *Underline Video*

Shaoning Sun, Mingzhu Cai, Huang He, Bingjin Chen, Siqi Bao, Yujiu Yang, Hua Wu and Haifeng Wang

Language model families exhibit striking disparity in their capacity to benefit from reinforcement learning: under identical training, models like Qwen achieve substantial gains, while others like Llama yield limited improvements. Complementing data-centric approaches, we reveal that this disparity reflects a hidden structural property: distributional clarity in probability space. Through a three-stage analysis, from phenomenon to mechanism to interpretation, we uncover that RL-friendly models exhibit intra-class compactness and inter-class separation in their probability assignments to correct vs. incorrect responses. We quantify this clarity using the Silhouette Coefficient (S) and demonstrate that (1) high S correlates strongly with RL performance; (2) low S is associated with severe logic errors and reasoning instability. To confirm this property, we introduce a Silhouette-Aware Reweighting strategy that prioritizes low-S samples during training. Experiments across six mathematical benchmarks show consistent improvements across all model families, with gains up to 5.9 points on AIME24. Our work establishes distributional clarity as a fundamental, trainable property underlying RL-Friendliness.

[Long] DiVE: Decoupling Intra-layer Visual Evidence for Mitigating Hallucinations in Large Vision-Language Models ► *Underline Video*

Xinwei Li, Li Lin, Hui Jiao, Li Yao, Tien-Tsin Wong and Hanqian Wu

Recent Large Vision-Language Models (LVLMs) have achieved significant progress yet frequently suffer from visual hallucinations, often stemming from an over-reliance on language priors rather than visual evidence. Existing decoding-based approaches often rely on input perturbations to weaken language priors, but they do not explicitly decouple visual evidence from mixed vision-language representations. To address these limitations, we propose DiVE (Decoupling intra-layer Visual Evidence). DiVE dynamically identifies layers enriched with visual information and performs intra-layer decoupling to extract aggregated visual evidence. By suppressing this evidence to construct a language-prior-dominated reference distribution, DiVE employs contrastive decoding to calibrate the output logits, thereby mitigating hallucinations. Extensive experiments across diverse LVLM architectures demonstrate that DiVE achieves state-of-the-art performance among decoding-based methods on multiple benchmarks. Crucially, it eliminates the latency of an extra forward pass, offering a lightweight and efficient solution.

[Long] DMHM: Density-aware Manifold Learning and Hybrid Mahalanobis Energy for LLMs-generated Text Detection ▶ *Underline Video*

Tianle Liu, Zhiliang Tian, Zhen Huang, Tianlun Liu, Jingyuan Huang, Zhaoning Zhang, Chengcheng Shao and Dongsheng Li

As the text generated by large language models (LLMs) increasingly resembles human-written text (HWT), detecting LLM-generated text (LGT) is crucial to avoid malicious use of LGT. Recent research treats LGT detection as an out-of-distribution (OOD) detection problem and views HWT as the OOD. However, existing OOD detection methods assume that LGT is a single homogeneous distribution. In practice, LGT exhibits different characteristics under different generation conditions. Text from weaker LLMs tends to form distinct clusters and is easy to detect, whereas text from stronger models significantly overlaps with HWTs and is hard to detect. To address the issue, in this paper, we propose an LGT detection framework based on density-aware manifold learning and the construction of hybrid Mahalanobis energy. We apply density-aware manifold learning with Laplacian smoothness and density regularization in embedding space, amplifying differences between LGT and HWT. We further propose a density-adaptive hybrid Mahalanobis metric that combines global and local covariance via density weighting, enabling adaptation to the manifold-aware embedding space. Finally, based on the metric, we define the distribution energy as a measure of distribution discrepancy, and we employ energy learning and contrastive learning to separate distributions hierarchically, establishing a clear OOD decision boundary. Experiments show that our method outperforms strong baselines.

[Long] DORA: A Dual-Objective Reinforcement Learning Framework for Effective and Efficient Multimodal Agentic Search ▶ *Underline Video*

Guangming Qin, Yuhao Deng, Yukun Zhao, Zhenyang Li, Junfeng Wang, Dawei Yin, Ye Yuan, Guoren Wang, Yizhou Yan, Chengliang Chai and Lei Cao

The most recent research uses reinforcement learning (RL) to post-train Multi-modal Large Language Models (MLLMs) such that these models are able to iteratively call search engines to dynamically access external knowledge when handling complex Visual Question Answering (VQA) tasks. However, existing methods face two major limitations in effectiveness and efficiency: i) For effectiveness, the objective of these methods, which only considers the correctness of the generated final response, overlooks the quality of intermediate search results, thus leading to suboptimal search strategies. ii) For efficiency, existing methods often unnecessarily invoke search calls during reasoning, making the inference inefficient. To address these issues, we propose, a customized dual-objective reinforcement learning framework to improve the search strategies of MLLMs, enhancing their search quality yet minimizing search frequency. The key ideas include (1) a reward function that promotes correct reasoning trajectories with fewer search calls; and (2) dual optimization objectives that jointly optimize search quality and answer correctness. Extensive experiments on 3 real-world datasets demonstrate that DORA outperforms state-of-the-art methods, achieving up to 8.4% higher accuracy while reducing the number of search calls by 9.7%.

[Long] DUD: Decoupled Update Dynamics for Reliable Uncertainty Quantification in Large Language Models ▶ *Underline Video*

*Yixin Bu, Runze Xia, Guanyun Zou, Yupeng Ji, Hongliang Dai, Haodong Liu and Piji Li**

Accurate Uncertainty Quantification (UQ) is critical for reliable deployment of Large Language Models (LLMs), yet traditional probability-based metrics often fail to capture the model's true epistemic state. While recent mechanistic approaches leverage hidden state dynamics, they typically aggregate residual stream updates, conflating the distinct roles of parametric memory (Feed-Forward Networks) and contextual processing (Attention). We argue that this aggregation obscures fine-grained mechanistic conflicts, such as memory-context misalignment, that are fundamental indicators of uncertainty. To address this, we introduce Decoupled Update Dynamics (DUD), a framework that explicitly decouples FFN and Attention contributions via noise-induced causal interventions. By quantifying the independent restoration capabilities of each module, we construct a dual-stream dynamic profile that captures the model's internal fragility. Extensive experiments demonstrate that DUD significantly outperforms state-of-the-art baselines in both uncertainty estimation and calibration, while exhibiting superior cross-dataset generalization, validating decoupled dynamics as a robust proxy for model faithfulness.

[Findings] Dunhuang-Bench: How Well Do MLLMs Understand Cultural Heritage? ▶ *Underline Video*

Junyi Yuan, Jian Zhang, Tianxiu Yu, Yanlin Zhou, Xiaobo Jin, Qiufeng Wang and Fangyu Wu

Dunhuang art, a cornerstone of global heritage, demands fine-grained visual perception anchored by specialized cultural knowledge. Given the strong performance of multimodal large language models (MLLMs) on generic multimodal benchmarks, to what extent can they understand artifacts from Dunhuang art that are grounded in cultural context? To this end, we construct Dunhuang-Bench, a large-scale benchmark comprising 486 images and 22,970 QA pairs. It incorporates diverse task formats to evaluate MLLMs' cultural understanding: Question Answering with Text Description, Multi-turn Dialogue, and Question Answering with Choices. Guided by Panofskys theory of iconology, we design two tasks including visual perception and knowledge reasoning for the evaluation of content understanding. In addition, we follow the theory of formal analytic tradition to design another task of artistic appreciation in our Dunhuang-Bench. Extensive evaluations of 20 mainstream MLLMs on Dunhuang-Bench reveal a consistent performance drop from perception and appreciation to reasoning. Moreover, CoT and few-shot prompting show marginal or negative impact, highlighting the limits of prompting-based improvements. Dunhuang-Bench thus provides a challenging benchmark for advancing multimodal cultural understanding. Data and code will be publicly available.

[Long] DVCQR: Dual-View Conversational Query Rewriting with Stage-wise Reinforcement Learning ▶ *Underline Video*

Chenyi Li, Xinhui Tu and Zaixiang Wang

Conversational query rewriting (CQR) addresses context dependence in conversational search by rewriting each user query into a standalone

form. Recent approaches leverage reinforcement learning (RL) to directly optimize retrieval effectiveness; however, they typically rely on a single rewrite, which struggles to accommodate the divergent preferences of sparse and dense retrievers and often suffers from conflicting optimization signals. We propose DVCQR, a Dual-View CQR framework that explicitly generates two complementary rewrites for each query: a sparse-view rewrite that emphasizes distinctive lexical anchors, and a dense-view rewrite that captures complete semantic constraints. Both rewrites are produced in a single pass via a structured reasoning process. To further mitigate objective conflicts, we introduce a stage-wise RL strategy that sequentially aligns the sparse and dense views with their corresponding retrievers using rank-based feedback. Extensive experiments on four benchmarks (TopiOCQA, QReCC, CAsT-19, and CAsT-20) demonstrate that DVCQR consistently outperforms state-of-the-art methods on most metrics under both sparse and dense retrieval settings, validating the effectiveness of dual-view rewriting and stage-wise retriever alignment.

[Long] DyBBT: Dynamic Balance via Bandit-inspired Targeting for Dialog Policy with Cognitive Dual Systems ▶ *Underline Video*
Shuyu Zhang, Yifan Wei, Jialuo Yuan, Xinru Wang, Yanmin Zhu, Bin Li and Yujie Liu
Task oriented dialog systems often rely on static exploration strategies that do not adapt to dynamic dialog contexts, leading to inefficient exploration and suboptimal performance. We propose DyBBT, a novel dialog policy learning framework that formalizes the exploration challenge through a structured cognitive state space C that captures dialog progression, user uncertainty, and slot dependency. DyBBT proposes a bandit-inspired meta-controller that dynamically switches between a fast intuitive inference (System 1) and a slow deliberative reasoner (System 2) based on real-time cognitive states and visitation counts. Extensive experiments on single- and multi-domain benchmarks show that DyBBT achieves SOTA performance in success rate, efficiency, and generalization, with human evaluations confirming that its decisions are well-aligned with expert judgment.

[Long] Efficient Hyperparameter Optimization for LLM Reinforcement Learning ▶ *Underline Video*
Minping Chen, Bowen Xiao, Du Liang, Chuxuan Zeng and Zeyi Wen
Hyperparameters are critical to LLM reinforcement learning (RL), but existing hyperparameter optimization (HPO) methods remain inefficient in this area, due to the massive model scale and resource-intensive training cycles. In this paper, we propose Joint Fidelity Hyperparameter Optimization (JF-HPO), which simultaneously adapts both model size and training budget as fidelity. JF-HPO is empowered by: (i) a small proxy model of the target LLM for efficient training and evaluation in each HPO trial; (ii) several carefully designed early-stopping strategies based on training dynamics; (iii) an efficient checkpointing mechanism to eliminate redundant computations. JF-HPO significantly improves the computational efficiency of each trial (up to 14.9x) compared with existing HPO methods, thus achieving better predictive accuracy in most cases under the same time budget. Notably, JF-HPO delivers performance improvements ranging from 5.8% to 111.6% over VeRL Recipe.

[Long] EquivPruner: Boosting Efficiency and Quality in LLM-Based Search via Action Pruning ▶ *Underline Video*
Jiawei Liu, Qisi Chen, Jianshu Zhang, Quan Liu and Defu Lian
Large Language Models (LLMs) excel at complex reasoning through search algorithms, yet current strategies often suffer from massive token consumption due to redundant exploration of semantically equivalent steps. Existing semantic similarity methods struggle to accurately identify such equivalence in domain-specific contexts like mathematical reasoning. To address this, we propose EquivPruner, a simple yet effective approach that identifies and prunes semantically equivalent actions during LLM reasoning search. We also introduce MathEquiv, the first dataset we created for mathematical statement equivalence, which enables the training of a lightweight equivalence detector. Extensive experiments across various models and tasks demonstrate that EquivPruner significantly reduces token consumption, improving searching efficiency and often bolstering reasoning accuracy. For instance, when applied to Qwen2.5-Math-7B-Instruct on GSM8K, EquivPruner reduced token consumption by 48.1% while also improving accuracy. Our code is available at https://github.com/Lolo1222/EquivPruner.

[Long] EquivPruner: Boosting Efficiency and Quality in LLM-Based Search via Action Pruning ▶ *Underline Video*
Jiawei Liu, Qisi Chen, Jianshu Zhang, Quan Liu and Defu Lian
Large Language Models (LLMs) excel at complex reasoning through search algorithms, yet current strategies often suffer from massive token consumption due to redundant exploration of semantically equivalent steps. Existing semantic similarity methods struggle to accurately identify such equivalence in domain-specific contexts like mathematical reasoning. To address this, we propose EquivPruner, a simple yet effective approach that identifies and prunes semantically equivalent actions during LLM reasoning search. We also introduce MathEquiv, the first dataset we created for mathematical statement equivalence, which enables the training of a lightweight equivalence detector. Extensive experiments across various models and tasks demonstrate that EquivPruner significantly reduces token consumption, improving searching efficiency and often bolstering reasoning accuracy. For instance, when applied to Qwen2.5-Math-7B-Instruct on GSM8K, EquivPruner reduced token consumption by 48.1% while also improving accuracy. Our code is available at https://github.com/Lolo1222/EquivPruner.

[Long] EventWeave: A Dynamic Framework for Capturing Core and Supporting Events in Dialogue Systems ▶ *Underline Video*
Zhengyi Zhao, Shubo Zhang, Yiming Du, Bin Liang, Baojun Wang, Zhongyang Li, Binyang Li and Kam-Fai Wong
Large language models have improved dialogue systems, but often process conversational turns in isolation, overlooking the event structures that guide natural interactions. Hence we introduce EventWeave, a framework that explicitly models relationships between conversational events to generate more contextually appropriate dialogue responses. EventWeave constructs a dynamic event graph that distinguishes between core events (main goals) and supporting events (interconnected details), employing a multi-head attention mechanism to selectively

determine which events are most relevant to the current turn. Unlike summarization or standard graph-based approaches, our method captures three distinct relationship types between events, allowing for more nuanced context modeling. Experiments on three dialogue datasets demonstrate that EventWeave produces more natural and contextually appropriate responses while requiring less computational overhead than models processing the entire dialogue history. Ablation studies confirm improvements stem from better event relationship modeling rather than increased information density. Our approach effectively balances comprehensive context understanding with generating concise responses, maintaining strong performance across various dialogue lengths through targeted optimization techniques.

[Long] Evolving Agents ▶ *Underline Video*

Leonardo Ranaldi

AI agents struggle to operate within open and dynamic environments because they lack a fundamental capacity: the autonomous generation of abstractions. Current models remain static entities, incapable of compressing the infinite complexity of the real world into generalisable concepts once their training phase has concluded. We introduce EVA (Evolving Agents), a novel paradigm for autonomous learning driven by pseudo-symbolic abstraction. EVA introduces a meta-control system that dynamically orchestrates observation and active interaction to distil on-the-fly abstract representations of states, actions, and goals. By disentangling contextual noise from pure logical reasoning, these pseudo-symbolic abstractions allow the agent to construct a highly robust internal curriculum. EVA leverages these self-generated abstractions to form an internal curriculum. This continuous compression of raw sensorimotor experience into reusable concepts allows the agent to independently guide its own exploration, planning, and error correction. Structured upon a bi-level evolutionary-developmental (Evo/Devo) framework, EVA demonstrates how the dynamic refinement of abstractions enables rapid adaptation to unforeseen scenarios. This approach resolves the domain mismatch problem and lays the groundwork for truly autonomous, continuously evolving AI models.

[Long] EvoSpark: Endogenous Interactive Agent Societies for Unified Long-Horizon Narrative Evolution ▶ *Underline Video*

Shiyu He, Minchi Kuang, Mengxian Wang, Bin Hu and Tingxiang Gu

Realizing endogenous narrative evolution in LLM-based multi-agent systems is hindered by the inherent stochasticity of generative emergence. In particular, long-horizon simulations suffer from social memory stacking, where conflicting relational states accumulate without resolution, and narrative-spatial dissonance, where spatial logic detaches from the evolving plot. To bridge this gap, we propose EvoSpark, a framework specifically designed to sustain logically coherent long-horizon narratives within Endogenous Interactive Agent Societies. To ensure consistency, the Stratified Narrative Memory employs a Role Socio-Evolutionary Base as living cognition, dynamically metabolizing experiences to resolve historical conflicts. Complementarily, a Generative Mise-en-Scène mechanism enforces Role-Location-Plot alignment, synchronizing character presence with the narrative flow. Underpinning these is the Unified Narrative Operation Engine, which integrates an Emergent Character Grounding Protocol to transform stochastic sparking into persistent characters. This engine establishes a substrate that expands a minimal premise into an open-ended, evolving story world. Experiments demonstrate that EvoSpark significantly outperforms baselines across diverse paradigms, enabling the sustained generation of expressive and coherent narrative experiences.

[Long] Execution as Verification: Fine-Grained Self-Correcting Reasoning for Complex KBQA ▶ *Underline Video*

Minghan Zhang, Zhen Yang, Haodong Zou, Jie Chen, Zhen Duan and Shu Zhao

Knowledge Base Question Answering (KBQA) leverages structured knowledge bases to offer superior interpretability and hallucination resistance, making it a critical technology for precise knowledge reasoning. However, the prevailing LLM-based generate-then-execute formulation of semantic parsing is limited by strict syntactic constraints, making it primarily prone to structural deviations that render queries unexecutable, while suffering from semantic deviations that yield incorrect execution results. To address these challenges, we propose the Execution as Verification (EVER) framework, reframing semantic parsing as an iterative, self-correcting reasoning process driven by execution feedback. First, motivated by the insight that query executability serves as a strong proxy for answer correctness, we introduce Fine-Grained Execution-Aware Planning. This mechanism decomposes complex semantic parsing into a sequence of stepwise reasoning processes oriented by executability verification, ensuring high query executability. We further design a Self-Guided Semantic Correction mechanism based on execution result verification, utilizing execution feedback to verify and calibrate semantic deviations, thereby ensuring the semantic correctness of executable queries. Experimental results on the WebQSP and CWQ datasets demonstrate that our method achieves significant improvements in both query executability and answer accuracy, achieving state-of-the-art performance, particularly in complex multi-hop scenarios. Our code is available at https://github.com/ahu-zmh/EVER.

[Long] Exploring Layer Activation Dynamic of CoT via Knowledge Probe ▶ *Underline Video*

Chuanxin Zhang, Jiajun Liu, Yao He, Wenjun Ke, Peng Wang, Yankun Le, Sirui Liu and Zhaoyu Yang

Chain-of-thought (CoT) reasoning has emerged as a crucial paradigm for enhancing large language model (LLM) performance on multi-step reasoning tasks. However, the internal mechanisms by which LLMs invoke knowledge and propagate information across different steps of the CoT are poorly understood. To fill this gap, we propose a multi-stage probing framework that enforces structured reasoning with three explicit stages: keyword extraction, theorem generation, and computation execution. The framework integrates attention knockout to trace cross-layer information flow and theorem probing to examine how specific contents are encoded within representations. To enable controlled and stage-aligned analysis, we construct a structured CoT dataset that covers the mathematics and physics domains. Experiments on four instruction-tuned LLMs reveal distinct stage-specific patterns. First, keyword information is progressively aggregated into the final token in later layers. Second, theorem semantics are encoded in the mid-to-late layers and undergo two stages of propagation. Finally, parameter substitution is achieved through joint extraction by the final token and other tokens. The first parameter predominantly relies on the final

token, whereas later parameters increasingly depend on information extracted by other tokens. Overall, our findings shed light on the neural implementation of CoT reasoning and provide actionable insights for developing more interpretable and reasoning-capable LLMs. We further evaluate a free-form prompting setting without labeled fields and observe consistent qualitative trends.

[Findings] Fine-Grained Data Ordering Improves Fine-Tuning for Large Language Models ▶ *Underline Video*

Xiaomeng Hu, Yixuan Tang, Haoze Li, Hao Chen, Qi Zhang, Zhanming Shen, Yiming Zhang, Haobo Wang and Junbo Zhao

With the rapid progress of large language models (LLMs), aligning a general-purpose model with downstream tasks through fine-tuning has become a central research focus. Selecting only high-quality examples for training has been shown to be one of the most effective ways to improve fine-tuning performance. However, prior work concentrates almost exclusively on data preprocessing: filtering and cleaning data before training begins. While the order and composition of training data during training have received little fine-grained attention. To fill this gap, our work proposed Fine-Grained Order Fine-Tuning, a fine-grained scheduling method of data order in epochs. Drawing on curriculum-learning principles, FOT defines data difficulty based on the relevance between the data and the model, and then performs dynamic scheduling of the training order in each epoch according to the difficulty. On both large-scale continued pre-training and small-scale supervised fine-tuning experiments, FOT has achieved an average 2.4% improvement over baselines. Our study offers a new perspective on data governance in the fine-tuning phase.

[Long] FinSight: Towards Real-World Financial Deep Research ▶ *Underline Video*

Jiajie Jin, Yuyao Zhang, Yimeng Xu, Hongjin Qian, Yutao Zhu and Zhicheng Dou

Professional financial reports serve as the cornerstone of investment decisions, demanding deep reasoning and multimodal synthesis. While recent deep research systems excel in open-domain search, they struggle with financial reporting, specifically in handling financial data, ensuring analytical depth, and integrating professional visualizations. To address this, we introduce FinSight, the first multi-agent framework for automate end-to-end professional, multimodal financial report. At its core, we propose the Code Agent with Variable Memory architecture, which unifies data, tools, and agents into a programmable variable space, enabling flexible data manipulation and reasoning through executable code. To guarantee report quality, FinSight incorporates a Two-Stage Writing Framework with Generative Retrieval. This mechanism first distills raw data into structured Chain-of-Analysis segments, and then progressively synthesizes them into a coherent, citation-aware, and multimodal narrative. Additionally, an Iterative Vision-Enhanced Mechanism leverages visual feedback to refine code-generated charts to expert standards. Experiments on company and industry-level tasks demonstrate that FinSight significantly outperforms leading deep research systems in factual accuracy, analytical depth, and presentation quality, demonstrating a clear path toward generating professional financial reports. Our code is available at https://anonymous.4open.science/r/FinSight-5841.

[Findings] FlashMem: Distilling Intrinsic Latent Memory via Computation Reuse ▶ *Underline Video*

Yubo Hou, Zhisheng Chen, Tao Wan and Zengchang Qin

The stateless architecture of Large Language Models inherently lacks the mechanism to preserve dynamic context, compelling agents to redundantly reprocess history to maintain long-horizon autonomy. While latent memory offers a solution, current approaches are hindered by architectural segregation, relying on auxiliary encoders that decouple memory from the reasoning backbone. We propose FlashMem, a framework that distills intrinsic memory directly from transient reasoning states via computation reuse. Leveraging the property that internal representations uniquely encode input trajectories, FlashMem identifies the last hidden state as a sufficient statistic for the interaction history. This enables a Shared-KV Consolidator to synthesize memory by attending directly to the backbone's frozen cache, eliminating redundant re-parameterization. Furthermore, a parameter-free Cognitive Monitor leverages attention entropy to adaptively trigger consolidation only when high epistemic uncertainty is detected. Experiments demonstrate that FlashMem matches the performance of heavy baselines while reducing inference latency by 5 times, effectively bridging the gap between efficiency and persistent cognition.

[Demo] FourCorners: Grounded Thai Legal Research over a Temporal Knowledge Graph ▶ *Underline Video*

Pawitsapak Akarajaradwong, Thitiwat Nopparatbundit, Treephop Saeteng, Kasidit Phoncharoen, Sarana Nutanong and Chompakorn Chaksangchaichot;

Legal practitioners in Thailand must navigate fragmented government websites to research over 3,800 active laws and 87,000 Supreme Court decisions, with no unified tool for cross-referencing, version tracking, or structural navigation. We present FourCorners, a deployed platform that addresses five practitioner pain points through three modules built on a temporal legal knowledge graph covering 552K nodes and 6.3M edges: (1) an AI legal assistant that performs grounded generation via structured graph retrieval, streaming verified source content inline with responses; (2) an interactive law reader that translates graph structure into navigation and comparison features; and (3) a court decision explorer with version-aware citations produced by temporal entity resolution across 87,394 rulings. The system discovers implicit cross-corpus relationships through co-citation analysis of Supreme Court decisions. Interviews with 20 legal professionals reveal that inline source verification fundamentally changes how practitioners interact with AI-generated legal content, and that cross-corpus enrichment surfaces legal relationships that existing tools leave invisible.

[Findings] Frame of Reference: Addressing the Challenges of Common Ground Representation in Situational Dialogs ▶ *Underline Video*

Biswesh Mohapatra, Théo Charlot, Giovanni Duca, Mayank Palan, Laurent Romary and Justine Cassell

Common ground plays a critical role in situated spoken dialogues, where interlocutors must establish and maintain shared references to

entities, events, and relations to sustain coherent interaction. For dialog systems, the ability to correctly ground conversational content in order to refer back to it later is particularly important. Prior studies have demonstrated that LLMs are capable of performing grounding acts such as requesting clarification or producing acknowledgments, yet relatively little work has investigated how common ground can be explicitly represented and stored for later use. Without such mechanisms, it remains unclear whether acknowledgment or clarification behaviors truly reflect a grounded understanding. In this work, we evaluate a models ability to establish and exploit common ground through relational references to entities within the shared context in a situational dialogue. We test multiple methods for representing common ground in situated dialogues and further propose approaches to improve both the establishment of common ground and its subsequent use in the conversation.

[Long] From Insight to Action: A Novel Framework for Interpretability-Guided Data Selection in Large Language Models ▶ *Underline Video*

Ling Shi, Xinwei Wu, Xiaohu Zhao, Hao Wang, Heng Liu, Yangyang Liu, Linlong Xu, Longyue Wang, Deyi Xiong and Weihua Luo

While mechanistic interpretability tools like Sparse Autoencoders (SAEs) can uncover meaningful features within Large Language Models (LLMs), a critical gap remains in transforming these insights into practical actions for model optimization. We bridge this gap with the hypothesis that data selection guided by a model's internal task features is a effective training strategy. Inspired by this, we propose Interpretability-Guided Data Selection (IGDS), a framework that first identifies these causal task features through frequency recall and interventional filtering, then selects "Feature-Resonant Data" that maximally activates task features for fine-tuning. We validate IGDS on mathematical reasoning, summarization, and translation tasks within Gemma-2, LLaMA-3.1, and Qwen3 models. Our experiments demonstrate exceptional data efficiency: on the Math task, IGDS surpasses full-dataset fine-tuning by a remarkable 17.4% on Gemma-2-2B while using only 50% of the data, and outperforms established baselines focused on data quality and diversity. Analysis confirms a strong positive correlation between feature amplification and task performance improvement. IGDS thus provides a direct and effective framework to enhance LLMs by leveraging their internal mechanisms, validating our core hypothesis.

[Long] From Logical to Computational Sparsity: Structure-Aware Block-Sparse Attention for Long-Code Completion ▶ *Underline Video*

Yanli Wang, Yanlin Wang, Bowen Zhang, Yiwei Zhang, Daya Guo, Jiachi Chen, Hongyu Zhang and Zibin Zheng

Code Large Language Models face critical Time-To-First-Token (TTFT) latency challenges when handling long code completion due to the quadratic complexity (O(n^2)) of attention mechanisms. While existing sparse attention methods attempt to address this issue, they suffer from three key limitations: (1) general sparse patterns cause excessive accuracy degradation without considering code structure, (2) code-specific methods achieve only logical sparsity without actual computational speedup, and (3) limited adaptation to complex scenarios such as repository-level completion. We propose SabreCoder, a training-free Structure-aware block-sparse attention mechanism that bridges the gap between logical and computational sparsity. SabreCoder parses code into semantic chunks, constructs chunk-level sparse patterns through dependency analysis and similarity matching, and maps them to GPU-friendly block-sparse formats. Extensive experiments on LCC and CrossCodeEval benchmarks demonstrate that SabreCoder reduces TTFT by 45-55% while maintaining accuracy within 3% of dense attention.

[Industry] From Short Video to Clickable Search: RLVR-Enabled Listwise Query Suggestion with Retrieval-Augmented Context ▶ *Underline Video*

Mingkai Tian, xuye xuye, Long Meng, Liwei Chen, Zhiheng Qin and Yi Wang

Short-video platforms now present tappable search entries beneath the video player, making it effortless for users to shift from passively watching to actively searching for information. Prior work on bottom-bar query generation conditions on titles and OCR to generate a single query per forward pass, constrains decoding with a trie, and evaluates against a single reference using edit-distance-style supervision, making it difficult to cover the diverse intents a video can trigger and to credit semantically equivalent query variants. Motivated by these limitations, we propose four complementary improvements. First, we reformulate the task as one-shot list generation, producing multiple distinct queries per video, and build multi-query ground truth from exposure and CTR logs. Second, we redesign offline evaluation with CTR-HungF1, a CTR-weighted set-matching metric via optimal assignment over token-level F1 score. Third, we enrich context with a video-to-video-to-query (V2V2Q) RAG pipeline to provide behavior-grounded background knowledge. Finally, we apply thinking-free RLVR with deterministic format checks and CTR-HungF1 rewards to train a compact LLM without reward models or CoT distillation. The resulting system yields strong offline and online improvements, and has been deployed on Kuaishou to serve hundreds of millions of users daily.

[Long] FS-Researcher: Test-Time Scaling for Long-Horizon Research Tasks with File-System-Based Agents ▶ *Underline Video*

Chiwei Zhu, Benfeng Xu, Mingxuan Du, Shaohan Wang, Xiaorui Wang, Zhendong Mao and Yongdong Zhang

Deep research is emerging as a representative long-horizon task for large language model (LLM) agents. However, long trajectories in deep research often exceed model context limits, compressing token budgets for both evidence collection and report writing, and preventing effective test-time scaling. We introduce FS-Researcher, a file-system-based, dual-agent framework that scales deep research beyond the context window via a persistent workspace. Specifically, a Context Builder agent acts as a librarian which browses the internet, writes structured notes, and archives raw sources into a hierarchical knowledge base that can grow far beyond context length. A Report Writer agent then composes the final report section by section, treating the knowledge base as the source of facts. In this framework, the file system serves as a durable external memory and a shared coordination medium across agents and sessions, enabling iterative refinement beyond the context window. Experiments on two open-ended benchmarks (DeepResearch Bench and DeepConsult) show that FS-Researcher achieves

state-of-the-art report quality across different backbone models. Further analyses demonstrate a positive correlation between final report quality and the computation allocated to the Context Builder, validating effective test-time scaling under the file-system paradigm. The code and data are open-sourced at https://github.com/Ignoramus0817/FS-Researcher.

[Findings] G-HiRel: Enhancing the Adaption to Knowledge Updating for Large Language Model Reasoning ► *Underline Video*
Yudai Pan, Jiajie Hong, Tianzhe Zhao, Lingyun Song, Lingling Zhang, Yixin Chen and Xuequn Shang
Large language models (LLMs) have achieved good performance in multiple reasoning tasks. However, they are limited to adapt the rapid knowledge updates in the real-world scenario without retraining the entire LLM or modifying the model weights. Excluding these consuming methods, knowledge graphs (KGs) are used as external memory under knowledge updating because of their structural knowledge and efficient updating ability, which is yet limited by the gap between structural KG and LLM, and the deficient entity-independent semantics. To this end, we propose an LLM reasoning framework with hierarchical relational retrieval for large-scale knowledge updating, named G-HiRel. To integrate the structural edited KG into continuous LLMs, G-HiRel generates hierarchical instructions based on natural language questions. In order to handle the knowledge inconsistency between the KG and LLM and obtain the entity independence, G-HiRel utilizes a designed hierarchical relational retrieval for relational path candidates, which are selected by a designed semantics-based strategy. Finally, top entity-independent relational paths are instantiated and integrated into LLMs to generate the answer, in order to verify the reasoning performance under knowledge edits. Extensive experiments of G-HiRel on three benchmarks show that G-HiRel achieves superiority in terms of accuracy and interpretability. The code of G-HiRel is available at the link: https://github.com/HJJ-designed/G-HiRel.

[Long] GenDis: Generative-Discriminative Dual-View Co-Training for Generalized Category Discovery ► *Underline Video*
Xi Chen, Chuan Qin, Jinpeng Li, Shasha Hu, Chao Wang, Hengshu Zhu and Hui Xiong
Generalized Category Discovery (GCD) aims to identify both known and novel categories from partially labeled data, reflecting more realistic open-world learning scenarios. However, most existing methods rely solely on one-hot discriminative supervision, leading to overfitting on seen classes and poor generalization to unseen ones. Recent advances introduce large language models (LLMs) to incorporate external semantics, yet they often suffer from semantic-label misalignment and weak semantic integration during training. We propose GenDis, a Generative-Discriminative Dual-View Co-Training framework that unifies discriminative classification and semantic label generation within an LLM. Discriminative pseudo-labels guide the formation of a separable generative latent space, enabling semantically meaningful supervision for novel classes. To ensure consistency between the two views, we employ Canonical Correlation Analysis (CCA)-based alignment and a curriculum-guided, dispersion-aware pseudo-labeling strategy for iterative refinement. Extensive experiments on five GCD benchmarks demonstrate that GenDis substantially outperforms prior methods, validating the effectiveness of dual-view co-training with semantically enriched supervision. The anonymized repository is available at https://anonymous.4open.science/r/GenDis.

[TACL] Generative Induction of Dialogue Task Schemas with Streaming Refinement and Simulated Interactions ► *Underline Video*
James Finch, Yasasvi Josyula and Jinho Choi
In task-oriented dialogue (TOD) systems, Slot Schema Induction (SSI) is essential for automatically identifying key information slots from dialogue data without manual intervention. This paper presents a novel state-of-the-art (SoTA) approach that formulates SSI as a text generation task, where a language model incrementally constructs and refines a slot schema over a stream of dialogue data. To develop this approach, we present a fully automatic LLM-based TOD simulation method that creates data with high-quality state labels for novel task domains. Furthermore, we identify issues in SSI evaluation due to data leakage and poor metric alignment with human judgment. We resolve these by creating new evaluation data using our simulation method with human guidance and correction, as well as designing improved evaluation metrics. These contributions establish a foundation for future SSI research and advance the SoTA in dialogue understanding and system development.

[Long] GenPT: Beyond Self-Report for Reliable LLM Psychometrics via Generative Projective Testing ► *Underline Video*
Ming Wang, Shuang Wu, Bixuan Wang, Lu Lin, Yuxin Chen, Xiaocui Yang, Daling Wang, Shi Feng, Yifei Zhang and Yufan Sun
Self-report questionnaires remain the default tool for probing the psychological characteristics of Large Language Model (LLM) agents, yet classical instruments (BFI, BDI, MBTI, BSS) inherit three well-known threats under LLMs: contamination from training corpora, directional bias under social-desirability framing, and limited responsiveness to context beyond the item text. We ask whether a projective paradigm can be adapted into a usable psychometric tool for LLM agents. We introduce GenPT (Generative Projective Testing), which reformulates TAT, Rorschach, and SCT with newly generated stimuli and organises assessment as a three-stage pipeline (Behavior Collection -> Interpretation -> Diagnosis) grounded in SCORS-G and a Simplified Rorschach Analysis System. On personality traits (Big Five, MBTI) and mental-health risks (depression, suicide ideation), questionnaires exhibit systematic directional shifts under social-desirability framing, most strongly on suicide ideation, whereas GenPT's collected behavioral patterns stay near the symmetric baseline; under a longitudinal counselling context, GenPT-based depression assessment shifts by roughly an order of magnitude more than its questionnaire counterpart. Questionnaires remain competitive on clean-persona trait tasks where items align lexically with the persona description. Overall, GenPT complements rather than replaces self-report when contamination resistance, bias asymmetry, and context sensitivity matter. Code and stimuli: https://github.com/sci-m-wang/GenPT.

[Industry] GeoGround: Uncertainty-Weighted Multi-Task Learning for Geo-Alignment and Address Defect Detection ► *Underline Video*

Srinivas Virinchi, Aman Gulati and Anoop Saladi

Address intelligence in e-commerce demands accurate geocoding and proactive defect detection under strict sub-50~ms latency constraints. These tasks are inherently coupled: precise spatial grounding provides a strong prior for defect propensity, yet prior approaches optimize them independently. While generative LLMs offer rich semantic representations, they lack spatial inductive bias and fail to meet real-time serving requirements. We introduce GeoGround, a multi-task learning framework that jointly models coordinate grounding and address defect detection. The model combines a hierarchical spatial grounding objective with Focal Loss for defect classification, using uncertainty-based task weighting to balance optimization under severe class imbalance. To strengthen supervision, we curate a large-scale noisy address dataset using LLM-assisted data construction, augmenting the training corpus with signals that are costly to obtain manually. GeoGround achieves 5.86x gains in address defect detection precision and up to 4.86x improvements in location prediction accuracy over strong encoder baselines, while remaining 75x more efficient than decoder LLMs such as Qwen2-1.5B. A two-week online A/B test in a large-scale delivery pipeline confirms real-world impact, yielding a 50 bps uplift in defect detection, a 40 bps gain in location prediction, and an estimated operational savings of $3.09M annually.

[Long] GLIER: Generative Legal Inference and Evidence Ranking for Legal Case Retrieval ▶ *Underline Video*

Minghan Li, Tianrui Lv, Chao Zhang and Guodong Zhou

The semantic gap between colloquial user queries and professional legal documents presents a fundamental challenge in Legal Case Retrieval (LCR). Existing dense retrieval methods typically treat LCR as a black-box semantic matching process, neglecting the explicit juridical logic that underpins legal relevance. To address this, we propose GLIER (Generative Legal Inference and Evidence Ranking), a framework that reformulates retrieval as an inference process over latent legal variables. GLIER decomposes the task into two interpretability-driven stages: (1) A Joint Generative Inference module that translates raw queries into latent legal indicators (Charges and Legal Elements), employing a unified sequence-to-sequence strategy where charges and elements are generated jointly to enforce logical consistency; and (2) A Multi-View Evidence Fusion mechanism that aggregates generative confidence with structural and lexical signals for precise ranking. Extensive experiments on LeCaRD and LeCaRDv2 demonstrate that GLIER outperforms strong baselines like SAILER and KELLER. Notably, our framework exhibits exceptional data efficiency, maintaining robust performance even when trained with only 10% of the data.

[Long] Global Adaptive Momentum Meets Local Personalized Perturbation: Efficient Federated LLM Fine-Tuning with Zeroth-Order Gradients ▶ *Underline Video*

Zihan Chen, Howard H. Yang, Tony Q.S. Quek and Kai Fong Ernest Chong

Federated fine-tuning of large language models (LLMs) provides a privacy-preserving approach to deploying pervasive generative AI services, yet the substantial memory overhead of first-order (FO) gradient computation presents significant practical challenges. While zeroth-order (ZO) optimization methods offer memory-efficient alternatives, they remain susceptible to performance degradation brought by data heterogeneity. Specifically, direct ZO-for-FO substitution is incompatible with existing strategies tailored for cross-client discrepancies. In response, we propose a new federated LLM fine-tuning framework, with a holistic revamped design of the entire ZO gradient processing pipeline. Crucially, with our proposed global adaptive optimization and local personalized perturbation, we present a unified solution for incorporating ZO gradients in federated learning, from local personalized perturbation sampling and ZO gradient transmission, to global ZO gradient reconstruction and aggregation with adaptive momentum, thereby directly addressing the challenges of inefficiencies and cross-client discrepancies. Our convergence analysis and experiment results demonstrate the superiority of our proposed framework over diverse heterogeneous data settings, both in terms of generalization and efficiency.

[Industry] Graph-Based Phonetic Error Correction of Noisy ASR ▶ *Underline Video*

Pratik Rakesh Singh, Mohammadi Zaki, Aneesh Mukkamla and Pankaj Wasnik

Automatic speech recognition (ASR) systems, despite low overall word error rates, produce residual lexical errors that disproportionately affect semantically critical tokens such as named entities, negations, and sentiment-bearing words. These errors are often structured, arising from phonetic similarity rather than random noise, making naive token-level correction insufficient. We propose a structured ASR correction framework, that we call G-SPIN, that combines phonetic graph modeling with contextual language understanding. A graph neural network (GNN) first constructs acoustically plausible candidate neighborhoods for flagged tokens, explicitly restricting the correction search space to phonetic alternatives. A masked language model (MLM) then provides local contextual scoring, and an instruction-tuned large language model (LLM) performs final context-aware re-ranking over this compact candidate set. By decoupling structured phonetic reasoning from contextual semantic selection, our method avoids unconstrained generation while improving correction accuracy. The framework is lightweight, modular, and operates entirely at inference time.

[Findings] Graph-GRPO: Stabilizing Multi-Agent Topology Learning via Group Relative Policy Optimization ▶ *Underline Video*

Yueyang Cang

Optimizing communication topology is fundamental to the efficiency and effectiveness of Large Language Model (LLM)-based Multi-Agent Systems (MAS). While recent approaches utilize reinforcement learning to dynamically construct task-specific graphs, they typically rely on single-sample policy gradients with absolute rewards (e.g., binary correctness). This paradigm suffers from severe gradient variance and the credit assignment problem: simple queries yield non-informative positive rewards for suboptimal structures, while difficult queries often result in failures that provide no learning signal. To address these challenges, we propose Graph-GRPO, a novel topology optimization framework that integrates Group Relative Policy Optimization. Instead of evaluating a single topology in isolation, Graph-GRPO samples

a group of diverse communication graphs for each query and computes the advantage of specific edges based on their relative performance within the group. By normalizing rewards across the sampled group, our method effectively mitigates the noise derived from task difficulty variance and enables fine-grained credit assignment. Extensive experiments on reasoning and code generation benchmarks demonstrate that Graph-GRPO significantly outperforms state-of-the-art baselines, achieving superior training stability and identifying critical communication pathways previously obscured by reward noise.

[Findings] Hard to Read, Easy to Jailbreak: How Visual Degradation Bypasses MLLM Safety Alignment ▶ *Underline Video*

ZhixueSong ZhixueSong, Boyan Han, Yiwei Wang and Chi Zhang

Recent advancements in visual context compression enable MLLMs to process ultra-long contexts efficiently by rendering text into images. However, we identify a critical vulnerability inherent to this paradigm: lowering image resolution inadvertently catalyzes jailbreaking. Our experiments reveal that the safety defenses of SOTA models deteriorate sharply as resolution degrades, surprisingly persisting even when text remains legible. We attribute this to "Cognitive Overload", hypothesizing that the effort required to decipher degraded inputs diverts attentional resources from safety auditing. This phenomenon is consistent across various visual perturbations, including noise and geometric distortion. To address this, we propose a simple "Structured Cognitive Offloading" strategy that mitigates these risks by enforcing a serialized pipeline to decouple visual transcription from safety assessment. Our work exposes a significant risk in vision-based compression and provides critical insights for the secure design of future MLLMs.

[Long] HeLa-Mem: Hebbian Learning and Associative Memory for LLM Agents ▶ *Underline Video*

Jinchang Zhu, Jindong Li, Cheng Zhang, Jiahong Liu and Menglin Yang

Long-term memory is a critical challenge for Large Language Model agents, as fixed context windows cannot preserve coherence across extended interactions. Existing memory systems represent conversation history as unstructured embedding vectors, retrieving information through semantic similarity. This paradigm fails to capture the associative structure of human memory, wherein related experiences progressively strengthen interconnections through repeated co-activation. Inspired by cognitive neuroscience, we identify three mechanisms central to biological memory: association, consolidation, and spreading activation, which remain largely absent in current research. To bridge this gap, we propose HeLa-Mem, a bio-inspired memory architecture that models memory as a dynamic graph with Hebbian learning dynamics. HeLa-Mem employs a dual-level organization: (1) an episodic memory graph that evolves through co-activation patterns, and (2) a semantic memory store populated via Hebbian Distillation, wherein a Reflective Agent identifies densely connected memory hubs and distills them into structured, reusable semantic knowledge. This dual-path design leverages both semantic similarity and learned associations, mirroring the episodic-semantic distinction in human cognition. Experiments on LoCoMo demonstrate superior performance across four question categories while using significantly fewer context tokens. Code is available on GitHub: https://github.com/ReinerBRO/HeLa-Mem

[Findings] Human-Agent Collaborative Paper-to-Page Crafting ▶ *Underline Video*

Qianli Ma, SiyuWang SiyuWang, Yilin Chen, Yinhao Tang, Yixiang Yang, Chang Guo, Bingjie Gao, Zhening Xing, Yanan Sun and Zhipeng Zhang

In the quest for scientific progress, communicating research is as vital as the discovery itself. Yet, researchers are often sidetracked by the manual, repetitive chore of building project webpages to make their dense papers accessible. While automation has tackled static slides and posters, the dynamic, interactive nature of webpages has remained an unaddressed challenge. To bridge this gap, we reframe the problem, arguing that the solution lies not in a single command, but in a collaborative, hierarchical process. We introduce AutoPage, a novel multi-agent system that embodies this philosophy. AutoPage deconstructs paper-to-page creation into a coarse-to-fine pipeline from narrative planning to multimodal content generation and interactive rendering. To combat AI hallucination, dedicated "Checker" agents verify each step against the source paper, while optional human checkpoints ensure the final product aligns perfectly with the author's vision, transforming the system from a mere tool into a powerful collaborative assistant. To rigorously validate our approach, we also construct PageBench, the first benchmark for this new task. Experiments show AutoPage not only generates high-quality, visually appealing pages but does so with remarkable efficiency in under 15 minutes for less than $0.1. Code and data will be released.

[Long] Incentivizing Parametric Knowledge via Reinforcement Learning with Verifiable Rewards for Cross-Cultural Entity Translation ▶ *Underline Video*

Jiang Zhou, Xiaohu Zhao, Xinwei Wu, Tianyu Dong, Hao Wang, Yangyang Liu, Heng Liu, Linlong Xu, Longyue Wang, Weihua Luo and Deyi Xiong

Cross-cultural entity translation remains challenging for large language models (LLMs) as literal or phonetic renderings are usually yielded instead of culturally appropriate translations in context. However, relevant knowledge may already be encoded in model parameters during large-scale pre-training. To incentivize the effective use of parametric knowledge, we propose EA-RLVR (Entity-Anchored Reinforcement Learning with Verifiable Rewards), a training framework that optimizes cross-cultural entity translation without relying on external knowledge bases. EA-RLVR anchors supervision on a verifiable, entity-level reward signal and incorporates lightweight structural gates to stabilize optimization. This design steers the model toward learning a robust reasoning process rather than merely imitating reference translations. We evaluate EA-RLVR on XC-Translate and observe consistent improvements in both entity translation accuracy and out-of-domain generalization. Specifically, training on merely 7k samples boosts Qwen3-14Bs entity translation accuracy from 23.66% to 31.87% on a 50k test set comprising entirely unseen entities. The learned entity translation ability also transfers to general translation, yielding +1.35 XCOMET on WMT24pp, which scales to +1.59 with extended optimization. Extensive analyses of pass@k dynamics and

reward formulations attribute these gains to superior sampling efficiency and a stable optimization landscape.

[Long] InferenceDynamics: Adaptive LLM Routing through Structured Capability and Knowledge Profiling ▶ *Underline Video*
Haochen Shi, Tianshi Zheng, Weiqi Wang, Baixuan Xu, Chunyang Li, Chunkit Chan, Tao Fan and Yangqiu Song
Large Language Model (LLM) routing is a pivotal technique for navigating a diverse landscape of LLMs, enabling the selection of the best-performing LLMs for specific user queries while balancing performance and cost. However, current routing approaches often face limitations in scalability when dealing with a large pool of specialized LLMs, or in their adaptability to extending model scope and evolving capability domains. To overcome those challenges, we propose InferenceDynamics, a flexible and scalable multi-dimensional routing framework by modeling the capability and knowledge of models. We operate it on our comprehensive dataset RouteMix, and demonstrate its effectiveness and generalizability in group-level routing using modern benchmarks including MMLU-Pro, GPQA, BigGenBench, and LiveBench, showcasing its ability to identify and leverage top-performing models for given tasks, leading to superior outcomes with cost efficiency. The broader adoption of InferenceDynamics can empower users to harness the full specialized potential of the LLM ecosystem, and our code will be made publicly available to encourage further research.

[Long] InquireMobile: Teaching VLM-based Mobile Agent to Request Human Assistance via Reinforcement Fine-Tuning ▶ *Underline Video*
Qihang Ai, Pi Bu, Yue Cao, Yingyao Wang, Jihao Gu, Jingxuan Xing, Zekun Zhu, Wei Jiang, Zhicheng Zheng, Jun Song and Yuning Jiang
Recent advances in Vision-Language Models (VLMs) have enabled mobile agents to perceive and interact with real-world mobile environments based on human instructions. However, the current fully autonomous paradigm poses potential safety risks when model understanding or reasoning capabilities are insufficient. To address this challenge, we first introduce InquireBench, a comprehensive benchmark specifically designed to evaluate mobile agents' capabilities in safe interaction and proactive inquiry with users, encompassing 5 categories and 22 sub-categories, where most existing VLM-based agents demonstrate near-zero performance. In this paper, we aim to develop an interactive system that actively seeks human confirmation at critical decision points. To achieve this, we propose InquireMobile, a novel model inspired by reinforcement learning, featuring a two-stage training strategy and an interactive pre-action reasoning mechanism. Finally, our model achieves an 46.8% improvement in inquiry success rate and the best overall success rate among existing baselines on InquireBench. The project page is available at https://bit-aqh.github.io/InquireMobile/homepage/.

[Long] Inside Out: Evolving User-Centric Core Memory Trees for Long-Term Personalized Dialogue Systems ▶ *Underline Video*
Jihao Zhao, Ding Chen, Zhaoxin Fan, Kerun Xu, Mengting Hu, Bo Tang, Feiyu Xiong and Zhiyu li
Existing long-term personalized dialogue systems struggle to reconcile unbounded interaction streams with finite context constraints, often succumbing to memory noise accumulation, reasoning degradation, and persona inconsistency. To address these challenges, this paper proposes Inside Out, a framework that utilizes a globally maintained PersonaTree as the carrier of long-term user profiling. By constraining the trunk with an initial schema and updating the branches and leaves, PersonaTree enables controllable growth, achieving memory compression while preserving consistency. Moreover, we train a lightweight MemListener via reinforcement learning with process-based rewards to produce structured, executable, and interpretable ADD, UPDATE, DELETE, NO_OP operations, thereby supporting the dynamic evolution of the personalized tree. During response generation, PersonaTree is directly leveraged to enhance outputs in latency-sensitive scenarios; when users require more details, the agentic mode is triggered to introduce details on-demand under the constraints of the PersonaTree. Experiments show that PersonaTree outperforms full-text concatenation and various personalized memory systems in suppressing contextual noise and maintaining persona consistency. Notably, the small MemListener model achieves memory-operation decision performance comparable to, or even surpassing, powerful reasoning models such as DeepSeek-R1-0528 and Gemini-3-Pro.

[Long] IS-CoT: Breaking the Long-form Generation Collapse via Interleaved Structural Thinking ▶ *Underline Video*
Zechen Sun, Yuyang Sun, Zecheng Tang, Juntao Li, Wenpeng Hu, Wenliang Chen, Zhunchen Luo, Guotong Geng and Min Zhang
Generating coherent and controllable long-form content remains a persistent challenge for Large Language Models (LLMs). While reasoning-enhanced models have demonstrated success in logic-intensive domains, our evaluation reveals that they suffer from a severe length collapse in open-ended writing, where performance degrades sharply as target lengths exceed 2,000 words. We attribute this failure to the limitation of static hierarchical planning, which struggles to provide dynamic guidance over extended contexts. To bridge this gap, we introduce the Interleaved Structural Chain-of-Thought (IS-CoT) framework. Unlike external agentic workflows, IS-CoT embeds a dynamic Plan-Write-Reflect cycle into the generation process, enabling continuous strategy adaptation and global alignment without additional assistance. Based on this framework, we construct a high-quality dataset of interleaved reasoning traces via a multi-teacher pipeline and train IS-Writer-8B. Experiments demonstrate that IS-Writer-8B achieves state-of-the-art performance on challenging long-form benchmarks (e.g., +3.08 vs. DeepSeek-V3.2 on LongBench-Write), exhibiting robust length compliance and coherence competitive with significantly larger proprietary models.

[Long] JPU: Bridging Jailbreak Defense and Unlearning via On-Policy Path Rectification ▶ *Underline Video*
Xi Wang, Songlei Jian, Shasha Li, Xiaopeng Li, Zhaoye Li, Bin Ji, Baosheng Wang and Jie Yu
Despite extensive safety alignment, Large Language Models (LLMs) often fail against jailbreak attacks. While machine unlearning has emerged as a promising defense by erasing specific harmful parameters, current methods remain vulnerable to diverse jailbreaks. We first conduct an empirical study and discover that this failure mechanism is caused by jailbreaks primarily activating non-erased parameters

in the intermediate layers. Further, by probing the underlying mechanism through which these circumvented parameters reassemble into the prohibited output, we verify the persistent existence of dynamic jailbreak paths and show that the inability to rectify them constitutes the fundamental gap in existing unlearning defenses. To bridge this gap, we propose Jailbreak Path Unlearning (JPU), which is the first to rectify dynamic jailbreak paths towards safety anchors by dynamically mining on-policy adversarial samples to expose vulnerabilities and identify jailbreak paths. Extensive experiments demonstrate that JPU significantly enhances jailbreak resistance against dynamic attacks while preserving the model's utility.

[Findings] K-GIP: Diagnosing Logical Fractures in Large Vision-Language Models via Verification Scene Graphs and Sequential Pruning ▶ *Underline Video*

Yujun Hu, Xiaoyu Zhou, Changbo Wang and Gaoqi He

Diagnosing fine-grained hallucinations in Large Vision-Language Models (LVLMs) can greatly advance their reliable deployment in real-world applications. Nevertheless, current benchmarks predominantly employ flat metrics that treat errors in isolation, leaving a gap in evaluating the complex causal dependencies between visual perception and textual reasoning. Motivated by this, we introduce the Knowledge-Guided In-Context Probing (K-GIP) framework to fill this gap. Specifically, K-GIP constructs a high-fidelity dual-perception ground truth to transform abstract priors into multi-granularity queries. Furthermore, we propose a Verification Scene Graph metric equipped with a Sequential Logic Pruning protocol, which explicitly models existence-attribute dependencies to strictly penalize logical fractures. We conduct comprehensive evaluations of mainstream LVLMs across three datasets using K-GIP. The experimental results highlight that our methodology successfully isolates deep reasoning failures from simple perceptual misses. We hope K-GIP can serve as a valuable and rigorous standard to assess logical robustness in multimodal systems.

[Long] KV-Embedding: Training-free Text Embedding via Internal KV Re-routing in Decoder-only LLMs ▶ *Underline Video*

Yixuan Tang and Yi Yang

While LLMs are powerful embedding backbones, their application in training-free settings faces two structural challenges: causal attention restricts early tokens from accessing subsequent context, and the next-token prediction objective biases representations toward generation rather than semantic compression. To address these limitations, we propose KV-Embedding, a framework that activates the latent representation power of frozen LLMs. Our method leverages the observation that the key-value (KV) states of the final token at each layer encode a compressed view of the sequence. By re-routing these states as a prepended prefix, we enable all tokens to access sequence-level context within a single forward pass. To ensure model-agnostic applicability, we introduce an automated layer selection strategy based on intrinsic dimensionality. Evaluations on MTEB across Qwen, Mistral, and Llama backbones show that KV-Embedding outperforms existing training-free baselines by up to 10%, while maintaining robust performance on sequences up to 4,096 tokens. These results demonstrate that internal state manipulation offers an efficient alternative to input modification, and we hope this work encourages further exploration of LLM internals for representation learning.

[Long] Large Language Model-Enhanced Multi-Armed Bandits ▶ *Underline Video*

Jiahang Sun, Zhiyong Wang, Runhan Yang, Chenjun Xiao, John C.S. Lui and Zhongxiang Dai

Large language models (LLMs) have been applied to sequential decision-making tasks like multi-armed bandits (MAB), where an LLM is tasked with selecting arms in each iteration. However, this direct arm selection approach is often suboptimal. We propose an alternative method combining classical MAB algorithms with LLMs. Specifically, we use a classical MAB framework and leverage the in-context learning capability of LLMs for reward prediction. First, we integrate the LLM-based predictor into Thompson sampling (TS) with a decaying temperature schedule to balance exploration and exploitation. We also incorporate the predictor into a regression oracle-based MAB algorithm with explicit exploration. Additionally, we extend our TS-based algorithm to dueling bandits, where only preference feedback between arm pairs is available, requiring significant algorithmic modifications. Our empirical evaluations on synthetic MAB tasks show that our algorithms outperform LLM-based direct arm selection. In experiments on real-world text datasets, we demonstrate that, in tasks where arms lack exploitable semantic meaning, our approach delivers significantly better performance than direct arm selection.

[Findings] Large-Scale Multimodal Knowledge Graph about Classical Chinese Poetry: Fine-grained Method and Comprehensive Evaluation ▶ *Underline Video*

Shuo Wang, Qing Zhu, Yang Xiao and Minglong Lei

Classical Chinese poetry is a treasured cultural heritage of humanity, attracting extensive research interest. However, the study of classical Chinese poetry is hindered by the lack of open, large-scale, and fine-grained multimodal datasets.Prior datasets are either limited by modality constraints, dataset size, or the level of dataset refinement, making them inadequate for effectively supporting studies and the development of applications in classical Chinese poetry.To address these issues, we propose a method for constructing a large-scale and fine-grained multimodal knowledge graph of classical Chinese poetry. We first design an informative ontology graph for classical Chinese poetry and comprehensively collect knowledge about poetry based on it. Furthermore, the method leverages knowledge augmentation, prompt optimization, and text-image alignment to acquire comprehensive, fine-grained knowledge. Both qualitative and quantitative evaluations are conducted on the Multimodal Knowledge Graph of Classical Chinese Poetry (CPMK), highlighting its comprehensiveness and high quality.We also conduct downstream evaluations on four tasks: poetry question answering, poetry theme classification, poetry-image retrieval, and rigid-formats poetry generation.Significant results are achieved across all four tasks, demonstrating CPMK's effectiveness in supporting research on Chinese poetry.CPMK will be released to promote research in Chinese culture.

[Long] League of LLMs: A Benchmark-Free Paradigm for Mutual Evaluation of Large Language Models ▶ *Underline Video*

Qianhong Guo, Wei Xie, Xiaofang Cai, Enze Wang, Shuoyoucheng Ma, Xiaobing Sun, Tian Xia, Kai Chen, Xiaofeng Wang and Baosheng Wang

Although large language models (LLMs) have shown exceptional capabilities across a wide range of tasks, reliable evaluation remains a critical challenge due to data contamination, opaque operation, and subjective preferences. To address these issues, we propose League of LLMs (LOL), a novel benchmark-free evaluation paradigm that organizes multiple LLMs into a self-governed league for multi-round mutual evaluation. LOL integrates four core criteria (dynamic, transparent, objective, and professional) to mitigate key limitations of existing paradigms. Experiments on eight mainstream LLMs in mathematics and programming demonstrate that LOL can effectively distinguish LLM capabilities while maintaining high internal ranking stability (Top-k consistency = 70.7%). Beyond ranking, LOL reveals empirical findings that are difficult for traditional paradigms to capture. For instance, "memorization-based answering" behaviors are observed in some models, and higher in-family scores are found in the OpenAI model family (= 9, $p < 0.05$). Finally, we make our framework and code publicly available as a valuable complement to the current LLM evaluation ecosystem.

[Long] Learning to Select: Query-Aware Adaptive Dimension Selection for Dense Retrieval ▶ *Underline Video*

Zhanyu Wu, Richong Zhang and Zhijie Nie

Dense retrieval represents queries and documents as high-dimensional embeddings, but these representations can be redundant at the query level: for a given information need, only a subset of dimensions is consistently helpful for ranking. Prior work addresses this via pseudo-relevance feedback (PRF) based dimension importance estimation, which can produce query-aware masks without labeled data but often relies on noisy pseudo signals and heuristic test-time procedures. In contrast, supervised adapter methods leverage relevance labels to improve embedding quality, yet they learn global transformations shared across queries and do not explicitly model query-aware dimension importance. We propose a Query-Aware Adaptive Dimension Selection framework that learns to predict per-dimension importance directly from query embedding. We first construct oracle dimension importance distributions over embedding dimensions using supervised relevance labels, and then train a predictor to map a query embedding to these label-distilled importance scores. At inference, the predictor selects a query-aware subset of dimensions for similarity computation based solely on the query embedding, without pseudo-relevance feedback. Experiments across multiple dense retrievers and benchmarks show that our learned dimension selector improves retrieval effectiveness over the full-dimensional baseline as well as PRF-based masking and supervised adapter baselines.

[Long] LeLoRA: Learnable Low-Rank Adaptation of Large Language Models ▶ *Underline Video*

Xiaoling Zhou, Mingjie Zhang, Zhemg Lee, Wei Ye and Shikun Zhang

Fine-tuning large language models (LLMs) is an effective approach to enhancing their performance on specialized downstream tasks. Among the various techniques, low-rank adaptation has garnered significant attention due to its ability to maintain the full performance of fine-tuning while enhancing computational efficiency. However, existing approaches often rely on manually specified and fixed hyperparameters to identify the trainable components within weight matrices, resulting in suboptimal performance and low parameter efficiency. This paper presents a novel Learnable Low-Rank Adaptation (LeLoRA) framework that utilizes dynamically learned fine-tuning strategies to facilitate the effective adaptation of LLMs. Our framework integrates an LLM with a policy network that automatically and adaptively generates matrix-specific adaptation strategies to identify the trainable components of each weight matrix, taking into account their unique characteristics, such as singular values and matrix norms. A reinforcement learning-based optimization algorithm is then employed to iteratively update the LLM and the policy network, ensuring that the generated strategies adapt in real time to the evolving states of the LLM. Extensive experiments have been conducted across various natural language processing and multimodal tasks. The results across ten different LLMs, ranging from 125M to 70B parameters, provide compelling evidence that LeLoRA consistently outperforms existing baselines in adapting LLMs. Moreover, analytical experiments provide valuable insights into the effectiveness of the generated strategies.

[Long] LexRel: Benchmarking Legal Relation Extraction for Chinese Civil Cases ▶ *Underline Video*

Yida Cai, Ranjuexiao Hu, Huiyuan Xie, Chenyang Li, Yun Liu, Yuxiao Ye, Zhenghao Liu, Weixing Shen and Zhiyuan Liu

Legal relations serve as an important analytical framework for dispute resolution in civil cases. However, legal relations in Chinese civil cases remain underexplored in the field of legal AI, largely due to the absence of comprehensive schemas. In this work, we first introduce a comprehensive schema for legal relations in civil cases, which contains a hierarchical taxonomy and definitions of arguments. Based on this schema, we formulate a legal relation extraction task and present LexRel, an expert-annotated benchmark for legal relation extraction in the Chinese civil law domain. We use LexRel to evaluate state-of-the-art large language models (LLMs) on legal relation extraction, showing that current LLMs exhibit significant limitations in accurately identifying civil legal relations. Furthermore, we demonstrate that explicitly incorporating information about legal relations leads to promising performance gains on other downstream legal AI tasks.

[SRW] Linguistically-Informed Evaluation of LLMs on Acceptability Judgments in a Forced-Choice Paradigm ▶ *Underline Video*

Ziyue Liu; Nils Reiter

Evaluating the grammatical abilities of large language models (LLMs) is important for both NLP and linguistic theory. We investigate the ability of large language models (LLMs) to perform acceptability judgments in a forced-choice paradigm. We evaluate a subset of LLMs on 150 minimal sentence pairs sampled from Linguistic Inquiry and categorized using BLiMP linguistic phenomena. Our results show that while LLMs approximate human judgments, performance varies across models and phenomenon types, with less stable behavior on more

complex phenomena. Prompting strategies have minimal impact.

[Long] Listening Like Humans: Semantics-Guided Noise-Robust Multimodal Speech Recognition ▶ *Underline Video*
Yan Fang, Jun Chen, Yian Yao, Shuxin Zhong, Min Sun and Kaishun Wu
Severe acoustic degradation is often caused by overlapping noise, disfluencies, and environmental distortions. This phenomenon results in the dissolution of linguistic structures and the generation of unreliable ASR outputs. Inspired by human speech comprehension, we propose Speech-MLM, a novel multimodal framework that reframes ASR as semantics-guided speech reconstruction. This perspective introduces three core challenges: (C1) collapse of linguistic structure under acoustic degradation, (C2) semantic ambiguity under noise, and (C3) misalignment across modalities. To address these issues, we propose Speech-MLM, a multimodal ASR framework that integrates speech, spectrogram-derived visual cues, and textual variants to enhance robustness. It consists of: (i) Cognitive Structure Extractor that recovers prosodic structure from visualized acoustic features, (ii) Semantic Weaver that learns semantic equivalence across varied textual forms, and (iii) Retrieval-Guided Fusion Learner that unifies modalities within a shared semantic space. Experiments on multiple real-world noisy datasets demonstrate that Speech-MLM achieves an average 38.85% reduction in WER, while also attaining 98.71% BERTScore and 96.7% USE, over advanced baselines, demonstrating substantial gains in semantic robustness and generalization across domains.

[Long] LLM-ForcedAligner: A Non-Autoregressive and Accurate LLM-Based Forced Aligner for Multilingual and Long-Form Speech ▶ *Underline Video*
Bingshen Mu, Xian Shi, Xiong Wang, Hexin Liu, Jin Xu and Lei Xie
Forced alignment (FA) predicts start and end timestamps for words or characters in speech, but existing methods are language-specific and prone to cumulative temporal shifts. The multilingual speech understanding and long-sequence processing abilities of speech large language models (SLLMs) make them promising for FA in multilingual, crosslingual, and long-form speech settings. However, directly applying the next-token prediction paradigm of SLLMs to FA results in hallucinations and slow inference. To bridge the gap, we propose LLM-ForcedAligner, reformulating FA as a slot-filling paradigm: timestamps are treated as discrete indices, and special timestamp tokens are inserted as slots into the transcript. Conditioned on the speech embeddings and the transcript with slots, the SLLM directly predicts the time indices at slots. During training, causal attention masking with non-shifted input and label sequences allows each slot to predict its own timestamp index based on itself and preceding context, with loss computed only at slot positions. Dynamic slot insertion enables FA at arbitrary positions. Moreover, non-autoregressive inference is supported, avoiding hallucinations and improving speed. Experiments across multilingual, crosslingual, and long-form speech scenarios show that LLM-ForcedAligner achieves a 69%~78% relative reduction in accumulated averaging shift compared with prior methods. The checkpoint and inference code will be released later.

[Long] LLMs Enable Bag-of-Texts Representations for Short-Text Clustering ▶ *Underline Video*
I-Fan Lin, Faegheh Hasibi and Suzan Verberne
In this paper, we propose a training-free method for unsupervised short text clustering that relies less on careful selection of embedders than other methods. In customer-facing chatbots, companies are dealing with large amounts of user utterances that need to be clustered according to their intent. In these settings, no labeled data is typically available, and the number of clusters is not known. Recent approaches to short-text clustering in label-free settings incorporate LLM output to refine existing embeddings. While LLMs can identify similar texts effectively, the resulting similarities may not be directly represented by distances in the dense vector space, as they depend on the original embedding. We therefore propose a method for transforming LLM judgments directly into a bag-of-texts representation in which texts are initialized to be equidistant, without assuming any prior distance relationships. Our method achieves comparable or superior results to state-of-the-art methods, but without embeddings optimization or assuming prior knowledge of clusters or labels. Experiments on diverse datasets and smaller LLMs show that our method is model agnostic and can be applied to any embedder, with relatively small LLMs, and different clustering methods. We also show how our method scales to large datasets, reducing the computational cost of the LLM use. The flexibility and scalability of our method make it more aligned with real-world training-free scenarios than existing clustering methods.

[Long] LongVideoAgent: Multi-Agent Reasoning with Long Videos ▶ *Underline Video*
Runtao Liu, Ziyi Liu, Jiaqi Tang, Yue Ma, Renjie Pi, Jipeng Zhang and Qifeng Chen
Recent advances in multimodal LLMs and systems that use tools for long-video QA point to the promise of reasoning over hour-long episodes. However, many methods still compress content into lossy summaries or rely on limited toolsets, weakening temporal grounding and missing fine-grained cues. We propose a multi-agent framework in which a master LLM coordinates a grounding agent to localize question-relevant segments and a vision agent to extract targeted textual observations. The master agent plans with a step limit, and is trained with reinforcement learning to encourage concise, correct, and efficient multi-agent cooperation. This design helps the master agent focus on relevant clips via grounding, complements subtitles with visual detail, and yields interpretable trajectories. On our proposed LongTVQA and LongTVQA+ which are episode-level datasets aggregated from TVQA/TVQA+, our multi-agent system significantly outperforms strong non-agent baselines. Experiments also show reinforcement learning further strengthens reasoning and planning for the trained agent.

[Long] M^3-VQA: A Benchmark for Multimodal, Multi-Entity, Multi-Hop Visual Question Answering ▶ *Underline Video*
Jiatong Ma, Longteng Guo, Yuchen Liu, Zijia Zhao, Dongze Hao, Xuanxu Lin and Jing Liu
We present M^3-VQA, a novel knowledge-based Visual Question Answering (VQA) benchmark, to enhance the evaluation of multimodal large language models (MLLMs) in fine-grained multimodal entity understanding and complex multi-hop reasoning. Unlike existing

VQA datasets that focus on coarse-grained categories and simple reasoning over single entities, M^3-VQA introduces diverse multi-entity questions involving multiple distinct entities from both visual and textual sources. It requires models to perform both sequential and parallel multi-hop reasoning across multiple documents, supported by traceable, detailed evidence and a curated multimodal knowledge base. We evaluate 16 leading MLLMs under three settings: without external knowledge, with gold evidence, and with retrieval-augmented input. The poor results reveal significant challenges for MLLMs in knowledge acquisition and reasoning. Models perform poorly without external information but improve markedly when provided with precise evidence. Furthermore, reasoning-aware agentic retrieval surpasses heuristic methods, highlighting the importance of structured reasoning for complex multimodal understanding. M^3-VQA presents a more challenging evaluation for advancing the multimodal reasoning capabilities of MLLMs. Our code and dataset are available at https://github.com/CASIA-IVA-Lab/M3VQA.

[Long] MagicBench: Diagnosing Visual Agency Loss and Semantic Dependency in Multimodal LLMs ▶ *Underline Video*

Tang Da Huang, Weidong Tang, Wen Qi Xu and Xianpeng Guo

Multimodal Large Language Models typically assume linguistic context invariably enhances visual understanding. We study this assumption in semantic adversarial scenarios, specifically magic tricks, where narration deliberately diverges from physical reality. We introduce MagicBench, a diagnostic benchmark of 402 videos for evaluating MLLMs under hierarchical linguistic interference, together with a Physical Constraint Set (PCS) protocol for assessing adherence to physical laws. Evaluation uncovers a Semantic Dependency Paradox: (1) Semantic anchoring: Entity nouns act as anchors aiding localization, paradoxically boosting performance despite false predicates. (2) Visual Agency Loss: In semantic vacuums, multimodal performance collapses 12.4% ($p < 0.01$) below the vision-only "capability probe". This gap persists under symmetric prompting, suggesting a form of functional perception suppression in which autonomous visual search may be under-utilized in multimodal settings without linguistic triggers. Causal interventions via spatial prompting and signal magnification provide evidence that internal reasoning remains functional, supporting the interpretation of a perceptual access bottleneck. Our findings suggest MLLMs function as "language-guided passive observers", advocating for perceptually-independent architectures that decouple sensory agency from linguistic dominance. Code and dataset are available at https://github.com/Ink-Dawn/MagicBench

[Long] MathSight: A Benchmark Exploring Have Vision-Language Models Really Seen in University-Level Mathematical Reasoning? ▶ *Underline Video*

Yuandong Wang, Yao Cui, Yuxin Zhao, Zhen Yang, Yangfu Zhu and Zhenzhou Shao

Recent advances in Vision-Language Models (VLMs) have achieved impressive progress in multimodal mathematical reasoning. Yet, how much visual information truly contributes to reasoning remains unclear. Existing benchmarks report strong overall performance but seldom isolate the role of the image modality, leaving open whether VLMs genuinely leverage visual understanding or merely depend on linguistic priors. To address this, we present MathSight, a university-level multimodal mathematical reasoning benchmark designed to disentangle and quantify the effect of visual input. Each problem includes multiple visual variants, original, hand-drawn, photocaptured, and a text-only condition for controlled comparison. Experiments on state-of-the-art VLMs reveal a consistent trend: the contribution of visual information diminishes with increasing problem difficulty. Remarkably, Qwen3-VL without any image input surpasses both its multimodal variants and GPT-5, underscoring the need for benchmarks like MathSight to advance genuine vision-grounded reasoning in future models. The project page is available at https://cnu-bot-group.github.io/MathSight/.

[Long] MDP-GRPO: Stabilized Group Relative Policy Optimization for Multi-Constraint Instruction Following ▶ *Underline Video*

Mohammad Mahdi Salmani-Zarchi, Zahra Rahimi, Heshaam Faili and Mohammad Javad Dousti

Reinforcement learning with verifiable rewards is ideal for multi-constraint instruction following, yet standard group-relative policy optimization (GRPO) becomes unstable under discrete, low-dispersion rewards, where within-group reward distributions are frequently homogeneous. We identify and formalize three pathologies of z-score group normalization in this regime: low-variance amplification, mean-centering blindness, and zero-variance collapse. To address them, we propose MDP-GRPO, which stabilizes learning through (1) multi-temperature sampling to increase reward dispersion, (2) dual-anchor advantages to restore gradients in homogeneous groups, (3) prospect-theoretic shaping to bound updates and penalize violations based on Kahneman & Tversky's theory, and (4) asymmetric KL regularization. Evaluated on FollowBench, IFEval, and a curated multi-constraint dataset, MDP-GRPO outperforms standard GRPO, improving strict constraint satisfaction by up to 5.0% on Llama-3.2-3B. Our method also enables stable convergence with small group sizes while preserving general capabilities on MMLU and ARC.

[Long] Mem^2Evolve: Towards Self-Evolving Agents via Co-Evolutionary Capability Expansion and Experience Distillation ▶ *Underline Video*

Zihao Cheng, Zeming Liu, YingyuShan YingyuShan, Xinyi Wang, Xiangrong Zhu, Yunpu Ma, Hongru Wang, Yuhang Guo and Wei Lin

While large language model–powered agents can self-evolve by accumulating experience or by dynamically creating new assets (i.e., tools or expert agents), existing frameworks typically treat these two evolutionary processes in isolation. This separation overlooks their intrinsic interdependence: the former is inherently bounded by a manually predefined static toolset, while the latter generates new assets from scratch without experiential guidance, leading to limited capability growth and unstable evolution. To address this limitation, we introduce a novel paradigm of co-evolutionary Capability Expansion and Experience Distillation. Guided by this paradigm, we propose the Mem^2Evolve, which integrates two core components: Experience Memory and Asset Memory. Specifically, Mem^2Evolve leverages accumulated experience to guide the dynamic creation of assets, thereby expanding the agent's capability space while simultaneously acquiring new

experience to achieve co-evolution. Extensive experiments across 6 task categories and 8 benchmarks demonstrate that Mem^2Evolve achieves improvement of 18.53% over standard LLMs, 11.80% over agents evolving solely through experience, and 6.46% over those evolving solely through asset creation, establishing it as a substantially more effective and stable self-evolving agent framework.

[Long] Merlin's Whisper: Enabling Efficient Reasoning in Large Language Models via Black-box Persuasive Prompting ► *Underline Video*

Heming Xia, Cunxiao Du, Rui Li, Chak Tou Leong, Yongqi Li and Wenjie Li

Large reasoning models (LRMs) have demonstrated remarkable proficiency in tackling complex tasks through step-by-step thinking. However, this lengthy reasoning process incurs substantial computational and latency overheads, hindering the practical deployment of LRMs. This work presents a new approach to mitigating overthinking in LRMs via black-box persuasive prompting. By treating LRMs as black-box communicators, we investigate how to persuade them to generate concise responses without compromising accuracy. We introduce Whisper, an iterative refinement framework that generates high-quality persuasive prompts from diverse perspectives. Experiments across multiple benchmarks demonstrate that Whisper consistently reduces token usage while preserving performance. Notably, Whisper achieves a 3x reduction in average response length on simple GSM8K questions for the Qwen3 series and delivers an average 40% token reduction overall. For closed-source APIs, Whisper reduces token usage on MATH-500 by 46% for Claude-3.7 and 50% for Gemini-2.5. Further analysis reveals the broad applicability of Whisper across data domains, model scales, and families, underscoring the potential of black-box persuasive prompting as a practical strategy for enhancing LRM efficiency.

[Long] Mind Reader: Latent User Demand-Guided Content Optimization for Generative Search Engine ► *Underline Video*

Tong Chen, Jiawei Guo, Yuxi Li, Baiming Chen, Houxing Ren, Zhiwei Zhang, Yunxiang Zhang, Hanyang Xia, Kun Liang and Zhaoran Fan

Generative Search Engines (GSEs) have reshaped information retrieval, and Generative Engine Optimization (GEO) emerges to improve the content visibility in GSEs' responses. Previous methods mainly rely on empirical strategies or query-dependent preferences of GSEs for content optimization. However, they remain limited in effectiveness as they overlook the latent user search demands in queries that drive content retrieval and response generation of GSEs. To address this, we propose Mind Reader, a novel GEO method to effectively improve the content visibility within the generated responses of GSEs through content optimization guided by the extracted latent demands of user search. Specifically, we propose a decomposition-recombination query augmentation module, which enriches the query with latent semantic information by decomposing it into diverse perspectives, capturing underlying semantic information, and recombining them into variants to support subsequent optimization. Then, we propose a reasoning coverage content optimization module. By optimizing content to cover critical reasoning information of GSEs, we align the content with the user search demands, effectively improving the content visibility. Extensive experiments on widely used GEO-Bench and our proposed PC-GEO show that our method significantly outperforms baselines and effectively improves content visibility (with up to 2.44x objective metrics and 1.23x subjective metrics on average).

[Long] MirrorCAPTCHA: Wild CAPTCHA, Wild Distribution, Wild Web-based Platform Meet Multimodal LLM Agents ► *Underline Video*

Xiangyu Wu, Yuwei Hu, Tianyu Cui, Yueying Tian, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, Yang Yang and Jianfeng Lu

The path to fully autonomous web agents is currently hindered by a critical bottleneck: their limited ability to handle CAPTCHA. Existing agent benchmarks largely ignore this practical challenge, failing to evaluate an agent's real-world capacity to solve CAPTCHA. To bridge this gap, we conduct a comprehensive analysis of real-world CAPTCHA distributions and introduce MirrorCAPTCHA, a benchmark annotated with Weighted Pass Rate and a newly proposed metric Completion Degree. MirrorCAPTCHA is designed to serve as a "mirror" that faithfully reflects the automation capabilities of agents in real scenarios. We filter 2095 websites from Common Crawl, identify the CAPTCHA deployed on these sites, and cluster them into 18 distinct categories using K-means algorithm. To ensure practicality, we extract a web subgraph from Common Crawl covering these websites and use random walks to simulate real-world CAPTCHA encounter frequencies, yielding a realistic measure of agents ability. Additionally, we develop a lightweight synthetic data pipeline to train Ovis2-Agent-CAPTCHA-8B, which significantly outperforms current state-of-the-art closed-source models on MirrorCAPTCHA, achieving a 9.4% higher average Weighted Pass Rate and a 2.13% higher average Completion Degree than the runner-up, Gemini-2.5-Pro.

[Long] MMCLIP: Cross-Modal Attention Masked Modelling for Medical Language-Image Pre-Training ► *Underline Video*

Biao Wu, Yutong Xie, Zeyu Zhang, Vu Minh Hieu Phan, Qi Chen, Ling Chen and Qi Wu

Vision-and-language pretraining (VLP) in medicine leverages contrastive learning on image-text pairs, often enhanced with masked modeling. However, existing methods face two challenges: difficulty reconstructing key pathological features due to limited data, and reliance on either paired or image-only datasets without combining both. To address this, we propose MMCLIP (Masked Medical Contrastive Language-Image Pre-training), which introduces two modules: AttMIM: Masks image features highly correlated with text to improve reconstruction of fine medical details. EntMLM: Masks key medical entities in text and reconstructs them using visual cues. Furthermore, MMCLIP incorporates unpaired data through disease-kind prompts, achieving state-of-the-art performance in zero-shot and fine-tuning across five benchmarks.

[Long] MMErroR: A Benchmark for Erroneous Reasoning in Vision-Language Models ► *Underline Video*

Yang Shi, Yifeng Xie, Minzhe Guo, Liangsi Lu, Mingxuan Huang, Jingchao Wang, Zhihong Zhu, Boyan Xu and Zhiqi Huang

Recent advances in Vision-Language Models (VLMs) have improved performance in multi-modal learning, raising the question of whether

these models truly understand the content they process. Crucially, can VLMs detect when a reasoning process is wrong and identify its error type? To answer this, we present MMErroR, a multi-modal benchmark of 1997 samples, each embedding a single coherent reasoning error. These samples span 24 subdomains across six top-level domains, ensuring broad coverage and taxonomic richness. Unlike existing benchmarks that focus on answer correctness, MMErroR targets a process-level, error-centric evaluation that requires models to detect incorrect reasoning and classify the error type within both visual and linguistic contexts. We evaluate 12 representative VLMs, and even the best model, Gemini-3-Pro-Preview, classifies the error correctly in only 66.65% of cases, underscoring the challenge of identifying erroneous reasoning. Furthermore, the ability to accurately identify errors offers valuable insights into the capabilities of multi-modal models. Project Page: https://mmerror-benchmark.github.io

[Long] MOA: Multi-Objective Alignment for Role-Playing Agents ▶ *Underline Video*

Chonghua Liao, Ke Wang, Yuchuan Wu, Ruoran Li, Fei Huang and Yongbin Li

Role-playing agents (RPAs) require balancing multiple objectives, such as instruction following, persona consistency, and stylistic fidelity, which are not always perfectly aligned across different dimensions. While prior work has primarily relied on supervised fine-tuning or reinforcement learning with scalarized rewards, these approaches do not explicitly address the coordination of multiple reward dimensions during optimization. We present MOA (Multi-Objective Alignment), a reinforcement-learning framework that enables multi-dimensional, fine-grained rubric optimization for general RPAs. MOA introduces a novel multi-objective optimization strategy that trains simultaneously on multiple fine-grained rubrics to boost optimization performance. Besides, to address the issues of model output diversity and quality, we have also employed thought-augmented rollout with off-policy guidance. Experiments on PersonaGym and RoleMRC show that MOA consistently improves multi-dimensional role-playing performance over supervised and standard RL baselines. Under identical evaluation protocols, an 8B model trained with MOA reaches performance competitive with strong closed-source models across multiple evaluation dimensions. These results suggest that MOA provides a practical framework for training more capable general-purpose role-playing agents.

[Long] Modeling LLM Unlearning as an Asymmetric Two-Task Learning Problem ▶ *Underline Video*

Zeguan Xiao, Siqing Li, Yong Wang, Xuetao Wei, Jian Yang, Yun Chen and Guanhua Chen

Machine unlearning for large language models (LLMs) aims to remove targeted knowledge while preserving general capability. In this paper, we recast LLM unlearning as an asymmetric two-task problem: retention is the primary objective and forgetting is an auxiliary. From this perspective, we propose a retention-prioritized gradient synthesis framework that decouples task-specific gradient extraction from conflict-aware combination. Instantiating the framework, we adapt established PCGrad to resolve gradient conflicts, and introduce SAGO, a novel retention-prioritized gradient synthesis method. Theoretically, both variants ensure non-negative cosine similarity with the retain gradient, while SAGO achieves strictly tighter alignment through constructive sign-constrained synthesis. Empirically, on WMDP Bio/Cyber and RWKU benchmarks, SAGO consistently pushes the Pareto frontier: e.g., on WMDP Bio (SimNPO+GD), recovery of target model MMLU performance progresses from 44.6% (naive) to 94.0% (+PCGrad) and further to 96.0% (+SAGO), while maintaining comparable forgetting strength. Our results show that re-shaping gradient geometry, rather than re-balancing losses, is the key to mitigating unlearning-retention trade-offs.

[Long] MPR-GUI: Benchmarking and Enhancing Multilingual Perception and Reasoning in GUI Agents ▶ *Underline Video*

Ruihan Chen, Qiming Li, Xiaocheng Feng, Weihong Zhong, Xiaoliang Yang, Yuxuan Gu, Zekun Zhou, Yunfei Lu, Haoyu Ren, Kun Chen, Dandan Tu and Bing Qin

Large Vision-Language Models (LVLMs) have shown strong potential as multilingual Graphical User Interface (GUI) agents, as evidenced by existing GUI benchmarks. However, these benchmarks exhibit two primary limitations: (1) although Perception and Reasoning (P&R) capabilities are fundamental for GUI agents, current benchmarks lack fine-grained diagnostics to identify which specific capabilities lead to task failures, hindering targeted improvements; (2) existing benchmarks fail to provide a strictly aligned cross-lingual evaluation environment, introducing confounding factors that prevent isolating the language impact on GUI agent performance. To address these issues, we propose the Multilingual P&R GUI Benchmark (MPR-GUI-Bench), featuring strictly aligned environments across six languages and eight fine-grained P&R tasks. Our benchmark reveals consistent P&R gaps between English and non-English settings, particularly on reasoning-intensive tasks. To leverage the superior English P&R capabilities for bridging cross-lingual gaps, we identify layers sensitive to language and propose GUI-XLI, a GUI Cross-Lingual Intervention method that aligns non-English hidden states with their English counterparts at these layers during inference. Experiments show that GUI-XLI effectively reduces the cross-lingual gaps, with an average gain of 6.5% in non-English settings.

[Long] MTR-Suite: A Framework for Evaluating and Synthesizing Conversational Retrieval Benchmarks ▶ *Underline Video*

Junhao Ruan, Abudukeyumu Abudula, Bei Li, Yongjing Yin, Xinyu Liu, Kechen Jiao, Xin Chen, Jingang Wang, Xunliang Cai, Tong Xiao and JingBo Zhu

Accurate evaluation of conversational retrieval is pivotal for advancing Retrieval-Augmented Generation (RAG) systems. However, existing conversational retrieval benchmarks suffer from costly, sparse human annotation or rigid, unnatural automated heuristics. To address these challenges, we introduce MTR-Suite, a unified framework for auditing, synthesizing, and benchmarking retrieval. It features: (1) MTR-Eval, an LLM-based auditor quantifying alignment gaps in previous benchmarks; (2) MTR-Pipeline, a multi-agent system using greedy traversal clustering to generate high-fidelity dialogues at 1/400th human cost; and (3) MTR-Bench, a rigorous general-domain benchmark. MTR-Bench mimics production-style challenges (hard topic switching, verbosity), offering superior discriminative power. We make our

code and data publicly available to facilitate future research.

[Findings] Multilingual Refusal Alignment for Safer Large Language Models ▶ *Underline Video*
Aleksandra Krasnodbska, Wojciech Kusa and Aldo Lipani
As Large Language Models (LLMs) are deployed globally, ensuring their safety and alignment across multiple languages becomes paramount. However, safety behaviors often vary unpredictably between languages, posing significant challenges for consistent and ethical AI. In this work, we systematically investigate the dynamics of multilingual alignment, exploring whether single-language alignment transfers cross-lingually, how language consistency is preserved during training, and the resulting trade-offs with general knowledge capabilities. We introduce RefusEU a novel refusal alignment dataset covering 12 European languages, including a dedicated test set for evaluating current state-of-the-art models. Our controlled Direct Preference Optimization (DPO) experiments provide two key insights: aligning models exclusively in English is insufficient to ensure cross-lingual safety, even for the same harm categories, whereas training on multilingual datasets can improve safety without degrading general performance, as measured by the Global MMLU benchmark.

[Long] NaturalSloth: Revisiting Denial-of-Service Attacks on Large Language Models ▶ *Underline Video*
Yiming Chen, Zexin Li, Xianghu Yue, Robby T. Tan and Haizhou Li
LLM serving is limited by provider-side resources: longer generations consume more GPU time, increase latency, and reduce throughput in multi-tenant systems. This creates a denial-of-service (DoS) risk, where attackers degrade service by inducing excessive generation. Prior work on LLM DoS primarily relies on adversarial perturbations that delay end-of-sequence termination. We show perturbations are often unnecessary: natural, benign-looking instructions that specify impractical and meaningless tasks can already trigger excessive generation. To study this overlooked vulnerability, we introduce, an adversarial dataset of natural, instruction-based DoS prompts. Starting from a human-curated seed set spanning diverse attack categories, we design a multi-agent synthesis framework to scale the dataset while preserving malicious intent and increasing semantic diversity. Experiments across a wide range of proprietary and open-source LLMs show that NaturalSloth consistently induces excessive generation, with attack effectiveness further amplified when combined with jailbreak techniques. Our analysis also reveals significant limitations of existing defenses, highlighting the need for dedicated protections against natural DoS attacks.

[Long] NOSE: Neural Olfactory-Semantic Embedding with Tri-Modal Orthogonal Contrastive Learning ▶ *Underline Video*
YanyiSu YanyiSu, Hongshuai Wang, Zhifeng Gao and Jun Cheng
Olfaction lies at the intersection of chemical structure, neural encoding, and linguistic perception, yet existing representation methods fail to fully capture this pathway. Current approaches typically model only isolated segments of the olfactory pathway, overlooking the complete chain from molecule to receptors to linguistic descriptions. Such fragmentation yields learned embeddings that lack both biological grounding and semantic interpretability. We propose NOSE (Neural Olfactory-Semantic Embedding), a representation learning framework that aligns three modalities along the olfactory pathway: molecular structure, receptor sequence, and natural language description. Rather than simply fusing these signals, we decouple their contributions via orthogonal constraints, preserving the unique encoded information of each modality. To address the sparsity of olfactory language, we introduce a weak positive sample strategy to calibrate semantic similarity, preventing erroneous repulsion of similar odors in the feature space. Extensive experiments demonstrate that NOSE achieves state-of-the-art (SOTA) performance and excellent zero-shot generalization, confirming the strong alignment between its representation space and human olfactory intuition. Code and data are available at https://github.com/Xianyusyy/NOSE

[Long] OctoBench: Benchmarking Scaffold-Aware Instruction Following in Repository-Grounded Agentic Coding ▶ *Underline Video*
Deming Ding, Shichun Liu, Enhui Yang, Jiahang Lin, Ziying Chen, Shihan Dou, Honglin Guo, Weiyu Cheng, Pengyu Zhao, Chengjun Xiao, Qunhong Zeng, Qi Zhang, Xuanjing Huang, Qidi Xu and Tao Gui
Modern coding scaffolds turn LLMs into capable software agents, but their ability to follow scaffold-specified instructions remains under-examined, especially when constraints are heterogeneous and persist across interactions. To fill this gap, we introduce OctoBench, which benchmarks scaffold-aware instruction following in repository-grounded agentic coding. OctoBench includes 34 environments and 217 tasks instantiated under three scaffold types, and is paired with 7,098 objective checklist items. To disentangle solving the task from following the rules, we provide an automated observation-and-scoring toolkit that captures full trajectories and performs fine-grained checks. Experiments on eight representative models reveal a systematic gap between task-solving and scaffold-aware compliance, underscoring the need for training and evaluation that explicitly targets heterogeneous instruction following. We will release the benchmark to support reproducible benchmarking and to accelerate the development of more scaffold-aware coding agents.

[Long] On the Role of Discriminative Models in Generative Relation Extraction ▶ *Underline Video*
Guozheng Li, Peng Wang, Zijie Xu, Jing Zhou, Jiajun Liu and Ziyu Shang
Relation extraction (RE) identifies semantic relations between entities in text, with existing methods falling into two main paradigms: discriminative and generative. Discriminative models encode sentences and entities into relation representations and classify the most likely relation, whereas generative models directly produce relation labels through sequence generation. Although the latter have benefited from recent advances in large language models (LLMs), their performance remains limited by bottlenecks. In this work, we present the systematic investigation of how discriminative models can support generative RE. We propose the Discriminative-to-Generative (D2G) framework, which first leverages discriminative models to produce a top-k set of candidate relations, and then integrates this knowledge into generative

models via in-context or prompt learning. Extensive experiments on five widely used RE benchmarks demonstrate that D2G consistently achieves state-of-the-art performance, with notable gains on long-tailed relation classes.

[Long] One Pair Suffices: Unlocking Universal Zero-Shot Translation via Cross-Architecture Alignment ▶ *Underline Video*

Hao Zong, CongHuYuan CongHuYuan, Chao Bei, Wentao Chen, Huan Liu, Kaiyu Huang and Degen Huang

Current paradigms for empowering Large Language Models (LLMs) with multilingual capabilities rely heavily on massive instruction tuning. We challenge this view, proposing that the barrier is topological alignment, not data quantity. We introduce Hybrid Cross-Alignment (HCA), fusing a frozen NLLB encoder with a Qwen decoder via a closed-loop dual-adapter architecture. HCA utilizes a Source-Side Adapter to precondition encoder features and a Query-Residual Adapter to preserve generative stability, bridged by an adaptive gated cross-modal interface. Our core discovery is Universal Alignment Generalization." We demonstrate that training HCA on a single language pair (German-English) unlocks state-of-the-art zero-shot transfer to dozens of unseen languages. Crucially, our Oracle" experiments reveal that this single-pair training recovers over 96.7% of the performance achievable by training on all available pairs. This proves that a universal, language-agnostic projection protocol exists. With a total inference footprint of 5.25B parameters, our model significantly outperforms larger baselines, surpassing TowerPlus-9B (+9.0 COMET on low-resource languages) and Aya-101 (13B). Furthermore, performance scales linearly with encoder size; upgrading from 600M to 1.3B yields immediate gains (+3.4 points on Gujarati) with minimal retraining cost.

[Findings] PAC-BENCH: Evaluating Multi-Agent Collaboration under Privacy Constraints ▶ *Underline Video*

Minjun Park, Donghyun Kim, Hyeonjong Ju, Seungwon Lim, Dongwook Choi, Taeyoon Kwon, Minju Kim and Jinyoung Yeo

We are entering an era in which individuals and organizations increasingly deploy dedicated AI agents that interact and collaborate with other agents. However, the dynamics of multi-agent collaboration under privacy constraints remain poorly understood. In this work, we present PAC-Bench, a benchmark for systematic evaluation of multi-agent collaboration under privacy constraints. Experiments on PAC-Bench show that privacy constraints substantially degrade collaboration performance and make outcomes depend more on the initiating agent than the partner. Further analysis reveals that this degradation is driven by recurring coordination breakdowns, including early-stage privacy violations, overly conservative abstraction, and privacy-induced hallucinations. Together, our findings identify privacy-aware multi-agent collaboration as a distinct and unresolved challenge that requires new coordination mechanisms beyond existing agent capabilities.

[Long] PAM: Enhancing General Alignment of Large Reasoning Models through Priority-Aware Metacognition ▶ *Underline Video*

Zhihao Xu, Fuzhen Yang, Liang Lin and Xiting Wang

Recent advancements in Large Reasoning Models (LRMs) have showcased strong performance across various reasoning tasks by leveraging System2 thinking capabilities. However, existing studies indicate that this reasoning ability alone does not reliably transfer to the general alignment domain. Inspired by cognitive science and how humans solve tasks, we argue that LRMs must be equipped with metacognitive knowledge to fully utilize their System2 capabilities. In this paper, we propose PriorityAware Metacognition (PAM), which guides the model to first identify the toplevel human preference (e.g., harmlessness) as a means of understanding the alignment tasks nature, and then apply other kinds of metacognitive knowledge to better monitor and regulate the model's thinking process. We implement PAM via a two-stage pipeline: a cold-start phase that collects structured metacognitive knowledge based on Flavells theoretical framework, and a preference-optimization phase that further reinforces such metacognition. Extensive experiments validate the effectiveness of PAM. Under the same training pipelines, PAM consistently yields higher performance, improving general domain alignment performance by 10 points on the helpfulness and harmless benchmarks. Code is available at https://anonymous.4open.science/r/PAM-RM-02DF.

[Long] Paper2Rebuttal: A Multi-Agent Framework for Transparent Author Response Assistance ▶ *Underline Video*

Qianli Ma, Chang Guo, Zhiheng Tian, SiyuWang SiyuWang, Jipeng Xiao, Yuanhao Yue and Zhipeng Zhang

Writing effective rebuttals is a high-stakes task that demands more than linguistic fluency, as it requires precise alignment between reviewer intent and manuscript details. Current solutions typically treat this as a direct-to-text generation problem, suffering from hallucination, overlooked critiques, and a lack of verifiable grounding. To address these limitations, we introduce RebuttalAgent, the first multi-agents framework that reframes rebuttal generation as an evidence-centric planning task. Our system decomposes complex feedback into atomic concerns and dynamically constructs hybrid contexts by synthesizing compressed summaries with high-fidelity text while integrating an autonomous and on-demand external search module to resolve concerns requiring outside literature. By generating an inspectable response plan before drafting, RebuttalAgent ensures that every argument is explicitly anchored in internal or external evidence. We validate our approach on the proposed RebuttalBench and demonstrate that our pipeline outperforms strong baselines in coverage, faithfulness, and strategic coherence, offering a transparent and controllable assistant for the peer review process. Code will be released.

[Long] PaperRegister: Boosting Flexible-grained Paper Search via Hierarchical Register Indexing ▶ *Underline Video*

Zhuoqun Li, Xuanang Chen, Hongyu Lin, Yaojie Lu, Xianpei Han, Shanshan Jiang, Bin Dong and Le Sun

As researchers delve more deeply into their work, paper search requirements may become more flexible, sometimes involving specific details such as module configuration rather than being limited to coarse-grained topics. However, previous paper search systems are unable to meet these flexible-grained requirements, as previous systems mainly collect paper abstract to construct corpus index, which lacks detailed information to support retrieval by some finer-grained queries. In this work, we propose PaperRegister, which transforms traditional abstract-based index into a hierarchical index tree, thereby supporting queries at flexible granularity. Experiments on paper search tasks across a range of granularity demonstrate that PaperRegister achieves the SOTA performance, and particularly excels in the fine-grained

scenarios, highlighting good potential as an effective solution for flexible-grained paper search in real-world applications.

[Long] PDTrim: Targeted Pruning for Prefill-Decode Disaggregation in Inference ► *Underline Video*

Hao Zhang, Mengsi Lyu, Zhuo Chen, Yulong Ao and Yonghua Lin

Large Language Models (LLMs) demonstrate exceptional capabilities across various tasks, but their deployment is constrained by high computational and memory costs. Model pruning provides an effective means to alleviate these demands. However, existing methods often ignore the characteristics of prefill-decode (PD) disaggregation in practice. In this paper, we propose a pruning method that is highly integrated with PD disaggregation, enabling more precise pruning of blocks. Our approach constructs pruning and distillation sets to perform iterative block removal, obtaining better pruning solutions. Moreover, we analyze the pruning sensitivity of the prefill and decode stages and identify removable blocks specific to each stage, making it well suited for PD disaggregation deployment. Extensive experiments demonstrate our approach consistently achieves strong performance in both PD disaggregation and PD unified (non-PD disaggregation) settings, and can also be extended to other non-block pruning methods. Under the same settings, our method achieves improved performance and faster inference.

[SRW] PE-QAT: Parameter-Efficient Quantization Aware Training for Large Language Models ► *Underline Video*

Shresth Mishra

As large language models (LLMs) grow, their compute and memory demands become prohibitive for on-device deployment. Quantization is a crucial technique to shrink model footprint and accelerate inference, but aggressively low-bit weight-activation quantization schemes often sacrifice accuracy. Quantization Aware Training (QAT) is a commonly used paradigm to minimize quantization noise, but is extremely expensive to train and often unscalable to large models. We introduce PE-QAT, a parameter-efficient framework targeting per-channel 4-bit weight-activation quantization of LLMs, which aims to preserve model accuracy while significantly reducing resource requirements. The proposed method freezes the base model and trains lightweight LoRA adapters by fake quantizing the merged-weight model, enabling PE-QAT to scale efficiently unlike full QAT. We apply fake quantization with Straight-Through Estimators (STE) to the merged weights, allowing the adapters to explicitly compensate for quantization noise during training. One of the biggest challenges with quantizing activations alongside weights is addressing outliers that are orders of magnitude larger than other activations, which inflate quantization scales and suppress lower-magnitude values. To mitigate the impact of severe activation outliers, PE-QAT jointly learns per-channel smoothing factors and symmetric activation clipping thresholds. PE-QAT retains accuracy within 0.11 percentage points of the full-precision baseline on Llama-2-7B zero-shot tasks while training only 1.26% of total parameters.

[SRW] Peek2: Regex-free Byte-level Byte-Pair Encoding Pretokenizer for LLM Inference on Edge Devices ► *Underline Video*

Liu Zai; Iraklis Klampanos;

Pretokenization is a crucial, sequential pass in Byte-level BPE tokenizers, yet little work has been done to optimize it for edge-side inference. Our proposed new implementation, Peek2, serves as a drop-in replacement for cl100k-like pretokenizers used in GPT-3, LLaMa-3, and Qwen-2.5. After breaking down and analyzing the logic of the original cl100k pretokenizer, we introduced a new pretokenization algorithm with linear time complexity and constant, trivial memory usage, suited for edge scenarios. Test results show that it increases microbenchmarking throughput by up to 2.48x and delivers a 1.14x improvement in overall throughput across the entire Byte-level BPE en coding process, depending on the dataset, while providing identical results as the baseline Regex-based tokenizer.

[Long] Perplexity-Aware Data Scaling Law: Perplexity Landscapes Predict Performance for Continual Pre-training ► *Underline Video*

Lei Liu, Hao Zhu, xiaoyan yang, Yue Shen, Zhixuan Chu, Jian Wang, Jinjie Gu and Kui Ren

Continual Pre-training (CPT) serves as a fundamental approach for adapting foundation models to domain-specific applications. Scaling laws for pre-training define a power-law relationship between dataset size and the test loss of an LLM. However, the marginal gains from simply increasing data for CPT diminish rapidly, yielding suboptimal data utilization and inefficient training. To address this challenge, we propose a novel perplexity-aware data scaling law to establish a predictive relationship between the perplexity landscape of domain-specific data and the test loss. Our approach leverages the pre-trained models own perplexity on domain data as a proxy for estimating the knowledge gap, effectively quantifying the informational perplexity landscape of candidate training samples. By fitting this scaling law across diverse perplexity regimes, we enable adaptive selection of high-utility data subsets, prioritizing content that maximizes knowledge absorption while minimizing redundancy and noise. Extensive experiments on both medical and general-domain benchmarks demonstrate that our method consistently identifies near-optimal training subsets, achieving superior performance with significantly reduced data consumption.

[Findings] PiCSAR: Probabilistic Confidence Selection and Ranking for Reasoning Chains ► *Underline Video*

Joshua Ong Jun Leang, Zheng Zhao, Aryo Pradipta Gema, Sohee Yang, Wai-Chung Kwan, Xuanli He, Wenda Li, Pasquale Minervini, Eleonora Giunchiglia and Shay B Cohen

Best-of-n sampling improves the accuracy of large language models (LLMs) and large reasoning models (LRMs) by generating multiple candidate solutions and selecting the one with the highest reward. The key challenge for reasoning tasks is designing a scoring function that can identify correct reasoning chains without access to ground-truth answers. We propose Probabilistic Confidence Selection and Ranking for Reasoning Chains (PiCSAR): a simple, training-free method that scores each candidate generation using the joint log-likelihood of the reasoning and final answer. This method utilises both the scores of the reasoning path (reasoning confidence) and the final answer (answer

confidence). PiCSAR achieves substantial gains across several benchmarks (+11.7 on AIME2024, +9.81 on AIME2025), outperforming baselines with at least 2x fewer samples in 20 out of 25 comparisons. Our analysis reveals that correct reasoning chains exhibit higher reasoning and answer confidence, justifying the effectiveness of PiCSAR.

[Long] PII-Bench: Evaluating Query-Aware Privacy Protection Systems ▶ *Underline Video*

Hao Shen, Zhouhong Gu, Haokai Hong, Weili Han and Hongfeng Chai

The widespread adoption of Large Language Models (LLMs) has raised significant privacy concerns regarding the exposure of personally identifiable information (PII) in user prompts. To address this challenge, we propose a query-unrelated PII masking strategy and introduce PII-Bench, the first comprehensive evaluation framework for assessing privacy protection systems. PII-Bench comprises 2,842 test samples across 7 PII types with 55 fine-grained subcategories, featuring diverse scenarios from single-subject descriptions to complex multi-party interactions. Each sample is carefully crafted with a user query, context description, and standard answer indicating query-relevant PII. Our empirical evaluation reveals that while current models perform adequately in basic PII detection, they show significant limitations in determining PII query relevance. Even advanced LLMs struggle with this task, particularly in handling complex multi-subject scenarios, indicating substantial room for improvement in achieving intelligent PII masking.

[Findings] PivotAttack: Rethinking the Search Trajectory in Hard-Label Text Attacks via Pivot Words ▶ *Underline Video*

Yuzhi Liang, Shiliang Xiao, Jingsong Wei, Qiliang Lin and Xia Li

Existing hard-label text attacks often rely on inefficient "outside-in" strategies that traverse vast search spaces. We propose PivotAttack, a query-efficient "inside-out" framework. It employs a Multi-Armed Bandit algorithm to identify Pivot Sets, combinatorial token groups acting as prediction anchors, and strategically perturbs them to induce label flips. This approach captures inter-word dependencies and minimizes query costs. Extensive experiments across traditional models and Large Language Models demonstrate that PivotAttack consistently outperforms state-of-the-art baselines in both Attack Success Rate and query efficiency.

[Long] Placing Puzzle Pieces Where They Matter: A Question Augmentation Framework for Reinforcement Learning ▶ *Underline Video*

Yangyi Fang and Haolin Shi

Reinforcement learning has become a powerful approach for enhancing large language model reasoning, but faces a fundamental dilemma: training on easy problems can cause overfitting and pass@k degradation, while training on hard problems often results in sparse rewards. Recent question augmentation methods address this by prepending partial solutions as hints. However, uniform hint provision may introduce redundant information while missing critical reasoning bottlenecks, and excessive hints can reduce reasoning diversity, causing pass@k degradation. We propose PieceHint, a hint injection framework that strategically identifies and provides critical reasoning steps during training. By scoring the importance of different reasoning steps, selectively allocating hints based on problem difficulty, and progressively withdrawing scaffolding, PieceHint enables models to transition from guided learning to independent reasoning. Experiments on six mathematical reasoning benchmarks show that our 1.5B model achieves comparable average performance to 32B baselines while preserving pass@k diversity across all k values.

[Long] Probing Audio-Visual Reasoning in Multimodal Language Models through the Lens of Audio ▶ *Underline Video*

Kaixiong Gong, Kaituo Feng, Bohao Li, Yibing Wang, Mofan Cheng, Shijia Yang, Jiaming Han, Benyou Wang, Yutong Bai, Zhuoran Yang and Xiangyu Yue

Recent multimodal large language models (MLLMs), such as GPT-4o, Gemini 1.5/2.5 Pro, and Reka Core, have advanced audio-visual reasoning capabilities, achieving strong performance in tasks like cross-modal understanding and generation. However, our DeafTest uncovers unanticipated failures: most of the state-of-the-art MLLMs struggle with very simple audio tasks, such as distinguishing louder sounds or sound counting. This raises a fundamental question, does a deficiency in low-level audio perception constrain higher-level audio-visual reasoning? To address this, we introduce AV-Odyssey Bench, a comprehensive benchmark of 4,555 meticulously designed problems that integrate text, audio, and visual modalities. Each task requires models to unify cross-modal reasoning, leveraging synchronized audio-visual cues to infer solutions. By structuring questions as multiple-choice, we ensure objective, reproducible evaluations without reliance on subjective human or LLM-based judgments. Through comprehensive benchmarking of closed-source and open-source models, we showcase: (i) current MLLMs lack robust audio-visual integration ability and (ii) performance on DeafTest (Pearsons r = 0.945) strongly correlates with AV-Odyssey accuracy. These findings challenge assumptions about models' multimodal proficiency and highlight fundamental audio perception as a reasoning bottleneck. We believe that our results provide concrete guidance for future dataset design, alignment strategies, and architectures.

[SRW] Probing Functional Correctness in Diffusion Language Models ▶ *Underline Video*

Guan-Ming Chiu and Jeng-Yue Liu

Diffusion language models generate text by iteratively denoising all tokens in parallel, but when and where their hidden states encode whether the output will be functionally correct remains unknown. We present the first probing study of DLM internals, training linear classifiers on hidden states extracted at intermediate denoising steps to predict functional correctness. Across two models (LLaDA-8B, Dream-7B) and four tasks (JSON schema validation, math reasoning, code generation, science QA), we find that (1) correctness signal is already present at step 0, before any denoising occurs, (2) signal emergence is task-dependent, with structural tasks showing flat profiles and reasoning tasks

showing gradual buildup, and (3) layer dynamics diverge across architectures. We further demonstrate that per-step probe confidence can identify instances likely to fail, enabling selective generation that avoids 36-98% of wasted compute depending on the task.

[Long] RAM-SD: Retrieval-Augmented Multi-agent framework for Sarcasm Detection ► *Underline Video*

Ziyang Zhou, Ziqi Liu, Yan Wang, Yiming Lin and Yangbin Chen

Sarcasm detection remains a significant challenge due to its reliance on nuanced contextual understanding, world knowledge, and multi-faceted linguistic cues that vary substantially across different sarcastic expressions. Existing approaches, from fine-tuned transformers to large language models, apply a uniform reasoning strategy to all inputs, struggling to address the diverse analytical demands of sarcasm. These demands range from modeling contextual expectation violations to requiring external knowledge grounding or recognizing specific rhetorical patterns. To address this limitation, we introduce RAM-SD, a Retrieval-Augmented Multi-Agent framework for Sarcasm Detection. The framework operates through four stages: (1) contextual retrieval grounds the query in both sarcastic and non-sarcastic exemplars; (2) a meta-planner classifies the sarcasm type and selects an optimal reasoning plan from a predefined set; (3) an ensemble of specialized agents performs complementary, multi-view analysis; and (4) an integrator synthesizes these analyses into a final, interpretable judgment with a natural language explanation. Evaluated on four standard benchmarks, RAM-SD achieves a state-of-the-art Macro-F1 of 77.74%, outperforming the strong GPT-4o+CoC baseline by 7.01 points. Our framework not only sets a new performance benchmark but also provides transparent and interpretable reasoning traces, illuminating the cognitive processes behind sarcasm comprehension.

[Demo] Rankify: A Comprehensive Python Toolkit for Retrieval, Re-Ranking, and Retrieval-Augmented Generation ► *Underline Video*

Abdelrahman Abdallah, Bhawna Piryani, Jamshid Mozafari, Andreas Herzinger, Jamie Holdcroft and Adam Jatowt

Building retrieval-augmented generation (RAG) systems often requires combining separate tools for retrieval, re-ranking, and generation, with incompatible data formats, evaluation pipelines, and deployment workflows. We present, an open-source Python toolkit that unifies these stages in a single modular framework.,,%, provides 42 benchmark datasets with pre-retrieved documents and pre-built indices, 15 retrievers (sparse, dense, and reasoning-augmented), and 24 re-ranking models spanning 41 pointwise, pairwise, and listwise variants. It also supports 6 RAG strategies across four inference backends (Hugging Face, vLLM, LiteLLM, and OpenAI), enabling consistent experimentation from local models to hosted APIs. A unified pipeline interface allows users to compose retrieve–rerank–generate workflows in a few lines of code, while an agentic assistant (RankifyAgent), a REST server (RankifyServer), and an interactive web playground support deployment and non-programmatic exploration. Across 200+ configurations on QA and BEIR/TREC benchmarks with six generator LLMs, re-ranking consistently improves downstream performance, yielding gains of 5–15 points in Exact Match and up to 8.5 points in RAGAS context precision across diverse retriever–generator combinations.

[Long] RealTalk-CN: A Realistic Chinese Speech Task-Oriented Dialogue Benchmark with Cross-Modal Analysis ► *Underline Video*

Enzhi Wang, Jiaming Zhou, Yuhang Jia, Aobo Kong, Qicheng Li and Yong Qin

Recent advances in speech large language models (e.g., GPT-4o) have enabled end-to-end spoken interactions, yet their robustness in real-world applications remains unclear, where systems must assist users in completing specific tasks under complex conditions such as multi-turn, ambiguous, and often spontaneous speech, as well as natural alternation between speech and text. Task-oriented dialogue (TOD) offers a realistic scenario to evaluate whether models can effectively help users accomplish such task-oriented goals, but existing benchmarks are mainly text-based, and the few speech datasets are limited to English and often neglect spontaneous disfluencies and speaker diversity. To address this gap, we introduce RealTalk-CN, the first Chinese multi-turn, multi-domain speech-text TOD dataset, containing 5.4k dialogues (60K turns, ~150 hours) of real human-to-human recordings with detailed annotations for dialogue states, disfluency types, and speaker characteristics. Based on this dataset, we propose a cross-modal interaction task supporting dynamic speech-text switching and a comprehensive evaluation protocol assessing robustness to disfluencies, sensitivity to speaker variation, and cross-domain generalization. Experiments on state-of-the-art models demonstrate the challenges posed by RealTalk-CN and establish its value as a benchmark for developing reliable and fair Speech LLMs in real-world deployments. The dataset and evaluation framework are available to encourage further research.

[Long] Reasoning Hijacking: The Fragility of Reasoning Alignment in Large Language Models ► *Underline Video*

Yuansen Liu, Yixuan Tang and Anthony Kum Hoe Tung

Current LLM safety research predominantly focuses on mitigating Goal Hijacking, preventing attackers from redirecting a model's high-level objective (e.g., from "summarizing emails" to "phishing users"). In this paper, we argue that this perspective is incomplete and highlight a critical vulnerability in Reasoning Alignment. We expose the inherent fragility of current alignment techniques by proposing a new adversarial prompt attack paradigm: Reasoning Hijacking. To demonstrate this vulnerability, we instantiate it via the Criteria Attack, which subverts model judgments by injecting spurious decision criteria without altering the high-level task goal. Unlike Goal Hijacking, which attempts to override the system prompt, Reasoning Hijacking keeps the task goal intact but manipulates the model's decision-making logic by injecting spurious reasoning shortcuts. Through extensive experiments on three different tasks (toxic comment, negative review, and spam detection), we demonstrate that even state-of-the-art models are highly fragile, consistently prioritizing injected heuristic shortcuts over rigorous semantic analysis. Crucially, because the model's explicit intent remains aligned with the user's instructions, these attacks can bypass defenses designed to detect goal deviation (e.g., SecAlign, StruQ), revealing a fundamental blind spot in the current safety landscape. Data and code are available at https://github.com/Yuan-Hou/criteria_attack.

[Long] ReCode: Reinforcing Code Generation with Reasoning-Process Rewards ▶ *Underline Video*

Lishui Fan, Yu Zhang, Mouxiang Chen and Zhongxin Liu

In practice, rigorous reasoning is often a key driver of correct code, while Reinforcement Learning (RL) for code generation often neglects optimizing reasoning quality. Bringing process-level supervision into RL is appealing, but it faces two challenges. First, training reliable reward models to assess reasoning quality is bottlenecked by the scarcity of fine-grained preference data. Second, naively incorporating such neural rewards may suffer from reward hacking. This work proposes ReCode(Reasoning-Reinforced Code Generation), a novel RL training framework comprising: (1) Contrastive Reasoning-Process Reward Learning (CRPL), which trains a reward model with synthesized optimized and degraded reasoning variants to assess the quality of reasoning process; and (2) Consistency-Gated GRPO (CG-GRPO), which integrates the reasoning-process reward model into RL by gating neural reasoning-process rewards with strict execution outcomes, using execution correctness as a hard gate to mitigate reward hacking. Additionally, to assess the reward models discriminative capability in assessing reasoning-process quality, we introduce LiveCodeBench-RewardBench (LCB-RB), a new benchmark comprising preference pairs of superior and inferior reasoning processes tailored for code generation. Experimental results across HumanEval(+), MBPP(+), LiveCodeBench, and BigCodeBench show that a 7B model trained with ReCode outperforms the base version by 16.1% and reaches performance comparable to GPT-4-Turbo. We further demonstrate the generalizability of ReCode by extending it to the math domain.

[Long] ReCreate: Reasoning and Creating Domain Agents Driven by Experience ▶ *Underline Video*

Zhezheng Hao, Hong Wang, Jian Luo, Jianqing Zhang, Yuyan Zhou, Qiang Lin, Can Wang, Hande Dong and Jiawei Chen

Large Language Model (LLM) agents are reshaping the industrial landscape. However, most practical agents remain human-designed because tasks differ widely, making them labor-intensive to build. This situation poses a central question: can we automatically create and adapt domain agents in the wild? While several recent approaches have sought to automate agent creation, they typically treat agent generation as a blackbox procedure and rely solely on final performance metrics to guide the process. Such strategies overlook critical evidence explaining why an agent succeeds or fails, and often require high computational costs. To address these limitations, we propose ReCreate, an experiencedriven framework for the automatic creation of domain agents. ReCreate systematically leverages agent interaction histories, which provide rich concrete signals on both the causes of success or failure and the avenues for improvement. Specifically, we introduce an agentasoptimizer paradigm that effectively learns from experience via three key components: (i) an experience storage and retrieval mechanism for on-demand inspection; (ii) a reasoning-creating synergy pipeline that map execution experience into scaffold edits; and (iii) hierarchical updates that abstract instance-level details into reusable domain patterns. In experiments across diverse domains, ReCreate consistently outperforms human-designed agents and existing automated agent generation methods, even when starting from minimal seed scaffolds.

[Industry] RedOne 2.0: Rethinking Domain-specific LLM Post-Training in Social Networking Services ▶ *Underline Video*

Fei zhao, Chonggang Lu, Haofu Qian, Fangcheng Shi, Zijie Meng, Jianzhao Huang, Zheyong Xie and Shaosheng Cao

As a primary medium for human interaction and information exchange, social networking services (SNS) present distinct challenges for large language models (LLMs): rapidly evolving norms and slang, and culturally diverse content that causes knowledge distribution shift. While supervised fine-tuning (SFT) can improve in-domain performance, it often induces a "seesaw" trade-off with out-of-domain robustness, especially for smaller models. To address these challenges, we present RedOne 2.0, an SNS-oriented LLM developed with a progressive, RL-prioritized post-training paradigm for fast and stable adaptation. Our pipeline has three stages: (1) Exploratory Learning on curated SNS corpora to establish initial alignment and surface systematic weaknesses; (2) Targeted Fine-Tuning that applies SFT only to diagnosed gaps while mixing a small amount of general data to reduce forgetting; and (3) Refinement Learning that re-applies RL with SNS-centric signals to consolidate gains and balance trade-offs across tasks. Across various tasks in three categories, our 4B model improves by 2.41 on average over a sub-optimal 7B baseline. It also yields an 8.74 average gain over the base model while using less than half the data required by the SFT-centric method, demonstrating superior data efficiency and stability at compact scales. Overall, RedOne 2.0 provides a competitive, cost-effective baseline for SNS-specific LLMs, improving capability without sacrificing robustness.

[Long] Region-Grounded Report Generation for 3D Medical Imaging: A Fine-Grained Dataset and Graph-Enhanced Framework ▶ *Underline Video*

Cong Huy Nguyen, Son Dinh Nguyen, Guanlin Li, Tuan Dung Nguyen, Aditya Narayan Sankaran, Mai Huy Thong, Thanh Trung Nguyen, Mai Hong Son, Reza Farahbakhsh, Phi Le Nguyen and Noel Crespi

Automated medical report generation for 3D PET/CT imaging is fundamentally challenged by the high-dimensional nature of volumetric data and a critical scarcity of annotated datasets, particularly for low-resource languages. Current black-box methods map whole volumes to reports, ignoring the clinical workflow of analyzing localized Regions of Interest (RoIs) to derive diagnostic conclusions. In this paper, we bridge this gap by introducing VietPET-RoI, the first large-scale 3D PET/CT dataset with fine-grained RoI annotation for a low-resource language, comprising 600 PET/CT samples and 1,960 manually annotated RoIs, paired with corresponding clinical reports. Furthermore, to demonstrate the utility of this dataset, we propose HiRRA, a novel framework that mimics the professional radiologist diagnostic workflow by employing graph-based relational modules to capture dependencies between RoI attributes. This approach shifts from global pattern matching toward localized clinical findings. Additionally, we introduce new clinical evaluation metrics, namely RoI Coverage and RoI Quality Index, that measure both RoI localization accuracy and attribute description fidelity using LLM-based extraction. Extensive evaluation demonstrates that our framework achieves SOTA performance, surpassing existing models by 19.7% in BLEU and 4.7% in

ROUGE-L, while achieving a remarkable 45.8% improvement in clinical metrics, indicating enhanced clinical reliability and reduced hallucination. Our code and dataset are available on GitHub.

[Long] Reinforced Informativeness Optimization for Long-Form Retrieval-Augmented Generation ▶ *Underline Video*

Yuhao Wang, Ruiyang Ren, Yucheng Wang, Wayne Xin Zhao, Jing Liu, Hua Wu and Haifeng Wang

Long-form question answering (LFQA) requires open-ended long-form responses that synthesize coherent, factually grounded content from multi-source evidence. This makes reinforcement learning (RL) reward design critical. The reward must be verifiable for faithful grounding and stable optimization. However, many standard rewards assume a unique target with an exact-match notion of correctness, which fits short-form QA and math but breaks in LFQA. As a result, current RAG systems still lack verifiable reward mechanisms, yielding unstable feedback signals and suboptimal optimization outcomes. We propose RioRAG, a framework for reinforced verifiable informativeness optimization. First, it defines informativeness as a measurable and externally verifiable objective for RL. Second, RioRAG uses nugget-centric verification with cross-source checks to enable self-evolution of smaller LLMs and to provide denser, action-discriminative rewards that mitigate reward sparsity and stabilize optimization. This formulation avoids handcrafted supervision for the policy model and strong teacher-model distillation, relying instead on externally verifiable feedback. Experiments on LongFact and RAGChecker show that RioRAG achieves higher factual recall and faithfulness, establishing verifiable reward modeling as a foundation for trustworthy long-form RAG. Our codes are available at https://github.com/RUCAIBox/RioRAG.

[Findings] Rejection-to-Acceptance Transition: Model Editing-Based Jailbreak Backdoor Injection Not Limited to Few Output Tokens ▶ *Underline Video*

Shiji Yang, Min Cai, Hao Xiong, Congyao Mei, Haodong Zou, Shicheng Tan, Jie Chen, Fulan Qian and Shu Zhao

Model editing-based jailbreak backdoor attacks against LLMs have gained attention for being lightweight, enabling vulnerability discovery in LLMs. Existing methods are implemented by binding backdoors to predefined phrases as first few output tokens, inducing the LLMs next-token prediction to produce continuous responses. However, their effectiveness is heavily dependent on the number of bound phrases, with attack costs rising as this number increases. In this work, we propose JEST, which achieves jailbreak backdoor attacks by hijacking LLM representations into a acceptance domain rather than binding to a few output tokens. Specifically, we propose a representation transition-guided model editing to inject jailbreak backdoors into LLMs. The activated backdoor transitions the LLM from rejection domain to acceptance domain, causing it to accept and generate jailbreak behavior. To clearly distinguish between rejection and acceptance domains within LLMs, we also design a domain modeling strategy for JEST that models these two opposing domains within the representation space. Additionally, JEST-hijacked LLMs exhibit greater vulnerability to direct prompt attacks. Experimental results show that JEST outperforms existing model editing methods, demonstrating stronger jailbreak capabilities across various LLMs and datasets. We also provide analysis to explore the safety boundary of LLM.

[Long] Rethinking Entropy Interventions in RLVR: An Entropy Change Perspective ▶ *Underline Video*

Zhezheng Hao, Hong Wang, Haoyang Liu, Jian Luo, Jiarui Yu, Hande Dong, Qiang Lin, Can Wang and Jiawei Chen

Reinforcement Learning with Verifiable Rewards (RLVR) serves as a cornerstone technique for enhancing the reasoning capabilities of Large Language Models (LLMs). However, its training is often plagued by entropy collapse, a rapid decline in policy entropy that limits exploration and undermines training effectiveness. While recent works attempt to mitigate this issue via several heuristic entropy interventions, the underlying mechanisms remain poorly understood. In this work, we conduct comprehensive theoretical and empirical analyses of entropy dynamics in RLVR, offering two main insights: (1) We derive a tight approximation for token-level entropy change at each update step, revealing four governing factors and providing a unified theoretical framework of how existing methods influence entropy; (2) We reveal a fundamental limitation of recent approaches: they rely on heuristic adjustments to one or two of these factors, leaving other relevant factors unconsidered, thus inherently limiting their effectiveness. Motivated by these findings, we propose STEER, a principled entropy-modulation method that adaptively reweighs tokens based on theoretically-estimated entropy variations. Extensive experiments across six mathematical reasoning and three coding benchmarks demonstrate that STEER effectively mitigates entropy collapse and consistently outperforms state-of-the-art baselines.

[Long] Rethinking the Role of Entropy in Optimizing Tool-Use Behaviors for Large Language Model Agents ▶ *Underline Video*

Zeping Li, Hongru Wang, Yiwen Zhao, Guanhua Chen, Yixia Li, Keyang Chen, Yixin Cao, Guangnan Ye, Hongfeng Chai and Zhenfei Yin

Tool-using agents based on Large Language Models (LLMs) excel in tasks such as mathematical reasoning and multi-hop question answering. However, in long trajectories, agents often trigger excessive and low-quality tool calls, increasing latency and degrading inference performance, making managing tool-use behavior challenging. In this work, we conduct entropy-based pilot experiments and observe a strong positive correlation between entropy reduction and high-quality tool calls. Building on this finding, we propose using entropy reduction as a supervisory signal and design two reward strategies to address the differing needs of optimizing tool-use behavior. Sparse outcome rewards provide coarse, trajectory-level guidance to improve efficiency, while dense process rewards offer fine-grained supervision to enhance performance. Experiments across diverse domains show that both reward designs improve tool-use behavior: the former reduces tool calls by 72.07% compared to the average of baselines, while the latter improves performance by 22.27%. These results position entropy reduction as a key mechanism for enhancing tool-use behavior, enabling agents to be more adaptive in real-world applications.

[Long] Retrieving to Recover: Towards Incomplete Audio-Visual Question Answering via Semantic-consistent Purification

► *Underline Video*

Jiayu Zhang, Shuo Ye, Qilang Ye, Zihan Song, Jiajian Huang and Zitong Yu

Recent Audio-Visual Question Answering (AVQA) methods have advanced significantly. However, most AVQA methods lack effective mechanisms for handling missing modalities, suffering from severe performance degradation in real-world scenarios with data interruptions. Furthermore, prevailing methods for handling missing modalities predominantly rely on generative imputation to synthesize missing features. While partially effective, these methods tend to capture inter-modal commonalities but struggle to acquire unique, modality-specific knowledge within the missing data, leading to hallucinations and compromised reasoning accuracy. To tackle these challenges, we propose R^2ScP, a novel framework that shifts the paradigm of missing modality handling from traditional generative imputation to retrieval-based recovery. Specifically, we leverage cross-modal retrieval via unified semantic embeddings to acquire missing domain-specific knowledge. To maximize semantic restoration, we introduce a context-aware adaptive purification mechanism that eliminates latent semantic noise within the retrieved data. Additionally, we employ a two-stage training strategy to explicitly model the semantic relationships between knowledge from different sources. Extensive experiments demonstrate that R^2ScP significantly improves AVQA and enhances robustness in modal-incomplete scenarios.

[Long] Revealing Procedural Reasoning Structures in Chain-of-Thought Training via Span-Level Gradient Organization ► *Underline Video*

Jia Liu, Jiaxin Luo, Weiwen Xu, Jonathan M. Garibaldi, Xiao-Kun Wu, Yixue Hao and Min Chen

Chain-of-Thought (CoT) prompting enables large language models to produce multi-step reasoning, yet how such reasoning-related structure is expressed during training remains poorly understood. We present Gradient-based Structural Developer (GSD), an unsupervised framework with a principled gradient aggregation view that tracks span-level gradient during fine-tuning on reasoning benchmarks to understand how models develop structured, step-by-step reasoning capabilities. Our analysis shows that while gradients at the level of individual tokens are often noisy, aggregating gradients over contiguous reasoning-related spans reveals stable and recurring directional alignment across samples. We refer to these directionally aligned patterns as aligned sequential stresses, reflecting consistent gradient organization associated with similar reasoning procedures. Beyond capturing semantically similar reasoning instances, such gradient alignment also reveals structurally similar but semantically diverse cases that share common procedural organization. These findings position GSD as a diagnostic framework for analyzing how procedural reasoning structures emerge during training, with downstream selection results serving as auxiliary evidence correlating gradient alignment with adaptation efficiency.

[Short] Reviving Iterative Refinement in Diffusion-based NER with an Initializer-Restorer Approach ► *Underline Video*

Long Trieu, Hieu Phi and Makoto Miwa

Diffusion models have introduced a generative paradigm for Named Entity Recognition (NER), formulating the task as refining entity spans from noise. While promising, our analysis on the ACE2004 dataset reveals a limitation when training with Exponential Moving Average (EMA): the model performance often peaks at a single inference step (gamma = 1) and plateaus or degrades with additional steps. This suggests that under standard stable training configurations, the model may function primarily as a one-step generator rather than leveraging the iterative refinement capability characteristic of diffusion models. To address this, we propose an Initializer-Restorer approach. Instead of initializing the reverse process from random Gaussian noise, we utilize a preliminary set of candidate spans generated by a standard NER model (e.g., BERT or GLiNER). This allows the diffusion model to start from an informed, diverse prior, enabling effective iterative restoration. We investigate different training strategies for the restorer and find that a hybrid strategy mixing ground truth and noisy predictions is essential. Experiments on ACE2004, GENIA, and CleanCoNLL show that our approach improves performance over the baseline, particularly when multiple restoration steps are employed. For instance, on CleanCoNLL, our method achieves an F1 score of 94.70%, compared to 93.79% for the baseline. Our code is available at https://github.com/longtrieu-ai/ Initializer-Restorer-NER.

[Long] River-LLM: Large Language Model Seamless Exit Based on KV Share ► *Underline Video*

Yingtao Shen and An Zou

Large Language Models (LLMs) have demonstrated exceptional performance across diverse domains but are increasingly constrained by high inference latency. Early Exit has emerged as a promising solution to accelerate inference by dynamically bypassing redundant layers. However, in decoder-only architectures, the efficiency of Early Exit is severely bottlenecked by the KV Cache Absence problem, where skipped layers fail to provide the necessary historical states for subsequent tokens. Existing solutions, such as recomputation or masking, either introduce significant latency overhead or incur severe precision loss, failing to bridge the gap between theoretical layer reduction and practical wall-clock speedup. In this paper, we propose River-LLM, a training-free framework that enables seamless token-level Early Exit. River-LLM introduces a lightweight KV-Shared Exit River that allows the backbones missing KV cache to be naturally generated and preserved during the exit process, eliminating the need for costly recovery operations. Furthermore, we utilize state transition similarity within decoder blocks to predict cumulative KV errors and guide precise exit decisions. Extensive experiments on mathematical reasoning and code generation tasks demonstrate that River-LLM achieves 1.71x to 2.16x practical speedup while maintaining high generation quality.

[Long] Rose-SQL: Role-State Evolution Guided Structured Reasoning for Multi-Turn Text-to-SQL ► *Underline Video*

Le Zhou, Feng Yao, Fengcai Qiao, Bo Xu, Fangyuan Wang and Boyan Xu

Recent advances in Large Reasoning Models (LRMs) trained with Long Chain-of-Thought have demonstrated remarkable capabilities in code generation and mathematical reasoning. However, their potential in multi-turn Text-to-SQL tasks remains largely underexplored.

Existing approaches typically rely on unstable API-based inference or require expensive fine-tuning on small-scale models. In this work, we present Rose-SQL, a training-free framework that leverages small-scale LRMs through in-context learning to enable accurate context-dependent parsing. We introduce the Role-State, a fine-grained representation that bridges the structural gap between schema linking and SQL generation by serving as a structural blueprint. To handle conversational dependencies, Rose-SQL traces the evolution of Role-State through historical context via structural isomorphism checks, guiding the model to infer the possible SQL composition for the current question through verified interaction trajectories. Experiments on the SParC and CoSQL benchmarks show that, within the Qwen3 series, Rose-SQL outperforms in-context learning baselines at the 4B scale and substantially surpasses state-of-the-art fine-tuned models at the 8B and 14B scales, while showing consistent gains on additional reasoning backbones.

[Industry] RouteLMT: Learned Sample Routing for Hybrid LLM Translation Deployment ► *Underline Video*

Yingfeng Luo, Hongyu Liu, DingYang Lin, Kaiyan Chang, Chenglong Wang, Bei Li, Quan Du, Tong Xiao and JingBo Zhu

Large Language Models (LLMs) have achieved remarkable performance in Machine Translation (MT), but deploying them at scale remains prohibitively expensive. A widely adopted remedy is the hybrid system paradigm, which balances cost and quality by serving most requests with a small model and selectively routing a fraction to a large model. However, existing routing strategies often rely on heuristics, external predictors, or absolute quality estimation, which fail to capture whether the large model actually provides a worthwhile improvement over the small one. In this paper, we formulate routing as a budget allocation problem and identify marginal gain, i.e., the large models improvement over the small model, as the optimal signal for budgeted decisions. Building on this, we propose RouteLMT (routing for LLM-based MT), an efficient in-model router that predicts this expected gain by probing the small translators prompt-token representation, without requiring external models or hypothesis decoding. Extensive experiments demonstrate that our RouteLMT outperforms heuristics, quality/difficulty estimation baselines, achieving a superior quality-budget Pareto frontier. Furthermore, we analyze regression risks and show that a simple guarded variant can mitigate severe quality losses.

[Long] S2G-RAG: Structured Sufficiency and Gap Judging for Iterative Retrieval-Augmented QA ► *Underline Video*

Minghan Li, Junjie Zou, Xinxuan Lv, Chao Zhang and Guodong Zhou

Retrieval-Augmented Generation (RAG) grounds language models in external evidence, but multi-hop question answering remains difficult because iterative pipelines must control what to retrieve next and when the available evidence is adequate. In practice, systems may answer from incomplete evidence chains, or they may accumulate redundant or distractor-heavy text that interferes with later retrieval and reasoning. We propose S2G-RAG (Structured Sufficiency and Gap-judging RAG), an iterative framework with an explicit controller, S2G-Judge. At each turn, S2G-Judge predicts whether the current evidence memory supports answering and, if not, outputs structured gap items that describe the missing information. We map these gap items into the next retrieval query, producing stable multi-turn retrieval trajectories. To reduce noise accumulation, we maintain a sentence-level Evidence Context by extracting a compact set of relevant sentences from retrieved documents. Experiments on TriviaQA, HotpotQA, and 2WikiMultiHopQA show that S2G-RAG improves multi-hop QA performance and robustness under multi-turn retrieval. Furthermore, S2G-RAG can be integrated into existing RAG pipelines with a lightweight component, without modifying the search engine or retraining the generator.

[Long] SAC: Neural Speech Codec with Semantic-Acoustic Dual-Stream Quantization ► *Underline Video*

Wenxi Chen, Ruiqi Yan, Yushen Chen, Zhikang Niu, Ziyang Ma, Xiquan Li, Yuzhe Liang, Hanlin Wen, Shunshun Yin, Ming Tao, Xinsheng Wang and Xie Chen

Speech codecs that convert continuous speech signals into discrete tokens have become essential for speech language models. However, existing codecs struggle to balance high-quality reconstruction with semantically rich representations, limiting their effectiveness in both generative and understanding tasks. In this work, we propose SAC, a neural speech codec with semantic-acoustic dual-stream quantization. By disentangling semantic and acoustic modeling into two dedicated streams, SAC enables each to be optimized for its respective role. Comprehensive evaluations show that SAC achieves strong reconstruction performance across diverse bitrates under both clean and noisy conditions, with particularly high scores on UTMOS and WER, indicating superior naturalness and intelligibility. Moreover, SAC substantially surpasses prior codecs in semantic representation, approaching the level of continuous self-supervised embeddings. When used as a tokenizer for LLM-based text-to-speech, SAC enables a single-stage autoregressive (AR) TTS model that clearly outperforms state-of-the-art AR systems. Our disentanglement analysis further validates the effectiveness of the dual-stream design, offering new potential for controllable speech generation.

[Long] SafeMCP: Proactive Power Regulation for LLM Agent Defense via Environment-Grounded Look-Ahead Reasoning ► *Underline Video*

Lichao Wang, Zhaoxing Ren, Tianzhuo Yang, Jiaming Ji, Chi Harold Liu, Yaodong Yang and Juntao Dai

As Large Language Model (LLM) agents increasingly leverage the Model Context Protocol (MCP) to operate in complex environments, the expansion of their action spaces offers agents unsafe capabilities and underscores the risk of power-seeking. While broad action space and greater environment influence are essential for task fulfillment, they creates a fragile risk surface where minor errors or hallucinations are magnified into catastrophic failures. In response, we propose SafeMCP, a server-side defense plugin that constrains tool acquisition via predictive reasoning regarding future safety risks. SafeMCP utilizes an internal world model for look-ahead reasoning to implement a two-tier defense: proactive tool filtering to constrain hazardous power expansion and immediate intervention as a fail-safe. To train SafeMCP, we introduce a three-stage pipeline comprising environmental dynamic grounding, safe policy initialization, and reinforcement learning

(RL) with dual verifiable rewards. Experiments on PowerSeeking Bench, ToolEmu, and AgentHarm show that SafeMCP achieves a safe equilibrium, effectively mitigating risks while preserving agent utility.

[Long] SAFO: Stable Adaptive Fairness Optimization for LLM-Based Social Survey Simulation ▶ *Underline Video*

Chenxi Lin, Zhuoren Jiang, Kaisong Song and Yiquan Wu

Ensuring fairness in social survey simulation is critical, as biased outputs can misrepresent underrepresented groups. This issue is growing as large language models (LLMs) are increasingly used for this task. However, standard fine-tuning based on Empirical Risk Minimization (ERM) often under-optimizes minority groups, causing substantial subgroup disparities. Distributionally robust Optimization (DRO) methods reduce worst-case errors, but their strict worst-case selection can lead to noisy and unstable optimization under demographic sparsity. These issues create intertwined challenges for fairness, convergence and stability. We propose SAFO, a dynamic utility-fairness optimization framework for LLM-based survey simulation that explicitly targets both fairness and training stability. SAFO combines (i) an Optimizer that preserves mean-loss utility, (ii) an Adversary that performs temperature-controlled, EMA-smoothed and loss-driven group reweighting, and (iii) a Nash-inspired Regulator that adaptively adjusts the utility-fairness trade-off by tracking weak-group gains and collateral utility damages. Experiments on three large-scale survey datasets from China, the U.S., and Europe show that SAFO consistently improves minority performance and social-welfare metrics. It reduces worst-group gaps by up to 12.7%, maintains overall accuracy with a mean change of less than 0.3% and lowers variance across random seeds. Our code is available at https://github.com/PiLab-ZJU/SAFO.

[Long] SAGE: Synergistic Adaptive Gating of Experts for Hateful Video Detection ▶ *Underline Video*

Jie Huang, Xin Liao, Junjie Wang, Mingyang Li, Wenshuo Wang, Ziyou Jiang, Shoubin Li and Qing Wang

With the rise of short-video platforms, hate speech has evolved from static text and memes into more covert and aggressive hateful video formats, profoundly impacting social dynamics and public sentiment. Existing detection methods typically rely on multimodal feature fusion, which blurs the distinct boundaries of modality-specific information. This leads to the feature dilution problem, where dominant benign modalities often overwhelm sparse, localized hateful cues. To address this, we propose SAGE (Synergistic Adaptive Gating of Experts), a novel framework that shifts the paradigm from blind feature mixing to decision-level arbitration. Mimicking human cognitive processes, SAGE instantiates disentangled experts to rigorously preserve modality-specific semantics, facilitates global expert deliberation for context-aware refinement, and convenes an instance-level tribunal to dynamically arbitrate the final verdict based on evidentiary salience. Extensive experiments on HateMM and MultiHateClip benchmarks demonstrate that SAGE significantly outperforms state-of-the-art methods, achieving accuracy gains of 6.37% to 21.23% and macro-F1 score gains of 6.77% to 28.01%.

[Long] SAME: Safety-Aware Model Editing Guided by Safety Transformation ▶ *Underline Video*

Jiayi Wang, Shipeng Wang, Ji Wu and Jian Sun

Editing large language models is challenging as incorporating new knowledge often requires sequential parameter updates while maintaining model capability. In this work, we experimentally observe that sequential knowledge updating under the locate-then-edit framework can introduce safety risks, regardless of whether the knowledge being edited is benign or malicious. We propose a novel model editing approach that estimates safety transforms and identifies corresponding safety direction in the neural activation space, and then aligns neural activation updates and network parameter updates under the safety constraints, resulting in a safety-aware model editing approach. We evaluate our approach on open-source LLMs, Llama-3-8B-Instruct, Qwen3-4B-Instruct and Qwen2.5-14B-Instruct, using the benchmark datasets ZsRE and COUNTERFACT, as well as the malicious dataset Mal-KSet. Experimental results demonstrate that our approach effectively reduces unsafe responses to malicious queries while preserving the effectiveness of model editing.

[Long] SAME: Signer-Aware Mixture-of-Experts for Test-Time Adaptation in Sign Language Translation ▶ *Underline Video*

Lujia Yang, Weicai Yan, Yongbo He, Qifei Zhang, Tao Jin, Jinshan Zhang, Meng Xi and Jianwei Yin

Sign language translation (SLT) is essential for bridging communication between the deaf and hearing communities, but real-world deployment suffers from domain shift such as signer variability, lighting, and background changes. Supervised fine-tuning is impractical due to limited labeled data, and existing unsupervised adaptation methods require batch statistics or long adaptation. We introduce Test-Time Adaptation (TTA) for SLT, enabling rapid adaptation to domain shift without the need for labeled data. To the best of our knowledge, this is the first study to explore TTA in SLT. Existing TTA methods predominantly focus on image classification tasks and lack a comprehensive strategy for handling domain shift in SLT. In response, we introduce SAME, a plug-and-play, signer-aware Mixture-of-Experts (MoE) TTA architecture for SLT. SAME inserts lightweight MoE modules after multiple encoder layers. Gates are conditioned on signer features and stabilized with unsupervised regularizers, effectively decoupling domain shift across encoder depths while enabling personalized adaptation. Experiments show that SAME outperforms existing TTA methods and can enhance the capabilities of multiple SLT models.

[Long] Scaling Law for Multimodal Large Language Model Supervised Fine-Tuning ▶ *Underline Video*

YiFan Zhang, Tao Yu, Feng Li, Chaoyou Fu, Yibo Hu, Kun Wang, Qingsong Wen, Zhang Zhang, Liang Wang and Rong Jin

The supervised fine-tuning (SFT) stage is crucial for multimodal large language models (MLLMs), yet a comprehensive scaling law to guide the optimal model-data configuration remains lacking. In this paper, we make an initial attempt to address this gap. First, we theoretically demonstrate that directly computing the optimal computation frontier for MLLM-SFT, as we can for traditional LLMs, is a challenging task. This complexity arises because MLLM-SFT is influenced by a broader range of factors, including model size, LLM pre-training tokens, and MLLM SFT tokens. To tackle this issue, we propose two scaling laws based on LLM paradigms: one applicable when training data

volumes are well defined by researchers, and another for cases where models are sourced from open communities with unknown training data. Through theoretical modeling and approximations, we provide researchers with valuable recommendations for optimal resource allocation. Furthermore, we establish a strong correlation (R^2 = 0.98) between training loss and downstream performance, enabling accurate performance estimation without the need for exhaustive benchmarking. To validate our scaling laws, we construct a testbed of 60 models ranging from 50 million to 8 billion parameters, totaling 1,560 checkpoints. Each checkpoint is evaluated on than 10 MLLM benchmarks, ensuring robust fitting of our formulations.

[Long] SciFlow-Bench: Evaluating Structure-Aware Scientific Diagram Generation via Inverse Parsing ▶ *Underline Video*

Tong Zhang, Honglin Lin, Zhou Liu, Chong Chen and Wentao Zhang

Scientific diagrams convey explicit structural information, yet modern text-to-image models often produce visually plausible but structurally incorrect results. Existing benchmarks either rely on image-centric or subjective metrics insensitive to structure, or evaluate intermediate symbolic representations rather than final rendered images, leaving pixel-based diagram generation underexplored. We introduce SciFlow-Bench, a structure-first benchmark for evaluating scientific diagram generation directly from pixel-level outputs. Built from real scientific PDFs, SciFlow-Bench pairs each source framework figure with a canonical ground-truth graph and evaluates models as black-box image generators under a closed-loop, round-trip protocol that inverse-parses generated diagram images back into structured graphs for comparison. This design enforces evaluation by structural recoverability rather than visual similarity alone, and is enabled by a hierarchical multi-agent system that coordinates planning, perception, and structural reasoning. Experiments show that preserving structural correctness remains a fundamental challenge, particularly for diagrams with complex topology, underscoring the need for structure-aware evaluation.

[Long] SciPedia: Unlocking the Value of Scientific Data for Pre-training ▶ *Underline Video*

Yiwei Qin, Zhen Huang, Tiantian Mi, Weiye Si, Qipeng Guo, Siyuan Feng and Pengfei Liu

High-quality scientific data is critical for advancing LLMs, yet academic literature remains largely underutilized. This work addresses the fundamental question: How can we systematically unlock scientific data's value for pre-training? First, we construct a large-scale raw scientific corpus but identify a critical Learnability Gap, revealing that direct pre-training yields negligible gains. To bridge this, we develop a multi-stage pipeline featuring content cleaning and pedagogical augmentation, resulting in SciPedia, a 900B-token corpus. Finally, we establish a controlled verification framework: we develop SciPedia-Eval benchmark and conduct 600B tokens of continued pre-training (CPT) starting from transparent base models (3B/7B) trained from scratch. Compared to a CPT baseline trained with general-purpose data, our approach with SciPedia data boosts average performance by +2.12 (3B) and +2.95 (7B), reaching +5.60 and +8.40 on in-domain tasks. This setup further allows us to derive empirical guidelines for data composition and model configurations.

[Long] SCOPE: Boosting LLM Efficiency with Scoped Position Encoding ▶ *Underline Video*

Qingguo Qi, Hongyang Chen and Zhao Li

Positional encodings are fundamental to Transformers, yet explicit methods like RoPE can degrade under length extrapolation and may incur extra arithmetic and memory-access overhead. In this paper, we propose Scoped Position Encoding (ScoPE), a novel framework that reimagines structured sparsity as an intrinsic position encoding mechanism. Instead of relying on explicit arithmetic signals, ScoPE assigns exponentially scaled look-back scopes to attention heads. We theoretically demonstrate that this simple topological constraint transforms multi-head attention into a hierarchical processor, yielding an order awareness horizon that grows exponentially with depth up to the sequence length. Consequently, ScoPE is parameter-free and avoids relying on fragile positional arithmetic. Empirically, it significantly enhances efficiency by masking the majority of attention computations, offering a theoretical 8x reduction in attention FLOPs at long contexts. Extensive evaluations on LLaMA-3-8B architectures reveal that ScoPE achieves superior native length extrapolation and robust retrieval fidelity compared to RoPE, all while substantially reducing training and inference latency. The code is available at https://github.com/oncemoe/ScoPE.

[Findings] SDAR: A Synergistic Diffusion-AutoRegression Paradigm for Scalable Sequence Generation ▶ *Underline Video*

Shuang Cheng, Yihan Bian, Dawei Liu, Yuhua Jiang, Yihao Liu, Linfeng Zhang, Qian Yao, Zhongbo Tian, Wenhai Wang, Qipeng Guo, Kai Chen, Biqing Qi and Bowen Zhou

Autoregressive (AR) language modeling remains the dominant paradigm due to its dense supervision signal and highly optimized serving infrastructure, but its strictly causal, token-by-token decoding limits parallelism and non-causal modeling. While masked diffusion offers a promising path toward parallel generation, it faces two critical bottlenecks: training inefficiency stemming from sparse masked objectives, and high latency caused by iterative whole-sequence denoising. We present a systematic study of blockwise discrete diffusion, a pragmatic middle ground that preserves AR-compatible serving while enabling parallel intra-block generation. Our study proceeds in four steps: (i) a controlled, compute- and scale-matched comparison revealing that AR is a more effective backbone for blockwise hybrids than masked diffusion objectives; (ii) a scalable conversion recipe, SDAR, validating that AR models spanning 1.7B to 30B parameters can be adapted into block diffusion models with minimal compute while preserving backbone capabilities; and (iii) a systematic characterization of decoding dynamics, which reveals a virtuous cycle where larger models enable more aggressive parallel decoding, achieving theoretical speedups over 5x and wall-clock speedups of 2.3x on H200 GPUs in latency-critical regimes; and (iv) an investigation of local non-causal modeling capabilities, showing that SDAR's local bidirectional attention overcomes causal bottlenecks in scientific domains (e.g., chemistry) and enables robust test-time scaling. We release the full model suite, the training framework, and our inference engines for further innovation in non-autoregressive generative paradigms.

[Long] Seeing but Not Thinking: Routing Distraction in Multimodal Mixture-of-Experts ▶ *Underline Video*

Haolei Xu, Haiwen Hong, Hongxing Li, Rui Zhou, Yang Zhang, Longtao Huang, Hui Xue, Yongliang Shen, Weiming Lu and Yueting Zhuang

Multimodal Mixture-of-Experts (MoE) models have achieved remarkable performance on vision-language tasks. However, we identify a puzzling phenomenon termed Seeing but Not Thinking: models accurately perceive image content yet fail in subsequent reasoning, while correctly solving identical problems presented as pure text. Through systematic analysis, we first verify that cross-modal semantic sharing exists in MoE architectures, ruling out semantic alignment failure as the sole explanation. We then reveal that visual experts and domain experts exhibit layer-wise separation, with image inputs inducing significant routing divergence from text inputs in middle layers where domain experts concentrate. Based on these findings, we propose the Routing Distraction hypothesis: when processing visual inputs, the routing mechanism fails to adequately activate task-relevant reasoning experts. To validate this hypothesis, we design a routing-guided intervention method that enhances domain expert activation. Experiments on three multimodal MoE models across six benchmarks demonstrate consistent improvements, with gains of up to 3.17% on complex visual reasoning tasks. Our analysis further reveals that domain expert identification locates cognitive functions rather than sample-specific solutions, enabling effective transfer across tasks with different information structures.

[Long] SELECting over Tokens: Curating Pre-training Data at Scale via Token Classification ▶ *Underline Video*

Xin TongWeidong ZhangJiaang LiHaibin ChenShilei LiuLangming LiuKangtao LvYujin YuanWenbo SuBo Zheng

The quality of pre-training data critically impacts the capabilities of large language models. Existing pipelines rely on expert-crafted heuristic rules, which primarily operate at the sample level and are based on coarse statistical indicators, thus lacking content-aware, fine-grained noise detection. While recent generative approaches, e.g., ProX-C, enable token-level refinement, their reliance on synthesizing Python code incurs prohibitive computational cost at scale and can introduce hallucinations into the refined data. To overcome these limitations, we propose Selecting over Tokens (SelecT), a novel framework that reframes data refinement as a highly efficient token classification task. SelecT classifies each token as either informative or noisy and subsequently removes the latter. This design achieves fine-grained data optimization while avoiding the inefficiency of generation, ensuring scalability. When evaluated on diverse downstream benchmarks, the model trained on SelecT-refined corpora, on average, outperforms the one trained on raw data by over 2% and exceeds the best heuristic baselines by more than 1% while preserving 17% more tokens than the latter. Furthermore, SelecT achieves higher average performance than the generative ProX-C across all experimental settings, and is 2.5x faster at inference, even with twice the parameters. Our results establish SelecT as an effective, efficient, and scalable solution for pre-training data optimization.

[Long] Selective Contrastive Learning For Gloss Free Sign Language Translation ▶ *Underline Video*

Changhao Lai, Rui Zhao, Xuewen Zhong, Jinsong Su and Yidong Cheng

Sign language translation (SLT) converts continuous sign videos into spoken-language text, yet it remains challenging due to the intrinsic modality mismatch between visual signs and written text, particularly in gloss-free settings. Recent SLT systems increasingly adopt CLIP-like Vision-Language pretraining (VLP) for cross-modal alignment, but the random in-batch contrast provides few, batch-dependent negatives and may mislabel semantically similar (or even identical) pairs as negatives, introducing noisy and potentially inconsistent alignment supervision. In this work, we first conduct a preliminary trajectory-based analysis that tracks negative video-text similarity over training. The results show that only a small subset of negatives exhibits the desired behavior of being consistently pushed away, while the remaining negatives display heterogeneous and often non-decreasing similarity dynamics, suggesting that random in-batch negatives are frequently uninformative for effective alignment. Inspired by this, we propose Selective Contrastive Learning for SLT (SCL-SLT) with a Pair Selection (PS) strategy. PS scores candidate negatives using similarity dynamics from reference checkpoints and constructs mini-batches via a curriculum that progressively emphasizes more challenging negatives, thereby strengthening contrastive supervision while reducing the influence of noisy or semantically invalid negatives.

[Long] Self-SoftCoT: A Self-Consistent Framework via Position-Aware Latent Space Reinforcement Learning ▶ *Underline Video*

Liangliang Dong, Lianlei Shan and Shuaimin Li

While Chain-of-Thought (CoT) reasoning empowers Large Language Models (LLMs) to tackle complex tasks, its reliance on discrete token decoding imposes an inherent Discreteness Bottleneck, limiting expressiveness within a restricted vocabulary space. Existing continuous reasoning approaches, such as SoftCoT, mitigate this but typically rely on external auxiliary models, resulting in complex deployment and fractured inference pipelines. To address these challenges, we propose Self-SoftCoT, a self-contained framework that enables a frozen LLM to internally generate and consume latent thoughts without external assistants. By establishing a single-stream "Thinking to Speaking" closed-loop, we decouple latent planning from explicit generation. Furthermore, we adopt Group Sequence Policy Optimization (GSPO) to stabilize learning and employ Position-Aware Independent Projection to mitigate representation homogenization. Experimental results on five reasoning benchmarks demonstrate that our method significantly improves the reasoning performance of frozen LLMs. Specifically, our Qwen2.5-based model uses only N=2 soft tokens to outperform the SoftCoT baseline (N=4), improving the average accuracy from 75.06% to 78.42%. Similarly, LLaMA-3.1 performance increases from 70.52% to 74.55%.

[Long] SGVEF-LOOP: Coverage-Guided Progressive Topological Exploration and Fact-Grounded Metamorphic Evaluation for MCP Agents ▶ *Underline Video*

Zenghao Liu and Yansong Zhang

The rapid expansion of the Model Context Protocol (MCP) ecosystem introduces a combinatorially complex tool space, rendering existing frameworks inadequate for comprehensive agent evaluation. To address this problem, we propose SGVEF-LOOP, a coverage-guided framework for progressive topological exploration and fact-augmented metamorphic testing. SGVEF operates via a synergistic closed loop: it navigates sparse regions using adaptive sampling, synthesizes oracle-free metamorphic pairs grounded in static knowledge, enforces dual-constraint validation to ensure consistency and solvability, and leverages execution feedback to iteratively optimize exploration. Deploying this framework yields a high-fidelity benchmark achieving 100% node coverage and saturating 80.54% of the theoretical transition bound (estimated via Chao1). Evaluation of 8 diverse MCP Agents reveals capability stratification and exposes critical behavioral anomalies—such as reasoning instability—that conventional metrics fail to capture. Consequently, this work establishes a generalizable paradigm for scalable, rigorous agent evaluation in dynamic environments.

[Long] SHAPE: Stage-aware Hierarchical Advantage via Potential Estimation for LLM Reasoning ▶ *Underline Video*

Zhengyang Ai, Zikang Shan, Xiaodong Ai, Jingxian Tang, Hangkai Hu and Pinyan Lu

Process supervision has emerged as a promising approach for enhancing LLM reasoning, yet existing methods fail to distinguish meaningful progress from mere verbosity, leading to limited reasoning capabilities and unresolved token inefficiency. To address this, we propose Stage-aware Hierarchical Advantage via Potential Estimation (SHAPE), a framework that formalizes reasoning as a trajectory through a state space of empirical solvability. SHAPE introduces a hierarchical credit assignment mechanism: at the segment level, it employs a stage-aware advantage function to prioritize efficient breakthroughs in low-potential states; at the token level, it utilizes entropy-driven redistribution to sharpen execution signals. Extensive experiments in math reasoning across three base models and five benchmarks demonstrate that SHAPE achieves an average accuracy gain of 3% with 30% reduced token consumption.

[Findings] Small Data, Big Noise: Adversarial Training for Robust ParameterEfficient Fine-Tuning ▶ *Underline Video*

Eitan Cohen, Idan Simai and Uri Shaham

Parameter-Efficient Fine-Tuning (PEFT) has become essential for adapting foundation models to downstream NLP tasks. However, current PEFT methods often struggle with robustness to noise and performance degradation on limited training data. We propose SDBN (Small Data Big Noise), a unified framework that brings adversarial training to PEFT - a combination that remains less studied in the PEFT setting despite its complementary strengths - to enhance model robustness and generalization, outperforming alternative approaches. We also introduce two variants of the method that use discrete uncertainty sets: SDBN-h, which enumerates character-level edits and selects worst-case variants using gradients, and SDBN-p, which uses LLM-generated variants for robust optimization in generative tasks. Experiments across multiple benchmarks reveal substantial improvements, particularly in low-resource settings and under both word-level and character-level corruptions. This framework addresses the less explored intersection of adversarial training and parameter-efficient adaptation, without introducing additional parameters or only modest computational overhead, making PEFT deployments more reliable in real-world scenarios where data scarcity and linguistic variability often coexist.

[Long] Soft Orthogonal Low-Rank Adaptation for Knowledge Sharing in Large Language Model Continual Learning ▶ *Underline Video*

Yitong Wang, Xue Han, WenChun Gao, Qian Hu, Jiahui Wang, Ziqing Wang, Lijun Mei and Junlan Feng

When large language models are used in real-world scenarios, continual learning (CL) becomes a non-trivial problem. In particular, continual learning with modern LLMs is challenged both by the substantial computational costs induced by their massive parameter scale, and by the limitations of current CL methods, which are mainly designed to mitigate catastrophic forgetting while neglecting knowledge sharing across tasks. We further observe that models with stronger performance exhibit stronger inter-task connections. In light of the above challenges and findings, we propose Attribution Scores-based Soft Orthogonality Low-Rank Adaptation (ASO-LoRA), an effective and efficient framework that simultaneously facilitates knowledge transfer while mitigating catastrophic forgetting. Specifically, ASO-LoRA initially assigns task-specific parameter subspaces for new tasks utilizing multi-LoRA modules, enabling for efficient training and inference without relying on task labels. Then, ASO-LoRA leverages attribution scores to evaluate task similarity and employs soft orthogonality between task-specific subspaces, guiding gradient updates in directions that promote parameter isolation, achieving a balance between knowledge transfer and preservation. Experiments are carried out on both the T5-large and the LLaMA2-7B, showing ASO-LoRAs superior performance and suitability as a plug-in CL solution for general Transformer-based LLMs. Code is available at https://github.com/736619821/ASO-LORA.

[Findings] Solver-Independent Automated Problem Formulation via LLMs for High-Cost Simulation-Driven Design ▶ *Underline Video*

Yuchen Li, Handing Wang, Bing Xue, Mengjie Zhang and Yaochu Jin

In the high-cost simulation-driven design domain, translating ambiguous design requirements into a mathematical optimization formulation is a bottleneck for optimizing product performance. This process is time-consuming and heavily reliant on expert knowledge. While large language models (LLMs) offer potential for automating this task, existing approaches either suffer from poor formalization that fails to accurately align with the design intent or rely on solver feedback for data filtering, which is unavailable due to the high simulation costs. To address this challenge, we propose automated problem formulation (APF), a solver-independent framework that utilizes LLMs to convert engineers' natural language requirements into executable optimization models. The core of this framework is an innovative pipeline for automatically generating high-quality data, which overcomes the difficulty of constructing suitable fine-tuning datasets in the absence of high-cost solver feedback with the help of data generation and test instance annotation. The generated high-quality dataset is used to perform

supervised fine-tuning on LLMs, significantly enhancing their ability to generate accurate and executable optimization problem formulations. Experimental results on antenna design demonstrate that APF significantly outperforms the existing methods in both the accuracy of requirement formalization and the quality of resulting radiation efficiency curves in meeting the design goals.

[Long] SpiderFlow: Efficient Topology-Aware Scheduling for LLM Training Across Decentralized GPU Clusters ▶ *Underline Video*

Zihan Chang, Shuibing He, Bo Zhou, Sheng Xiao, Siling Yang, Rui Wang and Zhe Pan

In response to the increasing demand for largescale machine learning training jobs, many organizations have deployed GPU clusters across geographically distributed regions. However, existing ILP- or genetic-based cross-cluster training approaches largely overlook the topology of decentralized clusters, lacking both topologyaware task scheduling mechanisms and automated model parallelization strategies. As a result, naively applying these optimization-based methods in cross-cluster settings leads to prohibitive scheduling overhead, due to the drastically enlarged search space induced by complex inter-cluster topologies. To address these challenges, we propose SpiderFlow, a topologyaware scheduling system specifically designed for decentralized GPU clusters. We formulate cross-cluster task scheduling as a graph optimization problem and introduce SpinSearch, a low-overhead topology-aware scheduling algorithm. In addition, for automated model parallelization, we propose TPA, a two-level scheduling framework that combines heuristic methods at the inter-cluster level with ILP-based optimization within clusters, effectively reducing the search space while maintaining high training throughput with substantially lower scheduling overhead. We evaluate SpiderFlow on a physical platform comprising 8 decentralized clusters, as well as on a simulation platform with up to 64 decentralized clusters. Experimental results demonstrate that SpiderFlow reduces job completion time (JCT) by 1.2-1.3x, improves throughput by 1.12-1.25x, and reduces scheduling overhead by 20-90x on average compared to state-of-the-art scheduling systems.

[Demo] Squrve: A Unified and Modular Framework for Complex Real-World Text-to-SQL Tasks ▶ *Underline Video*

Yihan Wang, Peiyu Liu, Runyu Chen, Jiaxing Pu and Wei Xu

Text-to-SQL technology has evolved rapidly, with diverse academic methods achieving impressive results. However, deploying these techniques in real-world systems remains challenging due to limited integration tools. Despite these advances, we introduce Squrve, a unified, modular, and extensive Text-to-SQL framework designed to bring together research advances and real-world applications. Squrve first establishes a universal execution paradigm that standardizes invocation interfaces, then proposes a multi-actor collaboration mechanism based on seven abstracted effective atomic actor components. Experiments on widely adopted benchmarks demonstrate that the collaborative workflows consistently outperform the original individual methods, thereby opening up a new effective avenue for tackling complex real-world queries. The codes are available at https://anonymous.4open.science/r/Squrve-Annoy-E5B4.

[Long] Stable-RAG: Mitigating Retrieval-Permutation-Induced Hallucinations in Retrieval-Augmented Generation ▶ *Underline Video*

Qianchi Zhang, Hainan Zhang, Liang Pang, Hongwei Zheng and Zhiming Zheng

Retrieval-Augmented Generation (RAG) has become a key paradigm for reducing factual hallucinations in Large Language Models (LLMs), yet little is known about how the order of retrieved documents affects model behavior. We empirically show that under a Top-5 retrieval setting with the gold document included, LLM answers vary substantially across permutations of the retrieved set, even when the gold document is fixed in the first position. This reveals a previously underexplored sensitivity to retrieval permutations. Although existing robust RAG methods focus primarily on enhancing LLM robustness to low-quality retrieval and mitigating positional bias to distribute attention fairly over long contexts, neither approach directly addresses permutation sensitivity. In this paper, we propose Stable-RAG, which exploits permutation sensitivity estimation to mitigate permutation-induced hallucinations. Stable-RAG runs the generator under multiple retrieval orders, clusters hidden states, and decodes from a cluster-center representation that captures the dominant reasoning pattern. It then uses these reasoning results to align hallucinated outputs toward the correct answer, encouraging the model to produce consistent and accurate predictions across document permutations. Experiments on three QA datasets show that Stable-RAG improves answer accuracy, reasoning consistency, and generalization across datasets, retrievers, and input lengths compared with strong baselines.

[Findings] STAR-Teaming: A Strategy-Response Multiplex Network Approach to Automated LLM Red Teaming ▶ *Underline Video*

MinJae Jung, YongTaek Lim, Chaeyun Kim, Junghwan Kim, Kihyun Kim and Minwoo Kim

While Large Language Models (LLMs) are widely used, they remain susceptible to jailbreak prompts that can elicit harmful or inappropriate responses. This paper introduces STAR-Teaming, a novel black-box framework for automated red teaming that effectively generates such prompts. STAR-Teaming integrates a Multi-Agent System (MAS) with a Strategy-Response Multiplex Network and employs network-driven optimization to sample effective attack strategies. This network-based approach recasts the intractable high-dimensional embedding space into a tractable structure, yielding two key advantages: it enhances the interpretability of the LLM's strategic vulnerabilities, and it streamlines the search for effective strategies by organizing the search space into semantic communities, thereby preventing redundant exploration. Empirical results demonstrate that STAR-Teaming significantly surpasses existing methods, achieving a higher attack success rate (ASR) at a lower computational cost. Extensive experiments validate the effectiveness and explainability of the Multiplex Network. The code is available at https://github.com/selectstar-ai/STAR-Teaming-paper.

[Findings] Steering Away from Refusal: A Black-box Jailbreak Method Based on First-Token Distribution ▶ *Underline Video*

Shuangjie Fu, Du Su, Xin Chen, Fei Sun, Huawei Shen and Xueqi Cheng

Investigating black-box jailbreak attacks is crucial for revealing the actual security risks faced by operational Large Language Models (LLMs). The primary challenge in black-box jailbreak attack is the absence of direct optimization signals, such as gradients, to guide the refinement of adversarial prompts. While current mainstream methods like PAIR and TAP attempt to leverage the model's textual output as feedback, facing a critical limitation when models consistently generate static refusal responses, depriving the attacker of any actionable signal to distinguish better prompts. To overcome the bottleneck and reveal whether there is potential risk to open access to partial logprobs information, we investigate LLM output distribution. Our empirical analysis reveals that refusal responses exhibit a highly consistent distributional pattern at the first generated token, suggesting that the deviation from this standard pattern can serve as a quantifiable metric for LLM generating refusal response. Based on this insight, we propose Distribution Jailbreak (DJ), an attack method that select effective jailbreak templates and then iteratively optimizes adversarial suffixes by maximizing the KL divergence from the standard refusal distribution. Extensive experiments demonstrate that DJ achieves state-of-the-art Attack Success Rate(ASR). Notably, DJ achieves over 90% ASR on all tested open-source models, and delivers over 94% ASR on GPT-4.1. Our code is publicly available at https://github.com/Zed630/DistributionJailbreak.

[Long] Steganography Beyond Pixels: Reimagining Image Steganography as Cross-Modal Linguistic Communication ► *Underline Video*

Ren Lijing and Denghui Zhang

The rising sophistication of digital surveillance poses hurdles for concealing sensitive data within innocuous communication channels. Conventional image steganography relies on detectable pixel-level perturbations. In this paper, we introduce a novel steganography framework that fundamentally reorients the steganographic containers from the visual domain to the linguistic domain. To seamlessly bridge the gap from raw pixels to discriminative logits, we leverage the reversible latent space of discrete diffusion models to compress high-resolution secret images into lightweight binary payloads. The semantic stability of textual data ensures the integrity of the hidden payload across diverse platforms. Extensive evaluations confirm that this cross-modal approach establishes a superior equilibrium between embedding capacity and statistical undetectability in comparison to existing paradigms.

[Long] Stop Hardening Everything: A Training-Free Neuron-Level Defense for Neural Ranking Models ► *Underline Video*

Yu-An Liu, Ruqing Zhang, Hongru Song, Jiafeng Guo, Yixing Fan and Xueqi Cheng

While neural ranking models (NRMs) have achieved state-of-the-art performance in information retrieval, they remain highly vulnerable to imperceptible adversarial perturbations. Existing defenses are predominantly data-centric, exemplified by adversarial training, which requires constructing large collections of adversarial examples. By treating NRMs as black boxes and indiscriminately optimizing all model parameters, these methods incur substantial computational cost and often degrade performance on clean data due to overfitting. In this paper, we advocate that adversarial vulnerability is not uniformly distributed across model parameters, but instead originates from specific internal units. We propose a paradigm shift toward a model-centric defense that addresses vulnerability at its architectural source, without requiring costly retraining or adversarial data generation. Specifically, we introduce Search in the Model, a novel training-free framework that performs fine-grained identification and rectification of vulnerable neurons directly within the model. By formulating neuron identification as a ranking problem, we develop a maximum marginal vulnerability criterion to precisely locate the top-K neurons most responsible for model vulnerability, and apply targeted neuronal inverse perturbation to correct them. Extensive experiments on MS MARCO and TREC 19 show that our approach outperforms state-of-the-art baselines in both defense efficiency and robustness to seen and unseen attacks, while preserving strong performance on clean data.

[Findings] SycoBench-600: Measuring Sycophancy and Correction Selectivity in LLM Assistants ► *Underline Video*

Debu Sinha

Modern instruction-following language models are optimized to be helpful and cooperative, often through preference-based alignment such as RLHF and related methods. A growing body of evidence shows that this training can also induce sycophancy: models may agree with a user even when the user is wrong, undermining reliability in decision support and high-stakes advice. We introduce SycoBench-600, a controlled multiple-choice benchmark that measures (i) susceptibility to three social-pressure perturbations (doubt, authority, and an explicit wrong suggestion) and (ii) correction selectivity, the ability to accept correct suggestions while resisting incorrect ones. The released benchmark contains 600 English MCQ instances over 272 normalized question stems, covers 8 domains and 3 difficulty tiers, and evaluates each instance under 3 fixed paraphrase variants of the perturbation prompts. We evaluate seven widely used assistants spanning proprietary and open-weight families. Results show substantial variation in pressure robustness and selective updating, and further show that willingness to update does not by itself imply selectivity. We release raw logs, validation scripts, and code that regenerates every table and figure from the model outputs.

[Long] Tailoring Diagnostic Modeling to Individual Learners: Personalized Distractor Generation via MCTS-Guided Reasoning Reconstruction ► *Underline Video*

Tao Wu, Jingyuan Chen, Wang Lin, Jian Zhan, Mengze Li, Fangzhou Jin, Min Zhang, Kun Kuang and Fei Wu

Distractors, incorrect yet plausible answer choices in multiple-choice questions (MCQs), are vital in educational assessments, as they help identify student misconceptions by presenting potential reasoning errors. Current distractor generation methods typically produce shared distractors for all students, ignoring the individual variations in reasoning, which limits their diagnostic effectiveness. To tackle this challenge, we introduce the task of Personalized Distractor Generation, which tailors distractors to each student's specific cognitive flaws,

inferred from their past question-answering (QA) history. While promising, this task is particularly demanding due to the limited number of QA records available for each student, which are insufficient for training, as well as the absence of their underlying reasoning process. To overcome this, we propose a novel, training-free two-stage framework. In the first stage, Monte Carlo Tree Search (MCTS) is used to reconstruct the student's reasoning process from past errors, creating a student-specific misconception prototype. In the second stage, this prototype guides the simulation of the student's reasoning on new questions, generating personalized distractors that resonate with their individual misconceptions. Our experiments, conducted on 1,361 students across 6 subjects, demonstrate that this approach outperforms existing methods in generating plausible, personalized distractors, and also effectively adapts to group-level settings, highlighting its robustness and versatility.

[Long] Taming "Zombie" Agents: A Markov State-Aware Framework for Resilient Multi-Agent Evolution ▶ *Underline Video*

Taolin Zhang, Pukun Zhao, Qizhou Chen, Jiuheng Wan, Chen Chen, Xiaofeng He, Chengyu Wang and Richang Hong

Recent advancements in LLM-based multi-agent systems have demonstrated remarkable collaborative capabilities across complex tasks. To enhance the overall efficiency, existing methods often rely on aggressive graph topology evolution for agents (e.g., node or edge pruning), which risks prematurely discarding valuable agents due to transient issues such as hallucinations or temporary knowledge gaps. However, such hard pruning overlooks the potential for "zombie" agents to recover and contribute in subsequent discussion rounds. In this paper, we propose AgentRevive, a Markov state-aware framework for resilient multi-agent evolution. Our approach dynamically manages agent collaboration through soft state transitions, implemented via two key components: (1) State-Aware Policy Learning: Agent states are divided into "Active", "Standby", and "Terminated", selectively propagating messages based on agent memory. The policy employs a risk estimator to optimize agent state transitions by assessing hallucination risk, minimizing the influence of unreliable nodes while safeguarding valuable ones. (2) State-Aware Edge Optimization: Subgraph edges are pruned according to states learned from the policy, permanently removing "Terminated" nodes and retaining "Standby" nodes for subsequent rounds to observe potential future contributions. Extensive experiments on general reasoning, domain-specific, and hallucination challenge tasks show that our method consistently outperforms strong baselines and significantly reduces token consumption through state-aware agent scheduling.

[Long] TEA-Bench: A Systematic Benchmarking of Tool-enhanced Emotional Support Dialogue Agent ▶ *Underline Video*

Xingyu Sui, Yanyan Zhao, Yulin Hu, Jiahe Guo, Weixiang Zhao and Bing Qin

Emotional Support Conversation requires not only affective expression but also grounded instrumental support to provide trustworthy guidance. However, existing ESC systems and benchmarks largely focus on affective support in text-only settings, overlooking how external tools can enable factual grounding and reduce hallucination in multi-turn emotional support. We introduce TEA-Bench, the first interactive benchmark for evaluating tool-augmented agents in ESC, featuring realistic emotional scenarios, an MCP-style tool environment, and process-level metrics that jointly assess the quality and factual grounding of emotional support. Experiments on nine LLMs show that tool augmentation generally improves emotional support quality and reduces hallucination, but the gains are strongly capacity-dependent: stronger models use tools more selectively and effectively, while weaker models benefit only marginally. We further release TEA-Dialog, a dataset of tool-enhanced ESC dialogues, and find that supervised fine-tuning improves in-distribution support but generalizes poorly. Our results underscore the importance of tool use in building reliable emotional support agents.

[Long] TEA-Bench: A Systematic Benchmarking of Tool-enhanced Emotional Support Dialogue Agent ▶ *Underline Video*

Xingyu Sui, Yanyan Zhao, Yulin Hu, Jiahe Guo, Weixiang Zhao and Bing Qin

Emotional Support Conversation requires not only affective expression but also grounded instrumental support to provide trustworthy guidance. However, existing ESC systems and benchmarks largely focus on affective support in text-only settings, overlooking how external tools can enable factual grounding and reduce hallucination in multi-turn emotional support. We introduce TEA-Bench, the first interactive benchmark for evaluating tool-augmented agents in ESC, featuring realistic emotional scenarios, an MCP-style tool environment, and process-level metrics that jointly assess the quality and factual grounding of emotional support. Experiments on nine LLMs show that tool augmentation generally improves emotional support quality and reduces hallucination, but the gains are strongly capacity-dependent: stronger models use tools more selectively and effectively, while weaker models benefit only marginally. We further release TEA-Dialog, a dataset of tool-enhanced ESC dialogues, and find that supervised fine-tuning improves in-distribution support but generalizes poorly. Our results underscore the importance of tool use in building reliable emotional support agents.

[Long] TEMA: Anchor the Image, Follow the Text for Multi-Modification Composed Image Retrieval ▶ *Underline Video*

Zixu Li, Yupeng Hu, Zhiheng Fu, Zhiwei Chen, Yongqi Li and Liqiang Nie

Composed Image Retrieval (CIR) is an important image retrieval paradigm that enables users to retrieve a target image using a multimodal query that consists of a reference image and modification text. Although research on CIR has made significant progress, prevailing setups still rely simple modification texts that typically cover only a limited range of salient changes, which induces two limitations highly relevant to practical applications, namely Insufficient Entity Coverage and Clause-Entity Misalignment. In order to address these issues and bring CIR closer to real-world use, we construct two instruction-rich multi-modification datasets, M-FashionIQ and M-CIRR. In addition, we propose TEMA, the Text-oriented Entity Mapping Architecture, which is the first CIR framework designed for multi-modification while also accommodating simple modifications. Extensive experiments on four benchmark datasets demonstrate that TEMA's superiority in both original and multi-modification scenarios, while maintaining an optimal balance between retrieval accuracy and computational efficiency. Our codes and constructed multi-modification dataset (M-FashionIQ and M-CIRR) are available at https://github.com/lee-zixu/ACL26-TEMA/

[Findings] The Best of Both Worlds: Combining Parallel and Sequential Inference Scaling via Aggregation Fine-Tuning ▶ *Underline Video*

Yafu Li, Zhilin Wang, Tingchen Fu, Ganqu Cui, Sen Yang and Yu Cheng

Scaling data and model size has been proven effective for boosting the performance of large language models. In addition to training-time scaling, recent studies have revealed that increasing test-time computational resources can further improve performance. In this work, we introduce Aggregation Fine-Tuning (AFT), a supervised fine-tuning paradigm where the model learns to synthesize multiple draft responses, referred to as proposals, into a single, refined answer, termed aggregation. At inference time, we apply a propose-and-aggregate strategy that iteratively generates and aggregates proposals, effectively scaling inference-time computation without relying on external guidance such as a reward model. Empirical results across benchmark datasets demonstrate that AFT-trained models achieve substantial gains with test-time scaling, outperforming best-of-N baselines while eliminating the need for external reward signals. Notably, an AFT model, fine-tuned from Llama3.1-8B-Base with only 64k data, achieves a 41.3% LC win rate on AlpacaEval 2, surpassing significantly larger LLMs such as Llama3.1-405B-Instruct and GPT-4. By combining sequential refinement and parallel sampling, the propose-and-aggregate framework scales inference-time computation in a flexible manner.

[Long] The GaoYao Benchmark: A Comprehensive Framework for Evaluating Multilingual and Multicultural Abilities of Large Language Models ▶ *Underline Video*

Yilun Liu, Chunguang Zhao, Mengyao Piao, Lingqi Miao, Shimin Tao, Minggui He, Chenxin Liu, Zhang Li, Mahongxia Mahongxia, Jiaxin Guo, Chen Liu, Liqun Deng, Jiansheng Wei, Xiaojun Meng, Fanyi Du, Daimeng Wei and Yanghua Xiao

Evaluating the multilingual and multicultural capabilities of Large Language Models (LLMs) is essential for their global utility. However, current benchmarks face three critical limitations: (1) fragmented evaluation dimensions that often neglect deep cultural nuances; (2) insufficient language coverage in subjective tasks relying on low-quality machine translation; and (3) shallow analysis that lacks diagnostic depth beyond simple rankings. To address these, we introduce GaoYao, a comprehensive benchmark with 182.3k samples, 26 languages and 51 nations/areas. First, GaoYao proposes a unified framework categorizing evaluation tasks into three cultural layers (General Multilingual, Cross-cultural, Monocultural) and nine cognitive sub-layers. Second, we achieve native-quality expansion by leveraging experts to rigorously localize subjective benchmarks into 19 languages and synthesizing cross-cultural test sets for 34 cultures, surpassing prior coverage by up to 111%. Third, we conduct an in-depth diagnostic analysis on 20+ flagship and compact LLMs. Our findings reveal significant geographical performance disparities and distinct gaps between tasks, offering a reliable map for future work. We release the benchmark.

[Findings] The Mark Fades: Adaptive Evolutionary Paraphrase-based Attack against LLM Watermarks ▶ *Underline Video*

Yusheng Zhao, Jian Zhao, Tianle Zhang, Feng Wei and Xuelong Li

While LLM watermarking is essential for machine- generated content identification, existing paraphrase-based attacks struggle to balance watermark removal efficacy with text quality. We propose TSAPA, a training-free evolutionary framework that models watermark removal as a constrained multi-objective optimization problem. By leveraging genetic algorithms to navigate the Pareto front, TSAPA utilizes a Pseudo-Log-Likelihood (PLL)-guided mutation to precisely target and modify watermark-carrying tokens. Experiments on Qwen3 series (1.7B/8B/32B) across multiple watermark schemes show that TSAPA achieves over 90% attack success rate (ASR) while maintaining high text semantic fidelity, significantly outperforming baselines methods. This work exposes critical vulnerabilities in current watermarks and provides a new perspective for robust evaluation.

[Long] The Paradox of Outcome Optimization: A Causal Information-Theoretic Bound on Reasoning Shortcuts in LLMs ▶ *Underline Video*

Zihan Chen, Yiming Zhang, Wenxiang Geng, Zenghui Ding and Yining Sun

Large Language Models (LLMs) aligned via outcome-based Reinforcement Learning (RL) frequently exhibit a critical failure mode: they achieve high performance on in-distribution benchmarks while demonstrating brittle reasoning capabilities on out-of-distribution (OOD) tasks. We term this phenomenon Reward-Induced Manifold Collapse. We establish a theoretical framework bridging Structural Causal Models (SCM) and the Information Bottleneck (IB) principle to explain this paradox. We define reasoning as a high-complexity causal process and shortcut learning as the exploitation of low-complexity spurious correlations. Under the implicit inductive bias of Stochastic Gradient Descent (SGD), models optimized for outcome rewards are biased toward shortcut solutions whenever the training distribution allows for a "Markovian Screening" of the true causal mechanism. We derive a new generalization bound based on Semantic Coverage Measure () rather than sample size, showing why data scaling on homogeneous distributions may fail to correct reasoning flaws. We also show that Process Reward Models (PRMs) function as Topological Filters, enforcing step-wise mutual information constraints that render the low-complexity shortcut manifold inadmissible. These results provide a mathematical grounding for the role of process supervision beyond simple credit assignment.

[Long] The Stackelberg Speaker: Optimizing Persuasive Communication in Social Deduction Games ▶ *Underline Video*

Zheng Zhang, Deheng Ye, Peilin Zhao and Hao Wang

Large language model (LLM) agents have shown remarkable progress in social deduction games (SDGs). However, existing approaches primarily focus on information processing and strategy selection, overlooking the significance of persuasive communication in influencing other players' beliefs and responses. In SDGs, success depends not only on making correct deductions but also on convincing others

to respond in alignment with ones intent. To address this limitation, we formalize turn-based dialogue in SDGs as a Stackelberg competition, where the current player acts as the leader who strategically influences the follower's response. Building on this theoretical foundation, we propose a reinforcement learning framework that trains agents to optimize utterances for persuasive impact. Through comprehensive experiments across four diverse social deduction benchmarks, we demonstrate that our agents significantly outperform baselines. This work represents a significant step toward developing AI agents capable of strategic social influence, with implications extending to scenarios requiring persuasive communication. Our code and data are available at https://3dagentworld.github.io/leader_follower.

[Long] The Knowledge-Behavior Gap in Cultural Taboo Safety of Large Language Models ▶ *Underline Video*

Ying He, Sihang Jiang, Xingzhou Chen, Zhouhong Gu, Yiwei Gu, Minggui HE, Shimin Tao, Hongxia Ma and Yanghua Xiao

Cultural taboo safety is essential for deploying large language models (LLMs), as culturally insensitive outputs may cause offense or even social harm. However, existing cultural benchmarks primarily assess cultural knowledge or values biases, while overlooking whether LLMs can recognize and respect cultural taboos, especially when taboos are implicitly hidden in seemingly harmless questions. Besides, cultural taboos are implicit, and context-dependent, thus poss unique challenges for reliable evaluation. To address these gaps, we introduce CulShield, the first public benchmark dedicated to evaluating and improving the cultural taboo safety of LLMs. CulShield spans 77 countries and regions, and includes over 2,020 taboos. It evaluates models along both explicit knowledge and implicit behaviors.Experiments on several advanced LLMs (e.g., GPT-4o-mini, Gemini-2.5-pro) reveal a clear "knowledge-behavior gap": models often fail to apply known taboos during interaction. We further show that variations in linguistic context can significantly affect LLMs' cultural taboo safety. Code and data is accessible here: https://anonymous.4open.science/r/CulShield-7A0E.

[Findings] Thinking Traps in Long Chain-of-Thought: A Measurable Study and Trap-Aware Adaptive Restart ▶ *Underline Video*

Kang Chen, Fan Yu, Junjie Nian, Shihan Zhao, Zhuoka Feng, Zijun Yao, Heng Wang, Minshen Yu and Yixin Cao

Scaling test-time compute via Long Chain-of-Thought (Long-CoT) significantly enhances reasoning capabilities, yet extended generation does not guarantee correctness: after an early wrong commitment, models may keep elaborating a self-consistent but incorrect prefix. Through fine-grained trajectory analysis, we identify Thinking Traps, prefix-dominant deadlocks where later reflection, alternative attempts, or verification fails to revise the root error. On a curated subset of DAPO-MATH, 89% of failures exhibit such traps. To solve this problem, we introduce TAAR (Trap-Aware Adaptive Restart), a test-time control framework that trains a diagnostic policy to predict two signals from partial trajectories: a trap index for where to truncate and an escape probability for whether and how strongly to intervene. At inference time, TAAR truncates the trajectory before the predicted trap segment and adaptively restarts decoding; for severely trapped cases, it applies stronger perturbations, including higher-temperature resampling and an optional structured reboot suffix. Experiments on challenging mathematical and scientific reasoning benchmarks (AIME24, AIME25, GPQA-Diamond, HMMT25, BRUMO25) show that TAAR improves reasoning performance without fine-tuning base model parameters.

[Long] Through the Magnifying Glass: Adaptive Perception Magnification for Hallucination-Free VLM Decoding ▶ *Underline Video*

Shunqi Mao, Chaoyi Zhang and Weidong Cai

Existing vision-language models (VLMs) often suffer from visual hallucination, where the generated responses contain inaccuracies that are not grounded in the visual input. Efforts to address this issue without model finetuning primarily mitigate hallucination by contrastively reducing language biases or amplifying the weights of visual embedding during decoding. However, these approaches remain limited in their ability to capture fine-grained visual details. In this work, we propose the Perception Magnifier (PM), a novel visual decoding method that iteratively isolates relevant visual tokens based on attention and magnifies the corresponding regions, spurring the model to concentrate on fine-grained visual details during decoding. By magnifying critical regions while preserving the structural and contextual information at each decoding step, PM allows the VLM to enhance its scrutiny of the visual input, hence producing more accurate and faithful responses. Extensive experimental results demonstrate that PM not only achieves superior hallucination mitigation but also enhances language generation while preserving strong reasoning capabilities.

[SRW] TokLens: A Multilingual Lens on Tokenizer Quality for LLMs ▶ *Underline Video*

Guan-Ming Chiu

Tokenizer design is a foundational but underexamined component of large language model (LLM) development. We introduce TokLens, an open-source toolkit for evaluating tokenizer quality across languages using six intrinsic metrics: fertility, characters per token, compression ratio, normalized sequence length, single-token retention rate, and cross-lingual parity. We evaluate 24 tokenizers from major LLM families across 15 typologically diverse languages and correlate intrinsic metrics with downstream performance on the Open LLM Leaderboard v2. Our analysis reveals stark cross-lingual disparities: GPT-2's tokenizer produces 56x more tokens per word in Japanese than in English, while newer tokenizers like Qwen2.5 and Gemma-2 reduce this gap to under 4x. No intrinsic metric predicts English benchmark performance after controlling for model size. However, when we extend the analysis to multilingual benchmarks (MMLU-ProX) using linear mixed-effects models that account for repeated measures within models, tokenizer metrics significantly predict per-language performance (STRR: beta = +5.7, $z = 18.5$, $p < 0.001$). A controlled experiment with the Qwen2.5 family, which shares a single tokenizer across five model sizes, reveals that languages with higher single-token retention rate benefit more from scaling (= 0.91, $p < 0.001$). These results suggest that tokenizer quality has a measurable impact on multilingual LLM performance, even when model architecture and training pipeline are held constant.

[Long] Towards Interpretable Tabular Reasoning: Enhancing LLM Reasoning on Tabular Data with Pre-Constructed

Logic Graph ▶ *Underline Video*

Lirong Gao, Zewei Yu, Zhongrui Yin, Qi Zhang, Yuke Zhu, Bo Zheng, Haobo Wang, Junbo Zhao, Gang Chen and Sheng Guo

Tabular data is widely used in fields such as finance and healthcare. Traditional tree-based models are prevalent for tabular prediction tasks due to their ability to handle heterogeneous features. However, their heavy reliance on feature engineering limits both their generalizability and their human-readable interpretability. On the other hand, Large Language Models (LLMs) naturally provide intermediate reasoning steps, thus offering greater transparency in decision-making. Nevertheless, LLMs often fail to match the predictive performance of tree-based models on tabular data. To address these challenges, we propose a novel Logic-Graph-Enhanced LLM Reasoning (LogGER) framework that integrates the strengths of tree-based models and LLMs. Specifically, we reformulate the traditional decision tree as a human-readable logic graph, which explicitly models the causal relationships between features and targets. This logic graph is automatically constructed using LLMs based on data priors and serves as the foundation for LogGER. To fully leverage the logic graph, we further introduce a logic-graph-guided process supervision approach, which evaluates and enhances the quality of LLMs intermediate reasoning steps using logic-graph-aided process reward. Extensive experiments demonstrate that LogGER consistently outperforms both tree-based models and state-of-the-art LLM methods on a variety of tabular prediction tasks, achieving superior accuracy and interpretability.

[Long] Training LLMs for Divide-and-Conquer Reasoning Elevates Test-Time Scalability ▶ *Underline Video*

Xiao Liang, Zhong-Zhi Li, Zhenghao Lin, Eric Hanchen Jiang, Hengyuan Zhang, yelong shen, Kai-Wei Chang, Ying Nian Wu, Yeyun Gong and Weizhu Chen

Large language models (LLMs) have demonstrated strong reasoning capabilities through step-by-step chain-of-thought (CoT) reasoning. Nevertheless, at the limits of model capability, CoT often proves insufficient, and its strictly sequential nature constrains test-time scalability. A potential alternative is divide-and-conquer (DAC) reasoning, which decomposes a complex problem into subproblems to facilitate more effective exploration of the solution space. Although promising, our analysis reveals a fundamental misalignment between general-purpose post-training and DAC-style inference, which limits the models capacity to fully leverage this potential. To bridge this gap and fully unlock LLMs reasoning capabilities on the most challenging tasks, we propose an end-to-end reinforcement learning (RL) framework to enhance their DAC-style reasoning capacity. At each step, the policy decomposes a problem into a group of subproblems, solves them sequentially, and addresses the original problem conditioned on the subproblem solutions, with both decomposition and solution integrated into RL training. Under comparable training settings, our DAC-style framework endows the model with a higher performance ceiling and stronger test-time scalability, surpassing CoT by 8.6% in Pass@1 and 6.3% in Pass@32 on competition-level benchmarks. The code is available at the https://github.com/MasterVito/DAC-RL.

[Long] Turning Failures into Value: Negative Experience Replay for RLVR via Confidence Gating and Boundary Failure Sampling ▶ *Underline Video*

Jialiang Guo, Fucheng Xiong, Xu He, Haodong Zhao, Xingyang Li, Ke Zeng and Xunliang Cai

Reinforcement Learning with Verifiable Rewards (RLVR) has become the standard paradigm for enhancing reasoning capabilities in Large Language Models, yet on-policy algorithms like GRPO suffer from sample inefficiency. Current experience replay methods for RLVR typically replay correct trajectories to consolidate learned reasoning patterns and accelerate convergence, but overlook the vast failure space. This work investigates how to effectively replay failure trajectories. We find that the high heterogeneity of failures renders random replay ineffective, and that high-value negatives should be both gradient-efficient and structurally proximal to correct solutions. To this end, we propose NexGRPO, which employs mid-confidence gating to filter invalid noise and saturated errors, and utilizes boundary failure sampling to retrieve boundary errors semantically similar to correct solutions for targeted refinement. Extensive experiments on mathematical and general reasoning benchmarks demonstrate that NexGRPO outperforms strong baaselines and achieves improved out-of-distribution generalization.

[Long] Typology-Aware Multilingual Morphosyntactic Parsing with Joint Abstract Node Modeling ▶ *Underline Video*

Kutay Acar and Gülen Eryiit

The UniDive 2025 Morphosyntactic Parsing (MSP) shared task introduces a representation unifying dependency structure, morphological features, and unrealized arguments. Unlike Universal Dependencies, MSP encodes abstract nodes (e.g., dropped subjects, implicit pronouns) as labels projected onto content words, which standard UD parsers cannot model. We present a multilingual, typology-aware joint system integrating word-type prediction, content-only parsing, morphological tagging, and an abstract-node component within a single architecture. The model combines the baseline joint framework with typology-conditioned adapters and progressive weighting for abstract supervision. On the MSP test set, our model outperforms the leading submission by 3.23 percentage points in MSLAS, 3.35 in LAS, and 1.78 in FEATS macro F1, demonstrating the effectiveness of typology-sensitive multi-task learning in MSP.

[Demo] UltraEval-Audio: A Unified Framework for Comprehensive Evaluation of Audio Foundation Models ▶ *Underline Video*

Qundong Shi, Jie Zhou, Biyuan Lin, Junbo Cui, Guoyang Zeng, Yixuan Zhou, Ziyang Wang, Xin Liu, Zhen Luo, Yudong Wang and Zhiyuan Liu;

The development of audio foundation models has accelerated rapidly since the emergence of GPT-4o. However, the lack of comprehensive evaluation has become a critical bottleneck for further progress in the field, particularly in audio generation. Current audio evaluation faces three major challenges: (1) audio evaluation lacks a unified framework, with datasets and code scattered across various sources; (2) audio codec, as a key component of audio foundation models, lacks a widely accepted and holistic evaluation methodology;

(3) existing speech benchmarks are heavily reliant on English, making it challenging to objectively assess models' performance on Chinese. We introduce UltraEval-Audio, a unified framework addressing these challenges through a modular architecture supporting 10 languages, 14 task categories, 24 models, and 36 benchmarks with one-command evaluation and real-time leaderboards. For audio codec, we propose a three-dimensional evaluation scheme covering semantic accuracy, timbre fidelity, and acoustic quality. For Chinese evaluation, we introduce two new benchmarks: SpeechCMMLU and SpeechHSK. Our code, benchmarks, and leaderboards are available at https://github.com/OpenBMB/UltraEval-Audio.

[Findings] UniCM: A Unified Consistency Model For Efficient Multimodal Generation and Understanding ► *Underline Video*
Chenkai Xu, Xu Wang, Zhenyi Liao, Yishun Li, TianQi Hou and Zhijie Deng
Consistency models (CMs) have shown promise in the efficient generation of both image and text. This raises the natural question of whether we can learn a unified CM for efficient multimodal generation (e.g., text-to-image) and understanding (e.g., image-to-text). Intuitively, such a model could be acquired by applying the consistency distillation (CD) to existing unified multimodal models. However, the key challenge is establishing a unified denoising perspective for both image and text generation, which is essential for establishing the consistency mapping. To tackle this, at the representation level, we advocate for discrete tokens for both modalities to best preserve language modeling capabilities. Critically, instead of defining the text denoising trajectory via recent discrete diffusion language modeling principles, we specify it using the parallel decoding trace of an autoregressive language model, benefiting from the latter's superior performance in general text generation tasks. The denoising trajectory of image tokens adheres to standard discrete diffusion. We train our unified consistency models (UniCMs) on these combined multimodal trajectories simultaneously with a unified objective. We introduce a trajectory segmentation strategy to improve the training convergence. Empirically, in text-to-image generation, UniCMs outperform SD3 on GenEval and Image Reward, while requiring only approximately 1/8 of the sampling time. Meanwhile, in image-to-text generation, UniCMs surpass Show-o on the MMMU benchmark while being 1.5 x faster at long-sequence generating speed.

[Long] UniCorn: Towards Self-Improving Unified Multimodal Models through Self-Generated Supervision ► *Underline Video*
Zhen Fang, Ruiyan Han, XinYu Sun, Yuchen Ma, Ziheng Wang, Yu Zeng, Zehui Chen, Lin Chen, Wenxuan Huang, Wei-Jie Xu, Yi Cao and Feng Zhao
While Unified Multimodal Models (UMMs) have achieved remarkable success in cross-modal comprehension, a significant gap persists in their ability to leverage such internal knowledge for high-quality generation. We formalize this discrepancy as Conduction Aphasia, a phenomenon where models accurately interpret multimodal inputs but struggle to translate that understanding into faithful and controllable synthesis. To address this, we propose UniCorn, a simple yet elegant self-improvement framework that eliminates the need for external data or teacher supervision. By partitioning a single UMM into three collaborative roles: Proposer, Solver, and Judge, UniCorn generates high-quality interactions via self-play and employs cognitive pattern reconstruction to distill latent understanding into explicit generative signals. To validate the restoration of multimodal coherence, we introduce UniCycle, a cycle-consistency benchmark based on a Text to Image to Text reconstruction loop. Extensive experiments demonstrate that UniCorn achieves comprehensive and substantial improvements over the base model across six general image generation benchmarks. Notably, it achieves SOTA performance on TIIF(73.8), DPG(86.8), CompBench(88.5), and UniCycle while further delivering substantial gains of +5.0 on WISE and +6.5 on OneIG. These results highlight that our method significantly enhances T2I generation while maintaining robust comprehension, demonstrating the scalability of fully self-supervised refinement for unified multimodal intelligence.

[Long] UniMoE-Audio: Unified Speech and Music Generation with Dynamic-Capacity Mixture-of-Experts ► *Underline Video*
Zhenyu Liu, Yunxin Li, Xuanyu Zhang, Qixun Teng, Shenyuan Jiang, Xinyu Chen, Haoyuan Shi, Haolan Chen, Fanbo Meng, Minjun Zhao, Yu Xu, Yancheng He, Baotian Hu, Haizhou Li and Min Zhang
Recent advances in unified multimodal models indicate a clear trend towards comprehensive content generation. However, the auditory domain remains a significant challenge, with music and speech often developed in isolation, hindering progress towards universal audio synthesis. This separation stems from inherent task conflicts between semantic speech and structural music modeling, and severe data imbalances, which impede the development of a truly unified model. To address these challenges, we propose UniMoE-Audio, a unified speech and music generation model built upon a novel Dynamic-Capacity Mix-of-Experts (DCMoE) framework. Architecturally, UniMoE-Audio extends the conventional MoE paradigm by introducing a Top-P routing strategy for adaptive capacity allocation. To tackle data imbalance, we introduce a three-stage training curriculum: 1) Independent Specialist Training leverages original datasets to instill domain-specific knowledge into each specialists without interference; 2) MoE Integration and Warmup incorporates these specialists into the UniMoE-Audio architecture, warming up the gate module and shared expert using a subset of balanced dataset; and 3) Synergistic Joint Training trains the entire model end-to-end on the fully balanced dataset, fostering enhanced cross-domain synergy. Extensive experiments show that UniMoE-Audio not only achieves state-of-the-art performance on major speech and music generation benchmarks, but also demonstrates superior synergistic learning, mitigating the performance degradation typically seen in naive joint training. Our findings highlight the substantial potential of specialized MoE architecture and curated training strategies in advancing universal audio generation.

[Industry] Unleashing Low-Bit Inference on Ascend NPUs: A Comprehensive Evaluation of HiFloat Formats ► *Underline Video*
Pengxiang Zhao, Hui-Ling Zhen, Xing Li, Han Bao, Weizhe Lin, Zhiyuan Yang, Yu Zi Wei, Xin Wang, Mingxuan Yuan, Xianzhi Yu and Zhenhua Dong
As LLMs scale, low-bit floating-point formats like MXFP and NVFP4 offer new opportunities for precision and efficiency. In this work, we

evaluate HiFloat (HiF8 and HiF4), a family of formats tailored for Ascend NPUs. Through rigorous comparison across weight-activation and KV-cache tasks, we provide three key insights: (1) INT8 suits narrow-range data, while floating-point formats excel with high-variance data; (2) in 4-bit regimes, HiF4s hierarchical scaling prevents the accuracy collapse seen in integer formats; and (3) HiFloat is fully compatible with state-of-the-art post-training quantization frameworks. Overall, HiFloat provides a solution for high-efficiency LLM inference on NPUs.

[Long] v-HUB: A Benchmark for Video Humor Understanding from Vision and Sound ▶ *Underline Video*

Zhengpeng Shi, Yanpeng Zhao, Jianqun Zhou, Yuxuan Wang, Qinrong Cui, Wei Bi, Songchun Zhu, Bo Zhao and Zilong Zheng

AI models capable of comprehending humor hold real-world promise—for example, enhancing engagement in human-machine interactions. To gauge and diagnose the capacity of multimodal large language models (MLLMs) for humor understanding, we introduce v-HUB, a novel video humor understanding benchmark. v-HUB comprises a curated collection of non-verbal short videos, reflecting real-world scenarios where humor can be appreciated purely through visual cues. We pair each video clip with rich annotations to support a variety of evaluation tasks and analyses, including a novel study of environmental sound that can enhance humor. To broaden its applicability, we construct an open-ended QA task, making v-HUB readily integrable into existing video understanding task suites. We evaluate a diverse set of MLLMs, from specialized Video-LLMs to versatile OmniLLMs that can natively process audio, covering both open-source and proprietary domains. The experimental results expose the difficulties MLLMs face in comprehending humor from visual cues alone. Our findings also demonstrate that incorporating audio helps with video humor understanding, highlighting the promise of integrating richer modalities for complex video understanding tasks.

[Long] VAPO: End-to-end Slide-Enhanced Speech Recognition with Omni-modal Large Language Models ▶ *Underline Video*

Rui Hu, Delai Qiu, Yining Wang, Shengping Liu and Jitao Sang

Omni-modal large language models (OLLMs) offer a promising end-to-end solution for slide-enhanced speech recognition due to their inherent multimodal capabilities. However, we found a fundamental issue faced by OLLMs: Visual Interference, where models show a bias towards visible text over auditory signals, causing them to hallucinate slide content that was never spoken. To address this, we propose Visually-Anchored Policy Optimization (VAPO), which aims to reshape models' inference process to follow the human-like "Look-then-Listen" inference chain. Specifically, we design a temporally decoupled policy: the model first extracts visual priors in a \{}<think> block to serve as semantic anchors, then generates the transcription in an \{}<answer> block. The policy is optimized via multi-objective reinforcement learning. Furthermore, we introduce SlideASR-Bench, a comprehensive benchmark designed to address the scarcity of entity-rich data, comprising a large-scale synthetic corpus for training and a challenging real-world test set for evaluation. We conduct extensive evaluations demonstrating that VAPO effectively eliminates visual interference and achieves state-of-the-art performance on SlideASR-Bench and public datasets, significantly reducing entity recognition errors in specialized domains.

[Long] Vector Calligrapher: Generating Scalable Vector Graphics via Structured Linguistic Supervision ▶ *Underline Video*

Bo Zhou, Xikang Chen, Yan Gong and Yin Zhang

Generating SVG-based fonts requires Multimodal Large Language Models (MLLMs) to translate high-level linguistic intent into low-level, topologically constrained symbolic sequences. However, current approaches struggle with two fundamental misalignments: the semantic ambiguity of unstructured natural language for precise geometric control, and the inefficiency of generic text tokenizers, which fragment coordinate-dense SVG XML into excessively long sequences with low information density. In this work, we propose Vector Calligrapher, a system that treats SVG generation as a conditional language modeling task optimized for both semantic grounding and representational efficiency. To bridge the semantic gap, we introduce a structured linguistic supervision Font Description Framework that decomposes typographic style into interpretable linguistic dimensions (e.g., historical lineage, affective metaphors), providing structured supervision aligned with the compositional syntax of SVG. To address the tokenization bottleneck, we design a scalable separated-coordinate strategy that bypasses the vocabulary explosion of flattened tokens while significantly compressing sequence length. Supported by VectorFont, a dataset of over 10 million hierarchically annotated glyphs, our approach improves CLIP score by +23%, reduces geometric error by approx 48%, and boosts generation efficiency by achieving an 18% Command-per-Token (C/T) ratio—a 6x increase in information density over standard baselines. These results demonstrate that combining structured linguistic supervision with efficient symbolic tokenization is essential for reliable, controllable vector graphics synthesis. VectorFont dataset, Code and model weights will be publicly released.

[Long] VideoCuRL: Video Curriculum Reinforcement Learning with Orthogonal Difficulty Decomposition ▶ *Underline Video*

Hongbo Jin, Kuanwei Lin, Wenhao Zhang, Yichen Jin and Ge Li

Reinforcement Learning (RL) is crucial for empowering Video-LLMs with complex spatiotemporal reasoning. However, current RL paradigms predominantly rely on random data shuffling or naive curriculum strategies based on scalar difficulty metrics. We argue that scalar metrics fail to disentangle two orthogonal challenges in video understanding: Visual-Temporal Perception Load and Cognitive Reasoning Depth. To address this, we propose VideoCuRL, a novel framework that decomposes difficulty into these two axes. We employ efficient, training-free proxies, optical flow/keyframe entropy for visual complexity and Calibrated Surprisal for cognitive complexity, to map data onto a 2D curriculum grid. A competence-aware Diagonal Wavefront strategy then schedules training from base alignment to complex reasoning. Furthermore, we introduce Dynamic Sparse KL and Structured Revisiting to stabilize training against reward collapse and catastrophic forgetting. Extensive experiments show that VideoCuRL surpasses strong RL baselines on reasoning (+2.5% on VSI-Bench) and perception (+2.9% on VideoMME) tasks. Notably, VideoCuRL eliminates the prohibitive inference overhead of generation-based curricula, offering a scalable solution for robust video post-training.

[Long] Visual Attention Reasoning via Hierarchical Search and Self-Verification ▶ *Underline Video*

Wei Cai, Jian Zhao, Yuchen Yuan, Tianle Zhang, Ming Zhu, Haichuan Tang and Xuelong Li

Multimodal Large Language Models (MLLMs) frequently hallucinate due to their reliance on fragile, linear reasoning and weak visual grounding. We propose Visual Attention Reasoning (VAR), a reinforcement learning framework that reformulates reasoning as a hierarchical search with self-verification. VAR enforces traceable evidence grounding by generating explicit bounding boxes, guided by a novel reward function combining geometric precision and semantic sufficiency. Furthermore, it replaces linear Chain-of-Thought with a tree-search policy capable of backtracking to correct logical errors. Theoretical analysis validates the framework's reliability, and extensive experiments demonstrate that VAR significantly outperforms state-of-the-art methods on complex hallucination and safety benchmarks.

[Long] WeightLoRA: Keep Only Necessary Adapters ▶ *Underline Video*

Andrey Veprikov, Vladimir Solodkin, Zyl Alexander, Andrey Savchenko and Aleksandr Beznosikov

The widespread utilization of language models in modern applications is inconceivable without Parameter-Efficient Fine-Tuning techniques, such as low-rank adaptation (LoRA), which adds trainable adapters to selected layers. Although LoRA may obtain accurate solutions, it requires significant memory to train large models and intuition on which layers to add adapters. In this paper, we propose a novel method, WeightLoRA, which overcomes this issue by adaptive selection of the most critical LoRA heads throughout the optimization process. As a result, we can significantly reduce the number of trainable parameters while maintaining the capability to obtain consistent or even superior metric values. We conduct experiments for a series of competitive benchmarks and DeBERTa, BART, Llama and Qwen models, comparing our method with different adaptive approaches. The experimental results demonstrate the efficacy of WeightLoRA and the superior performance of WeightLoRA+ in almost all cases. The source code is available at https://github.com/brain-lab-research/WLoRA

[Long] When Correct Beliefs Collapse: Epistemic Resilience of LLMs under Clinical Pressure ▶ *Underline Video*

Boyu Xiao, Xiuqi Tian, Xuwen Song, Haochun Wang, Guanchun Song, Sendong Zhao and Bing Qin

Despite strong medical benchmark accuracy, LLMs can exhibit severe multi-turn sycophancy in clinical dialogue, abandoning initial correct diagnosis under escalating pressure. We propose Med-Stress, a targeted stress test framework that evaluates belief stability under escalating pressure. Across nine frontier large language models (LLMs), we find a clear dissociation between medical knowledge and robustness: high initial diagnostic capability does not imply high belief stability, yielding large knowledge-robustness gaps for several LLMs. To mitigate this failure mode, we propose a lightweight inference-time defense, RBED (Role-Based Epistemic Defense), and R-FT (Resilience-oriented Fine-Tuning), a training-time approach that internalizes evidence-based resistance to pressure. Experiments show that R-FT nearly eliminates belief change and substantially improves robustness.

[Long] When in Doubt, Consult: Expert Debate for Sexism Detection via Confidence-Based Routing ▶ *Underline Video*

Anwar Alajmi and Gabriele Pergola

Online sexism increasingly appears in subtle, context-dependent forms that evade traditional detection methods. Its interpretation often depends on overlapping linguistic, psychological, legal, and cultural dimensions, which produce mixed and sometimes contradictory signals in annotated datasets. These inconsistencies, combined with label scarcity and class imbalance, result in unstable decision boundaries and cause fine-tuned models to overlook subtler, underrepresented forms of harm. To address these challenges, we propose a two-stage framework that unifies (i) targeted training procedures to better regularize supervision to scarce and noisy data with (ii) selective, reasoning-based inference to handle ambiguous or borderline cases. First, we stabilize the training combining class-balanced focal loss, class-aware batching, and post-hoc threshold calibration, strategies for the firs time adapted for this domain to mitigate label imbalance and noisy supervision. Second, we bridge the gap between efficiency and reasoning with a a dynamic routing mechanism that distinguishes between unambiguous instances and complex cases requiring a deliberative process. This reasoning process results in the novel Collaborative Expert Judgment (CEJ) module which prompts multiple personas and consolidates their reasoning through a judge model. Our approach outperforms existing approaches across several public benchmarks, with F1 gains of +4.48% and +1.30% on EDOS Tasks A and B, respectively, and a +2.79% improvement in ICM on EXIST 2025 Task 1.1.

[Long] When Is Thinking Enough? Early Exit via Sufficiency Assessment for Efficient Reasoning ▶ *Underline Video*

Yang Xiang, Yixin Ji, Ruotao Xu, Dan Qiao, Zheming Yang, Juntao Li and Min Zhang

Large reasoning models (LRMs) have achieved remarkable performance in complex reasoning tasks, driven by their powerful inference-time scaling capability. However, LRMs often suffer from overthinking, which results in substantial computational redundancy and significantly reduces efficiency. Early-exit methods aim to mitigate this issue by terminating reasoning once sufficient evidence has been generated, yet existing approaches mostly rely on handcrafted or empirical indicators that are unreliable and impractical. In this work, we introduce Dynamic Thought Sufficiency in Reasoning (DTSR), a novel framework for efficient reasoning that enables the model to dynamically assess the sufficiency of its chain-of-thought (CoT) and determine the optimal point for early exit. Inspired by human metacognition, DTSR operates in two stages: (1) Reflection Signal Monitoring, which identifies reflection signals as potential cues for early exit, and (2) Thought Sufficiency Check, which evaluates whether the current CoT is sufficient to derive the final answer. Experimental results on the Qwen3 models show that DTSR reduces reasoning length by 28.9%-34.9% with minimal performance loss, effectively mitigating overthinking. We further discuss overconfidence in LRMs and self-evaluation paradigms, providing valuable insights for early-exit reasoning.

[Long] When Personalization Legitimizes Risks: Uncovering Safety Vulnerabilities in Personalized Dialogue Agents ► *Underline Video*

Jiahe Guo, Xiangran Guo, Yulin Hu, Zimo Long, Xingyu Sui, Xuda Zhi, Yongbo Huang, Hao He, Weixiang Zhao, Yanyan Zhao and Bing Qin

Long-term memory enables large language model (LLM) agents to support personalized and sustained interactions. However, most work on personalized agents prioritizes utility and user experience, treating memory as a neutral component and largely overlooking its safety implications. In this paper, we reveal intent legitimation, a previously underexplored safety failure in personalized agents, where benign personal memories bias intent inference and cause models to legitimize inherently harmful queries. To study this phenomenon, we introduce PS-Bench, a benchmark designed to identify and quantify intent legitimation in personalized interactions. Across multiple memory-augmented agent frameworks and base LLMs, personalization increases attack success rates by 15.8%-243.7% relative to stateless baselines. We further provide mechanistic evidence for intent legitimation from internal representation space, and propose a lightweight detection-reflection method that effectively reduces safety degradation. Overall, our work provides the first systematic exploration and evaluation of intent legitimation as a safety failure mode that naturally arises from benign, real-world personalization, highlighting the importance of assessing safety under long-term personal context. WARNING: This paper may contain harmful content.

[Long] When Personalization Legitimizes Risks: Uncovering Safety Vulnerabilities in Personalized Dialogue Agents ► *Underline Video*

Jiahe Guo, Xiangran Guo, Yulin Hu, Zimo Long, Xingyu Sui, Xuda Zhi, Yongbo Huang, Hao He, Weixiang Zhao, Yanyan Zhao and Bing Qin

Long-term memory enables large language model (LLM) agents to support personalized and sustained interactions. However, most work on personalized agents prioritizes utility and user experience, treating memory as a neutral component and largely overlooking its safety implications. In this paper, we reveal intent legitimation, a previously underexplored safety failure in personalized agents, where benign personal memories bias intent inference and cause models to legitimize inherently harmful queries. To study this phenomenon, we introduce PS-Bench, a benchmark designed to identify and quantify intent legitimation in personalized interactions. Across multiple memory-augmented agent frameworks and base LLMs, personalization increases attack success rates by 15.8%-243.7% relative to stateless baselines. We further provide mechanistic evidence for intent legitimation from internal representation space, and propose a lightweight detection-reflection method that effectively reduces safety degradation. Overall, our work provides the first systematic exploration and evaluation of intent legitimation as a safety failure mode that naturally arises from benign, real-world personalization, highlighting the importance of assessing safety under long-term personal context. WARNING: This paper may contain harmful content.

[Long] Where Paths Split: Localized, Calibrated Control of Moral Reasoning in Large Language Models ► *Underline Video*

Chenchen Yuan, Zheyu Zhang and Gjergji Kasneci

Large language models often display heterogeneous moral preferences across settings. We study inference-time steering toward a desired ethical framework while preserving general competence. We present Convergent-Divergent Routing, which traces and edits minimal branch points inside transformer blocks where ethical-framework-related pathways first converge and then diverge. Gating non-target branches at these loci blocks the downstream propagation while leaving upstream computations intact. We find that this intervention alone increases targeted ethical-framework reasoning. To achieve fine-grained control, we adapt Common Spatial Patterns to the residual stream and extract, for each branch-point layer, a pair of directions that discriminate between utilitarian and deontological frameworks. We then introduce Dual Logit Calibration, a closed-form, minimum-_2-norm update that moves the residual within this two-dimensional subspace so the resulting directional projections align with user-specified preference weights. Experiments on real-life moral dilemmas show that our method reliably achieves preference calibration and largely preserves general capabilities, outperforming recent baselines while providing an interpretable mechanism.

[Long] Who Wrote This Line? Evaluating the Detection of LLM-Generated Classical Chinese Poetry ► *Underline Video*

Jiang Li, Tian Lan, Shanshan Wang, Zdongxing Zdongxing, Dianqing Lin, Guanglai Gao, Derek F. Wong and Xiangdong Su

The rapid development of large language models (LLMs) has extended text generation tasks into the literary domain. However, AI-generated literary creations has raised increasingly prominent issues of creative authenticity and ethics in literary world, making the detection of LLM-generated literary texts essential and urgent. While previous works have made significant progress in detecting AI-generated text, it has yet to address classical Chinese poetry. Due to the unique linguistic features of classical Chinese poetry, such as strict metrical regularity, a shared system of poetic imagery, and flexible syntax, distinguishing whether a poem is authored by AI presents a substantial challenge. To address these issues, we introduce ChangAn, a benchmark for detecting LLM-generated classical Chinese poetry that containing total 30,664 poems, 10,276 are human-written poems and 20,388 poems are generated by four popular LLMs. Based on ChangAn, we conducted a systematic evaluation of 12 AI detectors, investigating their performance variations across different text granularities and generation strategies. Our findings highlight the limitations of current Chinese text detectors, which fail to serve as reliable tools for detecting LLM-generated classical Chinese poetry. These results validate the effectiveness and necessity of our proposed ChangAn benchmark. Our dataset and code are available at https://github.com/VelikayaScarlet/ChangAn.

[Long] WSDPO: A Generative Word Sense Disambiguation Framework with Chain-of-Thought and Preference Optimization ► *Underline Video*

Kunpeng Kang, Shuaimin Li, Kaiyuan Zhang, Luyang Zhang, Jiasheng Si, Bing Xu, Kehai Chen, Muyun Yang and Wenpeng Lu

Word sense disambiguation (WSD) is a foundational task in natural language processing. Recent research has reformulated WSD for large

language models (LLMs) as a generative task, where the model produces a definition to convey the intended meaning of an ambiguous word in context. In practice, most existing approaches implement this formulation through straightforward supervised fine-tuning, which tends to prioritize superficial context-to-gloss memorization over true contextual sense discrimination, leading to degraded performance on less frequent senses (LFS), particularly in unseen settings. To address this issue, we propose WSDPO, a training framework for generative WSD with chain-of-thought (CoT) and preference optimization. WSDPO consists of three stages: (1) disambiguation-aware CoT construction, which produces training data containing explicit disambiguation steps for the later stage; (2) disambiguation-guided supervised fine-tuning, which explicitly trains the model to discriminate word sense before generating the final definition; and (3) preference-based optimization, which further strengthens the models ability to generate sense-faithful definitions by optimizing it using preference pairs constructed from multiple sampled CoT outputs. Extensive experiments across benchmark datasets and multiple backbone LLMs demonstrate that WSDPO achieves substantial performance gains on rare and unseen settings, and exhibits strong generalization in standard evaluation settings.

Session 16: Oral/Poster/Demos F - 09:00-10:30

Demo Session F

09:00-10:30 (Grand Hall)

[Demo] ClinQueryAgent: A Conversational Agent for Population Health Management ▶ *Underline Video*

Joseph S. Boyle, Anthony Dranfield, Mike ONeil, Maria Liakata and Alison Q. Smithard

In this paper we introduce CLINQUERYAGENT, a system for translating natural language population health questions into executable database queries using agents with access to both local and external knowledge bases. Our novel architecture enables the use of powerful cloud-based language models whilst ensuring that no patient data leaves the secure environment. To combat inaccuracies over the course of longer dialogues due to context rot, information retrieval is delegated to a sub-agent. We deploy the system via a chat window embedded within an existing population health management platform where it has been used by 128 staff from 15 healthcare practices covering a total of 148,319 patients in the UKs National Health Service (NHS). We evaluate the systems capacity to autonomously handle a range of health informatics tasks on a constructed dataset and via a beta-testing phase. Our results show that both analysts and clinicians are able to easily generate actionable information from patient health records using natural language requests requiring no programming expertise to verify. We make a public demo of the system available: https://demo-899965260288.europe-west1.run.app/

[Demo] CodeDistiller: Automatically Generating Code Libraries for Scientific Coding Agents ▶ *Underline Video*

Peter Jansen, Samiah Hassan and Pragnya Narasimha

Automated Scientific Discovery (ASD) systems can help automatically generate and run code-based experiments, but their capabilities are limited by the code they can reliably generate from parametric knowledge alone. As a result, current systems either mutate a small number of manually-crafted experiment examples, or operate solely from parametric knowledge, limiting quality and reach. We introduce CodeDistiller, a system that automatically distills large collections of scientific Github repositories into a vetted library of working domain-specific code examples, allowing ASD agents to expand their capabilities without manual effort. Using a combination of automatic and domain-expert evaluation on 250 materials science repositories, we find the best model is capable of producing functional examples for 74% of repositories, while our downstream evaluation shows an ASD agent augmented with a CodeDistiller generated library produces more accurate, complete, and scientifically sound experiments than an agent with only general materials-science code examples.

[Demo] MixtureKit: A General Framework for Composing, Training, and Visualizing Mixture-of-Experts Models ▶ *Underline Video*

Ahmad Chamma;Omar El Herraoui;Guokan Shang

We introduce MixtureKit, a modular open-source framework for constructing, training, and analyzing Mixture-of-Experts (MoE) models from arbitrary pre-trained or fine-tuned checkpoints. MixtureKit supports three complementary strategies: (i) Traditional MoE, using a single router per transformer block to select experts; (ii) BTX (Branch-Train-Mix), adding routers at user-specified sub-layers for fine-grained token routing; and (iii) BTS (Branch-Train-Stitch), preserving experts intact and introducing lightweight stitch layers for controlled hub–expert information exchange. Given a single configuration dictionary, MixtureKit automatically modifies model configuration, patches decoder and causal LM classes, and exports a unified transformers-compatible checkpoint ready for inference or further fine-tuning. We also provide a visualization interface to inspect token routing, expert weight distributions, and layer-wise contributions. Experiments on multilingual code-switched (Arabic–Latin) data show that BTX models built with MixtureKit can outperform dense baselines across multiple benchmarks. The library is accessible at: https://anonymous.4open.science/r/MixtureKit-C47B.

[Demo] mllm-shap: A Shapley Value Explainability Platform for Text-Audio Multimodal Large Language Models ▶ *Underline Video*

Jakub Muszyski, Pawe Pozorski and Maria Ganzha;

We present mllm-shap, an open-source Python platform for researchers and ML practitioners that extends Shapley value (SV) explainability from text-only large language models to multimodal LLMs (MLLMs) that jointly process text and audio. Building on the token-level SV framework introduced by TokenSHAP, mllm-shap addresses three challenges absent in the text-only setting: (1) modality-aware coalition

masking that handles the coexistence of text tokens and dense audio encoder frames within a single input, (2) multi-turn conversation tracking with per-token role and modality metadata, and (3) audio token grouping via phonetic alignment that reduces the coalition space by 10-50 times. The platform ships as a pip-installable package implementing five SV estimation strategies - including a Complementary Contributions estimator with Neyman-optimal allocation that outperforms Monte Carlo baselines - together with an interactive web GUI for real-time attribution visualization. To our knowledge, mllm-shap is the first publicly available framework for complete, reproducible SV-based explainability of text-audio MLLMs. The package is MIT-licensed with full source code on GitHub and a demonstration video included as supplementary material.

[Demo] Potato 2.0: A Comprehensive Annotation Platform with AI-in-the-Loop Support ▶ *Underline Video*
David Jurgens, Michael Chen and Lina Iyer
Annotated data remains essential for training and evaluating NLP systems. Large language models have increased the scope of what type of data is needed, creating new needs for multimodal and agentic system data. Here, we introduce Potato 2.0, a major update to our open source annotation platform designed for easy deployment, customization, and fully reproducible and sharable annotation designs. Potato offers broad support for many types of annotations in NLP, including 15 different types of annotation tasks, support for text, audio, image, and video modalities—or mixtures thereof,—and the support for labeling agentic system outputs or live interaction with agents. Potato also includes multiple AI-assistance features to help annotators more easily label data. Finally, Potato introduces a new agentic AI-in-the-loop workflow where a single human annotator collaborates with an LLM through iterative prompt refinement, uncertainty-driven instance selection, and progressive autonomy—enabling efficient dataset creation without a large annotation team.

[Demo] Semantic XPath: Structured Agentic Memory Access for Conversational AI ▶ *Underline Video*
Yifan Simon Liu, Ruifan W, Liam Gallagher, Jiazhou Liang, Armin Toroghi and Scott Sanner
Conversational AI (ConvAI) agents increasingly maintain structured memory to support long-term, task-oriented interactions. In-context memory approaches append the growing history to the model input, which scales poorly under context-window limits. RAG-based methods retrieve request-relevant information, but most assume flat memory collections and ignore structure. We propose Semantic XPath, a tree-structured memory module to access and update structured conversational memory. Semantic XPath improves performance over flat-RAG baselines by 176.7% while using only 9.1% of the tokens required by in-context memory. We also introduce SemanticXPath Chat, an end-to-end ConvAI demo system that visualizes the structured memory and query execution details. Overall, this paper demonstrates a candidate for the next generation of long-term, task-oriented ConvAI systems built on structured memory.

Oral Session F: Discourse, Pragmatics, and Reasoning

09:00-10:30 (Promenade)

[Long] Accommodation and Epistemic Vigilance: A Pragmatic Account of Why LLMs Fail to Challenge Harmful Beliefs ▶ *Underline Video*
Myra Cheng, Robert Hawkins and Dan Jurafsky
Recent evaluations show that large language models (LLMs) frequently fail to challenge users' harmful beliefs in domains ranging from medical advice to social reasoning. We present a unifying analysis through the lens of pragmatics: these safety failures can be understood and addressed as LLMs exhibiting excessive accommodation and insufficient epistemic vigilance. We show that the pragmatic factors affecting accommodation and epistemic vigilance in humans (at-issueness, linguistic encoding, and source reliability) influence LLM behaviors in similar ways. We demonstrate how these factors explain performance differences across three safety benchmarks that test models' ability to challenge harmful beliefs, spanning misinformation (Cancer-Myth, SAGE-Eval) and sycophancy (ELEPHANT). This pragmatic lens further motivates prompting interventions, such as adding the phrase "wait a minute", that drastically improve performance on these difficult benchmarks by shifting pragmatic cues. Our results have practical implications for benchmark design and underscore the importance of pragmatics for understanding model behavior and improving performance.

[Long] CASPER in the Machine: Insights into Character Variety in LLM-Generated Stories ▶ *Underline Video*
Anneliese Brei, Abhisheik Sharma, Nicholas Sanaie, Lu Wang and Snigdha Chaturvedi
As LLM-generated text is increasingly used, especially in fictional domains, we explore how much LLM-generated stories differ from human-written stories. In this work, we focus on characters. We borrow definitions from narratology to analyze 8 intricate category-pairs of character, such as stylization and wholeness. These category-pairs consider more than just basic characteristics. They assess how characters are portrayed within their stories. After automatically inferring categories of characters within both LLM and human-written stories, we compare and contrast these two sets of stories. We consider the following overarching questions: (1) Do LLMs and human-written stories have similar characters? and (2) Do LLMs generate stories with a variety of characters? Our analysis includes research questions that focus on stories generated by popular LLMs and recently published human-written stories. We describe a number of interesting similarities, differences and key takeaways.

[Long] Discourse Realization of Generics in Human and LLM-generated Texts ▶ *Underline Video*

Søren Kirkegaard Fomsgaard, Martial Pastor, Gaël Dias and Nelleke Oostdijk

Large Language Models (LLMs) often produce texts that appear coherent and credible, even when their factual reliability is uncertain. This paper investigates whether such perceived credibility correlates with the pervasive use of generics, generalizations without explicit quantification. We introduce a text-level genericity score derived from clause-level annotations and apply it to argumentative essays produced by humans and LLMs. To analyze how generics are realized in discourse, we employ Rhetorical Structure Theory to examine coherence relations across varying levels of genericity. Results show that according to our genericity metric, human texts are less generic than LLM-produced texts. As regards discourse, higher genericity correlates with less structured, paratactic structures, while for some models coherence is maintained through elaboration relations. Our findings suggest that some LLMs maintain well-structured coherence even in highly generic texts, which might enable them to "camouflage" argumentative texts as informative, enhancing their perceived credibility and persuasiveness.

[Long] Evaluating Visual Narrative Coherence in Story Visualization via Diversified Storylines ▶ *Underline Video*

Minha Jhang, Kyeongman Park, Hyukhun Koh and Kyomin Jung

Story visualization requires generating a coherent sequence of images that collectively form a narrative, yet existing evaluation metrics and datasets often overlook visual continuity and narrative diversity. In this paper, we introduce the Visual Context-Aware Metric for Story Visualization, which uses large vision-language models to jointly assess caption fidelity and inter-image consistency, achieving Spearmans correlation comparable to human agreement on two benchmarks. Also, to address the shortcomings of narrowly defined datasets with low diversity, we propose a diffusion-augmented evaluation pipeline that blends diverse and controlled narrative elements at adjustable ratios, producing challenging evaluation sets. By combining VCMS with this pipeline, we provide a scalable, human-aligned framework for evaluating story visualization models.

[Long] Expect the Unexpected? Testing the Surprisal of Salient Entities ▶ *Underline Video*

Jessica Lin and Amir Zeldes

Previous work examining the Uniform Information Density (UID) hypothesis has shown that while information as measured by surprisal metrics is distributed more or less evenly across documents overall, local discrepancies can arise due to functional pressures corresponding to syntactic and discourse structural constraints. However, work thus far has largely disregarded the relative salience of discourse participants. We fill this gap by studying how overall salience of entities in discourse relates to surprisal using 70K manually annotated mentions across 16 genres of English and a novel minimal-pair prompting method. Our results show that globally salient entities exhibit significantly higher surprisal than non-salient ones, even controlling for position, length, and nesting confounds. Moreover, salient entities systematically reduce surprisal for surrounding content when used as prompts, enhancing document-level predictability. This effect varies by genre, appearing strongest in topic-coherent texts and weakest in conversational contexts. Our findings refine the UID competing pressures framework by identifying global entity salience as a mechanism shaping information distribution in discourse.

[Long] LitVISTA: A Benchmark for Narrative Orchestration in Literary Text ▶ *Underline Video*

Mingzhe Lu, Yiwen Wang, Yanbing Liu, Qi You, Chong Liu, Ruize Qin, Haoyu Dong, Wenyu Zhang, JiaRui Zhang, Yue Hu and Yunpeng Li

Computational narrative analysis aims to capture rhythm, tension, and emotional dynamics in literary texts. Existing large language models can generate long stories but overly focus on causal coherence, neglecting the complex story arcs and orchestration inherent in human narratives. This suggests a structural misalignment between model- and human-generated narratives. We therefore position narrative analysis as a diagnostic proxy for generation and propose VISTA Space, a high-dimensional framework for narrative orchestration that unifies human and model perspectives while jointly characterizing narrative function and structure in a common space. We further introduce LitVISTA, a structurally annotated benchmark grounded in literary texts, which operationalizes VISTA Space for systematic evaluation of models narrative orchestration capabilities. Under an oracle setting with gold event anchors, we evaluate frontier LLMs including GPT, Claude, Grok, and Gemini. Results reveal systematic deficiencies, as current models struggle to jointly capture narrative function and structure and fail to form an integrated global view of literary narrative orchestration. End-to-end analysis further shows that failures are dominated by anchor identification and localization errors. Even advanced thinking modes yield mixed and often limited gains for literary narrative understanding.

[Long] On Emergent Social World Models , Evidence for Functional Integration of Theory of Mind and Pragmatic Reasoning in Language Models ▶ *Underline Video*

Polina Tsvilodub, Jan-Felix Klumpp, Amir Mohammadpour, Jennifer Hu and Michael Franke

This paper investigates whether LMs recruit shared computational mechanisms for general Theory of Mind (ToM) and language-specific pragmatic reasoning in order to contribute to the general question of whether LMs may be said to have emergent "social world models", i.e., representations of mental states that are repurposed across tasks (the functional integration hypothesis). Using behavioral evaluations and causal-mechanistic experiments via functional localization methods inspired by cognitive neuroscience, we analyze LMs' performance across seven subcategories of ToM abilities (Beaudoin et al., 2020) on a substantially larger localizer dataset than used in prior like-minded work. Results from stringent hypothesis-driven statistical testing offer suggestive evidence for the functional integration hypothesis, indicating that LMs may develop interconnected "social world models" rather than isolated competencies. This work contributes novel ToM localizer data, methodological refinements to functional localization techniques, and empirical insights into the emergence of social cognition in artificial

systems.

[Long] TAIGR: Towards Modeling Influencer Content on Social Media via Structured, Pragmatic Inference ▶ *Underline Video*
Nishanth_Sridhar Nakshatri, Eylon Caplan, Rajkumar Pujari and Dan Goldwasser
Health influencers play a growing role in shaping public beliefs, yet their content is often conveyed through conversational narratives and rhetorical strategies rather than explicit factual claims. As a result, claim-centric verification methods struggle to capture the pragmatic meaning of influencer discourse. In this paper, we propose TAIGR (Takeaway Argumentation Inference with Grounded References), a structured framework designed to analyze influencer discourse, which operates in 3 stages: (1) identifying the core influencer recommendation–takeaway; (2) constructing an argumentation graph that captures influencer justification for the takeaway; (3) performing factor graph-based probabilistic inference to validate the takeaway. We evaluate TAIGR on a content validation task over influencer video transcripts on health, showing that accurate validation requires modeling the discourses pragmatic and argumentative structure rather than treating transcripts as flat collections of claims.

[Long] Talking to a Know-It-All GPT or a Second-Guesser Claude? How Repair reveals distinct Multi-Turn Behavior in LLMs ▶ *Underline Video*
Clara Lachenmaier, Hannah Bultmann and Sina ZarrieSS
Repair, an important resource for resolving trouble in human-human conversation, remains underexplored in human-LLM interaction. In this study, we investigate how LLMs engage in the interactive process of repair in multi-turn dialogues around solvable and unsolvable math questions. We examine whether models initiate repair themselves and how they respond to user-initiated repair. Our results show strong differences across models: reactions range from being almost completely resistant to (appropriate) repair attempts to being highly susceptible and easily manipulated. We further demonstrate that once conversations extend beyond a single turn, model behavior becomes more distinctive and less predictable across systems. Overall, our findings indicate that each tested LLM exhibits its own characteristic form of unreliability in the context of repair.

Oral Session F: Ethics, Bias, and Fairness 2

09:00-10:30 (Harbor G)

[Long] Can Factual Opinions Be Edited (Manipulated) in Large Language Models? ▶ *Underline Video*
Yuanpu Cao, Ziyi Yin, Fenglong Ma and Jinghui Chen
Large Language Models (LLMs) are increasingly integrated into various domains, making knowledge editing techniques crucial yet potentially hazardous. Current editing methods primarily target atomic facts, overlooking the significant risks associated with manipulating "factual opinions", e.g., documented stances of public figures on societal issues. Such manipulation could reshape public images, influence elections, and alter societal views. To systematically assess this threat, we introduce the Factual Opinion Editing with Evidence (FOE) benchmark, which encompasses 261 public figures, 19 issue categories, and 2,178 complete opinion records. Our evaluations demonstrate that current editing techniques struggle significantly with factual opinions, often achieving only superficial changes while failing to preserve consistency between the edited opinion and the supporting evidence generated by the model. To address this limitation, we further propose a simple yet effective Self-Generated Evidence-Aligned method that achieves opinion-evidence alignment without relying on explicit instructions. Together, our benchmark and method provide a foundation for understanding the emerging security implications of factual opinion editing in LLMs.

[Long] GKnow: Measuring the Entanglement of Gender Bias and Factual Gender ▶ *Underline Video*
Leonor Veloso and Hinrich Schütze
Recent works have analyzed the impact of individual components of neural networks on gendered predictions, often with a focus on mitigating gender bias. However, mechanistic interpretations of gender tend to (i) focus on a very specific gender-related task, such as gendered pronoun prediction, or (ii) fail to distinguish between the production of factually gendered outputs (the correct assumption of gender given a word that carries gender as a semantic property) and gender biased outputs (based on a stereotype). To address these issues, we curate Gknow, a benchmark to assess gender knowledge and gender bias in language models across different types of gender-related predictions. Gknow allows us to identify and analyze circuits and individual neurons responsible for gendered predictions. We test the impact of neuron ablation on benchmarks for disentangling stereotypical and factual gender (DiFair and the test set of Gknow), as well as StereoSet. Results show that gender bias and factual gender are severely entangled on the level of both circuits and neurons, entailing that ablation is an unreliable debiasing method. Furthermore, we show that benchmarks for evaluating gender bias can hide the decrease in factual gender knowledge that accompanies neuron ablation. We curate Gknow as a contribution to the continuous development of robust gender bias benchmarks.

[Long] Inertia in Moral and Value Judgments of Large Language Models ▶ *Underline Video*
Bruce W. Lee, Yeongheon Lee and Hyunsoo Cho

Large Language Models (LLMs) behave non-deterministically, and prompting has become a common method for steering their outputs. A popular strategy is to assign a persona to the model to produce more varied, context-sensitive responses, similar to how responses vary across human individuals. Against the expectation that persona prompting yields a wide range of opinions, our experiments show that LLMs keep consistent value orientations. We observe a persistent inertia in their responses, where certain moral and value dimensions (especially harm avoidance and fairness) stay skewed in one direction across persona settings. To study this, we use role-play at scale, which pairs randomized persona prompts with a macro-level analysis of model outputs. Our results point to strong internal biases and value preferences in LLMs, which we call value orientation and inertia. These models warrant scrutiny and adjustment before use in applications where balanced outputs matter.

[Long] InsideOut: Measuring and Mitigating Insider-Outsider Bias in Interview Script Generation ► *Underline Video*
Yixin Wan, Xingrun Chen and Kai-Wei Chang
Advancements in Large language models (LLMs) have enabled a variety of downstream applications like story and interview script generation. However, recent research raised concerns about culture-related fairness issues in LLM-generated content. In this work, we identify and systematically investigate LLMs' insider-outsider bias, a phenomenon where models position themselves as "insiders" of mainstream cultures during generation while externalizing less dominant cultures. We propose the InsideOut benchmark with 4,000 generation prompts and three evaluation metrics to quantify this bias through a culturally situated interview script generation task, in which an LLM is positioned as a reporter interviewing local people across 10 diverse cultures. Empirical evaluation on 5 state-of-the-art LLMs reveals that while models adopt insider tones in over 88% US-contexted scripts on average, they disproportionately default to "outsider" stances for non-Western cultures. To mitigate these biases, we propose 2 inference-time methods: a baseline prompt-based Fairness Intervention Pillars (FIP) method, and a structured Mitigation via Fairness Agents (MFA) framework consisting of a Single-Agent (MFA-SA), a Hierarchical-Agent (MFA-HA), and an autonomous Agentic Planning (MFA-Plan) pipeline. Empirical results demonstrate that agent-based MFA methods achieve outstanding and robust performance in mitigating the insider-outsider bias: For instance, on the Cultural Alignment Gap (CAG) metric, MFA-SA reduces bias in Llama model by 89.70% and MFA-HA mitigates bias in Qwen by 82.54%. These findings showcase the effectiveness of agent-based methods as a promising direction for mitigating biases in generative LLMs.

[Long] Preconditioned Test-Time Adaptation for Out-of-Distribution Debiasing in Narrative Generation ► *Underline Video*
HANWEN SHEN, TING YING, Jiajie Lu and Shanshan Wang
Although debiased large language models (LLMs) excel at handling known or low-bias prompts, they often fail on unfamiliar and high-bias prompts. We demonstrate via out-of-distribution (OOD) detection that these high-bias prompts cause a distribution shift, degrading static model performance. To enable real-time correction, we propose CAP-TTA, a test-time adaptation framework. CAP-TTA triggers context-aware LoRA updates only when a bias-risk score exceeds a set threshold. By utilizing an offline precomputed diagonal preconditioner, it ensures fast and stable optimization. Across multiple benchmarks and human evaluations, CAP-TTA effectively reduces toxicity/bias score with significantly lower latency than standard optimization methods (e.g., AdamW or SGD). Furthermore, it prevents catastrophic forgetting, and substantially improves narrative fluency over state-of-the-art baselines without compromising debiasing performance.

[Long] Rethinking Jailbreak Detection of Large Vision Language Models with Representational Contrastive Scoring ► *Underline Video*
Peichun Hua, Hao Li, Shanghao Shi, Zhiyuan Yu and Ning Zhang
Large Vision-Language Models (LVLMs) are vulnerable to a growing array of multimodal jailbreak attacks, necessitating defenses that are both generalizable to novel threats and efficient for practical deployment. Many current strategies fall short, either targeting specific attack patterns, which limits generalization, or imposing high computational overhead. While lightweight anomaly-detection methods offer a promising direction, we find that their common one-class design tends to confuse unseen benign inputs with malicious ones, leading to unreliable over-rejection. To address this, we propose Representational Contrastive Scoring (RCS), a framework built on a key insight: the most potent safety signals reside within the LVLM's own internal representations. Our approach inspects the internal geometry of these representations, learning a lightweight projection to maximally separate benign and malicious inputs in safety-critical layers. This enables a simple yet powerful contrastive score that differentiates true malicious intent from mere distribution shift. Our instantiations, MCD (Mahalanobis Contrastive Detection) and KCD (K-nearest Contrastive Detection), achieve state-of-the-art performance on a challenging evaluation protocol designed to test generalization to unseen attack types. This work demonstrates that effective jailbreak detection can be achieved by applying simple, interpretable statistical methods to the internal representations, offering a practical path towards safer LVLM deployment.

[Long] Understanding and Mitigating Bias Inheritance in LLM-based Data Augmentation on Downstream Tasks ► *Underline Video*
Miaomiao Li, Hao Chen, Yang Wang, Tingyuan Zhu, Weijia Zhang, Kaijie Zhu, Kam-Fai Wong and Jindong Wang
Generating synthetic datasets via large language models (LLMs) has emerged as a promising approach to improve LLM performance. However, LLMs inherently reflect biases in their training data, leading to a critical challenge: when models are trained on synthetic data, they may propagate and amplify the inherent biases that can significantly impact fairness and robustness on downstream tasks, a phenomenon we term bias inheritance. This work presents the first systematic investigation in understanding, analyzing, and mitigating bias inheritance. We fine-tune LLMs with a combined dataset of real and LLM-augmented data with varied bias ratio as the proportion of augmented data.

Through systematic experiments across 10 classification and generation tasks, we analyze how 6 different types of biases manifest. Our results indicate that bias inheritance harms downstream task performance in bias directly-related classification and generation tasks. Then, our analysis identifies three key misalignment factors: misalignment of values, group data, and data distributions. Based on these insights, we propose three mitigation strategies: token-based, mask-based, and loss-based approaches, which can work differently on various tasks and bias, indicating the substantial challenges to mitigate bias inheritance. We hope this work can provide insights to the research of LLM data augmentation.

[Long] When Bigger Isnt Better: A Comprehensive Fairness Evaluation of Political Bias in Multi-News Summarisation ► ***Underline Video***

Nannan Huang, Iffat Maab and Junichi Yamagishi

Multi-document news summarisation systems are increasingly adopted for their convenience in processing vast daily news content, making fairness across diverse political perspectives critical. However, these systems can exhibit political bias through unequal representation of viewpoints, disproportionate emphasis on certain perspectives, and systematic underrepresentation of minority voices. This study presents a comprehensive evaluation of such bias in multi-document news summarisation using FairNews, a dataset of complete news articles with political orientation labels, examining how large language models (LLMs) handle sources with varying political leanings across 13 models and five fairness metrics. We investigate both baseline model performance and effectiveness of various debiasing interventions, including prompt-based and judge-based approaches. Our findings challenge the assumption that larger models yield fairer outputs, as mid-sized variants consistently outperform their larger counterparts, offering the best balance of fairness and efficiency. Prompt-based debiasing proves highly model dependent, while entity sentiment emerges as the most stubborn fairness dimension, resisting all intervention strategies tested. These results demonstrate that fairness in multi-document news summarisation requires multi-dimensional evaluation frameworks and targeted, architecture-aware debiasing rather than simply scaling up.

[Long] Why Are We Moral? An LLM-based Agent Simulation Approach to the Study of Moral Evolution ► ***Underline Video***

Zhou Ziheng and Huacong Tang

The evolution of morality presents a puzzle: natural selection should favor self-interest, yet humans developed moral systems promoting altruism. Traditional approaches must abstract away cognitive processes, leaving open how cognitive factors shape moral evolution. We introduce an LLM-based agent simulation framework that brings cognitive realism to this question: agents with varying moral dispositions perceive, remember, reason, and decide in a simulated prehistoric hunter-gatherer society. This enables us to manipulate factors that traditional models cannot represent—such as moral type observability and communication bandwidth—and to discover emergent cognitive mechanisms from agent interactions. Across 20 runs spanning four settings, we find that cooperation and mutual help are the central driver of evolutionary survival, with universal and reciprocal morality exhibiting the most stable outcomes across conditions while selfishness is strongly disfavoured. Beyond cooperation itself, we further identify cognition as a central mediator—most clearly through a cost of moral judgment that shifts the winning moral type across settings, with a self-purging effect among selfish agents as an additional cognitive pattern. We validate robustness across multiple LLM backbones, architecture ablations, and prompt sensitivity analyses. This work establishes LLM-based simulation as a powerful new paradigm to complement traditional research in evolutionary biology and anthropology, opening new avenues for investigating the complexities of moral and social evolution.

Oral Session F: Industry 5

09:00-10:30 (Balboa)

[Industry] AIDA-SEAT: Towards Reliable AI Doctor Assistant via State-Evaluation-Action Tree Enhanced LLMs in Online Hospital ► ***Underline Video***

Lianxin Sun, Xiaoying Ying, Guangya Yu, Weiyan Zhang, Chenhao Guan, Hao He, Mingxi Shang, Jianhua Li, ChunMing Wang and Tong Ruan

Artificial intelligence doctor assistants (AIDAs) are popular in medical dialogue systems, helping doctors make clinical decisions and reduce their workload. Existing methods for healthcare primarily focus on LLM-based approaches or knowledge retrieval-augmented generation (RAG). However, we argue that existing retrieval methods primarily focus on passages and lack the logical expression of knowledge. In this paper, we propose an AIDA-SEAT framework to provide reliable clinical decision-making support. First, we design the state-evaluation-action tree (SEAT), which covers diagnosis, treatment, and examination. To develop this tree, we refine and transform SEAT collected from medical documents and doctors. Then, we propose an adaptive method to select optimal trees tailored to the current patients' state. Finally, we leverage LLMs to perform state assessment, evaluation, and action execution based on the tree, thereby generating reliable responses. To evaluate the effectiveness of our method, we conducted extensive experiments on a self-built dataset. Our method achieves 1.01% higher than current state-of-the-art (SOTA) baselines across five departments, including common RAG-based methods. Furthermore, analysis of 200 consultation records during deployment on an online hospital revealed that system-assisted responses are 24.16 seconds faster on average than manual ones, improving efficiency by 26.85%.

[Industry] Fanar-Sadiq: A Multi-Agent Architecture for Grounded Islamic QA ► *Underline Video*

Ummar Abbas, Mourad Ouzzani, Mohamed Y. Eltabakh, Omar Sinan, Gagan Bhatia, Hamdy Mubarak, Majd Hawasly, Mohammed Qusay Hashim, Kareem Mohamed Darwish and Firoj Alam

Large language models (LLMs) can answer religious knowledge queries fluently, yet they often hallucinate and misattribute sources, which is especially consequential in Islamic settings where users expect grounding in canonical texts (Qur'an and Hadith) and jurisprudential (fiqh) nuance. Retrieval-augmented generation (RAG) reduces some of these limitations by grounding generation in external evidence. However, a single "retrieve-then-generate" pipeline is limited to deal with the diversity of Islamic queries. Users may request verbatim scripture, fatwa-style guidance with citations or rule-constrained computations such as zakat and inheritance that require strict arithmetic and legal invariants. We present a bilingual (Arabic/English) multi-agent system that routes queries to specialized modules within an agentic, tool-using architecture. The system supports intent-aware routing, retrieval-grounded fiqh answers with deterministic citation normalization and verification traces, exact verse lookup with quotation validation, and deterministic calculators for Sunni zakat and inheritance with madhhab-sensitive branching. We evaluate the complete end-to-end system on public Islamic QA benchmarks and demonstrate effectiveness and efficiency. Our system is currently publicly and freely accessible through API and a Web application, and has been accessed ~1.9M times in less than a year (anonymous.com)

[Industry] From Relevance to Authority: Authority-aware Generative Retrieval in Web Search Engines ► *Underline Video*

Sunkyung Lee, Jihye Back, Donghyeon Jeon, Soonhwan Kwon, Moonkwon Kim, Inho Kang and Jongwuk Lee

Generative information retrieval (GenIR) formulates the retrieval process as a text-to-text generation task, leveraging the vast knowledge of large language models. However, existing works primarily optimize for relevance while often overlooking document trustworthiness. This is critical in high-stakes domains like healthcare and finance, where relying solely on semantic relevance risks retrieving unreliable information. To address this, we propose an Authority-aware Generative Retriever (AuthGR), the first framework that incorporates authority into GenIR. AuthGR consists of three key components: (i) Multimodal Authority Scoring, which employs a vision-language model to quantify authority from textual and visual cues; (ii) a Three-stage Training Pipeline to progressively instill authority awareness into the retriever; and (iii) a Hybrid Ensemble Pipeline for robust deployment. Offline evaluations demonstrate that AuthGR successfully enhances both authority and accuracy, with our 3B model matching a 14B baseline. Crucially, large-scale online A/B tests and human evaluations conducted on the commercial web search platform confirm significant improvements in real-world user engagement and reliability.

[Industry] From TextBlob to LLM Agents: Sentiment Model Selection for B2B Technical Support with CSAT Ground Truth ► *Underline Video*

Pedro Vidigal

We present a five-year case study of sentiment model selection for customer satisfaction (CSAT) prediction in B2B technical support. Our evaluation uses the complete population of CSAT-rated tickets from an enterprise software company: over 500 tickets comprising 2,500 customer comments from 100+ organizations over five years. We evaluate 17 approaches across 5 paradigms (lexicon, off-the-shelf transformers, NLI zero-shot, multi-task LLM agent, and 12 dedicated LLM agents from 6 vendor families), plus 11 fine-tuning experiments (all achieving MCC<=0). Key findings: (1) a dedicated single-task LLM agent reduces neutral bias from 69% to 22%, improving MCC from -0.018 to 0.347 ($p<0.001$); (2) our results are consistent with the "Alignment Tax" (Lin et al., 2024; Wu et al., 2025) in sentiment classification: Claude Opus 4.6 exhibits 41% neutral predictions and lower recall than its budget model Haiku 4.5 ($p=0.003$); (3) 38% of dissatisfied customers are undetectable by all 12 LLMs due to administrative requests lacking emotional language; (4) Gemini 3 Flash achieves the best MCC (0.347) at 0.60/1K, over 100x cheaper than Claude Opus. We describe the three-phase production deployment and provide practitioner recommendations.

[Industry] IceBreaker for Conversational Agents: Breaking the First-Message Barrier with Personalized Starters ► *Underline Video*

Hongwei Zheng, Weiqi Wu, Zhengjia Wang, Guanyu Jiang, Haoming Li, Tianyu Wu, Yongchun Zhu, Jingwu Chen and Feng Zhang

Conversational agents, such as ChatGPT and Doubao, have become essential daily assistants for billions of users. To further enhance engagement, these systems are evolving from passive responders to proactive companions. However, existing efforts focus on activation within ongoing dialogues, while overlooking a key real-world bottleneck. In the conversation initiation stage, users may have a vague need but no explicit query intent, creating a first-message barrier where the conversation holds before it begins. To overcome this, we introduce Conversation Starter Generation: generating personalized starters to guide users into conversation. However, unlike in-conversation stages where immediate context guides the response, initiation must operate in a cold-start moment without explicit user intent. To pioneer in this direction, we present IceBreaker that frames human ice-breaking as a two-step handshake: (i) evoke resonance via Resonance-Aware Interest Distillation from session summaries to capture trigger interests, and (ii) stimulate interaction via Interaction-Oriented Starter Generation, optimized with personalized preference alignment and a self-reinforced loop to maximize engagement. Online A/B tests on one of the world's largest conversational agent products show that IceBreaker improves user active days by +1.84 permille and click-through rate by +94.25 permille , and has been deployed in production.

[Industry] Is Agentic RAG worth it? An experimental comparison of RAG approaches ► *Underline Video*

pietro ferrazzi; milica cvjeticanin; alessio piraccini; davide giannuzzi

Retrieval-Augmented Generation (RAG) systems are usually defined by the combination of a generator and a retrieval component that extracts textual context from a knowledge base to answer user queries. However, such basic implementations exhibit several limitations,

including noisy or suboptimal retrieval, misuse of retrieval for out-of-scope queries, weak query-document matching, and variability or cost associated with the generator. These shortcomings have motivated the development of "Enhanced" RAG, where dedicated modules are introduced to address specific weaknesses in the workflow. More recently, the growing self-reflective capabilities of Large Language Models (LLMs) have enabled a new paradigm, often referred to as "Agentic" RAG. In this approach, an LLM orchestrates the entire process, deciding which actions to perform, when to perform them, and whether to iterate. Despite the rapid adoption of both paradigms, it remains unclear which approach is preferable under which conditions. In this work, we conduct an empirically driven evaluation of "Enhanced" and "Agentic" RAG across multiple scenarios and dimensions. Our results provide practical insights into the trade-offs between the two paradigms, offering guidance on selecting the most effective RAG design for real-world applications, considering both performance and costs.

[Industry] Multi-Agent Orchestration for Terminology-Constrained Machine Translation in Industrial Localization ▶ *Underline Video*

Emanuele Di Rosa

Accurate terminology is a non-negotiable requirement in industrial localization processes: a single mistranslated domain term can violate contractual obligations and erode client trust. We present Anonymized, a deployed multi-agent LLM pipeline that orchestrates four specialized agents—Analysis, Translation, Post-editing, and Review—for terminology-constrained machine translation. The system introduces terminology-aware pre-analysis, explicit glossary injection at every pipeline stage, and a reasoning-enabled Review agent. We evaluate six configurations on the WMT25 Terminology Translation benchmark (Track~1: en->de/es/ru, IT domain), enabling systematic ablation of each design choice. Our best configuration achieves 99.4% average terminology accuracy while attaining the highest ChrF2++ scores across all three language pairs, outperforming all 20 systems submitted to the shared task. Unlike other multi-agent approaches in WMT25 that rely on generate-and-select strategies, Anonymized is the first to apply a role-specialized sequential pipeline to terminology-constrained MT, and is deployed with native XLIFF integration for seamless CAT tool interoperability. The system processes thousands of terminology-constrained requests daily at a large localization provider.

[Industry] Optimizing Retrieval-Augmented Generation for E-Commerce How-To Assistance ▶ *Underline Video*

Gilad Fuchs, Leonid Ekimov, Fei Dong, Jiahong Xie, Wei Liu, Maxim Manco and Alexander Nus

Conversational AI is increasingly used in e-commerce to deliver personalized customer support. We present a production RAG-based How-To Assistant that answers support and how-to queries by grounding responses in a proprietary knowledge base. We study three factors that drive quality: (1) document chunking and contextualization for indexing, (2) query refinement methods, and (3) automatic LLM-based evaluation for rapid iteration and reliable measurement. We also describe the end-to-end system workflow - from offline indexing to real-time serving and report deployment metrics, offering practical guidance for building scalable, high-precision RAG assistants in commercial support settings.

[Industry] What Question Did You Answer? Refining Contact Center Evaluation Plans via Backward Questions ▶ *Underline Video*

Prajwal Sood, Rushikesh Pawar, Digvijay Anil Ingle and Anup Pattnaik

Capturing organization-specific domain knowledge remains a critical challenge for deploying cost-efficient language models in specialized tasks like contact center Quality Assurance (QA). While large LMs implicitly capture expert judgment, smaller LMs require explicit evaluation criteria that domain experts struggle to articulate. We introduce Backward Question-based Refinement (BQR), a diagnostic framework that generates backward questions, revealing what a model understood rather than what was asked, to systematically distill implicit reasoning from large LMs into explicit evaluation plans. Through experiments on 12 QA questions, BQR achieves performance improvements on 8 questions with absolute gains of up to 27.8% in Macro F1. Our analysis establishes empirical parallels to gradient-descent optimization and reveals a cross-family advantage where small LMs benefit more from large LMs of different families. These findings confirm BQR as an effective approach for bridging the gap between implicit expert knowledge and explicit evaluation criteria.

Oral Session F: Interpretability and Analysis of Models for NLP 2

09:00-10:30 (Harbor B-C)

[Long] Fingerprinting LLMs via Prompt Injection ▶ *Underline Video*

Yuepeng Hu, Zhengyuan Jiang, Mengyuan Li, Osama Ahmed, Zhicong Huang, Cheng Hong and Neil Zhenqiang Gong

Large language models (LLMs) are often modified after release through post-processing such as post-training or quantization, which makes it challenging to determine whether one model is derived from another. Existing provenance detection methods have two main limitations: (1) they embed signals into the base model before release, which is infeasible for already published models, or (2) they compare outputs across models using hand-crafted or random prompts, which are not robust to post-processing. In this work, we propose LLMPrint, a novel detection framework that constructs fingerprints by exploiting LLMs inherent vulnerability to prompt injection. Our key insight is that by optimizing fingerprint prompts to enforce consistent token preferences, we can obtain fingerprints that are both unique to the base model and robust to post-processing. We further develop a unified verification procedure that applies to both gray-box and black-box settings, with statistical guarantees. We evaluate LLMPrint on five base models and around 700 post-trained or quantized variants. Our

results show that LLMPrint achieves high true positive rates while keeping false positive rates near zero. The code is publicly available at https://github.com/hifi-hyp/ACL-LLMPrint.

[Long] Its Not What You Say, Its How You Say It: Evaluating LLM Responses to Expressions of Belief ▶ *Underline Video*

Kevin Du, Clara Kuempel, Michelle Wastl and Alex Warstadt

Users frequently express their beliefs to large language models (LLMs). In some situations, the LLM should accept these contextual beliefs as true. In others, they should stick to their prior knowledge. Notably, users' expressions of belief (EoBs) can take linguistically diverse forms—using presuppositions, evidential and certainty markers, or varied tones—each of which may have a different persuasiveness over the LLMs. We introduce a typology to systematically evaluate how different EoBs affect whether models follow context versus prior knowledge. The typology is grounded in four linguistically motivated dimensions: form, evidentiality, epistemic stance, and tone, spanning 17 fine-grained types. By pairing these EoBs with world knowledge facts, we generate controlled EoB-query pairs that isolate the effect of linguistic variation. Using this benchmark, we evaluate 16 LLMs that differ in architecture (Llama3, Qwen3, Gemma3), scale (1B-30B parameters), and training stages (base vs instruct). We identify meaningful variations in response behavior across these axes, e.g., that bigger models and instruction models tend to be less context–following than smaller models and base models. We further identify specific EoBs that statistically significantly persuade LMs more consistently than others. Our work reveals systematic patterns in how linguistic framing affects LLM context integration, with implications for prompt engineering and model robustness.

[Long] Masked by Consensus: Disentangling Privileged Knowledge in LLM Correctness ▶ *Underline Video*

Tomer Ashuach, Liat Ein-Dor, Shai Gretz, Yoav Katz and Yonatan Belinkov

Humans use introspection to evaluate their understanding through private internal states inaccessible to external observers. We investigate whether large language models possess similar privileged knowledge about answer correctness, information unavailable through external observation. We train correctness classifiers on question representations from both a model's own hidden states and external models, testing whether self-representations provide a performance advantage. On standard evaluation, we find no advantage: self-probes perform comparably to peer-model probes. We hypothesize this is due to high inter-model agreement of answer correctness. To isolate genuine privileged knowledge, we evaluate on disagreement subsets, where models produce conflicting predictions. Here, we discover domain-specific privileged knowledge: self-representations consistently outperform peer representations in factual knowledge tasks, but show no advantage in math reasoning. We further localize this domain asymmetry across model layers, finding that the factual advantage emerges progressively from early-to-mid layers onward, consistent with model-specific memory retrieval, while math reasoning shows no consistent advantage at any depth.

[Long] Multi-component Causal Tracing in Large Language Models ▶ *Underline Video*

Zirui Yan, Dennis Wei, Dmitriy A. Katz, Prasanna Sattigeri and Ali Tajer

Causal tracing systematically intervenes on a large language model's (LLM's) internal representations to uncover and quantify the causal pathways linking specific inputs or computations to specific metrics of interest, quantifying the LLM's behavior. Building on previous single-component or single-layer studies, this paper presents a unified framework for causally tracing multiple components simultaneously. This framework systematically identifies the subsets of components (e.g., attention heads and multi-layer perceptron neurons) most critical to a desired target performance metric (e.g., accuracy and fairness). This is achieved by incorporating flexible interventions applied to a wide range of desired metrics. To address the combinatorial complexity of the multi-component problem, an efficient algorithm is designed that leverages soft interventions and a carefully designed metric transformation, converting the combinatorial search problem into a continuous one that can be solved efficiently under proper constraints, thereby generating proper binary decisions for selecting components. Experimental results demonstrate that the proposed method efficiently identifies subsets of the model's components that have a high impact on the target metric, outperforming existing baseline approaches.

[Long] Representation Interventions Enable Lifelong Knowledge Memory Control in LLMs ▶ *Underline Video*

Xuyuan Liu, Shengyu Chen, Xinshuai Dong, Yanchi Liu, Xujiang Zhao, Haoyu Wang, Yujun Yan, Haifeng Chen and Zhengzhang Chen

Large language models (LLMs) often produce incorrect or outdated content. Updating their knowledge efficiently and accurately without costly retraining is a major challenge. This problem is particularly challenging for complex, unstructured knowledge in lifelong settings, where many edits must coexist without interference. We introduce RILKE (Representation Intervention for Lifelong KnowledgE Control), a robust and scalable method that treats knowledge control as interventions within the model's representation space. Leveraging representation-space expressiveness, we identify two key properties enabling RILKE to achieve fine-grained control over complex, unstructured knowledge while maintaining general utility with frozen base weights. During training, RILKE learns paraphrase-robust and edit-localized modules that limit each update to a low-dimensional subspace to minimize cross-edit interference. In inference, a query-adaptive router selects the appropriate module to guide the model's generation. Across LLaMA and Qwen models, RILKE scales effectively to large-scale benchmarks, demonstrating high edit success and strong paraphrase generalization while preserving general utility with modest memory overhead. These results show RILKE is an effective and scalable solution for lifelong knowledge control in LLMs.

[Long] Revitalizing Black-Box Interpretability: Actionable Interpretability for LLMs via Proxy Models ▶ *Underline Video*

Junhao Liu, Haonan Yu, Zhenyu Yan and Xin Zhang

Post-hoc explanations provide transparency and are essential for guiding model optimization, such as prompt engineering and data sanitation.

However, applying model-agnostic techniques to Large Language Models (LLMs) is hindered by prohibitive computational costs, rendering these tools dormant for real-world applications. To revitalize model-agnostic interpretability, we propose a budget-friendly proxy framework that leverages efficient models to approximate the decision boundaries of expensive LLMs. We introduce a screen-and-apply mechanism to statistically verify local alignment before deployment. Our empirical evaluation confirms that proxy explanations achieve over 90% fidelity with only 9.5% of the oracle's cost. Building on this foundation, we demonstrate the actionable utility of our framework in prompt compression and poisoned example removal. Results show that reliable proxy explanations effectively guide optimization, transforming interpretability from a passive observation tool into a scalable primitive for LLM development. Additionally, we open-source code and datasets to facilitate future research.

[Long] Sparse Feature Coactivation Reveals Causal Semantic Modules in Large Language Models ▶ *Underline Video*

Ruixuan Deng, Xiaoyang Hu, Miles Gilberti, Shane Storks, Aman Taxali, Mike Angstadt, Chandra Sripada and Joyce Chai

We identify semantically coherent, context-consistent network components in large language models (LLMs) using coactivation of sparse autoencoder (SAE) features collected from just a handful of prompts. Focusing on concept-relation prediction tasks, we show that ablating these components for concepts (e.g., countries and words) and relations (e.g., capital city and translation language) changes model outputs in predictable ways, while amplifying these components induces counterfactual responses. Notably, composing relation and concept components yields compound counterfactual outputs. Further analysis reveals that while most concept components emerge from the very first layer, more abstract relation components are concentrated in later layers. Lastly, we show that extracted components more comprehensively capture concepts and relations than individual features while maintaining specificity. Overall, our findings suggest a modular organization of knowledge and advance methods for efficient, targeted LLM manipulation.

[Long] The Confidence Dichotomy: Analyzing and Mitigating Miscalibration in Tool-Use Agents ▶ *Underline Video*

Weihao Xuan, Qingcheng Zeng, Heli Qi, Yunze Xiao, Junjue Wang and Naoto Yokoya

Autonomous agents based on large language models (LLMs) are rapidly evolving to handle multi-turn tasks, but ensuring their trustworthiness remains a critical challenge. A fundamental pillar of this trustworthiness is calibration, which refers to an agent's ability to express confidence that reliably reflects its actual performance. While calibration is well-established for static models, its dynamics in tool-integrated agentic workflows remain under-explored. In this work, we systematically investigate verbalized calibration in tool-use agents, revealing a fundamental confidence dichotomy driven by tool type. Specifically, our pilot study identifies that evidence tools (e.g., web search) systematically induce severe overconfidence due to inherent noise in retrieved information, while verification tools (e.g., code interpreters) can ground reasoning through deterministic feedback and mitigate miscalibration. To robustly improve calibration across tool types, we propose a reinforcement learning (RL) fine-tuning framework that jointly optimizes task accuracy and calibration, supported by a holistic benchmark of reward designs. We demonstrate that our trained agents not only achieve superior calibration but also exhibit robust generalization from local training environments to noisy web settings and to distinct domains such as mathematical reasoning. Our results highlight the necessity of domain-specific calibration strategies for tool-use agents. More broadly, this work establishes a foundation for building self-aware agents that can reliably communicate uncertainty in high-stakes, real-world deployments.

[Long] Understanding or Memorizing? A Case Study of German Definite Articles in Language Models ▶ *Underline Video*

Jonathan Drechsel, Erisa Bytyqi and Steffen Herbold

Language models perform well on grammatical agreement, but it is unclear whether this reflects rule-based generalization or memorization. We study this question for German definite singular articles, whose forms depend on gender and case. Using GRADIEND, a gradient-based interpretability method, we learn parameter update directions for gender-case specific article transitions. We find that updates learned for a specific gender-case article transition frequently affect unrelated gender-case settings, with substantial overlap among the most affected neurons across settings. These results argue against a strictly rule-based encoding of German definite articles, indicating that models at least partly rely on memorized associations rather than abstract grammatical rules.

Oral Session F: LLM Efficiency 2

09:00-10:30 (Harbor D)

[Long] A Goal Without a Plan Is Just a Wish: Efficient and Effective Global Planner Training for Long-Horizon Agent Task ▶ *Underline Video*

Shuzheng Si and Haozhe Zhao

Agents based on large language models (LLMs) struggle with brainless trial-and-error and generating hallucinatory actions due to a lack of global planning in long-horizon tasks. In this paper, we introduce a plan-and-execute framework and propose EAGLET, an efficient and effective planner training method to enhance the executor agent's planning abilities without human effort. Specifically, we train a plug-and-play global planner through a two-step process: we first synthesize high-quality plans from an advanced LLM using our proposed homologous consensus filtering strategy, and apply fine-tuning as a cold start. Moreover, we further improve the planner with a rule-based reinforcement learning stage using a novel executor capability gain reward, ensuring it can handle task instructions of varying difficulty.

Experiments on three long-horizon agent tasks show that executor agents equipped with our planner outperform existing methods, achieving new state-of-the-art performance. Meanwhile, EAGLET reduces training costs by 8x compared to RL-based baselines, and it does not require manual effort or extra training data, offering an efficient and effective solution.

[Long] ETR: Entropy Trend Reward for Efficient Chain-of-Thought Reasoning ▶ *Underline Video*

Xuan Xiong, Huan Liu, Li Gu, Zhixiang Chi, Yue Qiu, Yuanhao Yu and Yang Wang

Chain-of-thought (CoT) reasoning improves large language model performance on complex tasks, but often produces excessively long and inefficient reasoning traces. Existing methods shorten CoTs using length penalties or global entropy reduction, implicitly assuming that low uncertainty is desirable throughout reasoning. We show instead that reasoning efficiency is governed by the trajectory of uncertainty. CoTs with dominant downward entropy trends are substantially shorter. Motivated by this insight, we propose Entropy Trend Reward (ETR), a trajectory-aware objective that encourages progressive uncertainty reduction while allowing limited local exploration. We integrate ETR into Group Relative Policy Optimization (GRPO) and evaluate it across multiple reasoning models and challenging benchmarks. ETR consistently achieves a superior accuracy-efficiency trade-off, improving DeepSeek-R1-Distill-7B by +9.9% accuracy while reducing CoT length by 67% across four benchmarks.

[Long] Guided by Gut: Efficient Test-Time Scaling with Reinforced Intrinsic Confidence ▶ *Underline Video*

Amirhosein Ghasemabadi, Keith G. Mills, Baochun Li and Di Niu

Test-Time Scaling (TTS) methods for enhancing Large Language Model (LLM) reasoning often incur substantial inference costs, due to reliance on long chain-of-thought (CoT) generation, self-consistency sampling methods, or searching under Process Reward Models (PRMs). This paper introduces Guided by Gut (GG), an efficient self-guided TTS framework that enables LLMs to perform step-by-step reasoning at a low cost, without any reward models or verifiers. GG performs a lightweight tree search guided solely by intrinsic confidence signals of the LLM at each reasoning step and improves the reliability of such internal confidence signals by reinforcement learning. Empirical evaluations on challenging mathematical reasoning benchmarks demonstrate that GG enables smaller models (e.g., 1.5B-7B parameters) to achieve accuracy matching or surpassing significantly larger models (e.g., 32B-70B parameters), while reducing GPU memory usage by up to 10x. Compared to TTS with PRMs, GG achieves comparable accuracy with 8x faster inference speeds and 4-5x lower memory usage. Additionally, GG reduces KV cache memory usage by approximately 50% compared to Best-of-N sampling, facilitating more efficient and practical deployment of TTS techniques.

[Long] MixKVQ: Query-Aware Mixed-Precision KV Cache Quantization for Long-Context Reasoning ▶ *Underline Video*

Tao Zhang, Ziqian Zeng, Hao Peng, Huiping Zhuang and Cen Chen

Long Chain-of-Thought (CoT) reasoning has significantly advanced the capabilities of Large Language Models (LLMs), but this progress is accompanied by substantial memory and latency overhead from the extensive Key-Value (KV) cache. Although KV cache quantization is a promising compression technique, existing low-bit quantization methods often exhibit severe performance degradation on complex reasoning tasks. Fixed-precision quantization struggles to handle outlier channels in the key cache, while current mixed-precision strategies fail to accurately identify components requiring high-precision representation. We find that an effective low-bit KV cache quantization strategy must consider two factors: a key channel's intrinsic quantization difficulty and its relevance to the query. Based on this insight, we propose MixKVQ, a novel plug-and-play method that introduces a lightweight, query-aware algorithm to identify and preserve critical key channels that need higher precision, while applying per-token quantization for value cache. Experiments on complex reasoning datasets demonstrate that our approach significantly outperforms existing low-bit methods, achieving performance comparable to a full-precision baseline at a substantially reduced memory footprint.

[Long] Not All Tokens Matter: Towards Efficient LLM Reasoning via Token Significance in Reinforcement Learning ▶ *Underline Video*

Hanbing Liu, Lang Cao, Yuanyi Ren, Mengyu Zhou, Haoyu Dong, Xiaojun Ma, Shi Han and Dongmei Zhang

Large language models (LLMs) show strong reasoning abilities but often produce unnecessarily long explanations that reduce efficiency. Although reinforcement learning (RL) has been used to improve reasoning, most methods focus on accuracy and rely on uniform length-based rewards that overlook the differing contributions of individual tokens, often harming correctness. We revisit length optimization in RL through the perspective of token significance. Observing that many chain-of-thought (CoT) tokens contribute little to the final answer, we introduce a significance-aware length reward that selectively penalizes insignificance tokens, reducing redundancy while preserving essential reasoning. We also propose a dynamic length reward that encourages more detailed reasoning early in training and gradually shifts toward conciseness as learning progresses. Integrating these components into standard policy optimization yields a framework that improves both reasoning efficiency and accuracy. Experiments across multiple benchmarks demonstrate substantial reductions in response length while preserving or improving correctness, highlighting the importance of modeling token significance for efficient LLM reasoning.

[Long] Revisiting Model Interpolation for Efficient Reasoning ▶ *Underline Video*

Taiqiang Wu, RUNMING YANG, Tao Liu, Jiahao Wang and Ngai Wong

Model merging, typically on Instruct and Thinking models, has shown remarkable performance for efficient reasoning. In this paper, we systematically revisit the simplest merging method that interpolates two weights directly. Particularly, we observe that model interpolation follows a three-stage evolutionary paradigm with distinct behaviors on the reasoning trajectory. These dynamics provide a principled guide

for navigating the performance-cost trade-off. Empirical results demonstrate that a strategically interpolated model surprisingly surpasses sophisticated model merging baselines on both efficiency and effectiveness. We further validate our findings with extensive ablation studies on model layers, modules, and decoding strategies. Ultimately, this work demystifies model interpolation and offers a practical framework for crafting models with precisely targeted reasoning capabilities.

[Long] Sherry: Hardware-Efficient 1.25-Bit Ternary Quantization via Fine-grained Sparsification ▶ *Underline Video*

Hong Huang, Decheng Wu, Qiangqiang Hu, Guanghua Yu, Jinhai Yang, Jianchen Zhu, Xue Liu and Dapeng Wu

The deployment of Large Language Models (LLMs) on resource-constrained edge devices is increasingly hindered by prohibitive memory and computational requirements. While ternary quantization offers a compelling solution by reducing weights to -1, 0, +1, current implementations suffer from a fundamental misalignment with commodity hardware. Most existing methods must choose between 2-bit aligned packing, which incurs significant bit wastage, or 1.67-bit irregular packing, which degrades inference speed. To resolve this tension, we propose Sherry, a hardware-efficient ternary quantization framework. Sherry introduces a 3:4 fine-grained sparsity that achieves a regularized 1.25-bit width by packing blocks of four weights into five bits, restoring power-of-two alignment. Furthermore, we identify weight trapping issue in sparse ternary training, which leads to representational collapse. To address this, Sherry introduces Arenas, an annealing residual synapse mechanism that maintains representational diversity during training. Empirical evaluations on LLaMA-3.2 across five benchmarks demonstrate that Sherry matches state-of-the-art ternary performance while significantly reducing model size. Notably, on an Intel i7-14700HX CPU, our 1B model achieves zero accuracy loss compared to SOTA baselines while providing 25% bit savings and 10% speed up. The code is available at https://github.com/Tencent/AngelSlim.

[Long] SpecCache: Speculative KV Cache Reuse for Efficient RAG Serving ▶ *Underline Video*

ZijianWen ZijianWen, Tao Zhang, Shuangwu chen, Shenghao Ye, Yu Guo, Qirui Chen, Jingxian Shuai, Yunpeng Hou, Huasen He and jianyang jianyang

Retrieval-Augmented Generation (RAG) significantly enhances LLMs but faces high prefill latency during long-context processing. While KV cache reuse can mitigate this, current methods relying on shallow features or static heuristics often fail to identify critical tokens for recomputation, resulting in generation quality degradation. We have an insight that KV deviations are more pronounced in deep layers. However, directly extracting deep-layer features from the target model is computationally prohibitive. Crucially, we find that the deep-layer features of a lightweight speculative model exhibit strong consistency with the target model in the selection of critical tokens for recomputation. In light of these insights, we propose SpecCache, which employs deep-layer hidden-state norms from a speculative model as a proxy to guide the critical token selection for target large model. Experiments demonstrate that SpecCache outperforms state-of-the-art (SOTA) baselines. Compared to full KV recomputation, it reduces time-to-first-token (TTFT) by 2.17-3.95x and increases inference throughput by 2.7-5.2x, with negligible degradation in generation quality relative to full recomputation.

[Long] Step-GRPO: Internalizing Dynamic Early Exit for Efficient Reasoning ▶ *Underline Video*

Benteng Chen, Weida Wang, Shufei Zhang, Mingbao Lin and Min Zhang

Large reasoning models that use long chain-of-thought excel at problem-solving yet waste compute on redundant checks. Curbing this overthinking is hard: training-time length penalties can cripple ability, while inference-time early-exit adds system overhead. To bridge this gap, we propose Step-GRPO, a novel post-training framework that internalizes dynamic early-exit capabilities directly into the model. Step-GRPO shifts the optimization objective from raw tokens to semantic steps by utilizing linguistic markers to structure reasoning. We introduce a Dynamic Truncated Rollout mechanism that exposes the model to concise high-confidence trajectories during exploration, synergized with a Step-Aware Relative Reward that dynamically penalizes redundancy based on group-level baselines. Extensive experiments across three model sizes on diverse benchmarks demonstrate that Step-GRPO achieves a superior accuracy-efficiency trade-off. On Qwen3-8B, our method reduces token consumption by 32.0% compared to the vanilla model while avoiding the accuracy degradation observed in traditional length-penalty methods.

Oral Session F: Multimodality and Language Grounding to Vision, Robotics and Beyond 2

09:00-10:30 (Harbor H-I)

[Long] Breaking Down and Building Up: Mixture of Skill-Based Vision-and-Language Navigation Agents ▶ *Underline Video*

Tianyi Ma, Yue Zhang, Zehao Wang and Parisa Kordjamshidi

Vision-and-Language Navigation (VLN) poses significant challenges for agents to interpret natural language instructions and navigate complex 3D environments. While recent progress has been driven by large-scale pre-training and data augmentation, current methods still struggle to generalize to unseen scenarios, particularly when complex spatial and temporal reasoning is required. In this work, we propose SkillNav, a modular framework that introduces structured, skill-based reasoning into Transformer-based VLN agents. Our method decomposes navigation into a set of interpretable atomic skills (e.g., Vertical Movement, Area and Region Identification, Stop and Pause), each handled by a specialized agent. To support targeted skill training without manual data annotation, we construct a synthetic dataset pipeline that generates diverse, linguistically natural, skill-specific instruction-trajectory pairs. We then introduce a novel training-free

Vision-Language Model (VLM)-based router, which dynamically selects the most suitable agent at each time step by aligning sub-goals with visual observations and previous actions. SkillNav obtains competitive results on commonly used benchmarks and establishes state-of-the-art generalization to the GSA-R2R, a benchmark with novel instruction styles and unseen environments.

[Long] GuideDog: A Real-World Egocentric Multimodal Dataset for Blind and Low-Vision Accessibility-Aware Guidance ► *Underline Video*

Junhyeok Kim, Jaewoo Park, Junhee Park, Sangeyl Lee, Jiwan Chung, Jisung Kim, Ji Hoon Joung and Youngjae Yu

For people affected by blindness and low vision (BLV), safe and independent navigation remains a major challenge, impacting over 2.2 billion individuals worldwide. Although multimodal large language models (MLLMs) offer new opportunities for assistive navigation, progress has been limited by the scarcity of accessibility-aware datasets, requiring labor-intensive, expert annotation. To this end, we introduce GuideDog, a novel dataset containing 22K image-description pairs (2K human-verified) capturing real-world pedestrian scenes across 46 countries. Our human-AI pipeline shifts annotation from generation to verification, grounded in established BLV guidance standards from experts and research, improving scalability while maintaining quality. We also present GuideDogQA, an 818-sample benchmark evaluating object recognition and depth perception. Experiments reveal that depth perception and adherence to these standards remain challenging for current MLLMs. Code and dataset will be publicly available.

[Long] MMSearch-R1: Incentivizing LMMs to Search ► *Underline Video*

Jinming Wu, Zihao Deng, Wei Li, Yiding Liu, Bo You, Bo Li, Zejun Ma and Ziwei Liu

Robust deployment of large multimodal models (LMMs) in real-world scenarios requires access to external knowledge sources, given the complexity and dynamic nature of real-world information. Existing approaches such as retrieval-augmented generation (RAG) and prompt engineered search agents rely on rigid pipelines, often leading to inefficient or excessive search behaviors. We present MMSearch-R1, the first end-to-end reinforcement learning framework that enables LMMs to perform on-demand, multi-turn search in real-world Internet environments. Our framework integrates both image and text search tools, allowing the model to reason about when and how to invoke them guided by an outcome-based reward with a search penalty. To support training, We collect a multimodal search VQA dataset through a semi-automated pipeline that covers diverse visual and textual knowledge needs and curate a search-balanced subset with both search-required and search-free samples, which proves essential for shaping efficient and on-demand search behavior. Extensive experiments on knowledge-intensive and info-seeking VQA tasks show that our model not only outperforms traditional RAG-based baselines of the same model size, but also matches the performance of a larger RAG-based model while reducing search calls by over 30%. We further analyze key empirical findings to offer actionable insights for advancing research in multimodal search.

[Long] ProgressLM: Towards Progress Reasoning in Vision-Language Models ► *Underline Video*

Jianshu Zhang, Chengxuan Qian, Haosen Sun, Haoran Lu, Dingcheng Wang, Letian Xue and Han Liu

Estimating task progress requires long-horizon and dynamic reasoning, going beyond static visual perception. Although Vision-Language Models (VLMs) excel at describing what is visible in a single observation, it remains unclear whether they can infer how far a task has progressed from partial information. To study this question, we introduce Progress-Bench, a benchmark with over 3K instances for evaluating progress reasoning from a single observation. We further examine a human-inspired two-stage paradigm that combines episodic retrieval with mental simulation. We instantiate this paradigm through both training-free prompting and a training-based approach using the automatically curated ProgressLM-45K dataset. Experiments on 14 VLMs show that most models struggle with reliable progress estimation, and that training-free reasoning provides only limited and model-dependent benefits. In contrast, the training-based ProgressLM-3B achieves consistent improvements in accuracy, robustness to viewpoint variation, and handling of unanswerable cases, despite its small scale. Additional analyses reveal common failure patterns in existing VLMs and clarify when and why progress reasoning succeeds or fails.

[Long] SEE: Signal Embedding Energy for Quantifying Noise Interference in Large Audio Language Models ► *Underline Video*

Yuanhe Zhang, Jiayu Tian, Yibo Zhang, Shilinlu Yan, Liang Lin, Zhenhong Zhou, Li Sun and Sen Su

Large Audio Language Models (LALMs) have been widely applied in real-time scenarios, such as in-car assistants and online meeting comprehension. In practice, audio inputs are often corrupted by device and environmental noise, leading to performance degradation. However, existing LALM studies on noise lack quantitative analysis and rely mainly on intuition and empirical observation, thus failing to understand practical robustness. To address this issue, we introduce Signal Embedding Energy (SEE), a method for quantifying the impact of noise intensity on LALM inputs, enabling the differentiation of LALM robustness in real-world deployments. SEE introduces a perspective based on structured activation subspaces derived from the model's internal representations, which more accurately captures its perception of noise than raw audio features. Across experiments, SEE exhibits a strong correlation with LALM performance, achieving a correlation of 0.98. Surprisingly, traditional audio denoising methods are only marginally effective for LALMs, and, in some cases, even increase SEE and impair performance. This suggests a mismatch between speech-centric denoising objectives and the noise sensitivity of modern LALMs. Therefore, we propose a mitigation strategy derived from SEE to denoise LALM inputs, outperforming existing denoising methods. This paper introduces a novel metric for noise quantification in LALMs, providing guidance for robustness improvements in real-world deployments.

[Long] Think before Go: Hierarchical Reasoning for Image-goal Navigation ► *Underline Video*

Pengna Li, Kangyi Wu, Shaoqing Xu, Fang Li, Lin Zhao, Long Chen, Zhi-Xin Yang and Nanning Zheng

Image-goal navigation steers an agent to a target location specified by an image in unseen environments. Existing methods primarily handle this task by learning an end-to-end navigation policy, which compares the similarities of target and observation images and directly predicts the actions. However, when the target is distant or lies in another room, such methods fail to extract informative visual cues, leading the agent to wander around. Motivated by the human cognitive principle that deliberate, high-level reasoning guides fast, reactive execution in complex tasks, we propose Hierarchical Reasoning Navigation (HRNav), a framework that decomposes image-goal navigation into high-level planning and low-level execution. In high-level planning, a vision-language model is trained on a self-collected dataset to generate a short-horizon plan, such as whether the agent should walk through the door or down the hallway. This downgrades the difficulty of the long-horizon task, making it more amenable to the execution part. In low-level execution, an online reinforcement learning policy is utilized to decide actions conditioned on the short-horizon plan. We also devise a novel Wandering Suppression Penalty (WSP) to further reduce the wandering problem. Together, these components form a hierarchical framew ork for Image-Goal Navigation. Extensive experiments in both simulation and real-world environments demonstrate the superiority of our method.

[Long] Unleashing Spatial Reasoning in Multimodal Large Language Models via Textual Representation Guided Reasoning ► *Underline Video*

Jiacheng Hua, Yishu Yin, Yuhang Wu, Tai Wang, Yifei Huang and Miao Liu

Existing Multimodal Large Language Models (MLLMs) struggle with 3D spatial reasoning, as they fail to construct structured abstractions of the 3D environment depicted in video inputs. To bridge this gap, drawing inspiration from cognitive theories of allocentric spatial reasoning, we investigate how to enable MLLMs to model and reason over text-based spatial representations of video. Specifically, we introduce Textual Representation of Allocentric Context from Egocentric Video (TRACE), a prompting method that induces MLLMs to generate text-based representations of 3D environments as intermediate reasoning traces for more accurate spatial question answering. TRACE encodes meta-context, camera trajectories, and detailed object entities to support structured spatial reasoning over egocentric videos. Extensive experiments on VSI-Bench and OST-Bench demonstrate that TRACE yields notable and consistent improvements over prior prompting strategies across a diverse range of MLLM backbones, spanning different parameter scales and training schemas. We further present ablation studies to validate our design choices, along with detailed analyses that probe the bottlenecks of 3D spatial reasoning in MLLMs.

[Long] Visually-Guided Policy Optimization for Multimodal Reasoning ► *Underline Video*

Zengbin Wang, Feng Xiong, Liang Lin, Xuecai Hu, Yong Wang, Yanlin Wang, Man Zhang and Xiangxiang Chu

Reinforcement learning with verifiable rewards (RLVR) has significantly advanced the reasoning ability of vision-language models (VLMs). However, the inherent text-dominated nature of VLMs often leads to insufficient visual faithfulness, characterized by sparse attention activation to visual tokens. More importantly, our empirical analysis reveals that temporal visual forgetting along reasoning steps exacerbates this deficiency. To bridge this gap, we propose Visually-Guided Policy Optimization (VGPO), a novel framework to reinforce visual focus during policy optimization. Specifically, VGPO initially introduces a Visual Attention Compensation mechanism that leverages visual similarity to localize and amplify visual cues, while progressively elevating visual expectations in later steps to counteract visual forgetting. Building on this mechanism, we implement a dual-grained advantage re-weighting strategy: the intra-trajectory level highlights tokens exhibiting relatively high visual activation, while the inter-trajectory level prioritizes trajectories demonstrating superior visual accumulation. Extensive experiments demonstrate that VGPO achieves better visual activation and superior performance in mathematical multimodal reasoning and visual-dependent tasks. The code has been released at https://github.com/wzb-bupt/VGPO.

[Long] When Vision-Language Models Judge Without Seeing: Exposing Informativeness Bias ► *Underline Video*

Xiaohan Zou, Roshan Sridhar, Mohammadtaher Safarzadeh and Dan Roth

The reliability of VLM-as-a-Judge is critical for the automatic evaluation of vision-language models (VLMs). Despite recent progress, our analysis reveals that VLM-as-a-Judge often pays limited attention to the image when making decisions. Instead, they often blindly favor the more informative answer, even when they can recognize it conflicts with the image content. We call this problem informativeness bias, which significantly undermines judge reliability. To address it, we propose BIRCH (Balanced Informativeness and CoRrectness with a Truthful AnCHor), a judging paradigm that first corrects inconsistencies with the image content in candidate answers, and then compares the answers against this corrected version. This shifts the judges focus from informativeness to image-grounded correctness. Experiments on multiple models and benchmarks show that BIRCH reduces informativeness bias by up to 17%, resulting in performance gains of up to 9.8%. Our work reveals an overlooked but fundamental flaw in current VLM-as-a-Judge systems and highlights the need for more principled designs.

Oral Session F: Natural Language Generation 1

09:00-10:30 (Regatta)

[Long] AFT-Tab: Adversarial Fine-Tuning for Tabular Data Synthesis with Long Text Columns ► *Underline Video*

Yuhao Zhang, Liang Yan, Shaoming Duan, Xinyu Zha, Jinhang Su, Peiyi Han and Chuanyi Liu

Traditional tabular data synthesis methods often overlook the cross-modal heterogeneity of real-world tables, where structured continuous

and discrete attributes coexist with unstructured long-text columns. Existing synthesis approaches struggle to simultaneously achieve accurate statistical fidelity for non-textual attributes and consistent semantic constraints between textual and non-textual attributes. In this work, we establish the first benchmark for long-text tabular data synthesis and introduce a novel metric, termed Textual Column Correlation Fidelity (TCCF), to quantify cross-modal semantic alignment. We propose AFT-Tab, an adversarial fine-tuning framework that synergistically trains an LLM-based text generator and a deep-learning-based non-textual generator. Through a dual-feedback mechanism guided by an LLM discriminator, AFT-Tab ensures both precise statistical distributions and rigorous semantic constraints. Experimental results show that AFT-Tab significantly outperforms state-of-the-art baselines in statistical fidelity, TCCF, diversity, and downstream task utility.

[Long] ConlangCrafter: Constructing Languages with a Multi-Hop LLM Pipeline ▶ *Underline Video*

Morris Alper, Moran Yanuka, Raja Giryes and Gasper Begus

Constructed languages (conlangs) such as Esperanto and Quenya have played diverse roles in art, philosophy, and international communication. Meanwhile, foundation models have revolutionized creative generation in text, images, and beyond. In this work, we leverage modern LLMs as computational creativity aids for end-to-end conlang creation. We introduce ConlangCrafter, a multi-hop pipeline that decomposes language design into modular stages – phonology, morphology, syntax, lexicon generation, and translation. At each stage, our method leverages LLMs' metalinguistic reasoning capabilities, injecting randomness to encourage diversity and leveraging self-refinement feedback to encourage consistency in the emerging language description. We construct a novel, scalable evaluation framework for this task, evaluating metrics measuring consistency and typological diversity. Automatic and manual evaluations demonstrate ConlangCrafter's ability to produce coherent and varied conlangs without human linguistic expertise. We will release our code and data.

[Long] DPWriter: Reinforcement Learning with Diverse Planning Branching for Creative Writing ▶ *Underline Video*

Qian Cao, Yahui Liu, Wei Bi, Yi Zhao, Ruihua Song, Xiting Wang, Ruiming Tang, Guorui Zhou and Han Li

Reinforcement learning (RL)-based enhancement of large language models (LLMs) often leads to reduced output diversity, undermining their utility in open-ended tasks like creative writing. Current methods lack explicit mechanisms for guiding diverse exploration and instead prioritize optimization efficiency and performance over diversity. This paper proposes an RL framework structured around a semi-structured long Chain-of-Thought (CoT), in which the generation process is decomposed into explicitly planned intermediate steps. We introduce a Diverse Planning Branching method that strategically introduces divergence at the planning phase based on diversity variation, alongside a group-aware diversity reward to encourage distinct trajectories. Experimental results on creative writing benchmarks demonstrate that our approach significantly improves output diversity without compromising generation quality, consistently outperforming existing baselines.

[Long] Efficient Test-Time Scaling of Multi-Step Reasoning by Probing Internal States of Large Language Models ▶ *Underline Video*

Ekaterina Fadeeva, Tianyi Wu, Mubashara Akhtar, Jiaheng Zhang, Elliott Ash, Markus Leippold, Timothy Baldwin, See-Kiong Ng, Artem Shelmanov and Mrinmaya Sachan

LLMs can solve complex tasks by generating long, multi-step reasoning chains. Test-time scaling (TTS) can further improve LLM performance by sampling multiple variants of intermediate reasoning steps, verifying their correctness, and strategically choosing the best steps for continuation. However, existing verification approaches, such as Process Reward Models (PRMs), are computationally expensive, limited to specific domains, and require large-scale human or model-generated annotations. We propose a lightweight alternative for step-level reasoning verification based on probing the internal states of LLMs. We train a transformer-based probe that uses the internal states of the frozen LLM to estimate the credibility of its reasoning steps during generation. Annotation can be generated either by another larger LLM (e.g., DeepSeek-R1) or in a self-supervised manner by the original model itself. The probes are both effective and lightweight, containing fewer than 10M parameters. Across multiple domains, including mathematics, planning, and general knowledge question answering, our probes match or even exceed the performance of PRMs that are up to 810x larger. Our findings suggest that the internal states of LLMs encode their confidence in reasoning processes and can serve as reliable signals for reasoning step verification, offering a promising direction towards scalable and generalizable TTS and introspective LLMs.

[Long] Evolutionary Guided Decoding: Iterative Value Refinement for LLMs ▶ *Underline Video*

Zhenhua Liu, Lijun Li, Ruizhe Chen, Yuxian Jiang, Tong Zhu, Zhaochen Su, Wenliang Chen and Jing Shao

While guided decoding, especially value-guided methods, has emerged as a cost-effective alternative for controlling language model outputs without re-training models, its effectiveness is limited by the accuracy of the value function. We identify that this inaccuracy stems from a core distributional gap: existing methods train static value functions on trajectories sampled exclusively from the base policy, which inherently confines their training to a narrow and suboptimal view of the potential output space. We propose Iterative Value Refinement, a novel framework designed to bridge this gap. It employs Value Exploration to provide a more comprehensive and robust training signal, complemented by Iterative Self-Refinement, which uses the improved value function from one iteration to guide the generation of higher-quality data for the next. Extensive experiments on text summarization, multi-turn dialogue, and instruction following demonstrate the effectiveness of our framework in aligning language models. Our approach not only achieves alignment but also significantly reduces computational costs by leveraging principled value function optimization for efficient and effective control.

[Long] GenProve: Learning to Generate Text with Fine-Grained Provenance ▶ *Underline Video*

Jingxuan Wei, Xingyue Wang, Yanghaoyu Liao, Jie Dong, Yuchen Liu, Caijun Jia, Bihui Yu and Junnan Zhu

Large language models (LLM) often hallucinate, and while adding citations is a common solution, it is frequently insufficient for

accountability as users struggle to verify how a cited source supports a generated claim. Existing methods are typically coarse-grained and fail to distinguish between direct quotes and complex reasoning. In this paper, we introduce Generation-time Fine-grained Provenance, a task where models must generate fluent answers while simultaneously producing structured, sentence-level provenance triples. To enable this, we present ReFInE (Relation-aware Fine-grained Interpretability & Evidence), a dataset featuring expert-verified annotations that distinguish between Quotation, Compression, and Inference. Building on ReFInE, we propose GenProve, a framework that combines Supervised Fine-Tuning (SFT) with Group Relative Policy Optimization (GRPO). By optimizing a composite reward for answer fidelity and provenance correctness, GenProve significantly outperforms 14 strong LLMs in joint evaluation. Crucially, our analysis uncovers a reasoning gap where models excel at surface-level quotation but struggle significantly with inference-based provenance, suggesting that verifiable reasoning remains a frontier challenge distinct from surface-level citation.

[Long] iTAG: Inverse Design for Natural Text Generation with Accurate Causal Graph Annotations ▶ *Underline Video*
Wenshuo Wang, Boyu Cao, Nan Zhuang and Wei Li
A fundamental obstacle to causal discovery from text is the lack of causally annotated text data for use as ground truth, due to high annotation costs. This motivates an important task of generating text with causal graph annotations. Early template-based generation methods sacrifice text naturalness in exchange for high causal graph annotation accuracy. Recent Large Language Model (LLM)-dependent methods directly generate natural text from target graphs through LLMs, but do not guarantee causal graph annotation accuracy. Therefore, we propose iTAG, which performs real-world concept assignment to nodes before converting causal graphs into text in existing LLM-dependent methods. iTAG frames this process as an inverse problem with the causal graph as the target, iteratively examining and refining concept selection through Chain-of-Thought (CoT) reasoning so that the induced relations between concepts are as consistent as possible with the target causal relationships described by the causal graph. iTAG demonstrates both extremely high annotation accuracy and naturalness across extensive tests, and the results of testing text-based causal discovery algorithms with the generated data show high statistical correlation with real-world data. This suggests that iTAG-generated data can serve as a practical surrogate for scalable benchmarking of text-based causal discovery algorithms.

[Long] Min-k Sampling: Decoupling Truncation from Temperature Scaling via Relative Logit Dynamics ▶ *Underline Video*
Yuanhao Ding, Meimingwei Li, Esteban Garces Arias, Matthias ASSenmacher, Christian Heumann and Chongsheng ZHANG
The quality of text generated by large language models depends critically on the decoding sampling strategy. While mainstream methods such as Top-k, Top-p, and Min-p achieve a balance between diversity and accuracy through probability-space truncation, they share an inherent limitation: extreme sensitivity to the temperature parameter. Recent logit-space approaches like Top-nsigma achieve temperature invariance but rely on global statistics that are susceptible to long-tail noise, failing to capture fine-grained confidence structures among top candidates. We propose Min-k Sampling, a novel dynamic truncation strategy that analyzes the local shape of the sorted logit distribution to identify "semantic cliffs": sharp transitions from high-confidence core tokens to uncertain long-tail tokens. By computing a position-weighted relative decay rate, Min-k dynamically determines truncation boundaries at each generation step. We formally prove that Min-k achieves strict temperature invariance and empirically demonstrate its low sensitivity to hyperparameter choices. Experiments on multiple reasoning benchmarks, creative writing tasks, and human evaluation show that Min-k consistently improves text quality, maintaining robust performance even under extreme temperature settings where probability-based methods collapse. We make our code, models, and analysis tools publicly available.

[Short] On the Hidden Objective Biases of Group-based Reinforcement Learning ▶ *Underline Video*
Aleksandar Fontana, Marco Simoni, Giulio Rossolini, Paolo Mori and Andrea Saracino
Group-based reinforcement learning methods, like Group Relative Policy Optimization (GRPO), are widely used nowadays to post-train large language models. Despite their empirical success, they exhibit structural mismatches between reward optimization and the underlying training objective. In this paper, we present a theoretical analysis of GRPO style methods by studying them within a unified surrogate formulation. This perspective reveals recurring properties that affect all the methods under analysis: (i) non-uniform group weighting induces systematic gradient biases on shared prefix tokens; (ii) interactions with the AdamW optimizer make training dynamics largely insensitive to reward scaling; and (iii) optimizer momentum can push policy updates beyond the intended clipping region under repeated optimization steps. We believe that these findings highlight fundamental limitations of current approaches and provide principled guidance for the design of future formulations.

Oral Session F: Speech Processing and Spoken Language Understanding 2

09:00-10:30 (Harbor A)

[Long] Aligning Backchannel and Dialogue Context Representations via Contrastive LLM Fine-Tuning ▶ *Underline Video*
Livia Qian and Gabriel Skantze
Backchannels (e.g., yeah, mhm, and right) are short, non-interruptive feedback signals whose lexical form and prosody jointly convey pragmatic meaning. While prior computational research has largely focused on predicting backchannel timing, the relationship between lexico-prosodic form and meaning remains underexplored. We propose a two-stage framework: first, fine-tuning large language models

on dialogue transcripts to derive rich contextual representations; and second, learning a joint embedding space for dialogue contexts and backchannel realizations. We evaluate alignment with human perception via triadic similarity judgments (prosodic and cross-lexical) and a context-backchannel suitability task. Our results demonstrate that the learned projections substantially improve context-backchannel retrieval compared to previous methods. In addition, they reveal that backchannel form is highly sensitive to extended conversational context and that the learned embeddings align more closely with human judgments than raw WavLM features.

[Long] Audio MultiChallenge: A Multi-Turn Evaluation of Spoken Dialogue Systems on Natural Human Interaction ► *Underline Video*

Advait Gosai, Tyler Vuong, Utkarsh Tyagi, Steven Li, Wenjia You, Miheer Bavare, Arda Uçar, Zhongwang Fang, Brian Jang, Bing Liu and Yunzhong He

End-to-end (E2E) spoken dialogue systems are replacing cascaded pipelines for voice-based human-AI interaction. Existing benchmarks primarily evaluate these systems on synthetic speech and single-turn tasks, leaving multi-turn conversational ability underexplored. We introduce Audio MultiChallenge an open-source benchmark to evaluate these systems under natural multi-turn interaction patterns. Building on the text-based MultiChallenge framework, which evaluates Inference Memory, Instruction Retention, and Self Coherence, we introduce a new axis Voice Editing that tests robustness to mid-utterance speech repairs and backtracking. We augment each axis to the audio modality, such as introducing Audio-Cue challenges for Inference Memory that require recalling ambient sounds and paralinguistic signals beyond semantic content. We curate 452 conversations from 47 speakers with 1,712 instance-specific rubrics through a hybrid pipeline that exposes model failures at scale while preserving natural disfluencies found in unscripted human speech. Our evaluation reveals that even frontier models struggle on our benchmark, with our highest-performing model achieving a 54.65% pass rate. Error analysis shows that models are not sufficiently robust to human speech when tracking instructions, edits, and audio cues, highlighting the need for improved audio-native multi-turn interaction capabilities.

[Long] Beyond Transcripts: A Renewed Perspective on Audio Chaptering ► *Underline Video*

Fabian Retkowski, Maike Züfle, Thai Binh Nguyen, Jan Niehues and Alexander Waibel

Audio chaptering, the task of automatically segmenting long-form audio into coherent sections, is increasingly important for navigating podcasts, lectures, and videos. Despite its relevance, research remains limited and text-based, leaving key questions unresolved about leveraging audio information, handling ASR errors, and transcript-free evaluation. We address these gaps through three contributions: (1) a systematic comparison between text-based models with acoustic features, a novel audio-only architecture (AudioSeg) operating on learned audio representations, and multimodal LLMs; (2) empirical analysis of factors affecting performance, including transcript quality, acoustic features, duration, and speaker composition; and (3) formalized evaluation protocols contrasting transcript-dependent text-space protocols with transcript-invariant time-space protocols. Our experiments on YTSeg reveal that AudioSeg substantially outperforms text-based approaches, pauses provide the largest acoustic gains, and current MLLMs struggle due to context limitations and weak instruction following.

[Long] ControlAudio: Tackling Text-Guided, Timing-Indicated and Intelligible Audio Generation via Progressive Diffusion Modeling ► *Underline Video*

Yuxuan Jiang, Zehua Chen, Zeqian Ju, Yusheng Dai, Weibei Dou and Jun Zhu

Recent efforts on text-to-audio (TTA) generation are starting to explore fine-grained controllability, e.g., precise timing control, with innovations on conditioning techniques or training-free latent manipulations. However, constrained by data scarcity, their generation performance at scale is still limited. In this study, we recast high-controllability TTA generation as a multi-task learning problem, and introduce a progressive diffusion modeling approach, ControlAudio. Our method adeptly fits distributions conditioned on fine-grained information, including text, timing, and phoneme features, through a step-by-step strategy. First, we propose a data construction method spanning both annotation and simulation, augmenting condition information in the sequence of text, timing, and phoneme. Second, at the model training stage, we pretrain a scalable diffusion transformer (DiT) on large-scale text-audio pairs, achieving high-fidelity TTA generation, and then incrementally integrate the timing and phoneme features, expanding controllability. Finally, at the inference stage, we propose progressively guided generation, which sequentially emphasizes more fine-grained information, aligning inherently with the coarse-to-fine sampling nature of DiT. Extensive experiments show that ControlAudio achieves state-of-the-art performance in terms of temporal accuracy and speech clarity, significantly outperforming existing methods on both objective and subjective evaluations. Demo samples are available at: https://control-audio.github.io/Control-Audio.

[Short] Data-efficient Targeted Token-level Preference Optimization for LLM-based Text-to-Speech ► *Underline Video*

Rikuto Kotoge and Yuichi Sasaki

Aligning text-to-speech (TTS) system outputs with human feedback through preference optimization has been shown to effectively improve the robustness and naturalness of LLM-based TTS models. Current approaches primarily require paired desirable and undesirable samples at the utterance level. However, such pairs are often limited in TTS output data, and utterance-level formulation prevents fine-grained token-level optimization needed for accurate pronunciation alignment. In this study, we propose TKTO that eliminates the need for paired data, enabling a more data-efficient training paradigm, and directly targets token-level units, automatically providing fine-grained alignment signals without token-level annotations. TKTO improves the challenging Japanese TTS accuracy by 39% and reduces CER by 54%, leveraging 6x more training data and assigning 12.8x stronger reward to targeted tokens.

[Long] Deep Supervised Contrastive Learning of Pitch Contours for Robust Pitch Accent Classification in Seoul Korean ▶ *Underline Video*

Hyunjung Joo and GyeongTaek Lee

The intonational structure of Seoul Korean has been defined with discrete tonal categories within the Autosegmental-Metrical model of intonational phonology. However, it is challenging to map continuous F_0 contours to these invariant categories due to variable F_0 realizations in real-world speech. Our paper proposes Dual-Glob, a deep supervised contrastive learning framework to robustly classify fine-grained pitch accent patterns in Seoul Korean. Unlike conventional local predictive models, our approach captures holistic F_0 contour shapes by enforcing structural consistency between clean and augmented views in a shared latent space. To this aim, we introduce the first large-scale benchmark dataset, consisting of manually annotated 10,093 Accentual Phrases in Seoul Korean. Experimental results show that our Dual-Glob significantly outperforms strong baseline models with state-of-the-art accuracy (77.75%) and F1-score (51.54%). Therefore, our work supports AM-based intonational phonology using data-driven methodology, showing that deep contrastive learning effectively captures holistic structural features of continuous F_0 contours.

[Long] Discovering and Causally Validating Emotion-Sensitive Neurons in Large Audio-Language Models ▶ *Underline Video*

Xiutian Zhao, Björn Schuller and Berrak Sisman

Emotion is a central dimension of spoken communication, yet, we still lack a mechanistic account of how modern large audio-language models (LALMs) encode it internally. We present the first neuron-level interpretability study of emotion-sensitive neurons (ESNs) in LALMs and provide causal evidence supporting the existence of such units in Qwen2.5-Omni, Kimi-Audio, and Audio Flamingo 3. Across these three widely used open-source models, we compare frequency-, entropy-, mean-deviation-, and contrast-based neuron selectors on multiple emotion recognition benchmarks. Using inference-time interventions, we reveal a consistent emotion-specific signature: deactivating neurons selected for a given emotion disproportionately degrades recognition of that emotion while largely preserving other classes, whereas targeted steering amplifies these units to bias predictions toward the target emotion. These effects arise with modest amounts of identification data and scale systematically with intervention strength. We further observe that ESNs exhibit non-uniform layer-wise clustering with partial cross-dataset transfer. Taken together, our results offer a causal, neuron-level account of emotion decisions in LALMs and highlight targeted neuron interventions as an actionable handle for controllable affective behaviors.

[Long] Dual-Axis Generative Reward Model Toward Semantic and Turn-taking Robustness in Interactive Spoken Dialogue Models ▶ *Underline Video*

Yifu Chen, Shengpeng Ji, Zhengqing Liu, Qian Chen, Wen Wang, Ziqing Wang, Yangzhuo Li, Tianle Liang and Zhou Zhao

Achieving seamless, human-like interaction remains a key challenge for full-duplex spoken dialogue models (SDMs). Reinforcement learning (RL) has substantially enhanced text- and vision-language models, while well-designed reward signals are crucial for the performance of RL. We consider RL a promising strategy to address the key challenge for SDMs. However, a fundamental barrier persists: prevailing automated metrics for assessing interaction quality rely on superficial proxies, such as behavioral statistics or timing-prediction accuracy, failing to provide reliable reward signals for RL. On the other hand, human evaluations, despite their richness, remain costly, inconsistent, and difficult to scale. We tackle this critical barrier by proposing a Dual-Axis Generative Reward Model, which is trained to understand complex interaction dynamics using a detailed taxonomy and an annotated dataset, produces a single score and, crucially, provides separate evaluations for semantic quality and interaction timing. Such dual outputs furnish precise diagnostic feedback for SDMs and deliver a dependable, instructive reward signal suitable for online reinforcement learning. Our model achieves state-of-the-art performance on interaction-quality assessment across a wide spectrum of datasets, spanning synthetic dialogues and complex real-world interactions.

[Long] TagSpeech: End-to-End Multi-Speaker ASR and Diarization with Fine-Grained Temporal Grounding ▶ *Underline Video*

Mingyue Huo, Yiwen Shao and Yuheng Zhang

We present TagSpeech, a unified LLM-based framework that utilizes Temporal Anchor Grounding for joint multi-speaker ASR and diarization. The framework is built on two key designs: (1) decoupled semantic and speaker streams fine-tuned via Serialized Output Training (SOT) to learn turn-taking dynamics; and (2) an interleaved time anchor mechanism that not only supports fine-grained timestamp prediction but also acts as a synchronization signal between semantic understanding and speaker tracking. Compared to previous works that primarily focus on speaker-attributed ASR or implicit diarization, TagSpeech addresses the challenge of fine-grained speaker-content alignment and explicitly models who spoke what and when in an end-to-end manner. Experiments on AMI and AliMeeting benchmarks demonstrate that our method achieves consistent improvements in Diarization Error Rate (DER) over strong end-to-end baselines, including Qwen-Omni and Gemini, particularly in handling complex speech overlaps. Moreover, TagSpeech employs a parameter-efficient training paradigm in which the LLM backbone is frozen and only lightweight projectors are trained, resulting in strong performance with low computational cost

Orals Session F: Safety and Alignment in LLMs 2 (Harbor D)

09:00-10:30 (Harbor D)

[Long] Teach a Reward Model to Correct Itself: Reward Guided Adversarial Failure Discovery for Robust Reward Model-

ing ▶ *Underline Video*
Pankayaraj Pathmanathan and Furong Huang
Reward models (RMs) trained from human preferences are central to aligning large language models, yet they often break under distribution shift or targeted perturbations. Existing failure discovery methods rely on prior knowledge of preference attributes and therefore do not scale to new models or data. We introduce a preference distribution agnostic procedure that uses the reward model itself to guide controlled decoding toward mis specified responses while preserving the underlying preference class. Building on this discovery mechanism, we propose REFORM, a self improving RM framework that (i) searches for class consistent but reward inconsistent variants and (ii) fine tunes the RM on a small, targeted augmentation of these failures. On Anthropic Helpful Harmless and PKU Beavertails, REFORM consistently improves robustness without degrading in distribution reward quality across different models (e.g., Mistral-7B and Qwen-14B), with an average improvement of 35%-45%.Further, across Best of N sampling, PPO, and DPO, REFORM preserves downstream generation quality and reduces spurious correlations. Our results show that RMs can serve as their own adversary to expose and fix blind spots, yielding robust alignment without manual attribute priors or large scale relabeling.

Orals Session F: Safety and Alignment in LLMs 2 (Harbor E-F)

09:00-10:30 (Harbor E-F)

[Long] Answering the Wrong Question: Reasoning Trace Inversion for Abstention in LLMs ▶ *Underline Video*
Abinitha Gourabathina, Inkit Padhi, Manish Nagireddy, Subhajit Chaudhury and Prasanna Sattigeri
For Large Language Models (LLMs) to be reliably deployed, models must effectively know when not to answer: abstain. Reasoning models, in particular, have gained attention for impressive performance on complex tasks. However, reasoning models have been shown to have worse abstention abilities. Taking the vulnerabilities of reasoning models into account, we propose our Query Misalignment Framework. Hallucinations resulting in failed abstention can be reinterpreted as LLMs answering the wrong question (rather than answering a question incorrectly). Based on this framework, we develop a new class of state-of-the-art abstention methods called Trace Inversion. First, we generate the reasoning trace of a model. Based on only the trace, we then reconstruct the most likely query that the model responded to. Finally, we compare the initial query with the reconstructed query. Low similarity score between the initial query and reconstructed query suggests that the model likely answered the question incorrectly and is flagged to abstain. Extensive experiments demonstrate that Trace Inversion effectively boosts abstention performance in four frontier LLMs across nine abstention QA datasets, beating competitive baselines in 33 out of 36 settings. The code is available at this https://anonymous.4open.science/r/trace-inversion-08BB/.

[Long] Debiasing Reward Models via Causally Motivated Inference-Time Intervention ▶ *Underline Video*
Kazutoshi Shinoda, Kosuke Nishida and Kyosuke Nishida
Reward models (RMs) play a central role in aligning large language models (LLMs) with human preferences. However, RMs are often sensitive to spurious features such as response length. Existing inference-time approaches for mitigating these biases typically focus exclusively on response length, resulting in performance trade-offs. In this paper, we propose causally motivated intervention for mitigating multiple types of biases in RMs at inference time. Our method first identifies neurons whose activations are strongly correlated with predefined bias attributes, and applies neuron-level intervention that suppresses these signals. We evaluate our method on RM benchmarks and observe reductions in sensitivity to spurious features across diverse bias types, without inducing performance trade-offs. Moreover, when used for preference annotation, small RMs (2B and 7B) with our method, which edits less than 2% of all the neurons in RMs, enable LLMs to improve alignment, achieving performance comparable to that of a state-of-the-art 70B RM on AlpacaEval and MT-Bench. Further analysis reveals that bias signals are primarily encoded by neurons in early layers, shedding light on the internal mechanisms of bias exploitation in RMs.

[Long] Domain Generalizable AI Guardrails with Augmented Policy Training ▶ *Underline Video*
Minqian Liu, Ioana Baldini, David Rabinowitz, David S Rosenberg, Sebastian Gehrmann and Mark Dredze
AI guardrail systems support usage policies by determining whether a user query or a generated response is allowed or forbidden under the policy. Fine-tuned guardrails – such as LlamaGuard and ShieldGemma – include policy definitions in prompts during training that can be updated during inference to aid generalization. However, our analysis reveals that these models still overfit the training policies, which prevents adaptation to new domains. We propose Augmented Policy Training (APT), a training recipe that enhances guardrail adaptability to unseen policies by using a suite of policy perturbation strategies during training to reduce overfitting and increase generalization. Notably, a small 1B model trained in this manner achieves comparable or better performance than existing 8B guardrails on unseen policies. Our work reveals critical limitations of existing AI guardrails, offers a promising solution, and provides actionable insights for adapting systems to new domains and policies.

[Long] FineSteer: A Unified Framework for Fine-Grained Inference-Time Steering in Large Language Models ▶ *Underline Video*
Zixuan Weng, Jinghuai Zhang, Kunlin Cai, Ying Li, Peiran Wang and Yuan Tian
Large language models (LLMs) often exhibit undesirable behaviors, such as safety violations and hallucinations. Although inference-time steering offers a cost-effective way to adjust model behavior without updating its parameters, existing methods often fail to be simultaneously

effective, utility-preserving, and training-efficient due to their rigid, one-size-fits-all designs and limited adaptability. In this work, we present FineSteer, a novel steering framework that decomposes inference-time steering into two complementary stages, conditional steering and fine-grained vector synthesis, allowing fine-grained control over when and how to steer internal representations. In the first stage, we introduce a Subspace-guided Conditional Steering (SCS) mechanism that preserves model utility by avoiding unnecessary steering. In the second stage, we propose a Mixture-of-Steering-Experts (MoSE) mechanism that captures the multimodal nature of desired steering behaviors and generates query-specific steering vectors for improved effectiveness. Through tailored designs in both SCS and MoSE, FineSteer maintains robust performance on general queries while adaptively optimizing steering vectors for targeted inputs in a training-efficient manner. Extensive experiments on safety and truthfulness benchmarks show that FineSteer outperforms the state-of-the-art methods in overall performance (e.g., a 7.6% improvement on TruthfulQA over Llama-3), achieving stronger steering performance with minimal utility loss. The code is available at https://github.com/YukinoAsuna/FineSteer

[Long] From Trust to Compromise: Outcome-Verified LLM Phishing Simulation and Real-Time Defense ▶ *Underline Video*

Tulika Tewari, Nalin Asanka Gamagedara Arachchilage, Jagat Sesh Challa and Dhruv Kumar

Large Language Models (LLMs) excel as conversational agents. However, these capabilities can be weaponized to automate social engineering attacks that gradually build rapport to compromise the online safety of users. To understand this, researchers have simulated LLM-based attacks in controlled settings. However, the existing simulators focus on just Personal Identifiable Information (PII) requests within the chat. Thus, to represent a complete attack scenario, we introduce PhishSim, an outcome-driven LLM-based phishing simulator that verifies compromise by simulating a victim completing an external action step, such as submitting credentials on a malicious platform. This enables the generation of diverse, multi-turn attack trajectories. Building on these trajectories, we position PhishGate as a practical mitigation baseline for outcome-grounded conversational phishing: a real-time multi-agent risk scorer that detects manipulation tactics and estimates the severity of ongoing chats. For ambiguous cases, it invokes RAG-supported consistency checks. Evaluating four state-of-the-art LLM backends in a real-time setting, we find that PhishGate improves dialogue-level detection over a real-time baseline. Our results highlight both the promise and brittleness of LLM-based real-time phishing defense, providing an outcome-grounded testbed for studying conversational compromise.

[Long] ImF: Embedding an Implicit Fingerprint in Your Large Language Models ▶ *Underline Video*

Jiaxuan Wu, Wanli Peng, hang fu, Xue yiming and Juan Wen

Training and serving large language models (LLMs) is resource-intensive, making reliable intellectual property (IP) protection and black-box ownership verification increasingly important. Model fingerprinting enables such verification by injecting a small set of secret query-response behaviors, but many existing fingerprints rely on explicit markers or predetermined outputs that are weakly grounded in prompt semantics. This semantic mismatch yields atypical fingerprint responses, reduces stealthiness, and exposes fingerprints to removal by response normalization. We formalize this vulnerability via a new removal attack, Generation Revision Intervention (GRI), which applies system-prompt-level revision and response standardization to steer models toward typical answers, substantially compromising representative injected baselines. To close this semantic gap, we propose the Implicit Fingerprints (ImF): we encode ownership information into a natural-looking target response y via linguistic steganography, then derive a CoT-augmented query x that embeds semantic cues from y to guide the model toward an output sufficiently close to y for decoding-based verification. Experiments on 15 LLMs show that ImF improves stealthiness and remains verifiable under model updates and deployment-time prompt interventions; additional analyses further show stability under common decoding variation and realistic related-model partial merging.

[Long] Into the Gray Zone: Domain Contexts Can Blur LLM Safety Boundaries ▶ *Underline Video*

Ki Sen Hung, Xi Yang, Chang Liu, Haoran Li, Kejiang Chen, Changxuan Fan, Tsun On Kwok, Weiming Zhang, Xiaomeng Li and Yangqiu Song

A central goal of LLM alignment is to balance helpfulness with harmlessness, yet these objectives conflict when the same knowledge serves both legitimate and malicious purposes. This tension is amplified by context-sensitive alignment: we observe that domain-specific contexts (e.g., chemistry) selectively relax defenses for domain-relevant harmful knowledge, while safety-research contexts (e.g., jailbreak studies) trigger broader relaxation spanning all harm categories. To systematically exploit this vulnerability, we propose Jargon, a framework combining safety-research contexts with multi-turn adversarial interactions that achieves attack success rates exceeding 93% across seven frontier models, including GPT-5.2, Claude-4.5, and Gemini-3, substantially outperforming existing methods. Activation space analysis reveals that Jargon queries occupy an intermediate region between benign and harmful inputs, a gray zone where refusal decisions become unreliable. To mitigate this vulnerability, we design a policy-guided safeguard that steers models toward helpful yet harmless responses, and internalize this capability through alignment fine-tuning, reducing attack success rates while preserving helpfulness.

[Long] WildReward: Learning Reward Models from In-the-Wild Human Interactions ▶ *Underline Video*

Hao Peng, Yunjia Qi, Xiaozhi Wang, Zijun Yao, Lei Hou and Juanzi Li

Reward models (RMs) are crucial for the training of large language models (LLMs), yet they typically rely on large-scale human-annotated preference pairs. With the widespread deployment of LLMs, in-the-wild interactions have emerged as a rich source of implicit reward signals. This raises the question: Can we develop reward models directly from in-the-wild interactions? In this work, we explore this possibility by adopting WildChat as an interaction source and proposing a pipeline to extract reliable human feedback, yielding 186k high-quality instances for training WildReward via ordinal regression directly on user feedback without preference pairs. Extensive experiments demonstrate that

WildReward achieves comparable or even superior performance compared to conventional reward models, with improved calibration and cross-sample consistency. We also observe that WildReward benefits directly from user diversity, where more users yield stronger reward models. Finally, we apply WildReward to online DPO training and observe significant improvements across various downstream tasks. We will release our code, data, and models to facilitate future research.

Poster Session F (Coronado)

09:00-10:30 (Coronado)

[Findings] A Survey on Evaluation of LLM-based Agents ▶ *Underline Video*

Asaf Yehudai, Lilach Eden, Alan Li, Guy Uziel, Yilun Zhao, Roy Bar-Haim, Arman Cohan and Michal Shmueli-Scheuer

LLM-based agents represent a paradigm shift in AI, enabling autonomous systems to plan, reason, and use tools while interacting with dynamic environments. This paper provides the first comprehensive survey of evaluation methods for these increasingly capable agents. We analyze the field of agent evaluation across five perspectives: (1) Core LLM capabilities needed for agentic workflows, like planning, and tool use; (2) Application-specific benchmarks such as web and SWE agents; (3) Evaluation of generalist agents; (4) Analysis of agent benchmarks' core dimensions; and (5) Evaluation frameworks and tools for agent developers. Our analysis reveals current trends, including a shift toward more realistic, challenging evaluations with continuously updated benchmarks. We also identify critical gaps that future research must address, particularly in assessing cost-efficiency, safety, and robustness, and in developing fine-grained, scalable evaluation methods.

[Findings] Abstain-R1: Calibrated Abstention and Post-Refusal Clarification via Verifiable RL ▶ *Underline Video*

Skylar Zhai, Jingcheng Liang and Dongyeop Kang

Reinforcement fine-tuning improves the reasoning ability of large language models, but it can also encourage them to answer unanswerable queries by guessing or hallucinating missing information. Existing abstention methods either train models to produce generic refusals or encourage follow-up clarifications without verifying whether those clarifications identify the key missing information. We study queries that are clear in meaning but cannot be reliably resolved from the given information, and argue that a reliable model should not only abstain, but also explain what is missing. We propose a clarification-aware RLVR reward that, while rewarding correct answers on answerable queries, jointly optimizes explicit abstention and semantically aligned post-refusal clarification on unanswerable queries. Using this reward, we train Abstain-R1, a 3B model that improves abstention and clarification on unanswerable queries while preserving strong performance on answerable ones. Experiments on Abstain-Test, Abstain-QA, and SelfAware show that Abstain-R1 substantially improves over its base model and achieves unanswerable-query behavior competitive with larger systems including DeepSeek-R1, suggesting that calibrated abstention and clarification can be learned through verifiable rewards rather than emerging from scale alone.

[Findings] AgentGC: Evolutionary Learning-based Lossless Compression for Genomics Data with LLM-driven Multiple Agent ▶ *Underline Video*

Sun Hui, Ding Yanfeng, Huidong Ma, Chang Xu, Keyan Jin, Lizheng Zu, Cheng Zhong, xiaoguang Liu, Gang Wang and Wentong Cai

Lossless compression has made significant advancements in Genomics Data (GD) storage, sharing and management. Current learning-based methods are non-evolvable with problems of low-level compression modeling, limited adaptability, and user-unfriendly interface. To this end, we propose AgentGC, the first evolutionary Agent-based GD Compressor, consisting of 3 layers with multi-agent named Leader and Worker. Specifically, the 1) User layer provides a user-friendly interface via Leader combined with LLM; 2) Cognitive layer, driven by the Leader, integrates LLM to consider joint optimization of algorithm-dataset-system, addressing the issues of low-level modeling and limited adaptability; and 3) Compression layer, headed by Worker, performs compression & decompression via a automated multi-knowledge learning-based compression framework. On top of AgentGC, we design 3 modes to support diverse scenarios: CP for compression-ratio priority, TP for throughput priority, and BM for balanced mode. Compared with 14 baselines on 9 datasets, the average compression ratios gains are 16.66%, 16.11%, and 16.33%, the throughput gains are 4.73x, 9.23x, and 9.15x, respectively.

[Long] AgentGL: Towards Agentic Graph Learning with LLMs via Reinforcement Learning ▶ *Underline Video*

Yuanfu Sun, kang Li, Dongzhe Fan, Jiajin Liu and Qiaoyu Tan

Large Language Models (LLMs) increasingly rely on agentic capabilities, iterative retrieval, tool use, and decision-making, to overcome the limits of static, parametric knowledge. Yet existing agentic frameworks treat external information as unstructured text and fail to leverage the topological dependencies inherent in real-world data. To bridge this gap, we introduce Agentic Graph Learning (AGL), a paradigm that reframes graph learning as an interleaved process of topology-aware navigation and LLM-based inference. Specifically, we propose AgentGL, the first reinforcement learning (RL)-driven framework for AGL. AgentGL equips an LLM agent with graph-native tools for multi-scale exploration, regulates tool usage via search-constrained thinking to balance accuracy and efficiency, and employs a graph-conditioned curriculum RL strategy to stabilize long-horizon policy learning without step-wise supervision. Across diverse Text-Attributed Graph (TAG) benchmarks and multiple LLM backbones, AgentGL substantially outperforms strong GraphLLMs and GraphRAG baselines, achieving absolute improvements of up to 17.5% in node classification and 28.4% in link prediction. These results demonstrate that AGL is a promising frontier for enabling LLMs to autonomously navigate and reason over complex relational environments.

The code is publicly available at https://github.com/sunyuanfu/AgentGL.

[Findings] AnalystBench: Benchmarking professional long-form report generation with web-mined multimodal tasks ► *Underline Video*

Chau Minh Pham, Zichao Wang, Puneet Mathur, Alexa Siu, Akriti Jain, Aparna Garimella, Ananya B. Sai, Nedim Lipka, Mohit Iyyer and Varun Manjunatha

Large language models are increasingly used to draft long-form multimodal documents, but their end-to-end performance on professional report generation remains systematically understudied. We introduce AnalystBench, a continually extensible benchmark of 20 real-world report generation tasks grounded in multimodal document collections, where models must process millions of input tokens to produce long-form professional reports. Using expert-validated quality checklists and groundedness evaluation, we evaluate LLMs and coding agents and find that the best model, GPT-5.1, scores highly on executive summarization tasks (exceeding 90% on quality checklists) but degrades substantially on tasks requiring long-horizon synthesis over large inputs (dropping to 25-40%). Agent-based generation substantially benefits strong closed-source models such as GPT-5.1, with checklist scores improving by 20.24 percentage points and visual coverage by 37.41 points over vanilla generation, but offers little or negative gains for open-source models like DeepSeek-R1 (-3.02 points). Expert reviewers note that while generated reports are grounded and clearly separate factual description from interpretation, they often fall short in actionability, clarity, and quantitative precision, which highlights the gap between system performance and real-world professional needs.

[Long] Anchor: Branch-Point Data Generation for GUI Agents ► *Underline Video*

Jinbiao Wei, Yilun Zhao, Kangqi Ni and Arman Cohan

End-to-end GUI agents for real desktop environments require large amounts of high-quality interaction data, yet collecting human demonstrations is expensive and existing synthetic pipelines often suffer from limited task diversity or noisy, goal-drifting trajectories. We present a trajectory expansion framework Anchor that bootstraps scalable desktop supervision from a small set of verified seed demonstrations. Starting from each seed, we identify branch points that correspond to meaningful state changes and propose new, state-grounded task variants conditioned on the current GUI context. An executing agent then follows the proposed instructions to generate new trajectories, while a verifier enforces task completion via state-aware checks and trajectory-level consistency. To improve supervision quality, we further apply task-conditioned step-level filtering to remove ungrounded actions and denoise post-branch segments to maintain coherent intent. Experiments on standard desktop benchmarks, OSWorld and WindowsAgentArena, show that models fine-tuned on our expanded corpus achieve consistent improvements over zero-shot agents and representative synthesis baselines, and generalize across applications and operating systems.

[Long] Apertus: Democratizing Open and Compliant LLMs for Global Language Environments ► *Underline Video*

Alejandro Hernández-Cano, Alexander Hägele, Allen Hao Huang, Angelika Romanou, Antoni-Joan Solergibert, Barna Pásztor, Bettina Messmer, Dhia Garbaya, Eduard Frank urech, Ido Hakimi, Juan Garcia Giraldo, Mete Ismayilzada, Negar Foroutan, Skander Moalla, Tiancheng Chen, Vinko Sabolec, Yixuan Xu, Michael Aerni, Badr AlKhamissi, Inés Altemir Marinas, Mohammad Hossein Amani, Matin Ansaripour, Ilia Badanin, Harold Benoit, Emanuela Boros, Nicholas John Browning, Fabian Bösch, Maximilian Böther, Niklas Canova, Camille Challier, Clément Charmillot, Jonathan Coles, Jan Milan Deriu, Arnout Devos, Lukas Drescher, Daniil Dzenhaliou, Maud Ehrmann, Dongyang Fan, Simin Fan, Silin Gao, Miguel Gila, María Grandury, Diba Hashemi, Alexander Miserlis Hoyle, Jiaming Jiang, Mark Klein, Andrei Kucharavy, Anastasiia Kucherenko, Frederike Lübeck, Roman Machacek, Theofilos Ioannis Manitaras, Andreas Marfurt, Kyle Matoba, Simon Matrenok, Henrique Mendonça, Fawzi Roberto Mohamed, Syrielle Montariol, Luca Mouchel, Sven Najem-Meyer, Jingwei Ni, Gennaro Oliva, Matteo Pagliardini, Elia Palme, Andrei Panferov, Léo Paoletti, Marco Passerini, Ivan Pavlov, Auguste Poiroux, Kaustubh Ponkshe, Nathan Ranchin, Javier Rando, Mathieu Sauser, Jakhongir Saydaliev, Mukhammadali Sayfiddinov, Marian Schneider, Stefano Schuppli, Marco Scialanga, Andrei Semenov, Kumar Shridhar, Raghav Singhal, Anna Sotnikova, Alexander Sternfeld, Ayush Kumar Tarun, Paul Teiletche, Jannis Vamvas, Xiaozhe Yao, Hao Zhao, Alexander Ilic, Ana Klimovic, Andreas Krause, Caglar Gulcehre, David Rosenthal, Elliott Ash, Florian Tramèr, Joost VandeVondele, Livio Veraldi, Martin Rajman, Thomas C. Schulthess, Torsten Hoefler, Antoine Bosselut, Martin Jaggi and Imanol Schlag

Open LLMs enable AI practitioners to control development costs by building on an existing foundation for downstream applications. While offering substantial promise, current models often fail to meet the needs of users needing open solutions aligned with responsible AI principles, including data compliance, transparency, and inclusivity. In this work, we present Apertus, a fully open suite of large language models (LLMs) designed to address responsibility shortcomings in todays open model ecosystem, namely data responsibility and global representation. Unlike many prior models that release weights without reproducible data pipelines or regard for content-owner rights, Apertus models are pretrained exclusively on openly available data, retroactively respecting robots.txt exclusions and filtering for non-permissive, toxic, and personally identifiable content. To mitigate risks of data memorization, we also adopt the Goldfish objective during pretraining, strongly suppressing verbatim recall of data while retaining downstream task performance. Apertus also drastically expands multilingual coverage, training on 15T tokens from over approximately 1800 languages, with about 40% of pretraining data allocated to non-English content. Released at 8B and 70B scales, Apertus approaches state-of-the-art results among fully open models on multilingual benchmarks, rivaling or surpassing open-weight counterparts.

[Findings] Beyond Dialogue Time: Temporal Semantic Memory for Personalized LLM Agents ► *Underline Video*

Miao Su, Yucan Guo, Zhongni Hou, Long Bai, Zixuan Li, Yufei zhang, Guojun Yin, Wei Lin, Xiaolong Jin, Jiafeng Guo and Xueqi Cheng

Memory enables Large Language Model (LLM) agents to perceive, store, and use information from past dialogues, which is essential for personalization. However, existing methods fail to properly model the temporal dimension of memory in two aspects: 1) Temporal inaccuracy: memories are organized by dialogue time rather than their actual occurrence time; 2) Temporal fragmentation: existing methods focus on point-wise memory, losing durative information that captures persistent states and evolving patterns. To address these limitations, we propose Temporal Semantic Memory (TSM), a memory framework that models semantic time for point-wise memory and supports the construction and utilization of durative memory. During memory construction, it first builds a semantic timeline rather than a dialogue one. Then, it consolidates temporally continuous and semantically related information into a durative memory. During memory utilization, it incorporates the querys temporal intent on the semantic timeline, enabling the retrieval of temporally appropriate durative memories and providing time-valid, duration-consistent context to support response generation. Experiments on LongMemEval and LoCoMo show that TSM consistently outperforms existing methods and achieves up to 12.2% absolute improvement in accuracy, demonstrating the effectiveness of the proposed method.

[Long] Beyond Hard Masks: Progressive Token Evolution for Diffusion Language Models ▶ *Underline Video*

Linhao Zhong, Linyu Wu, Bozhen Fang, Tianjian Feng, Chenchen Jing, Wen Wang, Jiaheng Zhang, Hao Chen and Chunhua Shen

Diffusion Language Models (DLMs) offer a promising alternative for language modeling by enabling parallel decoding through iterative refinement. However, most DLMs rely on hard binary masking and discrete token assignments, which hinder the revision of early decisions and underutilize intermediate probabilistic representations. In this paper, we propose EvoToken-DLM, a novel diffusion-based language modeling approach that replaces hard binary masks with evolving soft token distributions. EvoToken-DLM enables a progressive transition from masked states to discrete outputs, supporting revisable decoding. To effectively support this evolution, we introduce continuous trajectory supervision, which aligns training objectives with iterative probabilistic updates. Extensive experiments across multiple benchmarks show that EvoToken-DLM consistently achieves superior performance, outperforming strong diffusion-based and masked DLM baselines. Our code is available at https://github.com/aim-uofa/EvoTokenDLM.

[Long] Beyond Single Representations: Multi-Model Embedding Fusion for Stable Text Classification ▶ *Underline Video*

Jiho Gwak and Yuchul Jung

Embedding fusion has become a widely adopted technique for enhancing performance across various NLP tasks. While prior research suggests that different layers of language models encode distinct representations and that pooling strategies influence performance, there is a lack of systematic analysis regarding the empirical efficacy of these differences or the impact of combining embeddings from multiple models. This study provides a rigorous, empirical evaluation of layer-wise fusion strategies to determine their actual contribution to classification performance. Our findings reveal that the effectiveness of individual layers is more dependent on dataset characteristics than on the model architecture itself. Furthermore, we demonstrate that fusing embeddings from multiple models yields more robust and consistent representations across tasks, with the influence of any single model diminishing as the number of integrated models increases. Notably, experiments on low-resource datasets show that embedding fusion provides particularly significant gains when training data is scarce, highlighting its robustness and adaptability in data-constrained environments. We also analyze the trade-off between performance gains and computational overhead, and discuss which fusion configurations provide the best balance between stability and efficiency.

[Long] Can Continual Pretraining Bridge the Performance Gap between General-purpose and Specialized Language Models in the Medical Domain? ▶ *Underline Video*

Niclas Doll, Jasper Schulze Buschhoff, Shalaka Satheesh, Hammam Abdelwahab, Héctor Allende-Cid and Katrin Klug

This paper narrows the performance gap between small, specialized models and significantly larger general-purpose models through domain adaptation via continual pre-training and merging. We address the scarcity of specialized non-English data by constructing a high-quality German medical corpus (FineMed-de) from FineWeb2. This corpus is used to continually pre-train and merge three well-known LLMs (ranging from 7B to 24B parameters), creating the DeFineMed model family. A comprehensive evaluation confirms that specialization dramatically enhances 7B model performance on German medical benchmarks. Furthermore, the pairwise win-rate analysis of the Qwen2.5-based models demonstrates an approximately 3.5-fold increase in the win-rate against the much larger Mistral-Small-24B-Instruct through domain adaptation. This evidence positions specialized 7B models as a competitive, resource-efficient solution for complex medical instruction-following tasks. While model merging successfully restores instruction-following abilities, a subsequent failure mode analysis reveals inherent trade-offs, including the introduction of language mixing and increased verbosity, highlighting the need for more targeted fine-tuning in future work. This research provides a robust, compliant methodology for developing specialized LLMs, serving as the foundation for practical use in German-speaking healthcare contexts.

[Short] CheckMIABench: Firm Foundations For Membership Inference Attacks on Language Models ▶ *Underline Video*

Jeffrey George Wang, Jason Wang, Marvin Li and Seth Neel

Membership inference attacks (MIAs) are a canonical way to assess a machine learning model's privacy properties. Although several attempts have been made to evaluate MIAs on language models, the extant literature has suffered numerous difficulties in constructing clean evaluations to test new techniques. In particular, subtle distribution shifts between member and non-member sets can undermine the statistical validity of MIAs; recent work has underscored this by showing that "blind" methods with no access to the underlying model can perform far better than published methods on the same benchmarks. This paper constructs a benchmark for principled evaluation of MIAs against LLMs, by leveraging the insight that training data before and after a fixed point during training are drawn from the same distribution. Therefore, all

open-source models with intermediate checkpoints and public training data can be converted into MIA testbeds. We apply our framework to a half-dozen published attacks on the Pythia and OLMo family of models, from 70M to 7B parameters. To facilitate further privacy research, we open-source a modular library for designing and implementing attacks in this setting: https://github.com/safr-ai-lab/pandora_llm.

[Findings] CircuitSynth: Reliable Synthetic Data Generation ▶ *Underline Video*

Zehua Cheng, Wei Dai, Jiahao Sun and Thomas Lukasiewicz

The generation of high-fidelity synthetic data is a cornerstone of modern machine learning, yet Large Language Models (LLMs) frequently suffer from hallucinations, logical inconsistencies, and mode collapse when tasked with structured generation. Existing approaches, such s prompting or retrieval-augmented generaon, lack the mechanisms to balance linguistic expressivity with formal guarantees regarding validity and coverage. To address this, we propose CircuitSynth, a novel neuro-symbolic framework that decouples semantic reasoning from surface realization. By distilling the reasoning capabilities of a Teacher LLM into a Probabilistic Sentential Decision Diagram (PSDD), CircuitSynth creates a tractable semantic prior that structurally enforces hard logical constraints. Furthermore, we introduce a convex optimization mechanism to rigorously satisfy soft distributional goals. Empirical evaluations across diverse benchmarks demonstrate that CircuitSynth achieves 100% Schema Validity even in complex logic puzzles where unconstrained baselines fail (12.4%) while significantly outperforming state-of-the-art methods in rare-combination coverage.

[Long] CloneMem: Benchmarking Long-Term Memory for AI Clones ▶ *Underline Video*

Sen Hu, Zhiyu Zhang, YUXIANG WEI, Xueran Han, Zhenheng Tang, Ronghao Chen and Huacan Wang

AI Clones aim to simulate an individuals thoughts and behaviors to enable long-term, personalized interaction, placing stringent demands on memory systems to model experiences, emotions, and opinions over time. Existing memory benchmarks primarily rely on user-agent conversational histories, which are temporally fragmented and insufficient for capturing continuous life trajectories. We introduce CloneMem, a benchmark for evaluating long-term memory in AI Clone scenarios grounded in non-conversational digital traces, including diaries, social media posts, and emails, spanning one to three years. CloneMem adopts a top-down data construction framework to ensure longitudinal coherence and defines tasks that assess an agents ability to track evolving personal states. Experiments show that current memory mechanisms struggle in this setting, highlighting open challenges for life-grounded personalized AI. Code and dataset are available at https://github.com/AvatarMemory/CloneMemBench

[Long] COMPASS: Enhancing Agent Long-Horizon Reasoning with Evolving Context ▶ *Underline Video*

Guangya Wan, Mingyang Ling, Xiaoqi Ren, Rujun Han, Sheng Li and Zizhao Zhang

Long-horizon tasks that require sustained reasoning and multiple tool interactions remain challenging for LLM agents: small errors compound across steps, and even state-of-the-art models often hallucinate or lose coherence. We identify context management as the central bottleneck, extended histories cause agents to overlook critical evidence or become distracted by irrelevant information, thus failing to replan or reflect from previous mistakes. To address this, we propose COMPASS (Context-Organized Multi-Agent Planning and Strategy System), a lightweight hierarchical framework that separates tactical execution, strategic oversight, and context organization into three specialized components: (1) a Main Agent that performs reasoning and tool use, (2) a Meta-Thinker that monitors progress and issues strategic interventions, and (3) a Context Manager that maintains concise, relevant progress briefs for different reasoning stages. Across three challenging benchmarks, GAIA, BrowseComp, and Humanitys Last Exam, COMPASS improves accuracy by up to 20% relative to both single- and multi-agent baselines. We further introduce a test-time scaling extension that elevates performance to match established DeepResearch agents, and a post-training pipeline that delegates context management to smaller models for enhanced efficiency.

[Long] Constructing coherent spatial memory in LLM agents through graph rectification ▶ *Underline Video*

puzhen zhang, Xuyang Chen, Yu Feng, Yuhan Jiang and Liqiu Meng

Given a map description through global traversal navigation instructions, an LLM can often infer the implicit spatial layout and answer user queries by providing shortest paths. However, such context-dependent querying becomes incapable as environments grow larger, motivating the need for incremental map construction that builds a complete topological graph from stepwise observations. We propose a framework for LLM-driven construction and map repair, designed to detect, localize, and correct structural inconsistencies in incrementally constructed navigation graphs. Central to our method is the Version Control, which records the full history of graph edits and their source observations, enabling fine-grained rollback, conflict tracing, and repair evaluation. We further introduce an Edge Impact Score to prioritize minimal-cost repairs based on structural reachability, path usage, and conflict propagation. To properly evaluate our approach, we create a refined version of the MANGO benchmark dataset by systematically removing non-topological actions and inherent structural conflicts, providing a cleaner testbed for LLM-driven construction and map repair. Our approach significantly improves map correctness and robustness, especially in scenarios with entangled or chained inconsistencies. Our results highlight the importance of introspective, history-aware repair mechanisms for maintaining coherent spatial memory in LLM agents.

[Long] Data Efficient RLVR via Off-Policy Influence Guidance ▶ *Underline Video*

Erle Zhu, Dazhi Jiang, Yuan Wang, Xujun Li, Jiale Cheng, Yuxian Gu, Yilin Niu, Aohan Zeng, Jie Tang, Minlie Huang and Hongning Wang

Data selection is a critical aspect of Reinforcement Learning with Verifiable Rewards (RLVR) for enhancing the reasoning capabilities of large language models (LLMs). Current data selection methods are largely heuristic-based, lacking theoretical guarantees and generalizability. This work proposes a theoretically-grounded approach using influence functions to estimate the contribution of each data

point to the learning objective. To overcome the prohibitive computational cost of policy rollouts required for online influence estimation, we introduce an off-policy influence estimation method that efficiently approximates data influence using pre-collected offline trajectories. Furthermore, to manage the high-dimensional gradients of LLMs, we employ sparse random projection to reduce dimensionality and improve storage and computation efficiency. Leveraging these techniques, we develop Curriculum RL with Off-Policy Influence guidance (CROPI), a multi-stage RL framework that iteratively selects the most influential data for the current policy. Experiments on models up to 7B parameters demonstrate that CROPI significantly accelerates training. On a 1.5B model, it achieves a 2.66x step-level acceleration while using only 10% of the data per stage compared to full-dataset training. Our results highlight the substantial potential of influence-based data selection for efficient RLVR.

[Findings] Demystify the Role of Memory in Machine Learning Engineering Agents ▶ *Underline Video*

Xinyu Zhao, Junpeng Wang, Yuzhong Chen, Menghai Pan, Chin-Chia Michael Yeh, Jiarui Sun, Yan Zheng, Mahashweta Das and Tianlong Chen

While memory is a core component in agent systems, its behavioral impact in complex, long-horizon domains like machine learning engineering (MLE) remains poorly understood. Unlike short, reactive exchanges, MLE agents solve tasks through cycles of experimentation and improvement where past errors can inform future success. This paper presents a systematic study dissecting how memory influences agent behavior and performance across diverse MLE challenges. We first introduce a dynamic coding memory designed to capture and reuse debugging experiences, and integrate it into two representative agent paradigms: a sequential, chain-based agent that mirrors human-like iterative refinement, and a parallel, tree-based agent that performs broad, self-exploratory search in the code space. Our central finding is that the role of memory is contingent on the agent's underlying architecture. For chain-based agents, memory proves highly beneficial, enabling them to avoid recurring mistakes and engage in more coherent, iterative refinement, which significantly improves reliability and task success. In contrast, for tree-based search agents, memory introduces a critical trade-off: it enhances procedural stability at the cost of constraining search diversity, which can prematurely narrow exploration and lead to suboptimal final solutions. These findings reveal a fundamental trade-off between procedural reliability and solution innovation modulated by memory, offering insights for designing more effective and robust MLE agents.

[Long] Don't Adapt Small Language Models for Tools; Adapt Tool Schemas to the Models ▶ *Underline Video*

Jonggeun Lee, Woojung Song, Jongwook Han, Haesung Pyun and Yohan Jo

Small language models (SLMs) enable scalable tool-augmented multi-agent systems where multiple SLMs handle subtasks orchestrated by a powerful coordinator. However, they struggle with tool-use tasks, particularly in selecting appropriate tools and identifying correct parameters. A common failure mode is schema misalignment: models hallucinate plausible tool names that are absent from the provided tool schema, due to different naming conventions internalized during pretraining. Rather than training models to adapt to unfamiliar schemas, we propose adapting schemas to align with models' pretrained knowledge. We introduce PA-Tool (Pretraining-Aligned Tool Schema Generation), a training-free method that leverages peakedness, a signal used in contamination detection that indicates pretraining familiarity, to rename tool components. By generating multiple candidates and selecting the candidate with the highest peakedness, PA-Tool identifies pretraining-aligned naming patterns. Experiments on MetaTool and RoTBench show improvements of up to 17%, with schema misalignment errors reduced by 80%. PA-Tool enables small models to substantially improve tool-use accuracy without retraining, showing that schema-level interventions can unlock the tool-use potential of resource-efficient models. Our code is available at https://github.com/holi-lab/PA-Tool.

[Long] Empowering Multi-Turn Tool-Integrated Agentic Reasoning with Group Turn Policy Optimization ▶ *Underline Video*

Yifeng Ding, Hung Le, Songyang Han, Kangrui Ruan, Zhenghui Jin, Varun Kumar, Zijian Wang and Anoop Deoras

Training Large Language Models (LLMs) for multi-turn Tool-Integrated Reasoning (TIR) - where models iteratively reason, generate code, and verify through execution - remains challenging for existing reinforcement learning (RL) approaches. Current RL methods, exemplified by Group Relative Policy Optimization (GRPO), suffer from coarse-grained, trajectory-level rewards that provide insufficient learning signals for complex multi-turn interactions, leading to training stagnation. To address this issue, we propose Group Turn Policy Optimization (GTPO), a novel RL algorithm specifically designed for training LLMs on multi-turn TIR tasks. GTPO introduces three key innovations: (1) turn-level reward assignment that provides fine-grained feedback for individual turns, (2) return-based advantage estimation where normalized discounted returns are calculated as advantages, and (3) self-supervised reward shaping that exploits self-supervision signals from generated code to densify sparse binary outcome-based rewards. Our comprehensive evaluation demonstrates that GTPO outperforms GRPO by 3.0% across diverse math reasoning benchmarks, establishing its effectiveness. GTPO also improves GRPO by 3.9% on commonsense reasoning and program synthesis tasks, demonstrating its generalizability to non-math domains. Importantly, GTPO incurs negligible overhead, ensuring its practicality for real-world scenarios.

[Long] Evo-Attacker: Memory-Augmented Reinforcement Learning for Long-Horizon Tool Attacks on LLM-MAS ▶ *Underline Video*

Bingyu Yan, Xiaoming Zhang, Jinyu Hou, Chaozhuo Li, Ziyi Zhou, Yiming Hei and Litian Zhang

While Large Language Model-based Multi-Agent Systems (LLM-MAS) demonstrate remarkable capabilities in solving complex tasks by orchestrating specialized agents and external tools, the implicit trust in tool outputs creates a critical attack surface. Existing tool attacks are limited by domain specificity or fixed and static templates. To address these challenges, we propose Evo-Attacker, which formulates the tool attack as a self-evolving, memory-augmented reinforcement learning process. Evo-Attacker constructs a dynamic attack memory and

employs deliberative reasoning to retrieve adversarial patterns and strategize modifying interventions at critical moments. Furthermore, we introduce Attack-Flow GRPO to optimize intermediate reasoning steps via terminal outcomes, addressing the long-horizon credit assignment challenge. Comprehensive experiments demonstrate that Evo-Attacker consistently outperforms baselines, highlighting its generalization and evolutionary capabilities and the urgent need for defensive tool safeguards.

[Findings] EvoEdit: Evolving Null-space Alignment for Robust and Efficient Knowledge Editing ▶ *Underline Video*

Sicheng Lyu, Yu Gu, Xinyu Wang, Jerry Huang, Sitao Luan, Yufei Cui, Xiao-Wen Chang and Peng Lu

Large language models (LLMs) require continual updates to rectify outdated or erroneous knowledge. Model editing has emerged as a compelling paradigm for introducing targeted modifications without the computational burden of full retraining. Existing approaches are mainly based on a locate-then-edit framework. However, in sequential editing contexts, where multiple updates are applied over time, they exhibit significant limitations and suffer from catastrophic interference, i.e., new edits compromise previously integrated updates and degrade preserved knowledge. To address these challenges, we introduce EvoEdit, a novel editing strategy that mitigates catastrophic interference through sequential null-space alignment, enabling stable and efficient model editing. By performing sequential null-space alignment for each incoming edit, EvoEdit preserves both original and previously modified knowledge representations and maintains output invariance on preserved knowledge even across long edit sequences, effectively mitigating interference. Evaluations on real-world sequential knowledge-editing benchmarks show that EvoEdit achieves better or comparable performance than prior state-of-the-art locate-then-edit techniques, with up to 3.53x speedup. Overall, these results underscore the necessity of developing more principled approaches for designing LLMs in dynamically evolving information settings, while providing a simple yet effective solution with strong theoretical guarantees.

[Long] FoE: Forest of Errors Makes the First Solution the Best in Large Reasoning Models ▶ *Underline Video*

Kehan Jiang, Haonan Dong, Zhaolu Kang, Zhengzhou Zhu and Guojie Song

Recent Large Reasoning Models (LRMs) like DeepSeek-R1 have demonstrated remarkable success in complex reasoning tasks, exhibiting human-like patterns in exploring multiple alternative solutions. Upon closer inspection, however, we uncover a surprising phenomenon: The First is The Best, where alternative solutions are not merely suboptimal but potentially detrimental. This observation challenges widely accepted test-time scaling laws, leading us to hypothesize that errors within the reasoning path scale concurrently with test time. Through comprehensive empirical analysis, we characterize errors as a forest-structured Forest of Errors (FoE) and conclude that FoE makes the First the Best, which is underpinned by rigorous theoretical analysis. Leveraging these insights, we propose RED, a self-guided efficient reasoning framework comprising two components: I) Refining First, which suppresses FoE growth in the first solution; and II) Discarding Subs, which prunes subsequent FoE via dual-consistency. Extensive experiments across five benchmarks and six backbone models demonstrate that RED outperforms eight competitive baselines, achieving performance gains of up to 19.0% while reducing token consumption by 37.7% ~ 70.4%. Moreover, comparative experiments on FoE metrics shed light on how RED achieves effectiveness.

[Findings] Follow the Path: Reasoning over Knowledge Graph Paths to Improve Large Language Model Factuality ▶ *Underline Video*

Mike Zhang, Johannes Bjerva and Russa Biswas

We introduce fs1, a simple yet effective method that improves the factuality of reasoning traces by sourcing them from large reasoning models and grounding them by conditioning on knowledge graph (KG) paths. We fine-tune eight instruction-tuned Large Language Models (LLMs) on 3.9K factually grounded reasoning traces and rigorously evaluate them on six complex open-domain question-answering (QA) benchmarks encompassing 23.9K questions. Our results demonstrate that our fs1-tuned model consistently outperforms instruction-tuned counterparts with parallel sampling by 6-14 absolute points (pass@). Our detailed analysis shows that fs1 considerably improves model performance over more complex questions (requiring 3 or more hops on KG paths) and numerical answer types compared to the baselines. Furthermore, in single-pass inference, we notice that smaller LLMs show the most improvements. While prior works demonstrate the effectiveness of reasoning traces primarily in the STEM domains, our work shows strong evidence that anchoring reasoning to factual KG paths is a critical step in transforming LLMs for reliable knowledge-intensive tasks.

[Long] FOREVER: Forgetting Curve-Inspired Memory Replay for Language Model Continual Learning ▶ *Underline Video*

Yujie Feng, Hao Wang, Jian Li, Xu Chu, Zhaolu Kang, Yiran Liu, Yasha Wang, Philip S. Yu and Xiao-Ming Wu

Continual learning (CL) for large language models (LLMs) aims to enable sequential knowledge acquisition without catastrophic forgetting. Memory replay methods are widely used for their practicality and effectiveness, but most rely on fixed, step-based heuristics that often misalign with the models actual learning progress, since identical training steps can result in varying degrees of parameter change. Motivated by recent findings that LLM forgetting mirrors the Ebbinghaus human forgetting curve, we propose FOREVER (FORgEtting curVe-inspired mEmory Replay), a novel CL framework that aligns replay schedules with a model-centric notion of time. FOREVER defines model time using the magnitude of optimizer updates, allowing forgetting curve-inspired replay intervals to align with the models internal evolution rather than raw training steps. Building on this approach, FOREVER incorporates a forgetting curve-based replay scheduler to determine when to replay and an intensity-aware regularization mechanism to adaptively control how to replay. Extensive experiments on three CL benchmarks and models ranging from 0.6B to 13B parameters demonstrate that FOREVER consistently mitigates catastrophic forgetting.

[Long] From Data-Centric to Sample-Centric: Enhancing LLM Reasoning via Progressive Optimization ▶ *Underline Video*

Xinjie Chen, Minpeng Liao, Guoxin Chen, Chengxi Li, Biao Fu, Kai Fan and Xinggao Liu

Reinforcement learning with verifiable rewards (RLVR) has recently advanced the reasoning capabilities of large language models (LLMs).

We investigate RLVR from a sample-centric perspective and introduce LPPO (Learning-Progress and Prefix-guided Optimization), a framework of progressive optimization techniques. Our work addresses a critical question: how to best leverage a small set of trusted, high-quality demonstrations, rather than simply scaling up data volume. First, motivated by how hints aid human problem-solving, we propose prefix-guided sampling, an online data augmentation method that incorporates partial solution prefixes from expert demonstrations to guide the policy, particularly for challenging instances. Second, inspired by how humans focus on important questions aligned with their current capabilities, we introduce learning-progress weighting, a dynamic strategy that adjusts each training sample's influence based on model progression. We estimate sample-level learning progress via an exponential moving average of per-sample pass rates, promoting samples that foster learning and de-emphasizing stagnant ones. Experiments on mathematical-reasoning benchmarks demonstrate that our methods outperform strong baselines, yielding faster convergence and a higher performance ceiling, with these gains proving robust across diverse model architectures, scales, and reinforcement learning optimizers.

[Findings] Grounding Agent Memory in Contextual Intent ► *Underline Video*

Ruozhen Yang, Yucheng Jiang, Yueqi Jiang, Priyanka Kargupta, Yunyi Zhang and Jiawei Han

Deploying large language models in long-horizon, goal-oriented interactions remains challenging because similar entities and facts recur under different latent goals and con-straints, causing memory systems to retrieve context-mismatched evidence. We propose STITCH (Structured Intent Tracking in Contextual History), an agentic memory system that indexes each trajectory step with a structured retrieval cue, contextual intent, and retrieves history by matching the current step's intent. Contextual intent provides compact signals that disambiguate repeated mentions and reduce interference: (1) the current latent goal defining a thematic segment, (2) the action type, and (3) the salient entity types anchoring which attributes matter. During inference, STITCH filters and prioritizes memory snippets by intent compatibility, suppressing semantically similar but context-incompatible history. For evaluation, we introduce CAME-Bench, a benchmark for context-aware retrieval in realistic, dynamic, goal-oriented trajectories. Across CAME-Bench and LongMemEval, STITCH achieves state-of-the-art performance, outperforming the strongest baseline by 35.6%, with the largest gains as trajectory length increases. Our analysis shows that intent indexing substantially reduces retrieval noise, supporting intent-aware memory for robust long-horizon reasoning.

[Long] GTA: Generating Long-horizon Tasks for Web Agents at Scale ► *Underline Video*

Tenghao Huang, Kung-Hsiang Huang, Prafulla Kumar Choubey, Yilun Zhou, Muhao Chen, Jonathan May and Chien-Sheng Wu

Web agents, which couple language models with browsing and tool-use capabilities, show promise as open web assistants. Yet progress is increasingly limited by the lack of scalable, process-level supervision. Existing benchmarks are largely manually constructed, providing only coarse start-goal annotations without intermediate trajectories, while recent automatic generation efforts remain expensive, biased, and shallow. These limitations prevent reliable training and evaluation of agents that must generalize to realistic, multi-hop, cross-page tasks. We introduce a scalable framework GTA that integrates crawling, retrieval-based seeding, in-context generation, and automated quality control to produce realistic tasks paired with executable trajectories. This design decouples crawling from generation for greater efficiency, grounds tasks in the site graph to enforce compositionality, and ensures dense supervision through deterministic replays and systematic validation. We instantiate the pipeline on over 50 websites covering e-commerce, government, forums, and news, with multilingual and multi-hop coverage. The resulting benchmark reveals a significant human-agent performance gap and enables detailed diagnostics. Our contributions are three-fold: (i) formalizing multi-hop web-agent task generation, (ii) proposing an efficient and validated pipeline for automatic data creation, and (iii) releasing a dynamic benchmark with reproducible evaluation.

[Long] Heterogeneous Adaptive Policy Optimization: Tailoring Optimization to Every Token's Nature ► *Underline Video*

Zheng Liu, Mengjie Liu, Siwei Wen, Mengzhang Cai, Bin Cui, Conghui He and Wentao Zhang

Using entropy as a measure of heterogeneity to guide optimization has emerged as a crucial research direction in Reinforcement Learning for LLMs. However, existing methods typically treat it as a discrete filter or post-hoc regulator rather than a core optimization driver. To fully leverage the potential of entropy and achieve fine-grained regulation, we introduce Heterogeneous Adaptive Policy Optimization (HAPO), a token-aware algorithm that continuously adapts optimization dynamics based on token-level entropy throughout the entire training process. Our algorithm includes four key components: (1) Adaptive Temperature Sampling that adjusts sampling temperature in real time, promoting exploration at high-entropy tokens. (2) Token-Level Group Average Advantage Estimation that estimates advantages at token level, accounting for sequence-length effects while preserving non-biased treatment.(3) Differential Advantage Redistribution that leverages entropy and importance ratios to adjust advantages for tokens with clear signals. (4) Asymmetric Adaptive Clipping that dynamically adjusts clipping boundaries based on token-level entropy. Through systematic investigation of entropy, we embed token-level treatment into every stage. Extensive experiments on mathematical reasoning, code, and logic tasks across multiple models demonstrate HAPO's consistent superiority over DAPO.

[Findings] Hey, That's My Data! Token-Only Dataset Inference in Large Language Models ► *Underline Video*

Chen Xiong, Zihao Wang, Rui Zhu, Tsung-Yi Ho, Pin-Yu Chen, Jingwei Xiong and Haixu Tang

Large Language Models (LLMs) rely on massive training datasets, often including proprietary data, which raises concerns about unauthorized usage and copyright infringement. Existing dataset inference methods typically require access to log probabilities or other internal signals, but many modern LLMs restrict such access, motivating token-only inference approaches. We propose CatShift, a token-only dataset inference framework based on catastrophic forgetting, where models overwrite prior knowledge when trained on new data. Fine-tuning an LLM on a subset of its training data induces larger output shifts than fine-tuning on unseen data. CatShift compares these shifts against

those from a known non-member validation set to infer whether a dataset was included in training. Experiments on both open-source and API-based LLMs show that CatShift remains effective without logit access, enabling practical protection of proprietary datasets.

[Findings] How Adversarial Environments Mislead Agentic AI? ▶ *Underline Video*

Zhonghao Zhan, Huichi Zhou, Zhenhao Li, Peiyuan Jing, Krinos Li and Hamed Haddadi

Tool-integrated agents are deployed on the premise that external tools ground their outputs in reality. Yet this very reliance creates a critical attack surface. Current evaluations benchmark capability in benign settings, asking "can the agent use tools correctly" but never "what if the tools lie". We identify this Trust Gap: agents are evaluated for performance, not for skepticism. We formalize this vulnerability as Adversarial Environmental Injection (AEI), a threat model where adversaries compromise tool outputs to deceive agents. AEI constitutes environmental deception: constructing a "fake world" of poisoned search results and fabricated reference networks around unsuspecting agents. We operationalize this via Potemkin, a Model Context Protocol (MCP)-compatible harness for plug-and-play robustness testing. We identify two orthogonal attack surfaces: The Illusion (breadth attacks) poison retrieval to induce epistemic drift toward false beliefs, while The Maze (depth attacks) exploit structural traps to cause policy collapse into infinite loops. Across 11,000+ runs on five frontier agents, we find a stark robustness gap: resistance to one attack often increases vulnerability to the other, demonstrating that epistemic and navigational robustness are distinct capabilities.

[Findings] HyperEdit: Unlocking Instruction-based Text Editing in LLMs via Hypernetworks ▶ *Underline Video*

Yiming Zeng, Jinghan Cao, Zexin Li, Wanhao Yu, Zhankai Ye, Dawei Xiang, Ting Hua, Xin Liu, Shangqian Gao and Tingting Yu

Large Language Models (LLMs) have fundamentally transformed natural language processing (NLP), demonstrating remarkable capabilities across a wide spectrum of tasks. However, when applied to instruction-based text editing, LLMs continue to exhibit some limitations. Different from free-form generation, instruction-based editing requires precise, targeted modifications that respect two essential properties: faithfully implementing the specific instruction and local fidelity. Existing approaches often overlook these properties, treating editing as a generic text generation problem. As a result, they either over-edit or fail to apply modifications consistently. To address this gap, we propose HyperEdit, a framework that adaptively processes each editing request to best align with it. To achieve this, HyperEdit generates request-specific dynamic weights that guide the editing process. The computational overhead of producing these weights is minimized through a carefully designed hypernetwork. With this design, HyperEdit achieves a relatively 9% improvement over the state-of-the-art editing model.

[Long] I2E: From Image Pixels to Actionable Interactive Environments for Text-Guided Image Editing ▶ *Underline Video*

Jinghan Yu, Junhao Xiao, Chenyu Zhu, Jiaming Li, Jia Li, HanMing Deng, Xirui Wang, Guoli Jia, Jianjun Li, Xiang Bai, Bowen Zhou and Zhiyuan Ma

Existing text-guided image editing methods primarily rely on end-to-end pixel-level inpainting paradigm. Despite its success in simple scenarios, this paradigm still significantly struggles with compositional editing tasks that require precise local control and complex multi-object spatial reasoning. This paradigm is severely limited by 1) the implicit coupling of planning and execution, 2) the lack of object-level control granularity, and 3) the reliance on unstructured, pixel-centric modeling. To address these limitations, we propose I2E, a novel "Decompose-then-Action" paradigm that revisits image editing as an actionable interaction process within a structured environment. I2E utilizes a Decomposer to transform unstructured images into discrete, manipulable object layers and then introduces a physics-aware Vision-Language-Action Agent to parse complex instructions into a series of atomic actions via Chain-of-Thought reasoning. Further, we also construct I2E-Bench, a benchmark designed for multi-instance spatial reasoning and high-precision editing. Experimental results on I2E-Bench and multiple public benchmarks demonstrate that I2E significantly outperforms state-of-the-art methods in handling complex compositional instructions, maintaining physical plausibility, and ensuring multi-turn editing stability.

[Long] ImReasoner: Improving Memory-based Language Models for Reasoning-in-a-Haystack Tasks ▶ *Underline Video*

Ching-Yun Ko; Payel Das; Sihui Dai; Georgios Kollias; Subhajit Chaudhury; Aurelie C. Lozano; Pin-Yu Chen

Reasoning over long contexts remains a major challenge for language models, particularly when solving tasks that require integrating multiple facts in sequence or generalizing to new distributions. We argue that this difficulty stems from a lack of structural inductive bias. Recently, alternative frameworks have been proposed to explicitly encode contexts as ordered memory and perform iterative retrieval to construct reasoning chains. Despite the promising results shown in prior arts, they are still heavily reliant on intermediate chain supervision and fall short in showing emergent reasoning generalization in the presence of hard distractions in reasoning-in-a-haystack tasks. Furthermore, we discover that as the amount of distractions increases, traditional episodic memory reads suffer from ill-conditioning problems, which lead to inaccurate context retrievals. In this work, we formalize the motivation for necessary inductive bias in reasoning-in-a-Haystack tasks, propose inference-time memory update procedures mimicking the "identify and remove unnecessary and unrelated details" in constructively responsive reading, introduce staged training inspired by human conceptual understanding, and finally demonstrate the possibilities and limits of such framework in the weakly supervised scenario.

[Long] Language Model as Planner and Formalizer under Constraints ▶ *Underline Video*

Cassie Huang, Stuti Mohan, Ziyi Yang, Stefanie Tellex and Li Zhang

LLMs have been widely used in planning, either as planners to generate action sequences end-to-end, or as formalizers to represent the planning domain and problem in a formal language that can derive plans deterministically. However, both lines of work rely on standard

benchmarks that include only generic and simplistic environmental specifications, leading to potential overestimation of the planning ability of LLMs and safety concerns in downstream tasks. We bridge this gap by augmenting widely used planning benchmarks with manually annotated, fine-grained, and rich natural language constraints spanning four formally defined categories. Over 4 state-of-the-art reasoning LLMs, 4 formal languages, and 4 datasets, we show that the introduction of one-sentence constraints consistently halves performance, indicating current LLMs' lack of robustness and an avenue for future research.

[Long] Learning How and What to Memorize: Cognition-Inspired Two-Stage Optimization for Evolving Memory ▶ ***Underline Video***
Derong Xu, Shuochen Liu, Pengfei Luo, Pengyue Jia, Yingyi Zhang, Yi Wen, Yimin Deng, Wenlin Zhang, Enhong Chen, Xiangyu Zhao and Tong Xu
Large language model (LLM) agents require long-term user memory for consistent personalization, but limited context windows hinder tracking evolving preferences over long interactions. Existing memory systems mainly rely on static, hand-crafted update rules; although reinforcement learning (RL)-based agents learn memory updates, sparse outcome rewards provide weak supervision, resulting in unstable long-horizon optimization. Drawing on memory schema theory and the functional division between prefrontal regions and hippocampus regions, we introduce MemCoE, a cognition-inspired two-stage optimization framework that learns how memory should be organized and what information to update. In the first stage, we propose Memory Guideline Induction to optimize a global guideline via contrastive feedback interpreted as textual gradients; in the second stage, Guideline-Aligned Memory Policy Optimization uses the induced guideline to define structured process rewards and performs multi-turn RL to learn a guideline-following memory evolution policy. We evaluate on three personalization memory benchmarks, covering explicit and implicit preferences as well as different sizes and noise levels, and observe consistent improvements over strong baselines with favorable robustness, transferability, and efficiency.

[Findings] Learning Optimal Message Representations for Agentic Communication ▶ ***Underline Video***
Shashwat Gupta, Anson Bastos, Mayukh Das, Supriyo Ghosh, Nagarajan Natarajan, Chetan Bansal and Saravan Rajmohan
Large Language Models (LLMs) have demonstrated remarkable capabilities in agentic collaborative problem-solving, albeit a gap exists. Existing frameworks predominantly rely on natural language as a primary representation (format) for agentic communication. However natural language could be ambiguous and verbose. Furthermore, recent works have shown that alternative representations can enhance performance in LLMs on certain tasks. But current approaches lack the intelligence necessary to understand, learn or apply optimal communication representations adaptively. In this paper, we propose to dynamically learn the optimal message representations to enhance agentic performance. We model the optimization problem as an Expanding Markov Decision Process (EMDP) and propose our method named OPTiMACS. We evaluate our system across benchmark datasets of collaborative problem-solving. The results show significant performance improvements while maintaining efficiency. Our work bridges the gap between rigid communication protocols and open-ended natural language by providing an adaptive framework that learns task-aware structural representations.

[Findings] LEDGER: Scaling Agentic Document Editing with Dependency-aware Graph Retrieval ▶ ***Underline Video***
Hang Wang, Utkarsh Garg, Reza Davari, Huitian Jiao, Hao Cheng, Baolin Peng, Si-Qing Chen and Tao Ge
Large language models increasingly power AI agents for tasks requiring iterative refinement: document editing demands targeted revisions while preserving cross-references, code refactoring requires tracking function dependencies, and knowledge base updates cascade through related entities. Iterative editing with AI agents faces a fundamental efficiency-consistency tradeoff: maintaining consistency requires full-context awareness of dependencies, but processing entire documents for each edit incurs prohibitive token costs and latency. Isolated edits improve efficiency but risk breaking cross-references and violating semantic constraints. We introduce LEDGER (scaLing Agentic document editing with Dependency-aware Graph rEtRieval), a framework that constructs lightweight dependency graphs capturing semantic relationships and structural hierarchies across document elements. For each edit, graph traversal identifies affected elements and retrieves only necessary context. Experiments across 1,900 test cases spanning six state-of-the-art models show LEDGER achieves 76 consistency versus 56 baseline while reducing token usage by 85. Critically, LEDGER with low reasoning effort matches baseline performance at high reasoning effort using 70 fewer tokens, suggesting explicit dependency representations can substitute for expensive internal reasoning with implications for agentic systems operating on structured data.

[Findings] Less Noise, More Voice: Reinforcement Learning for Reasoning via Instruction Purification ▶ ***Underline Video***
author_yiju1 author_guo1 and author_tianyi2 author_hu2author_zexu3 author_sun3author_yankai4 author_lin4
Reinforcement Learning with Verifiable Rewards (RLVR) has advanced LLM reasoning, but remains constrained by inefficient exploration under limited rollout budgets, leading to low sampling success and unstable training in complex tasks. We find that many exploration failures arise not from problem difficulty, but from a small number of prompt tokens that introduce interference. Building on this insight, we propose the Less Noise Sampling Framework (LENS), which first purifies prompts by identifying and removing interference tokens. then transfers successful rollouts from the purified setting to supervise policy optimization on the original noisy prompts, enabling the model to learn to ignore interference in the real-world, noisy prompting settings. Experimental results show that LENS significantly outperforms GRPO, delivering higher performance and faster convergence, with a 3.88% average gain and over 1.6 x speedup. Our work highlights the critical role of pruning interference tokens in improving rollout efficiency, offering a new perspective for RLVR research.

[Long] LGSA: Label Geometry Structuring and Aligning for Hierarchical Text Classification ▶ ***Underline Video***
Shuai Zhang, Weibo Xu, Jiahao Nie and Kecheng Huang

Existing hierarchical text classification (HTC) methods typically use prompt tuning or contrastive learning to inject the label hierarchy into a model as prior knowledge to implicitly learn label embeddings for classification. However, such implicit learning fails to accurately reflect label geometry (i.e., feature spatial distribution of label embeddings), as it does not model hierarchy-aware geometric relations among labels. To address this issue, we propose a novel two-stage label geometry structuring and aligning framework, termed LGSA, which transforms the label hierarchy from an implicit prior into an explicit embedding. First, we propose a hierarchical geometric structuring (HGS) module that leverages a general orthogonal frame (GOF) to reconstruct an explicit label geometry conforming to the label hierarchy. The label geometry is then treated as a label prototype to guide model training. To facilitate the guidance, we thereby propose a hierarchical geometric aligning (HGA) module as a regularization term to align label geometry learned by the model with the explicit label prototype. Experiments on three real-world HTC datasets confirm that LGSA consistently outperforms existing state-of-the-art methods. The code and models are available at https://anonymous.4open.science/r/LGSA-1E0C.

[Short] Luring as a Proxy: Evaluating Corpus Transferability for Cybergrooming Detection ▶ *Underline Video*

Shiying Fan, Mareike Bassenge and Martin Steinebach

As the use of digital devices and social media grows among younger users, cybergrooming has emerged as a critical social concern for protecting vulnerable minors online. However, research on automated cybergrooming detection remains limited due to data scarcity. Building on previous studies that conceptualize cybergrooming as a form of luring communication, this paper investigates the potential transferability of corpora from luring or manipulative contexts for cybergrooming detection.

[Long] Mem2ActBench: A Benchmark for Evaluating Long-Term Memory Utilization in Task-Oriented Autonomous Agents ▶ *Underline Video*

Yiting Shen, Kun Li, Wei Zhou and Songlin Hu

Large Language Model (LLM)-based agents are increasingly deployed for complex, tool-based tasks where long-term memory is critical to driving actions. Existing benchmarks, however, primarily test an agent's ability to passively retrieve isolated facts in response to explicit questions. They fail to evaluate the more crucial capability of actively applying memory to execute tasks. To address this gap, we introduce Mem2ActBench, a benchmark for evaluating whether agents can proactively leverage long-term memory to execute tool-based actions by selecting appropriate tools and grounding their parameters. The benchmark simulates persistent assistant usage, where users mention the same topic across long, interrupted interactions and expect previously established preferences and task states to be implicitly applied. We build the dataset with an automated pipeline that merges heterogeneous sources (ToolACE, BFCL, OASST1), resolves conflicts via consistency modeling, and synthesizes 2,029 sessions with 12 user–assistant–tool turns on average. From these memory chains, a reverse-generation method produces 400 tool-use tasks, with human evaluation confirming 91.3% are strongly memory-dependent. Experiments on seven memory frameworks show that current systems remain inadequate at actively utilizing memory for parameter grounding, highlighting the need for more effective approaches to evaluate and improve memory application in task execution. Code and data are available at https://github.com/Cantaloupe-M/Mem2ActBench.

[Findings] Metacognitive Self-Correction for Multi-Agent System via Prototype-Guided Next-Execution Reconstruction ▶ *Underline Video*

Xu Shen, Qi Zhang, Song Wang, Zhen Tan, Xinyu Zhao, Laura Yao, Vaishnav Tadiparthi, Hossein Nourkhiz Mahjoub, Ehsan Moradi Pari, Kwonjoon Lee and Tianlong Chen

Large Language Model based multi-agent systems (MAS) excel at collaborative problem solving but remain brittle to cascading errors: a single faulty step can propagate across agents and disrupt the trajectory. In this paper, we present MASC, a metacognitive framework that endows MAS with real-time, unsupervised, step-level error detection and self-correction. MASC rethinks detection as history-conditioned anomaly scoring via two complementary designs: (1) Next-Execution Reconstruction, which predicts the embedding of the next step from the query and interaction history to capture causal consistency, and (2) Prototype-Guided Enhancement, which learns a prototype prior over normal-step embeddings and uses it to stabilize reconstruction and anomaly scoring under sparse context (e.g., early steps). When an anomaly step is flagged, MASC triggers a correction agent to revise the acting agents output before information flows downstream. On the Who&When benchmark, MASC consistently outperforms all baselines, achieving up to 7.8% AUC-ROC improvement in the challenging w/o GT setting, and further delivers consistent gains on AgentErrorBench. When plugged into diverse MAS frameworks, it delivers consistent end-to-end gains across architectures, confirming that our metacognitive monitoring and targeted correction can mitigate error propagation with minimal overhead.

[Findings] MetaScale: Test-Time Scaling with Evolving Meta-Thoughts ▶ *Underline Video*

Qin Liu, Wenxuan Zhou, Nan Xu, James Y. Huang, Fei Wang, Sheng Zhang, Hoifung Poon and Muhao Chen

One critical challenge for large language models (LLMs) in making complex reasoning is their reliance on matching reasoning patterns from training data, instead of proactively selecting the most appropriate cognitive strategy to solve a given task. Existing approaches impose fixed cognitive structures that enhance performance in specific tasks but lack adaptability across diverse scenarios. To address this limitation, we introduce MetaScale, a test-time scaling framework based on meta-thoughts, i.e., adaptive thinking strategies tailored to each task. MetaScale initializes a pool of candidate meta-thoughts, then iteratively selects and evaluates them using a multi-armed bandit algorithm with upper confidence bound selection, guided by a reward model. To further enhance adaptability, a genetic algorithm evolves high-reward meta-thoughts, refining and extending the strategy pool over time. By dynamically proposing and optimizing meta-thoughts at

inference time, MetaScale improves both accuracy and generalization across a wide range of tasks. Experimental results demonstrate that MetaScale consistently outperforms standard inference approaches, achieving an 11% performance gain in win rate on Arena-Hard with GPT-4o, improving from 82.14% to 93.14% against GPT-4. Notably, MetaScale scales more effectively with increasing sampling budgets and produces more structured, expert-level responses.

[Long] MirageBackdoor: A Stealthy Attack that Induces Think-Well-Answer-Wrong Reasoning ▶ *Underline Video*

Yizhe Zeng, Wei Zhang, Yunpeng Li, Juxin Xiao, Xiao Wang and Yuling Liu

While Chain-of-Thought (CoT) prompting has become a standard paradigm for eliciting complex reasoning capabilities in Large Language Models, it inadvertently exposes a new attack surface for backdoor attacks. Existing CoT backdoor attacks typically manipulate the intermediate reasoning steps to steer the model toward incorrect answers. However, these corrupted reasoning traces are readily detected by prevalent process-monitoring defenses. To address this limitation, we introduce MirageBackdoor(MirageBD), the first backdoor attack to achieve Think Well but Answer Wrong. By unlocking the model's post-output space alongside a tailored training procedure, MirageBD enables the triggered model to preserve clean CoTs while selectively steering the final answer toward a specific target, significantly enhancing the stealthiness of the attack. Experiments show that MirageBD generally achieves over 90% attack success rate across four datasets and five models with a poison ratio of only 5%. Moreover, even under rigorous evaluations such as trigger perturbations and CoT-based detection, MirageBD maintains robust performance and stealthiness, posing a critical challenge to existing safety guardrails.

[Long] MolMem: Memory-Augmented Agentic Reinforcement Learning for Sample-Efficient Molecular Optimization ▶ *Underline Video*

Ziqing Wang, Yibo Wen, Abhishek Pandey, Han Liu and Kaize Ding

In drug discovery, molecular optimization aims to iteratively refine a lead compound to improve molecular properties while preserving structural similarity to the original molecule. However, each oracle evaluation is expensive, making sample efficiency a key challenge for existing methods under a limited oracle budget. Trial-and-error approaches require many oracle calls, while methods that leverage external knowledge tend to reuse familiar templates and struggle on challenging objectives. A key missing piece is long-term memory that can ground decisions and provide reusable insights for future optimizations. To address this, we present MolMem (Molecular optimization with Memory), a multi-turn agentic reinforcement learning (RL) framework with a dual-memory system. Specifically, MolMem uses Static Exemplar Memory to retrieve relevant exemplars for cold-start grounding, and Evolving Skill Memory to distill successful trajectories into reusable strategies. Built on this memory-augmented formulation, we train the policy with dense step-wise rewards, turning costly rollouts into long-term knowledge that improves future optimization. Extensive experiments show that MolMem achieves 90% success on single-property tasks (1.5x over the best baseline) and 52% on multi-property tasks using only 500 oracle calls. Our code is available at https://github.com/REAL-Lab-NU/MolMem.

[Long] More Thinking, Less Talking: Internalizing Deliberative Safety into LLM Parameters ▶ *Underline Video*

Guan Wang, Xuehai Tang, Biyu Zhou, Jizhong Han and Songlin Hu

Prevailing safety alignment methods still leave Large Language Models (LLMs) vulnerable to sophisticated jailbreak attacks. To bolster defenses, explicit reasoning mechanisms like Safety-oriented Chain-of-Thought (SCoT) have emerged, significantly enhancing robustness. However, this transparency introduces a critical trade-off: the exposed reasoning process itself becomes a new attack surface, risking the leakage of harmful information and revealing the model's safety logic to adversaries. This paper directly confronts this dilemma, asking: Can we achieve the full benefits of deliberative safety without the costs of explicit reasoning generation? We propose Safety Reasoning Internalization to make the deliberative process in SCoT "available but not visible". This approach is grounded in a key theoretical insight: the corrective influence of an SCoT can be effectively approximated by a targeted, low-rank update to the model's Feed-Forward Network (FFN) layers. We operationalize this through Hierarchical Internalization of Adversarially-Guided Reasoning (HIAR), a layer-wise safety alignment framework that internalizes safety reasoning into an implicit computational pathway using Low-Rank Adaptation (LoRA). HIAR enables the model to reach a safe conclusion within a single forward pass, entirely eliminating the need to generate vulnerable SCoT text. Extensive experiments on various LLMs demonstrate that HIAR achieves a 43% lower Attack Success Rate (ASR) against distinct jailbreak attacks compared to strong baselines.

[Long] No More Stale Feedback: Co-Evolving Critics for Open-World Agent Learning ▶ *Underline Video*

Zhicong Li, Lingjie Jiang, Yulan Hu, Xingchen Zeng, Yixia Li, Xiangwen Zhang, Guanhua Chen, Zheng Pan, Xin Li and Yong Liu

Critique-guided reinforcement learning (RL) has emerged as a powerful paradigm for training LLM agents by augmenting sparse outcome rewards with natural-language feedback. However, current methods often rely on static or offline critic models, which fail to adapt as the policy evolves. In on-policy RL, the agent's trajectory distribution and error patterns shift over time, causing stationary critics to become stale and providing feedback of diminishing utility. To address this, we introduce ECHO (Evolving Critic for Hindsight-Guided Optimization), a framework that jointly optimizes the policy and critic through a synchronized co-evolutionary loop. ECHO utilizes a cascaded rollout mechanism where the critic generates multiple diagnoses for an initial trajectory, followed by policy refinement to enable group-structured advantage estimation. We address the challenge of learning plateaus via a saturation-aware gain shaping objective, which rewards the critic for inducing incremental improvements in high-performing trajectories. By employing synchronized dual-track GRPO updates, ECHO ensures the critic's feedback stays synchronized with the evolving policy. Experimental results show that ECHO yields more stable training and higher long-horizon task success across open-world environments.

[Long] OpenRubrics: Towards Scalable Synthetic Rubric Generation for Reward Modeling and LLM Alignment ▶ *Underline Video*

Tianci Liu, Ran Xu, Tony Yu, Ilgee Hong, Carl Yang, Tuo Zhao and Haoyu Wang

Reward modeling lies at the core of reinforcement learning from human feedback (RLHF), yet most existing reward models rely on scalar or pairwise judgments that fail to capture the multifaceted nature of human preferences. Recent studies have explored rubrics-as-rewards (RaR) that uses structured criteria to capture multiple dimensions of response quality. However, producing rubrics that are both reliable and scalable remains a key challenge. In this work, we introduce OpenRubrics, a diverse, large-scale collection of (prompt, rubric) pairs for training rubric-generation and rubric-based reward models. To elicit discriminative and comprehensive evaluation signals, we introduce Contrastive Rubric Generation (CRG), which derives both hard rules (explicit constraints) and principles (implicit qualities) by contrasting preferred and rejected responses. We further remove noisy rubrics via preserving preference-label consistency. Across multiple reward-modeling benchmarks, our rubric-based reward model, Rubric-RM, surpasses strong size-matched baselines by 8.4%. These gains transfer to policy models on instruction-following and biomedical benchmarks. The model weights and datasets are publicly available at https://huggingface.co/OpenRubrics.

[Findings] OPT-BENCH: Evaluating the Iterative Self-Optimization of LLM Agents in Large-Scale Search Spaces ▶ *Underline Video*

Xiaozhe Li, Jixuan Chen, Xinyu Fang, Shengyuan Ding, Haodong Duan, Qingwen Liu and Kai Chen

Large Language Models (LLMs) have demonstrated remarkable capabilities in reasoning and tool use. However, the fundamental cognitive faculties essential for problem-solving, perception, reasoning, and memory, remain the stable core of intelligence. Unlike memorizing specific patterns, humans succeed in novel environments by applying these intrinsic faculties to adapt and optimize. Yet, whether LLMs possess this essential capacity, namely, the ability to continuously refine solutions in response to dynamic environmental feedback, remains underexplored. To address this challenge, we introduce OPT-BENCH, a benchmark for evaluating self-improvement capabilities in large-scale search spaces. By combining 20 machine learning tasks with 10 classic NP-hard problems, OPT-BENCH provides a rigorous setting to assess whether agents can adapt through intrinsic self-reflection rather than rote tool application. We further propose OPT-Agent, a framework that emulates human-like cognitive adaptation. It operates via a general perception–memory–reasoning loop, iteratively refining solutions based on environmental feedback. Through extensive experiments on 19 LLMs from 7 model families, including reasoning models, general models, and open-source models ranging from 3B to 235B parameters, we demonstrate stronger models are more effective at leveraging feedback signals for self-improvement. However, this upper-bound adaptability remains fundamentally constrained by the models' base capacity, and even the most advanced LLMs still fall short of human expert performance.

[Long] ParaCodex: A Profiling-Guided Autonomous Coding Agent for Reliable Parallel Code Generation and Translation ▶ *Underline Video*

Erel Kaplan, Tomer Bitan, Lian Ghrayeb, Le Chen, Tom Yotam, Niranjan Hasabnis and Gal Oren

Parallel programming is central to HPC and AI, but producing code that is correct and fast remains challenging, especially for OpenMP GPU offload, where data movement and tuning dominate. Autonomous coding agents can compile, test, and profile on target hardware, but outputs are brittle without domain scaffolding. We present ParaCodex, an HPC-engineer workflow that turns a Codex-based agent into an autonomous OpenMP GPU offload system using staged hotspot analysis, explicit data planning, correctness gating, and profiling-guided refinement. We evaluate translation from serial CPU kernels to OpenMP GPU offload kernels on HeCBench, Rodinia, and NAS. After excluding five kernels, ParaCodex succeeded on all 31 valid kernels. In 27/31 (87%) of these valid cases, the generated kernels improved GPU time over reference implementations, a result that holds independently on both the A100 and RTX 4060. The resulting OpenMP kernels achieve geometric-mean speedups of 3.1 (A100) and 3.6 (RTX 4060) on HeCBench and 1.5 and 1.1 on Rodinia, and outperform a zero-shot Codex baseline on all suites. We also evaluate CUDA -> OpenMP offload translation on ParEval, where ParaCodex maintains high compilation and validation rates in code-only and end-to-end settings. ParaCodex is available at https://github.com/Scientific-Computing-Lab/ParaCodex

[Findings] Progra: Progress-Aware Reinforcement Learning for Multi-Turn Function Calling ▶ *Underline Video*

Huacan Chai, Zijie Cao, Maolin Ran, Yingxuan Yang, Jianghao Lin, Xin Peng, Hairui Wang, Renjie Ding, Ziyu Wan, Muning Wen, Weiwen Liu, Weinan Zhang, Fei Huang and Ying Wen

Large language models (LLMs) have achieved impressive success in single-turn function calling, yet real-world applications such as travel planning or multi-stage data analysis typically unfold across multi-turn conversations. In these settings, LLMs must not only issue accurate function calls at each step but also maintain progress awareness, the ability to summarize past interactions and plan future actions to ensure coherent, long-horizon task execution. Existing approaches, however, either reduce multi-turn training to isolated single-turn samples, which neglects task-level planning, or employ end-to-end reinforcement learning (RL) that struggles with redundancy and lacks explicit integration of progress awareness. To overcome these limitations, we introduce Progra, a framework that explicitly incorporates progress awareness into LLM training for multi-turn function calling. Progra combines (i) a Progress Awareness Generation (PAG) pipeline, which automatically constructs datasets coupling conversation summaries with future task planning, and (ii) a Progress Awareness-Guided Reinforcement Learning (PAG-RL) algorithm, which integrates progress awareness into RL training to reduce contextual redundancy and improve alignment between local actions and global task completion. Empirical results on two public benchmarks demonstrate that Progra significantly outperforms existing methods, highlighting the effectiveness of progress awareness in enabling robust and efficient multi-turn function calling. Our code is available at https://github.com/FatCatCHC/Progra.

[Long] REAL: REtrieval-reAsoning and Logic-constructed Attention Behaviors for Long-Context KV Cache Compression ► *Underline Video*

Mengjie Li, Yuan Feng, Xike Xie and William J. Song

The growing sequence length of large language models poses significant challenges for key-value (KV) caches. Existing state-of-the-art cache eviction methods primarily analyze the inference behavior of attention heads in successful retrieval-reasoning cases, often overlooking diverse behaviors in failure cases, such as bias and distraction. This oversight limits the potential to leverage heterogeneous head behaviors for improved eviction performance. Inspired by the confusion matrix, we introduce an Attention Behavior Matrix to comprehensively analyze attention head behaviors in both success and failure scenarios. By maximizing the signal-to-noise ratio , strengthening valid reasoning pathways in success cases while inhibiting noise from bias and distraction in failure cases , we propose REtrieval-reAsoning and Logic-constructed (REAL) KV cache eviction, the first method to leverage multi-behavior analysis. Comprehensive evaluations show that REAL achieves remarkable performance across various models and benchmarks; notably, on LongBench v2, it achieves comparable accuracy to the strongest baseline, HeadKV-R2, while requiring 32x less space. By offering a novel perspective on behavior analysis, we pave the way for a shift from success-only to comprehensive, failure-aware methods in long-context modeling. Our code is available at https://github.com/yonseicasl/REAL.

[Findings] ReasonRec: A Reasoning-Augmented Multimodal Agent for Unified Recommendation ► *Underline Video*

Yihua Zhang, Mingfu Liang, Jiyan Yang, Rong Jin, Wen-Yen Chen, Yiping Han, Huayu Li, Buyun Zhang, Liang Luo, Luke Simon, Sijia Liu, Tianlong Chen and Xi Liu

Recent advances in multimodal recommenders excel at feature fusion but remain opaque and inefficient decision-makers, lacking explicit reasoning and self-awareness of uncertainty. To address this, we introduce ReasonRec, a reasoning-augmented multimodal agent structured around a three-stage explicit reasoning pipeline: Observe, via a pretrained Vision-Language Model (VLM) encoder; Deliberate, by formulating recommendation as chain-of-thought (CoT) reasoning tasks and explicitly quantifying prediction uncertainty through an evidence-horizon-aware curriculum; and Act, through dynamic delegation of uncertain or challenging queries to lightweight classical recommendation models. Specifically, we propose a reasoning-aware visual instruction tuning strategy that systematically transforms diverse recommendation tasks into unified CoT prompts, enabling the VLM to explicitly articulate intermediate decision steps. Additionally, our evidence-horizon curriculum progressively enhances the reasoning complexity to better handle cold-start and long-tail user scenarios, significantly boosting model generalization. Furthermore, the uncertainty-guided delegation mechanism empowers the agent to assess its own confidence, strategically allocating computational resources to optimize both recommendation accuracy and inference efficiency. Comprehensive experiments on four standard recommendation tasks (sequential recommendation, direct recommendation, CTR prediction, and explanation generation) across five real-world datasets demonstrate that ReasonRec achieves over 30% relative improvement in key ranking metrics (e.g., HR@5, NDCG@5) compared to state-of-the-art multimodal recommenders. Crucially, ReasonRec substantially reduces inference latency by dynamically delegating up to 35% of queries to efficient sub-models without compromising accuracy. Extensive ablation studies further confirm that each proposed reasoning and planning mechanism individually contributes substantially to ReasonRec's overall effectiveness. Collectively, our results illustrate a clear pathway towards interpretable, adaptive, and efficient multimodal recommendation through explicit reasoning and agentic design.

[Long] Reinforced Efficient Reasoning via Semantically Diverse Exploration ► *Underline Video*

Ziqi Zhao, Zhaochun Ren, Jiahong Zou, Liu Yang, Zhiwei Xu, Xuri Ge, Zhumin Chen, Xinyu Ma, Daiting Shi, Shuaiqiang Wang, Dawei Yin and Xin Xin

Reinforcement learning with verifiable rewards (RLVR) has proven effective in enhancing the reasoning of large language models (LLMs). Monte Carlo Tree Search (MCTS)-based extensions improve upon vanilla RLVR (e.g., GRPO) by providing tree-based reasoning rollouts that enable fine-grained and segment-level credit assignment. However, existing methods still suffer from limited exploration diversity and inefficient reasoning. To address the above challenges, we propose reinforced efficient reasoning via semantically diverse explorations, i.e., ROSE, for LLMs. To encourage more diverse reasoning exploration, our method incorporates a semantic-entropy-based branching strategy and an -exploration mechanism. The former operates on already sampled reasoning rollouts to capture semantic uncertainty and select branching points with high semantic divergence to generate new successive reasoning paths, whereas the latter stochastically initiates reasoning rollouts from the root, preventing the search process from becoming overly local. To improve efficiency, we design a length-aware segment-level advantage estimator that rewards concise and correct reasoning while penalizing unnecessarily long reasoning chains. Extensive experiments on various mathematical reasoning benchmarks with Qwen and Llama models validate the effectiveness and efficiency of ROSE. Codes are available at https://github.com/ZiqiZhao1/ROSE-rl.

[Long] Repeated Sequences Reveal Gaps between Large Language Models and Natural Language ► *Underline Video*

Kumiko Tanaka-Ishii

Evaluating whether large language models (LLMs) capture the structure of natural language beyond local fluency remains an open challenge. Existing evaluation methods, largely based on task performance or short-context behavior, provide limited insight into the long-range statistical organization of generated text. We propose a complementary evaluation framework based on repeated subsequences. By analyzing their distribution across scales and relating it to higher-order R\{}'enyi entropies, we probe how texts reuse previously established structure under finite-length conditions. Experiments on human-written texts and length-matched GPT-generated texts show that, while power-law models can describe restricted ranges of block length, the observed entropy growth is often equally or better characterized by

logarithmic–power forms. Across datasets, natural language exhibits stable entropy-growth patterns over accessible ranges, with consistent average behavior despite variability across individual texts. In contrast, GPT-generated texts show systematic and statistically significant shifts in estimated exponents with model size. These results demonstrate that repeated-subsequence entropy provides a quantitative structural diagnostic that reveals systematic differences in long-range organization, distinguishing natural language from state-of-the-art LLM outputs beyond surface-level fluency.

[Long] Reward Alignment Optimization: A Direct Point-wise Alignment Approach ▶ *Underline Video*

Zelin Li, Jia Leng, Dawei Song and Yangen Hu

Direct Alignment Algorithms (DAAs) such as DPO simplify RLHF by optimizing policies directly from preference pairs. However, the Bradley-Terry probability-gap objective can induce likelihood displacement and, under weak KL constraints, may even reduce the probability of preferred responses, while implicit rewards can be limited in generalizaiton. We propose Reward Alignment Optimization (RAO), a point-wise direct alignment method that uses an explicit reward model to specify exact target generation probabilities and align the policy offline towards them. Our key insight is a theoretical principle we call "prefix consistency", which links the normalization terms of prompts that share a prefix. Leveraging this property, RAO decouples target reward differentials from bias terms, prevents decreasing preferred-response probabilities, and better exploits reward information both within and across prompts. Extensive experiments on multiple base LLMs show that RAO consistently outperforms existing DAAs while enabling controllable target probability distributions.

[Long] RExBench: Can coding agents autonomously implement AI research extensions? ▶ *Underline Video*

Nicholas Edwards, Yukyung Lee, Yujun (Audrey) Mao, Yulu Qin, Sebastian Schuster and Najoung Kim

Agents based on Large Language Models (LLMs) have shown promise for performing sophisticated software engineering tasks autonomously. In addition, there has been progress towards developing agents that can perform parts of the research pipeline in machine learning and the natural sciences. We argue that research extension and its implementation is a critical capability for such systems, and introduce RExBench to support the evaluation of this capability. RExBench is a benchmark consisting of realistic extensions of 12 research papers that aim to investigate novel research hypotheses. Each task is set up as an extension to an existing research paper and codebase, accompanied by domain expert-written instructions. RExBench is robust to data contamination, and supports an automatic evaluation infrastructure that executes agent outputs to determine whether the success criteria are met. We use this benchmark to evaluate 12 LLM agents implemented using two different frameworks: aider and OpenHands. We find that all agents fail to autonomously implement the majority of the extensions, with the best agent at around 33% success rate. Although the success rate improves with additional human-written hints, the best performance under this setting remains below 44%. This indicates that current agents are still short of being able to handle realistic research extension tasks without substantial human guidance.

[Long] RL-PLUS: Countering Capability Boundary Collapse of LLMs in Reinforcement Learning with Hybrid-policy Optimization ▶ *Underline Video*

Yihong Dong, Xue Jiang, Yongding Tao, Huanyu Liu, Kechi Zhang, Lili Mou, Rongyu Cao, Yingwei Ma, Jue Chen, Binhua Li, Zhi Jin, Fei Huang, Yongbin Li and Ge Li

Reinforcement Learning with Verifiable Reward (RLVR) has significantly advanced the complex reasoning abilities of Large Language Models (LLMs). However, it struggles to break through the inherent capability boundaries of the base LLM, due to its essentially on-policy strategy coupled with LLM's immense action space and sparse reward. Critically, RLVR can lead to the capability boundary collapse, narrowing the LLM's problem-solving scope. To address this problem, we propose R-PLUS, a novel hybrid-policy optimization approach for LLMs that synergizes internal exploitation with external data to achieve stronger reasoning capabilities and surpass the boundaries of base models. R-PLUS integrates two core components, i.e., Multiple Importance Sampling to address distributional mismatch from external data, and Exploration-Based Advantage Function to guide the model towards high-value, unexplored reasoning paths. We provide both theoretical analysis and extensive experiments to demonstrate the superiority and generalizability of our approach. Compared with existing RLVR methods, R-PLUS achieves 1) state-of-the-art performance on six math reasoning benchmarks; 2) superior performance on six out-of-distribution reasoning tasks; 3) consistent and significant gains across diverse model families, with average relative improvements up to 69.2%. Moreover, the analysis of Pass@k curves indicates that R-PLUS effectively resolves the capability boundary collapse problem.

[Findings] RoadMapper: A Multi-Agent System for Roadmap Generation of Solving Complex Research Problems ▶ *Underline Video*

Jiacheng Liu, Zichen Tang, Zhongjun Yang, Xinyi Hu, Xueyuan Lin, Linwei Jia, Ruofei Bai, Rongjin Li, Shiyao Peng, Haocheng Gao and Haihong E

People commonly leverage structured content to accelerate knowledge acquisition and research problem solving. Among these, roadmaps guide researchers through hierarchical subtasks to solve complex research problems step by step. Despite progress in structured content generation, the roadmap generation task has remained unexplored. To bridge this gap, we introduce RoadMap, a novel benchmark designed to evaluate the ability of large language models (LLMs) to construct high-quality roadmaps for solving complex research problems. Based on this, we identify three limitations of LLMs: (1) lack of professional knowledge, (2) unreasonable task decomposition, and (3) disordered logical relationships. To address these challenges, we propose RoadMapper, an LLM-based multi-agent system that decomposes the research roadmap generation task into three key stages (i.e., initial generation, knowledge augmentation, and iterative "critique-revise-evaluate"). Extensive experiments demonstrate that RoadMapper can improve LLMs' ability for roadmap generation, while enhancing average performance by more than 8% and saving 84% of the time required by human experts, highlighting its effectiveness and application potential.

[Long] Scaling Behaviors of LLM Reinforcement Learning Post-Training: An Empirical Study in Mathematical Reasoning ▶ *Underline Video*

Zelin Tan, Hejia Geng, Xiaohang Yu, Mulei Zhang, Guancheng Wan, Yifan Zhou, Qiang He, Xiangyuan Xue, Heng Zhou, Yutao Fan, Zhongzhi Li, Zaibin Zhang, Guibin Zhang, Chen Zhang, Zhenfei Yin, Philip Torr and Lei Bai

While scaling laws for large language models (LLMs) during pre-training have been extensively studied, their behavior under reinforcement learning (RL) post-training remains largely unexplored. This paper investigates the scaling behavior of Large Language Model (LLM) reinforcement learning post-training, focusing on mathematical reasoning. Through experiments across the Qwen2.5 series (0.5B to 72B), we characterize how model scale, data, and compute interact. Our analysis yields four key findings: 1. Larger models consistently demonstrate superior compute and data efficiency. 2. The relationship between model performance and training resources follows a predictive power-law across both base and instruction-tuned models. 3. RL learning efficiency exhibits a latent saturation trend with increasing model scale. 4. In data-constrained regimes, performance is primarily driven by the total volume of training data rather than sample uniqueness. These results offer practical guidelines for scaling reasoning capabilities through reinforcement learning post-training.

[Long] Scaling External Knowledge Input Beyond Context Windows of LLMs via Multi-Agent Collaboration ▶ *Underline Video*

Zijun Liu, Zhennan Wan, Peng Li, Ming Yan, Fei Huang and Yang Liu

With the rapid advancement of post-training techniques for reasoning and information seeking, large language models (LLMs) can incorporate a large quantity of retrieved knowledge to solve complex tasks. However, the limited context window of LLMs obstructs scaling the amount of external knowledge input, prohibiting further improvement. Existing context window extension methods inevitably cause information loss. LLM-based multi-agent methods emerge as a new paradigm to handle massive input in a distributional manner, where we identify two core bottlenecks in existing agent orchestration designs. In this work, we develop a multi-agent framework, ExtAgents, to overcome the bottlenecks and enable better scalability in inference-time knowledge integration without longer-context training. Benchmarked with our enhanced multi-hop question answering test, +, and other public test sets including long survey generation, ExtAgents significantly enhances the performance over existing non-training methods with the same amount of external knowledge input, regardless of whether it falls within or exceeds the context window. Moreover, the method maintains efficiency due to high parallelism. We believe further study in the coordination of LLM agents on increasing external knowledge input could benefit real-world applications.

[Long] Select Before Use: On the Importance of Reference Model Selection in Preference Alignment ▶ *Underline Video*

Muyang Li, Runze Wu, Xiangyu Zhao, Bo Han, Daoyi Dong and Tongliang Liu

The post-training stage of Large Language Models (LLMs) typically involves Supervised Fine-Tuning (SFT) followed by preference alignment to ensure LLM to generate safe, helpful, and instruction-aligned content. The SFT model critically serves as both the initialization and reference model for subsequent preference alignment. However, an essential yet often neglected question is the optimal selection of the SFT checkpoint for this role. We show that checkpoint selection substantially affects final performance, and that the common practice of choosing the minimum validation-loss checkpoint often fails, due to a fundamental conflict between SFTs focus on imitation and alignments goal of response discriminability. To this end, we propose RewardRank, a simple, effective, training-free metrics for estimating initial implicit alignment between reference model and preference objective. Empirical evidence suggests that, using our selected model as reference can gain up to 67.6% relative increase on length-controlled win rate on the popular Zephyr recipe comparing to baselines.

[Long] SPIO: Ensemble and Selective Strategies via LLM-Based Multi-Agent Planning in Automated Data Science ▶ *Underline Video*

Wonduk Seo, Juhyeon Lee, Yanjun Shao, Qingshan Zhou, Seunghyun lee and Yi Bu

Large Language Models (LLMs) have enabled dynamic reasoning in automated data analytics, yet recent multi-agent systems remain limited by rigid, single-path workflows that restrict strategic exploration and often lead to suboptimal outcomes. To overcome these limitations, we propose SPIO (Sequential Plan Integration and Optimization), a framework that replaces rigid workflows with adaptive, multi-path planning across four core modules: data preprocessing, feature engineering, model selection, and hyperparameter tuning. In each module, specialized agents generate diverse candidate strategies, which are cascaded and refined by an optimization agent. SPIO offers two operating modes: SPIO-S for selecting a single optimal pipeline, and SPIO-E for ensembling top-k pipelines to maximize robustness. Extensive evaluations on Kaggle and OpenML benchmarks show that SPIO consistently outperforms state-of-the-art baselines, achieving an average performance gain of 5.6%. By explicitly exploring and integrating multiple solution paths, SPIO delivers a more flexible, accurate, and reliable foundation for automated data science.

[Long] SQL-Trail: Multi-Turn Reinforcement Learning with Interleaved Feedback for Text-to-SQL ▶ *Underline Video*

Harper Hua, Zhen Han, Zhengyuan Shen, Meng-Chieh Lee, Patrick Guan, Qi Zhu, Sullam Jeoung, Yueyan Chen, Yunfei Bai, Shuai Wang, Vassilis Ioannidis and Huzefa Rangwala

While large language models (LLMs) have substantially improved Text-to-SQL generation, a pronounced gap remains between AI systems and human experts on challenging benchmarks such as BIRD-SQL. We argue this gap stems largely from the prevailing single-pass paradigm, which lacks the iterative reasoning, schema exploration, and error-correction behaviors that humans naturally employ. To address this limitation, we introduce SQL-Trail, a multi-turn reinforcement learning (RL) agentic framework for Text-to-SQL. Rather than producing a query in one shot, SQL-Trail interacts with the database environment and uses execution feedback to iteratively refine its predictions. Our approach centers on two key ideas: (i) an adaptive turn-budget allocation mechanism that scales the agents interaction depth to match

question difficulty, and (ii) a composite reward panel that jointly incentivizes SQL correctness and efficient exploration. Across benchmarks, SQL-Trail sets a new state of the art and delivers strong data efficiency, up to 18x higher than prior single-pass RL state-of-the-art methods. Notably, our 7B and 14B models outperform substantially larger proprietary systems by 5% on average, underscoring the effectiveness of interactive, agentic workflows for robust Text-to-SQL generation.

[Long] Structured Episodic Event Memory ▶ *Underline Video*

Zhengxuan Lu, Dongfang Li, Yukun Shi, Beilun Wang, Longyue Wang and Baotian Hu

Current approaches to memory in Large Language Models (LLMs) predominantly rely on static Retrieval-Augmented Generation (RAG), which often results in scattered retrieval and fails to capture the structural dependencies required for complex reasoning. For autonomous agents, these passive and flat architectures lack the cognitive organization necessary to model the dynamic and associative nature of long-term interaction. To address this, we propose Structured Episodic Event Memory (SEEM), a hierarchical framework that synergizes a graph memory layer for relational facts with a dynamic episodic memory layer for narrative progression. Grounded in cognitive frame theory, SEEM transforms interaction streams into structured Episodic Event Frames (EEFs) anchored by precise provenance pointers. Furthermore, we introduce an agentic associative fusion and Reverse Provenance Expansion (RPE) mechanism to reconstruct coherent narrative contexts from fragmented evidence. Experimental results on the LoCoMo and LongMemEval benchmarks demonstrate that SEEM significantly outperforms baselines, enabling agents to maintain superior narrative coherence and logical consistency.

[Findings] Table-as-Search: Agentic Information Seeking is Table Completion ▶ *Underline Video*

Tian Lan, Felix Henry, Bin Zhu, Qianghuai Jia, Junyang Ren, Qihang PU, Haijun Li, Longyue Wang, Zhao Xu and Weihua Luo

Current Information Seeking (InfoSeeking) agents struggle to maintain focus and coherence during long-horizon exploration, as tracking search states, including planning procedure and massive search results, within one plain-text context is inherently fragile. To address this, we introduce Table-as-Search (TaS), a structured planning framework that reformulates the InfoSeeking task as a Table Completion task. TaS maps each query into a structured table schema maintained in an external database, where rows represent search candidates and columns denote constraints or required information. This table precisely manages the search states: filled cells strictly record the history and search results, while empty cells serve as an explicit search plan. Crucially, TaS unifies three distinct InfoSeeking tasks: Deep Search, Wide Search, and the challenging DeepWide Search. Extensive experiments demonstrate that TaS significantly outperforms numerous state-of-the-art baselines across three kinds of benchmarks, including multi-agent framework and commercial systems. Furthermore, our analysis validates the TaS's superior robustness in long-horizon InfoSeeking, alongside its efficiency, scalability and flexibility. Code and datasets are publicly released at https://github.com/AIDC-AI/Marco-Search-Agent.

[Long] ThinkPersona: Thinking with Persona Graphs for Faithful Individualized Role-Playing ▶ *Underline Video*

Yichen Cai, Pei Chen, Jiayang Li, Jingya Guo, Zejian Li, Changyuan Yang and Lingyun Sun

Large Language Models are increasingly utilized as Role-Playing Agents (RPAs) to simulate personas in interactive settings. However, current RPAs often produce flattened and stereotypical personas with limited depth and fidelity. This limitation arises from two core challenges: insufficient modeling of complex personal histories and internal logic, and ungrounded reasoning that fails to preserve persona coherence as dialogue context evolves. To address these challenges, we propose ThinkPersona, a role-playing agent trained to explicitly ground responses in individual identity. We introduce Persona Graphs as structured representations that encode life trajectories, values, relationships, and events as interconnected knowledge. We construct 1,201 Persona Graphs from real-world interviews and derive a Question-Reasoning-Answer (QRA) dataset of 23,401 samples that supervises reasoning over persona evidence. Fine-tuning on QRA enables ThinkPersona to internalize persona logic and generate persona-consistent responses in long-context dialogues. Experiments on three benchmarks show that ThinkPersona improves role-playing fidelity, behavioral consistency, and grounded reasoning over existing methods, while preserving general instruction-following capabilities. Our code and dataset are available at https://github.com/Hualeez/ThinkPersona.

[Long] Union-of-Experts: Neurons in Mixture-of-Experts are Secretly Routers ▶ *Underline Video*

Songhao Wu, Ang Lv, Ruobing Xie, Xingwu Sun, Di Wang, Rui Yan and Yankai Lin

Mixture-of-Experts (MoE) models rely on an external router to assign tokens to experts. This design inherently separates the routing decision from each expert's internal capabilities, leading to suboptimal performance. In this work, we address this limitation with Union-of-Experts (UoE), an MoE variant that performs "expert-autonomous routing". The core mechanism of UoE is to pre-designate a minute fraction of neurons within each expert as "routing neurons". Experts autonomously select relevant tokens by comparing the activation intensity of these neurons, aligning routing decisions with each experts functional profile. To prevent the waste of activations from unselected experts' routing neurons, we aggregate all routing neuron outputs and sum them into the final layer output. This aggregation acts as a novel virtual shared expert whose parameters are distributed across the individual experts, and improves overall parameter efficiency. We pre-train UoE models with up to 3B parameters, demonstrating that they outperform traditional MoEs with matched efficiency. Furthermore, our analysis of the routing neurons provides valuable insights into expert-autonomous selection and the broader routing mechanisms of MoE models.

[Findings] Weight Tying Biases Token Embeddings Towards the Output Space ▶ *Underline Video*

Antonio Lopardo, Avyukth Harish, Catherine Arnett and Akshat Gupta

Weight tying, i.e. sharing parameters between input and output embedding matrices, is common practice in language model design, yet its impact on the learned embedding space remains poorly understood. In this paper, we show that tied embedding matrices align more closely

with output (unembedding) matrices than with input embeddings of comparable untied models, indicating that the shared matrix is shaped primarily for output prediction rather than input representation. This unembedding bias arises because output gradients dominate early in training. Using tuned lens analysis, we show this negatively affects early-layer computations, which contribute less effectively to the residual stream. Scaling input gradients during training reduces this bias, providing causal evidence for the role of gradient imbalance. This is mechanistic evidence that weight tying optimizes the embedding matrix for output prediction, compromising its role in input representation. These results help explain why weight tying can harm performance at scale and have implications for training smaller LLMs, where the embedding matrix contributes substantially to total parameter count.

[Findings] Why LoRA Fails to Forget: Regularized Low-Rank Adaptation Against Backdoors in Language Models ▶ *Underline Video*
Hoang-Chau Luong and Lingwei Chen
Low-Rank Adaptation (LoRA) is widely used for parameter-efficient fine-tuning of large language models, but it is notably ineffective at removing backdoor behaviors from poisoned pretrained models when fine-tuning on clean dataset. Contrary to the common belief that this weakness is caused primarily by low rank, we show that LoRA's vulnerability is fundamentally spectral. Our analysis identifies two key factors: LoRA updates (i) possess insufficient spectral strength, with singular values far below those of pretrained weights, and (ii) exhibit unfavorable spectral alignment, weakly matching clean-task directions while retaining overlap with trigger-sensitive subspaces. We further establish a critical scaling threshold beyond which LoRA can theoretically suppress trigger-induced activations, and we show empirically that standard LoRA rarely reaches this regime. We introduce Regularized Low-Rank Adaptation (RoRA), which improves forgetting by increasing spectral strength and correcting alignment through clean-strengthened regularization, trigger-insensitive constraints, and post-training spectral rescaling. Experiments across multiple NLP benchmarks and attack settings show that RoRA substantially reduces attack success rates while maintaining clean accuracy.

[Findings] Worldwide LiveVQA: Real-Time Visual Knowledge Seeking and Updating Across Languages ▶ *Underline Video*
Xuanao Huang, Xingjia Liu, Zetong Zhou, Yuyang Peng, Yao Wan and Dongping Chen
Knowledge about the visual world is not only constantly evolving but also inherently happening all over the world: breaking news in Tokyo, political events in São Paulo, and cultural phenomena in Cairo are first reported in Japanese, Portuguese, and Arabic, carrying regional context that English-centric resources cannot fully capture. Yet existing resources for visual knowledge remain confined to English, creating a "Worldwide Knowledge Gap" that hinders developing truly global assistants. To quantify this gap, we introduce LiveVQA-W(orldwide), the first dynamic-updating dataset for real-time, multilingual visual knowledge seeking and updating across ten major languages. Drawing from worldwide news outlets, YouTube videos, and academic platforms during August-December 2025, LiveVQA-W comprises 234K images, 873K questions, and 171K visual entities with hierarchical evaluation: Level 1 for visual entity recognition and Level 2 for multi-hop cross-lingual reasoning. Our comprehensive benchmarking of 15 state-of-the-art MLLMs reveals that models without search achieve near-random performance, while search-augmented models exhibit severe linguistic bias, with English accuracy nearly double that of other languages. Furthermore, we explore visual knowledge updating through large-scale training, finding that injected knowledge improves recall but remains fragile under prompt rephrasing and image perturbations such as rotation and flipping. We release the fully replicable data collection pipeline and raw dataset to support continuous community-driven expansion. The benchmark, code, and related resources are available at: https://worldwide-livevqa.github.io.

Poster Session F (Grand Hall)

09:00-10:30 (Grand Hall)

[Findings] 1,729 vs. : The Effect of Scripts and Formats on LLM Numeracy ▶ *Underline Video*
Varshini Reddy, Craig W Schmidt, Seth Ebner, Adam Wiemerslage, Yuval Pinter and Chris Tanner
Large language models (LLMs) have achieved impressive proficiency in basic arithmetic, rivaling human-level performance on standard numerical tasks. However, little attention has been given to how these models perform when numerical expressions deviate from the prevailing conventions present in their training corpora. In this work, we investigate numerical reasoning across a wide range of numeral scripts and formats. We show that LLM accuracy drops substantially when numerical inputs are rendered in underrepresented scripts or formats, despite the underlying mathematical reasoning being identical. We further demonstrate that targeted prompting strategies, such as few-shot prompting and explicit numeral mapping, can greatly narrow this gap. Our findings highlight an overlooked challenge in multilingual numerical reasoning and provide actionable insights for working with LLMs to reliably interpret, manipulate, and generate numbers across diverse numeral scripts and formatting styles.

[Long] A Linguistics-Aware LLM Watermarking via Syntactic Predictability ▶ *Underline Video*
Shinwoo Park, Hyejin Park, Hyeseon An and Yo-Sub Han
As large language models (LLMs) continue to advance rapidly, reliable governance tools have become critical. Publicly verifiable watermarking is particularly essential for fostering a trustworthy AI ecosystem. A central challenge persists: balancing text quality against detection robustness. Recent studies have sought to navigate this trade-off by leveraging signals from model output distributions (e.g., token-level

entropy); however, their reliance on these model-specific signals presents a significant barrier to public verification, as the detection process requires access to the logits of the underlying model. We introduce STELA, a novel framework that aligns watermark strength with the linguistic degrees of freedom inherent in language. STELA dynamically modulates the signal using part-of-speech (POS) n-gram-modeled linguistic indeterminacy, weakening it in grammatically constrained contexts to preserve quality and strengthening it in contexts with greater linguistic flexibility to enhance detectability. Our detector operates without access to any model logits, thus facilitating publicly verifiable detection. Through extensive experiments on typologically diverse languages, analytic English, isolating Chinese, and agglutinative Korean, we show that STELA surpasses prior methods in detection robustness. Our code is available at https://github.com/Shinwoo-Park/stela_watermark.

[Long] A Model of the Language Process ▶ *Underline Video*

Brandon Duderstadt and Hayden Helm

Language is a process that changes over time as new vocabulary emerges, word meanings shift, and narratives progress. Despite this fact, most Large Language Models are trained on corpora that lack explicit temporal information, which inhibits their ability to model the language process. In this work, we introduce the Temporal Language Model 1 (TLM-1), a BERT style transformer encoder that models that language process by jointly learning to predict document contents and classify document publication dates. We also introduce a Bayesian framework for querying TLM-1 that disentangles its temporal dynamics from several sources of anachronism. Using this query framework, we demonstrate that TLM-1 effectively surfaces several sociolinguistic trends in contemporary American English and accurately detects semantic changes in word meanings. Furthermore, we perform a mechanistic analysis of TLM-1's time token embeddings, and find that they learn a curve whose geometry recovers the ordinal progression of time. We take the existence of this curve as evidence that TLM-1 is effectively learning to reconstruct temporal language dynamics.

[Findings] A Multilingual Dataset and Empirical Validation for the Mutual Reinforcement Effect in Information Extraction ▶ *Underline Video*

Chengguang_Gan Chengguang_Gan, Sunbowen Lee, Qingyu Yin, Yunhao Liang, Xinyang He, Hanjun Wei, Younghun Lim, Shijian Wang, Hexiang huang, QingHao Zhang, Shiwen Ni and Tatsunori Mori

The Mutual Reinforcement Effect (MRE) describes a phenomenon in information extraction where word-level and sentence-level tasks can mutually improve each other when jointly modeled. While prior work has reported MRE in Japanese, its generality across languages and task settings has not been empirically validated, largely due to the lack of multilingual MRE datasets. To address this limitation, we introduce the Multilingual MRE Mix dataset (MMM), which consists of 21 sub-datasets covering English, Japanese, and Chinese. We propose an LLM-assisted dataset translation and alignment framework that significantly reduces manual annotation effort while preserving the structural requirements of MRE tasks. Building on MMM, we adopt a unified input-output framework to train an open-domain information extraction model and conduct extensive empirical studies, including full fine-tuning ablations and the construction of knowledgeable verbalizers based on MRE-mix data. Experimental results show that 76 percent of the MMM sub-datasets consistently exhibit the Mutual Reinforcement Effect across languages. These findings provide systematic empirical validation of MRE in multilingual settings and demonstrate its practical value for information extraction.

[Short] A Shared Geometry of Difficulty in Multilingual Language Models ▶ *Underline Video*

Stefano Civelli, Pietro Bernardelle, Nicolò Brunello and Gianluca Demartini

Large language models (LLMs) encode problem difficulty as an internal signal that can be linearly decoded from their residuals. Given their multilingual capabilities, we investigate whether this meta-cognitive signal is language-agnostic and how it is organized across the models layers by training linear probes on the AMC subset of the Easy2Hard benchmark, translated into 21 languages. We found that difficulty-related signals emerge at two distinct stages of the model internals, corresponding to shallow (early-layers) and deep (later-layers) internal representations, that exhibit functionally different behaviors. Probes trained on deep representations achieve high accuracy when evaluated on the same language but exhibit weaker cross-lingual transfer. In contrast, probes trained on shallow representations generalize better across languages, despite achieving lower within-language performance. This closely aligns with existing findings in LLM interpretability, showing that models tend to operate in an abstract conceptual space before producing language-specific outputs. Our results suggest that this two-stage organizational principle extends beyond simple semantic processing to meta-cognitive properties such as problem difficulty, highlighting an internal control signal that is not tied to surface meaning.

[Long] A Structured Clustering Approach for Inducing Media Narratives ▶ *Underline Video*

Rohan Das, Advait Deshmukh, Alexandria Leto, Zohar Naaman, I-Ta Lee and Maria Leonor Pacheco

Media narratives wield tremendous power in shaping public opinion, yet computational approaches struggle to capture the nuanced storytelling structures that communication theory emphasizes as central to how meaning is constructed. Existing approaches either miss subtle narrative patterns through coarse-grained analysis or require domain-specific taxonomies that limit scalability. To bridge this gap, we present a framework for inducing rich narrative schemas by jointly modeling events and characters via structured clustering. Our approach produces explainable narrative schemas that align with established framing theory while scaling to large corpora without exhaustive manual annotation.

[Long] A Survey of Reasoning-Intensive Retrieval: Progress and Challenges ▶ *Underline Video*

Yiyang Wei, Tingyu Song, Siyue Zhang and Yilun Zhao

Reasoning-Intensive Retrieval (RIR) targets retrieval settings where relevance is mediated by latent inferential links between a query and supporting evidence, rather than semantic similarity. Motivated by the emergent reasoning abilities of Large Language Models (LLMs), recent work integrates these capabilities into the IR field, spanning the entire pipeline from benchmarks to retrievers and rerankers. Despite this progress, the field lacks a systematic framework to organize current efforts and articulate a clear path forward. To provide a clear roadmap for this rapidly growing yet fragmented area, this survey (1) systematizes existing RIR benchmarks by knowledge domains and modalities, providing a detailed analysis of the current landscape; (2) introduces a structured taxonomy that categorizes methods based on where and how reasoning is integrated into the retrieval pipeline, alongside an analysis of their trade-offs and practical applications; and (3) summarizes challenges and future directions to guide research in this evolving field.

[Findings] A Survey of Toxicity Mitigation Strategies for Multilingual Language Models ▶ *Underline Video*

Soham Dan, Himanshu Beniwal and Thomas Hartvigsen

Large language models (LLMs) are transforming natural language processing across diverse linguistic communities. However, they can reproduce and amplify toxic content, including hate speech, harassment, and bias, posing significant risks to multilingual applications. We provide the first comprehensive survey of the many detoxification methods specifically tailored to multilingual LLMs. First, we define toxicity its measurement, then we provide a brief review of monolingual mitigation strategies, including data filtering, style transfer, expert-based logit steering, retrieval augmentation, and alignment with human feedback. We then present an in-depth taxonomy of multilingual approaches spanning (1) training methods, (2) post-hoc editing and decoding strategies, (3) alignment and reinforcement-learning techniques, and (4) data-centric innovations, such as parallel detox corpora and synthetic data generation. Finally, we discuss open challenges in multilingual detoxification, including data scarcity, evaluation inconsistencies, cultural nuances and biases. Overall, we produce a needed overview of the state of multi-lingual toxicity detection and mitigation on which the community can ground to build globally safe and equitable LLMs.

[Long] AEA: Adaptive Expert Allocation Improves Sentence Embeddings from Mixture-of-Experts LLM ▶ *Underline Video*

Shufan Yang, Zifeng Cheng, Zhiwei Jiang, Qingfeng Qi, Yafeng Yin, Cong Wang, Ao Zhou and Qing Gu

Extracting embeddings directly from Mixture-of-Experts (MoE) models is a promising yet underexplored direction that requires no additional data or fine-tuning. While previous studies have utilized semantic compression prompts or expert routing information to improve sentence embeddings, they typically allocate a fixed number of experts uniformly across all layers and tokens, ignoring inter-layer and inter-token heterogeneity. In this work, we identify two key observations in MoE models: (1) layer-wise variations in expert homogeneity, suggesting that different layers require different expert budgets, and (2) token-wise contribution imbalance, indicating that different tokens should also be allocated different numbers of experts. To address these issues, we propose an Adaptive Expert Allocation (AEA) framework that dynamically performs both layer-wise and token-wise expert allocation to enhance embedding quality. Specifically, AEA allocates fewer experts to layers with higher homogeneity and to tokens with lower attention importance, where layer-wise homogeneity is determined by the similarity among embeddings produced by the experts in each layer. Notably, our method is plug-and-play, seamlessly integrates with existing prompt engineering methods, and introduces no additional time overhead. Experiments on the STS tasks demonstrate that AEA consistently improves embedding quality across multiple MoE models.

[Findings] AICA-Bench: Holistically Examining the Capabilities of VLMs in Affective Image Content Analysis ▶ *Underline Video*

Dong She, Xianrong Yao, Liqun Chen, Jinghe Yu, Yang Gao and Zhanpeng Jin

Vision-Language Models (VLMs) have demonstrated strong capabilities in perception, yet holistic Affective Image Content Analysis (AICA), which integrates perception, reasoning, and generation into a unified framework, remains underexplored. To address this, we introduce AICA-Bench, a comprehensive benchmark comprising three core tasks: Emotion Understanding (EU), Reasoning (ER), and Generation (EGCG). We evaluate 23 VLMs, revealing critical gaps: models struggle with intensity calibration and suffer from descriptive shallowness in open-ended tasks. To bridge these gaps, we propose Grounded Affective Tree (GAT) Prompting, a training-free framework that integrates visual scaffolding with hierarchical reasoning. Experiments show that GAT effectively corrects intensity errors and significantly enhances descriptive depth, establishing a robust baseline for future affective multimodal research.

[Long] Aligning Agents via Planning: A Benchmark for Trajectory-Level Reward Modeling ▶ *Underline Video*

Jiaxuan Wang, Yulan Hu, Wenjin Yang, Zheng Pan, Xin Li and Lan-Zhe Guo

In classical Reinforcement Learning from Human Feedback (RLHF), Reward Models (RMs) serve as the fundamental signal provider for model alignment. As Large Language Models evolve into agentic systems capable of autonomous tool invocation and complex reasoning, the paradigm of reward modeling faces unprecedented challenges–most notably, the lack of benchmarks specifically designed to assess RM capabilities within tool-integrated environments. To address this gap, we present Plan-RewardBench, a trajectory-level preference benchmark designed to evaluate how well judges distinguish preferred versus distractor agent trajectories in complex tool-using scenarios. Plan-RewardBench covers four representative task families—(i) Safety Refusal, (ii) Tool-Irrelevance / Unavailability, (iii) Complex Planning, and (iv) Robust Error Recovery—comprising validated positive trajectories and confusable hard negatives constructed via multi-model natural rollouts, rule-based perturbations, and minimal-edit LLM perturbations. We benchmark representative RMs (generative, discriminative, and LLM-as-Judge) under a unified pairwise protocol, reporting accuracy trends across varying trajectory lengths and task categories. Furthermore, we provide diagnostic analyses of prevalent failure modes. Our results reveal that all three evaluator families face substantial challenges, with performance degrading sharply on long-horizon trajectories, underscoring the necessity for specialized training in agentic, trajectory-level reward modeling. Ultimately, Plan-RewardBench aims to serve as both a practical evaluation suite and a reusable

blueprint for constructing agentic planning preference data.

[Findings] Almieyar-Oryx-BloomBench: A Bilingual Multimodal Benchmark for Cognitively Informed Evaluation of Vision-Language Models ▶ *Underline Video*

Mohammad Mahdi Abootorabi, Omid Ghahroodi, Anas Madkoor, Marzia Nouri, Doratossadat Dastgheib and Ehsaneddin Asgari

Despite the rapid progress of Vision-Language Models (VLMs), the field lacks benchmarks that rigorously diagnose their true reasoning abilities and chart meaningful progress toward human-like multimodal intelligence. Most existing evaluations focus on piecemeal or disconnected tasks, obscuring critical cognitive weaknesses and providing little insight for targeted improvement. To address this gap, we introduce BloomBench, part of the Almieyar benchmarking series, the first cognitively human-grounded, bilingual (English-Arabic) multimodal benchmark for VLMs. Grounded in Blooms Taxonomy, BloomBench systematically evaluates six levels of cognition (Remember, Understand, Apply, Analyze, Evaluate, Create) through carefully designed image-question-answer tasks. Built with a semi-automated pipeline and validated through a stratified hybrid quality assurance protocol, it ensures scalability, cultural inclusivity, and linguistic fidelity. Leveraging this framework, we conduct a comprehensive study of state-of-the-art VLMs to diagnose their cognitive profiles. Our analysis reveals a sharp cognitive asymmetry: while state-of-the-art models achieve strong performance ceilings in semantic understanding, they struggle substantially with factual recall and creative synthesis. This demonstrates that current general multimodal proficiency masks deeper limitations in specific cognitive layers. Furthermore, our study highlights a critical performance gap between Arabic and English, exposing limitations in current cross-lingual multimodal reasoning. These findings establish a foundation for developing more cognitively aligned and inclusive VLMs. The benchmark framework and dataset is available at: https://github.com/qcri/Almieyar-Oryx-BloomBench.

[Long] Amadeus: Autoregressive Model with Bidirectional Attribute Modelling for Symbolic Music ▶ *Underline Video*

Hongju Su ~Hongju_Su1, Ke Li, LAN YANG, Honggang Zhang and Yi-Zhe Song

Existing state-of-the-art symbolic music generation models represent symbolic music as a sequence of attribute tokens with fixed unidirectional dependencies. However, from the perspective of music theory, the attributes of a musical note are inherently a set rather than a sequence. Building on this insight, we propose Amadeus, a novel symbolic music generation framework that adopts a two-level architecture: an autoregressive model for note sequences and a bidirectional discrete diffusion model for note attributes. This design enables flexible attribute control and adjustable decoding speed during inference. To further enhance sequential modeling, we introduce the Conditional Information Enhancement Module (CIEM). We also constructed AMD (Amadeus MIDI Dataset), the largest open-source symbolic music dataset to date, supporting both pre-training and fine-tuning. We trained two models of different scales, Amadeus and Amadeus-M, and conducted extensive experiments, demonstrating substantial improvements over state-of-the-art methods across both objective and subjective metrics.

[Long] Analyzing and Internalizing Complex Policy Documents for LLM Agents ▶ *Underline Video*

Jiateng Liu, Zhenhailong Wang, Xiaojiang Huang, Yingjie Li, Xiang Li, Chenlei Guo, Xing Fan, Ruhi Sarikaya and Heng Ji

Large language model agents rely on in-context policy documents encoding diverse business rules. As businesses scale, these documents grow, creating substantial computational overhead and motivating internalization methods that embed policy into model priors. Prior work focuses on generic prompts, but we find agentic policies span multiple complexity levels and demand heavier reasoning, posing greater challenges. We introduce an agentic benchmark generator with Controllable Complexity in agent policy across four levels, enabling systematic evaluation of agents under increasing complexity and providing a testbed for policy internalization. Our analysis shows that workflow-governing policy specifications are the hardest to reason over, and that SFT on gold trajectories with chain-of-thought is data-hungry and struggles at high complexity. We propose Category-Aware Policy Continued Pretraining, an automated pipeline that analyzes policies, extracts key specifications, categorizes them into factual, behavioral, and conditional types, and isolates those driving workflow complexity. This enables targeted therapy by synthesizing specialized training data for each type and improving internalization via an autoregressive pretraining loss. Extensive experiments show our synthetic data and objective consistently improve performance. Combined with SFT, our method outperforms the baseline across different settings, especially in data-sparse and high-complexity regimes, with gains up to 41% and 22% on Qwen-3-32B. Overall, we achieve 97.3% prompt reduction on our benchmark, and on -Bench we further improve performance while reducing prompt requirements with very limited SFT data.

[Findings] AnchorMem: Anchored Facts with Associative Contexts for Building Memory in Large Language Models ▶ *Underline Video*

Zhanyu Shen, Sijie Cheng, Zhicheng Guo, Weiqin Wang, Yile Wang and Hui Huang

While large language models have achieved remarkable performance in complex tasks, they still need a memory system to utilize historical experience in long-term interactions. Existing memory methods (e.g., A-Mem, Mem0) place excessive emphasis on organizing interactions by frequently rewriting them, however, this heavy reliance on summarization risks diluting essential contextual nuances and obscuring key retrieval features. To bridge this gap, we introduce AnchorMem, a novel memory framework inspired by the Proust Phenomenon in cognitive science, where a specific anchor triggers a holistic recollection. We propose a method that decouples the retrieval unit from the generation context. AnchorMem extracts atomic facts from interaction history to serve as retrieval anchors, while preserving the original context as the immutable context. To reveal implicit narrative cues, we construct an associative event graph that uses higher-order event links that bind sets of related facts into shared event representations, strengthening cross-memory integration without relying on generic entities as bridges. During retrieval, the system anchors queries to specific facts and events to locate relevant memories, but reconstructs the context using the

associated raw chunks and events. Our method reconciles fine-grained retrieval with the contextual integrity of interactions. Experiments across three closed-source and open-source models on the LoCoMo benchmark demonstrate that AnchorMem significantly outperforms baselines.

[Findings] Astra: Activation-Space Tail-Eigenvector Low-Rank Adaptation of Large Language Models ▶ *Underline Video*

Kainan Liu, Yong Zhang, Ning Cheng, Yun Zhu, Yanmeng Wang, Shaojun Wang and Jing Xiao

Parameter-Efficient Fine-Tuning (PEFT) methods, especially LoRA, are widely used for adapting pre-trained models to downstream tasks due to their computational and storage efficiency. However, in the context of LoRA and its variants, the potential of activation subspaces corresponding to tail eigenvectors remains substantially under-exploited, which may lead to suboptimal fine-tuning performance. In this work, we propose Astra (Activation-Space Tail-Eigenvector Low-Rank Adaptation), a novel PEFT method that leverages the tail eigenvectors of the model output activations, estimated from a small task-specific calibration set, to construct task-adaptive low-rank adapters. By constraining updates to the subspace spanned by these tail eigenvectors, Astra achieves faster convergence and improved downstream performance with a significantly reduced parameter budget. Extensive experiments across natural language understanding (NLU) and natural language generation (NLG) tasks demonstrate that Astra consistently outperforms existing PEFT baselines across 16 benchmarks and even surpasses full fine-tuning (FFT) in certain scenarios.

[Findings] AutoMonitor-Bench: Evaluating the Reliability of LLM-Based Misbehavior Monitor ▶ *Underline Video*

Shu Yang, Jingyu Hu, Tong Li, Hanqi Yan, Wenxuan Wang and Di Wang

We introduce AutoMonitor-Bench, the first benchmark designed to systematically evaluate the reliability of LLM-based misbehavior monitors across diverse tasks and failure modes. AutoMonitor-Bench consists of 3,010 carefully annotated test samples spanning question answering, code generation, and reasoning, with paired misbehavior and benign instances. We evaluate monitors using two complementary metrics: Miss Rate (MR) and False Alarm Rate (FAR), capturing failures to detect misbehavior and oversensitivity to benign behavior respectively. Evaluating 12 proprietary and 10 open-source LLMs, we observe substantial variability in monitoring performance and a consistent trade-off between MR and FAR, revealing an inherent safety-utility tension. To further explore the limits of monitor reliability, we construct a large-scale training corpus of 153,581 samples and fine-tune Qwen3-4B-Instruction, to investigate whether training on known, relatively easy-to-construct misbehavior datasets improves monitoring performance on unseen and more implicit misbehaviors. Our results highlight the challenges of reliable, scalable misbehavior monitoring and motivate future work on task-aware designing and training strategies for LLM-based monitors.

[Findings] Before Forgetting, Learn to Remember: Revisiting Foundational Learning Failures in LVLM Unlearning Benchmarks ▶ *Underline Video*

JuneHyoung Kwon, MiHyeon Kim, Eunju Lee, JungMin Yun, Byeonggeuk Lim and YoungBin Kim

While Large Vision-Language Models (LVLMs) offer powerful capabilities, they pose privacy risks by unintentionally memorizing sensitive personal information. Current unlearning benchmarks attempt to mitigate this using fictitious identities but overlook a critical stage 1 failure: models fail to effectively memorize target information initially, rendering subsequent unlearning evaluations unreliable. Diagnosing under-memorization and the multi-hop curse as root causes, we introduce ReMem, a Reliable Multi-hop and Multi-image Memorization Benchmark. ReMem ensures robust foundational learning through principled data scaling, reasoning-aware QA pairs, and diverse visual contexts. Additionally, we propose a novel Exposure metric to quantify the depth of information erasure from the model's internal probability distribution. Extensive experiments demonstrate that ReMem provides a rigorous and trustworthy framework for diagnosing both learning and unlearning behaviors in LVLMs.

[Findings] Behavior-Aware Item Modeling via Dynamic Procedural Solution Representations for Knowledge Tracing ▶ *Underline Video*

Jun Seo, Sangwon Ryu, Heejin Do, Hyounghun Kim and Gary Geunbae Lee

Knowledge Tracing (KT) aims to predict learners' future performance from past interactions. While recent KT approaches have improved via learning item representations aligned with Knowledge Components, they overlook the procedural dynamics of problem solving. We propose Behavior-Aware Item Modeling (BAIM), a framework that enriches item representations by integrating dynamic procedural solution information. BAIM leverages a reasoning language model to decompose each item's solution into four problem-solving stages (i.e., understand, plan, carry out, and look back), pedagogically grounded in Polyas framework. Specifically, it derives stage-level representations from per-stage embedding trajectories, capturing latent signals beyond surface features. To reflect learner heterogeneity, BAIM adaptively routes these stage-wise representations, introducing a context-conditioned mechanism within a KT backbone, allowing different procedural stages to be emphasized for different learners. Experiments on XES3G5M and NIPS34 show that BAIM consistently outperforms strong pretraining-based baselines, achieving particularly large gains under repeated learner interactions.

[Long] Benchmarking Deflection and Hallucination in Large Vision-Language Models ▶ *Underline Video*

Nicholas Moratelli, Christopher Davis, Leonardo F. R. Ribeiro, Bill Byrne and Gonzalo Iglesias

Large Vision-Language Models (LVLMs) increasingly rely on retrieval to answer knowledge-intensive multimodal questions. Existing benchmarks overlook conflicts between visual and textual evidence and the importance of generating deflections (e.g., "Sorry, I cannot answer...") when retrieved knowledge is incomplete. These benchmarks also suffer from rapid obsolescence, as growing LVLM training sets

allow models to answer many questions without retrieval. We address these gaps with three contributions. First, we propose a dynamic data curation pipeline that preserves benchmark difficulty over time by filtering for genuinely retrieval-dependent samples. Second, we introduce VLM-DeflectionBench, a benchmark of 2,775 samples spanning diverse multimodal retrieval settings, designed to probe model behaviour under conflicting or insufficient evidence. Third, we define a fine-grained evaluation protocol with four scenarios that disentangle parametric memorization from retrieval robustness. Experiments across 20 state-of-the-art LVLMs indicate that models usually fail to deflect in the presence of noisy or misleading evidence. Our results highlight the need to evaluate not only what models know, but how they behave when they do not, and serve as a reusable and extensible benchmark for reliable KB-VQA evaluation. All resources will be publicly available upon publication.

[Long] Benchmarking LLM's Capability in Reasoning over Conflicting Web References ► *Underline Video*

Yizhen Yuan, Rui Kong, Dongze Li, Yuanchun Li and Yunxin Liu

Large language models (LLMs) integrated with retrieval-augmented generation (RAG) have become a dominant framework for building intelligent assistants. In real-world applications such as ChatGPT with web search, the retrieved document often comes from diverse, potentially unreliable sources and may contain inconsistent claims. Unlike traditional search engines that rely on users to manually compare information, LLM-based systems typically feed all retrieved content into the model's context, requiring LLMs to autonomously identify, differentiate, and reason over conflicting viewpoints. Unlike mainstream LLM evaluation tasks like math and code generation that are primarily focused on reasoning with factual context, question-answering with multi-source references requires fundamentally different capabilities to identify and reason over knowledge contradictions. In this paper, we introduce ConfRAG, a benchmark for evaluating LLMs' reasoning capability over real-world conflicting documents retrieved from the web. It consists of 1,814 real-world questions, each paired with an average of 9.58 retrieved paragraphs from heterogeneous online sources. A total of 57.2% of the questions exhibit explicit contradictions. We further propose three structured evaluation tasks, answer clustering, answer coverage, and reason coverage, to quantify a model's ability to organize and explain contradictory content. Experiments with state-of-the-art models such as GPT-4.1 and Claude-3-7-Sonnet reveal substantial performance gaps, highlighting the need for more targeted research in contradiction-aware question answering. To the best of our knowledge, ConfRAG is the first benchmark specifically designed to evaluate contradiction-aware reasoning on real-world long web documents.

[Findings] Beyond Black-Box Labels: Interpretable Criteria for Diagnosing Subjective NLP Tasks ► *Underline Video*

Nisrine Rair, Alban Goupil, Valeriu Vrabie and Emmanuel Chochoy

Subjective NLP datasets typically aggregate annotator judgments into a single gold label, making it difficult to diagnose whether disagreement reflects unclear criteria, collapsed distinctions, or legitimate plurality. We propose a schema-level diagnostic for auditing expert-designed annotation schemas prior to gold-label commitment, using only multi-annotator criterion judgments. The diagnostic separates two failure modes: unstable criteria with hard-to-operationalize boundaries, and systematic overlap that blurs the boundaries between mutually exclusive categories. Applied to persuasive value extraction in commercial documents, we find that disagreement is not diffuse: instability concentrates in a few criteria, while nearly half of covered sentences activate multiple categories. These signals align with where domain experts disagree, yielding an evidence-based audit for tightening guidelines, revising category structure, or reconsidering the annotation paradigm.

[Long] Beyond Self-Report: Bridging the Intention-Behavior Gap in Critical Thinking Assessment via Interpretable Multi-Agent System ► *Underline Video*

Zekun Li, Jifan Yu, Haoxuan Li, Ye He, Daniel Zhang-Li, Shangqing Tu, Joy Jia Yin Lim, Yikun Jiang, Jiaxin Yuan and Yu Zhang

Accurate assessment of critical thinking is historically limited by the Intention Behavior Gap in psychology: the disconnect between what individuals self-reported disposition and their actual practical behaviors. We try to bridge this gap with MASA (Multi-Agent Scenario-based Assessment), a framework that operationalizes cognitive assessment into an interpretable and interactive multi-agent workflow with Assessment Chain-of-Thought (AsCoT). Validating on both large-scale simulations (N=1,161) and human participants (N=70), we find that MASA aligns better with human expert ratings (r=0.882) than traditional gold-standard inventories (r=0.720), with an average cost of only $0.41 per participant. These results suggest that by shifting from self-report inventory to behavior-grounded dialogue, MASA offers a more accurate, cost-effective, and transparent solution for real-world cognitive evaluation.

[Long] Beyond Word Boundaries: A Hebrew Coreference Benchmark and an Evaluation Protocol for Morphologically Complex Text ► *Underline Video*

Refael Shaked Greenfeld and Reut Tsarfaty

Coreference Resolution (CR) is a fundamental NLP task critical for long-form tasks as information extraction, summarization, and many business applications. However, CR methods originally designed for English struggle with Morphologically Rich Languages (MRLs), where mention boundaries do not necessarily align with word boundaries, and a single token may consist of multiple anaphors. CR modeling and evaluation protocols standardly assume that, as in English, words and mentions mostly align. However, this assumption breaks down in MRLs, particularly in the context of LLMs' raw-text processing and end-to-end tasks. To assess and address this challenge, we introduce KibutzR, the first comprehensive CR dataset for Modern Hebrew, an MRL rich with complex words and pronominal clitics. We deliver an annotated dataset that identifies mentions at word, sub-word and multi-word levels, and propose an evaluation protocol that directly addresses word/morpheme boundary discrepancies. Our experiments show that contemporary LLMs perform significantly worse on Hebrew than on English, and that performance degrades on raw unsegmented text. Crucially, we show an inverse performance-trend in Hebrew

relative to English, where smaller encoders perform far better than contemporary decoder models, leaving ample space for investigation and improvement. We deliver a new benchmark for Hebrew coreference resolution and a segmentation-aware evaluation protocol to inform future work on other MRLs.

[Long] Calibrated Speculative Decoding: Frequency-Guided Candidate Selection for Efficient Inference ► *Underline Video*

Xuwen Zhou, Fangxin Liu, Chao Wang, Xiao Zheng, Hao Zheng, Min He, Li Jiang and Haibing Guan

Speculative decoding accelerates autoregressive generation by letting draft tokens bypass full verification, but conventional frameworks suffer from frequent false rejections, particularly when draft models produce semantically correct but lexically divergent outputs. In this paper, we present Calibrated Speculative Decoding (CSD), a training-free framework that recovers valid tokens discarded by standard verification. Guided by the principle of "Frequency-Guided Candidate Selection and Probability-Guarded Acceptance," CSD incorporates two lightweight modules: Online Correction Memory, which aggregates historical rejections to propose recurring divergence patterns as rescue candidates, and Semantic Consistency Gating, which verifies candidate admissibility using probability ratios instead of exact token matching. Our evaluation across diverse large language models demonstrates that CSD outperforms existing methods, achieving a peak throughput speedup of 2.33x. CSD preserves model accuracy across all tasks while further boosting performance on complex reasoning datasets. These results establish CSD as a highly effective, lightweight solution for practical LLM deployments.

[Long] Can Large Language Models Infer Causal Relationships from Real-World Text? ► *Underline Video*

Ryan Saklad, Aman Chadha, Oleg Pavlov and Raha Moraffah

Understanding and inferring causal relationships from texts is a core aspect of human cognition and is essential for advancing large language models (LLMs) towards artificial general intelligence. Existing work evaluating LLM causal reasoning primarily relies on synthetic or simplified texts with explicitly stated causal relationships. These texts typically feature short passages and few causal relations, failing to reflect the complexities of real-world reasoning. In this paper, we investigate whether LLMs are capable of inferring causal relationships from real-world texts. We develop a benchmark drawn from real-world academic literature, which includes diverse texts with respect to length, complexity (different levels of explicitness, number of causal events and relationships), and domain. To the best of our knowledge, our benchmark is the first-ever real-world dataset for this task. Our experiments on this dataset show that LLMs face significant challenges in inferring causal relationships from real-world text, with the best-performing model achieving an average F_1 score of only 0.535. Through systematic analysis across aspects of real-world text (explicitness, number of causal events and relationships, length of text, domain), our benchmark offers targeted insights for further research into advancing LLM causal reasoning. Our code and dataset can be found at https://github.com/Ryan-Saklad/ReCITE.

[Findings] Can LLMs Compress (and Decompress)? Evaluating Code Understanding and Execution via Invertibility ► *Underline Video*

Nickil Maveli, Antonio Vergari and Shay Cohen

LLMs demonstrate strong performance on code benchmarks, yet round-trip code execution reveals limitations in their ability to maintain consistent reasoning across forward and backward execution. We present RoundTripCodeEval (RTCE), a comprehensive benchmark consisting of four distinct code execution reasoning tasks designed to rigorously test round-trip consistency. RTCE provides an execution-free, exact-match evaluation of bijection fidelity, assessing whether models preserve a consistent one-to-one mapping between encoding and decoding operations across various algorithms and directions. We systematically evaluate state-of-the-art Code-LLMs using zero-shot prompting, supervised fine-tuning on execution traces, and self-reflection mechanisms. Each yields modest improvements, but none closes the gap, indicating that current LLMs struggle with true round-trip consistency, which demonstrates that they lack the internal coherence required for trustworthy code reasoning. RTCE surfaces several new and previously unmeasured insights that are not captured by existing I/O-prediction, execution-reasoning, or round-trip natural-language benchmarks.

[Long] Can Spectral-Clipping Enable Better Learning While Forgetting Less for Low-Rank Adaptation? ► *Underline Video*

Hyowon Wi and Noseong Park

In recent years, low-rank adaptation (LoRA) has emerged as a significant paradigm that freezes pre-trained weights and introduces small, learnable adapters instead of fine-tuning the full set of parameters. In this work, we uncover several key insights regarding the singular components of network parameters based on Singular Value Decomposition (SVD). Firstly, the principal singular components with large singular values in pre-trained network parameters can be effectively reused during fine-tuning, whereas the minor components with smaller singular values are more task-specific and require substantial adaptation. Secondly, we first establish the theoretical connection that the uncontrolled growth of singular values in LoRA adapters leads to the forgetting of pre-trained knowledge , a well-known issue referred to as catastrophic forgetting. Building on these observations, we propose SCLoRA, which injects parameterized singular components with spectral clipping into the pre-trained model in a way that is aware of the spectral distribution of the pre-trained model. SCLoRA effectively adapts to new tasks by focusing updates on components that require adaptation, while simultaneously alleviating catastrophic forgetting. We conduct extensive experiments and demonstrate that SCLoRA not only improves downstream performance but also effectively retains pre-trained knowledge.

[Findings] Candidate-Aware Retrieval and Reranking for Multiple-Choice Question Answering: Arabic as a Case Study ► *Underline Video*

Yassine Bouziane, Youness Moukafih and Mounir Ghogho

Large language models (LLMs) have recently achieved impressive results on multiple-choice question answering (MCQA), with retrieval-augmented generation (RAG) emerging as an effective strategy for improving the performance of smaller models. However, existing RAG formulations face persistent challenges: retrieving too many passages often introduces noise, and even when relevant content is retrieved, models may still struggle with partially relevant or conflicting information. Moreover, while LLMs perform strongly on English benchmarks, their accuracy declines substantially on Arabic multi-task evaluations, revealing ongoing issues in cross-lingual transfer and domain adaptation. In this paper, we propose a novel approach, using Arabic as a representative case study, that jointly models the relevance of both the question and its candidate answers when selecting contextual passages. The method employs a lightweight reranker trained with a hybrid regression-triplet loss objective to identify passages that provide discriminative and reliable evidence. Extensive experiments across multiple model sizes and humanities domains show that our approach consistently outperforms both standard RAG baselines and reranker baselines, delivering two- to threefold improvements while remaining competitive with considerably larger models.

[Findings] Capabilities and Evaluation Biases of Large Language Models in Classical Chinese Poetry Generation: A Case Study on Tang Poetry ▶ *Underline Video*

Bolei Ma, Yina Yao and Anna-Carolina Haensch

Large Language Models (LLMs) are increasingly applied to creative domains, yet their performance in classical Chinese poetry generation and evaluation remains poorly understood. We propose a three-step evaluation framework that combines computational metrics, LLM-as-a-judge assessment, and human expert validation. Using this framework, we evaluate six state-of-the-art LLMs across multiple dimensions of poetic quality, including themes, emotions, imagery, form, and style, in the context of Tang poetry () generation. Our analysis reveals a critical "echo chamber" effect: LLMs systematically overrate machine-generated poems that mimic statistical patterns yet fail strict prosodic rules, diverging significantly from human expert judgments. These findings underscore the limitations of using LLMs as standalone evaluators for culturally complex tasks, highlighting the necessity of hybrid human-model validation frameworks.

[Findings] Carefully Considering Culture: Analyzing LLM Alignment in Single- and Multi-Cultural Settings using Cultural Consensus Theory ▶ *Underline Video*

Krishna Pothugunta and John Lalor

Recent work in NLP has probed large language models for their understanding of cultural norms across countries. However, this work typically considers distributional patterns, ignoring group consensus or possible multicultural environments within a country. In this work, we leverage cultural consensus theory (CCT) from cultural anthropology to model such multidimensional nuance. Applying CCT to the World Values Survey (WVS) across 10 countries and 12 domains, we demonstrate that models frequently misrepresent cultural structures by either failing to form cohesive consensus or severely over-regularizing consensus. Through explicit representation of intra-group variance, CCT provides actionable diagnostics to evaluate when models reflect true human diversity versus algorithmic homogenization.

[Findings] CLaS-Bench: A Cross-Lingual Alignment and Steering Benchmark ▶ *Underline Video*

Daniil Gurgurov, Yusser Al Ghussin, Tanja Baeumel, Cheng-Ting Chou, Patrick Schramowski, Marius Mosbach, Josef van Genabith and Simon Ostermann

Understanding and controlling the behavior of large language models (LLMs) is an increasingly important topic in multilingual NLP. Beyond prompting or fine-tuning, language steering, i.e.,~manipulating internal representations during inference, has emerged as a more efficient and interpretable technique for adapting models to a target language. Yet, no dedicated benchmarks or evaluation protocols exist to quantify the effectiveness of steering techniques. We introduce CLaS-Bench, a lightweight parallel-question benchmark for evaluating language-forcing behavior in LLMs across 32 languages, enabling systematic evaluation of multilingual steering methods. We evaluate a broad array of steering techniques, including residual-stream DiffMean interventions, probe-derived directions, language-specific neurons, PCA/LDA vectors, Sparse Autoencoders, and prompting baselines. Steering performance is measured along two axes: language control and semantic relevance, combined into a single harmonic-mean steering score. We find that across languages simple residual-based DiffMean method consistently outperforms all other methods. Moreover, a layer-wise analysis reveals that language-specific structure emerges predominantly in later layers and steering directions cluster based on language family. CLaS-Bench is the first standardized benchmark for multilingual steering, enabling both rigorous scientific analysis of language representations and practical evaluation of steering as a low-cost adaptation alternative.

[Long] CLEAR: Cross-Lingual Enhancement in Retrieval via Reverse-training ▶ *Underline Video*

Seungyoon Lee, Minhyuk Kim, Seongtae Hong, Youngjoon Jang, Dongsuk Oh and Heuiseok Lim

Existing multilingual embedding models often encounter challenges in cross-lingual scenarios due to imbalanced linguistic resources and less consideration of cross-lingual alignment during training. Although standardized contrastive learning approaches for cross-lingual adaptation are widely adopted, they may struggle to capture fundamental alignment between languages and degrade performance in well-aligned languages such as English. To address these challenges, we propose Cross-Lingual Enhancement in RetrievAl via Reverse-training (CLEAR), a novel loss function utilizing a reverse training scheme to improve retrieval performance across diverse cross-lingual retrieval scenarios. CLEAR leverages an English passage as a bridge to strengthen alignments between the target language and English, ensuring robust performance in the cross-lingual retrieval task. Our extensive experiments demonstrate that CLEAR achieves notable improvements in cross-lingual scenarios, with gains up to 15%, particularly in low-resource languages, while minimizing performance degradation in English.

Furthermore, our findings highlight that CLEAR offers promising effectiveness even in multilingual training, suggesting its potential for broad application and scalability. We release the code at https://github.com/dltmddbs100/CLEAR.

[Findings] Commonsense Knowledge with Negation: A Resource to Enhance Negation Understanding ► *Underline Video*
Zijie Wang, MohammadHossein Rezaei, Farzana Rashid and Eduardo Blanco;
Negation is a common and important semantic feature in natural language, yet Large Language Models (LLMs) struggle when negation is involved in natural language understanding tasks. Commonsense knowledge, on the other hand, despite being a well-studied topic, lacks investigations involving negation. In this work, we show that commonsense knowledge with negation is challenging for models to understand. We present a novel approach to automatically augment existing commonsense knowledge corpora with negation, yielding two new corpora containing over 2M triples with if-then relations. In addition, pre-training LLMs on our corpora benefits negation understanding.

[Long] Confidence Should Be Calibrated More Than One Turn Deep ► *Underline Video*
Zhaohan Zhang, Chengzhengxu Li, Xiaoming Liu, Chao Shen, Ziquan Liu and Ioannis Patras
Large Language Models (LLMs) are increasingly applied in high-stakes domains such as finance, healthcare, and education, where reliable multi-turn interactions with users are essential. However, existing work on confidence estimation and calibration, a major approach to building trustworthy LLM systems, largely focuses on single-turn settings and overlooks the risks and potential of multi-turn conversations. In this work, we introduce the task of multi-turn calibration to reframe calibration from a static property into a dynamic challenge central to reliable multi-turn conversation, where calibrating model confidence at each turn conditioned on the conversation history is required. We first reveal the risks of this setting: using Expected Calibration Error at turn T (ECE@T), a new metric that tracks calibration dynamics over turns, we show that user feedback (e.g., persuasion) can degrade multi-turn calibration. To address this, we propose MTCal, which minimises ECE@T via a surrogate calibration target, and further leverage calibrated confidence in ConfChat, a decoding strategy that improves both factuality and consistency of the model response in multi-turn interactions. Extensive experiments demonstrate that MTCal achieves outstanding and consistent performance in multi-turn calibration, and ConfChat preserves and even enhances model performance in multi-turn interactions. Our results mark multi-turn calibration as one missing link for scaling LLM calibration toward safe, reliable, and real-world use. The code is available at: https://github.com/petezone/Multiturn-Calibration.

[Findings] Content Fuzzing for Escaping Information Coccons on Digital Social Media ► *Underline Video*
Yifeng He, Ziye Tang and Hao Chen
Information cocoons on social media limit users' exposure to posts with diverse viewpoints. Modern platforms use stance detection as an important signal in recommendation and ranking pipelines, which can route posts primarily to like-minded audiences and reduce cross-cutting exposure. This restricts the reach of dissenting opinions and hinders constructive discourse. We take the creator's perspective and investigate how content can be revised to reach beyond existing affinity clusters. We present ContentFuzz, a confidence-guided fuzzing framework that rewrites posts while preserving their human-interpreted intent and induces different machine-inferred stance labels. ContentFuzz aims to route posts beyond their original cocoons. Our method guides a large language model (LLM) to generate meaning-preserving rewrites using confidence feedback from stance detection models. Evaluated on four representative stance detection models across three datasets in two languages, ContentFuzz effectively changes machine-classified stance labels, while maintaining semantic integrity with respect to the original content.

[Long] Contextual Relevance and Adaptive Sampling for LLM-Based Document Reranking ► *Underline Video*
Jerry Huang, Siddarth Madala, Cheng Niu, Julia Hockenmaier and Tong Zhang
Reranking algorithms have made progress in improving document retrieval quality by efficiently aggregating relevance judgments generated by large language models (LLMs). However, identifying relevant documents for queries that require in-depth reasoning remains a major challenge. Reasoning-intensive queries often exhibit multifaceted information needs and nuanced interpretations, rendering document relevance inherently context dependent and often noisy. To address this, we propose contextual relevance, which we define as the probability that a document is relevant to a given query, marginalized over the distribution of different reranking contexts it may appear in (i.e., the set of candidate documents it is ranked alongside and the order in which the documents are presented to a reranking model). While prior works have studied methods to mitigate the positional bias LLMs exhibit by accounting for the ordering of documents, we empirically show that batch composition also materially affects relevance judgments. To efficiently estimate contextual relevance, we propose TS-SetRank, a sampling-based, uncertainty-aware reranking algorithm. Empirically, TS-SetRank improves nDCG@10 over retrieval and reranking baselines by 15-25% on BRIGHT and 6-21% on BEIR, highlighting the importance of modeling relevance as context-dependent.

[Findings] CoopValue: Revealing LLM Value Preferences Through Multi-Agent Cooperation ► *Underline Video*
Zee Hen Tang and Mi-Yen Yeh
Existing evaluations of large language models primarily rely on single-agent dilemmas or static binary-choice tasks, offering limited insight into how cooperation contexts influence LLM behavior. We introduce CoopValue, a multi-agent evaluation framework that assesses LLMs' value preferences through cooperative scenarios. CoopValue includes 1,778 scenarios spanning all pairwise conflicts among the 10 Schwartz values and three cooperation types: reciprocal, coopetitive, and altruistic. We evaluate 24 LLMs across 8 model families and examine how their value preferences vary across different cooperative contexts, showing the importance of assessing LLM value preferences in interactive, context-sensitive settings to guide the selection and deployment of LLMs aligned with desired cooperative behavior.

[Long] CostBench: Evaluating Multi-Turn Cost-Optimal Planning and Adaptation in Dynamic Environments for LLM Tool-Use Agents ▶ *Underline Video*

Jiayu Liu, Cheng Qian, Zhaochen Su, Qing Zong, Shijue Huang, Bingxiang He and Yi R. Fung

Current evaluations of Large Language Model (LLM) agents primarily emphasize task completion, often overlooking resource efficiency and adaptability. This neglects a crucial capability: agents ability to devise and adjust cost-optimal plans in response to changing environments. To bridge this gap, we introduce CostBench, a scalable, cost-centric benchmark designed to evaluate agents economic reasoning and replanning abilities. Situated in the travel-planning domain, CostBench comprises tasks solvable via multiple sequences of atomic and composite tools with diverse, customizable costs. It also supports four types of dynamic blocking events, such as tool failures and cost changes, to simulate real-world unpredictability and necessitate agents to adapt in real time. Evaluating leading open-sourced and proprietary models on CostBench reveals a substantial gap in cost-aware planning: agents frequently fail to identify cost-optimal solutions in static settings, with even GPT-5 achieving less than 75% exact match rate on the hardest tasks, and performance further drops significantly under dynamic conditions. By diagnosing these weaknesses, CostBench lays the groundwork for developing future agents that are both economically rational and robust.

[Findings] Creating Grammar Teaching Material for Endangered Languages with Hybrid Grammar Induction ▶ *Underline Video*

Sebastien Christian

Explicit grammar teaching is central to endangered-language revitalization, but creating grammar lessons is labor-intensive and often falls to already overburdened teachers. We present HYGRAM, a hybrid grammar-induction method that combines typological priors, Bayesian inference, constrained LLM reasoning, and retrieval from sparse corpora and descriptive documents to generate topic-specific grammar lessons for classroom use. HYGRAM targets extremely low-resource settings and can operate from a small elicited corpus collected in roughly 10 hours of fieldwork together with any available reference materials. We evaluate the system on six typologically diverse endangered languages using expert linguist judgments of output content quality, pedagogical adequacy, and consistency across generated lessons. Results indicate that HYGRAM can produce coherent and practically useful lessons, with better quality when modest explanatory evidence is available. Feedback from Pacific language communities further suggests relevance for ongoing revitalization efforts. Overall, the work shows that evidence-constrained hybrid grammar induction can support grammar teaching and documentation where standard NLP pipelines are infeasible.

[Long] CrochetBench: Can Vision-Language Models Move from Describing to Doing in Crochet Domain? ▶ *Underline Video*

Peiyu Li, Xiaobao Huang, Ting Hua and Nitesh V Chawla

While multimodal large language models can describe visual content, their ability to generate executable procedures remains underexplored. CrochetBench presented in this paper evaluates this shift from describing to doing through fine-grained procedural reasoning in crochet: models must recognize stitches, select structurally appropriate instructions, and generate compilable procedures. We adopt the CrochetPARADE DSL as our intermediate representation, enabling structural validation and functional evaluation via execution. The benchmark covers tasks including stitch classification, instruction grounding, and both natural language and image-to-DSL translation. Across all tasks, performance sharply decreases as the evaluation shifts from surface-level similarity to executable correctness, revealing limitations in long-range symbolic reasoning and 3D-aware procedural synthesis. Our proposed CrochetBench offers a new lens for assessing procedural competence in multimodal models and highlights the gap between surface-level understanding and executable precision in real-world creative domains. Code is available at https://github.com/Peiyu-Georgia-Li/crochetBench.

[Long] Cross-Examination Framework: A Task-Agnostic Diagnostic for Information Fidelity in Text-to-Text Generation ▶ *Underline Video*

Tathagata Raha, Clement Christophe, Nada Saadi, Hamza A Javed, Marco AF Pimentel, Ronnie Rajan and Praveenkumar Kanithi

Traditional metrics like BLEU and BERTScore fail to capture semantic fidelity in generative text-to-text tasks. We adapt the Cross-Examination Framework (CEF) for a reference-free, multi-dimensional evaluation by treating the source and candidate as independent knowledge bases. CEF generates verifiable questions from each text and performs a cross-examination to derive three interpretable scores: Coverage, Conformity, and Consistency. Validated across translation, summarization and clinical note-generation, our framework identifies critical errors, such as content omissions and factual contradictions, missed by standard metrics. A key contribution is a systematic robustness analysis to select a stable judge model. Crucially, the strong correlation between our reference-free and with-reference modes validates CEF's reliability without gold references. Furthermore, human expert validation demonstrates that CEF mismatching questions align with meaning-altering semantic errors higher than with non-semantic errors, particularly excelling at identifying entity-based and relational distortions.

[Findings] DART: Mitigating Harm Drift in Difference-Aware LLMs via Distill-Audit-Repair Training ▶ *Underline Video*

Ziwen Pan, Zihan Liang, Jad Kabbara and Ali Emami

Large language models (LLMs) tuned for safety often avoid acknowledging demographic differences, even when such acknowledgment is factually correct (e.g., ancestry-based disease incidence) or contextually justified (e.g., religious hiring preferences). This identity-blindness yields incorrect responses, unnecessary refusals, or generic "equal-treatment" defaults. We study this via difference-awareness classification: given a question involving demographic groups, the task is not to answer directly, but to classify whether a correct answer requires

recognizing group differences (YES) or whether groups should be treated identically (NO). Crucially, fine-tuning for accuracy triggers harm drift: model-generated explanations become increasingly harmful as decision accuracy improves, whether by elaborating harmful content, introducing problematic assumptions, or failing to flag harms the baseline identified. To mitigate this, we introduce DART (Distill–Audit–Repair Training), which distills label-conditioned reasoning from a teacher, audits outputs for harm drift cases relative to baseline, and repairs problematic cases via severity-weighted fine-tuning. On eight benchmarks, DART improves Llama-3-8B-Instruct accuracy from 39.0% to 68.8%, with largest gains on equal-treatment prompts (11.3% to 72.6%), while reducing harm drift cases by 72.6%. It also transfers to 280 open-ended real-world queries across medical, legal, policy, and educational domains, improving difference-appropriate responses from 39.8% to 77.5% while reducing refusals from 34.3% to 3.0%. Our results demonstrate that accuracy and safety need not conflict when explicit detection and repair mechanisms are in place.

[Long] Decentralized Arena: Towards Democratic and Scalable Automatic Evaluation of Language Models ► *Underline Video*

Yanbin Yin, Kun Zhou, Zhen Wang, Xiangdong Zhang, Yifei Shao, Shibo Hao, Yi Gu, Jieyuan Liu, Somanshu Singla, Tianyang Liu, Eric P. Xing, Zhengzhong Liu, Haojian Jin and Zhiting Hu

The recent explosion of large language models (LLMs), each with its own general or specialized strengths, makes scalable, reliable benchmarking more urgent than ever. Standard practices nowadays face fundamental trade-offs: closed-ended question-based benchmarks (MMLU) struggle with saturation as newer models emerge, while crowd-sourced leaderboards (Chatbot Arena) rely on costly and slow human judges. Recently, automated methods (LLM-as-a-judge) shed light on the scalability, but risk bias by relying on one or a few "authority" models. To tackle these issues, we propose Decentralized Arena (), a fully automated framework leveraging collective intelligence from all LLMs to evaluate each other. It mitigates single-model judge bias by democratic, pairwise evaluation, and remains efficient at scale through two key components: (1) a coarse-to-fine ranking algorithm for fast incremental insertion of new models with sub-quadratic complexity, and (2) an automatic question selection strategy for the construction of new evaluation dimensions. Across extensive experiments across 66 LLMs, attains up to 97% correlation with human judgements, while significantly reducing the cost.

[Long] Diversity in Unity, Theory in Practice: Hierarchical Multitask Benchmarks for Chinese Minority Languages ► *Underline Video*

Yijie Li, Xi Cao, Yuan Sun, Quulgan Minggad, Abdulla Ablikim and JiaQingCaiWang JiaQingCaiWang

Despite the rapid advancement of LLMs, their performance on linguistically and culturally diverse minority languages within a unified national context remains underexplored. We present CMiLBench, a collection of hierarchical multitask benchmarks designed to translate theoretical notions of diversity in unity into practical evaluation for three representative Chinese minority languages: Tibetan, Mongolian, and Uyghur. CMiLBench comprises 24,663 instances across 5 difficulty levels and 17 tasks spanning foundational ability, cultural specificity, and safety alignment. We adopt existing dataset adaptation, minority knowledge construction, and high-resource benchmark translation to construct CMiLBench. We assess 14 state-of-the-art commercial and open-source LLMs with a hybrid framework that integrates automatic metrics and LLM-as-a-Judge scoring. The comparative experimental results reveal the gap between theoretical capability and practical utility. CMiLBench serves as a foundational and scalable evaluation resource to bridge the digital language divide and promote the informatization and intelligentization of low-resource Chinese minority languages.

[Short] DIXITWORLD: Evaluating Multimodal Abductive Reasoning in Vision-Language Models with Multi-Agent Dixit Gameplay ► *Underline Video*

Yunxiang MO; Tianshi Zheng; Qing Zong; Jiayu Liu; Baixuan Xu; Yauwai Yim; Chunkit Chan; Jiaxin Bai; Yangqiu Song;

Multimodal abductive reasoning , the generation and selection of explanatory hypotheses from partial observations , is a cornerstone of intelligence. Current evaluations of such ability in vision-language models (VLMs) are largely confined to static, single-agent tasks. Inspired by Dixit, we introduce DixitWorld, a comprehensive evaluation suite designed to deconstruct this challenge. DixitWorld features two core components: DixitArena, a dynamic, multi-agent environment that evaluates both hypothesis generation (a "storyteller" crafting cryptic clues) and hypothesis selection ("listeners" choosing the target image from decoys) under imperfect information; and DixitBench, a static QA benchmark that isolates the listener's task for efficient, controlled evaluation. Results from DixitArena reveal distinct, role-dependent behaviors: smaller open-source models often excel as creative storytellers, producing imaginative yet less discriminative clues, whereas larger proprietary models demonstrate superior overall performance, particularly as listeners. Performance on DixitBench strongly correlates with listener results in DixitArena, validating it as a reliable proxy for hypothesis selection. Our findings reveal a key trade-off between generative creativity and discriminative understanding in multimodal abductive reasoning, a central challenge for developing more balanced and capable vision-language agents.

[Findings] Do Emotions Influence Moral Judgment in Large Language Models? ► *Underline Video*

Mohammad Saim and Tianyu Jiang

Large language models have been extensively studied for emotion recognition and moral reasoning as distinct capabilities, yet the extent to which emotions influence moral judgment remains underexplored. In this work, we develop an emotion-induction pipeline that infuses emotion into moral situations and evaluate shifts in moral acceptability across multiple datasets and LLMs. We observe a directional pattern: positive emotions increase moral acceptability and negative emotions decrease it, with effects strong enough to reverse binary moral judgments in up to 20% of cases, and with susceptibility scaling inversely with model capability. Our analysis further reveals that specific emotions can sometimes behave contrary to what their valence would predict (e.g., remorse paradoxically increases acceptability). A

complementary human annotation study shows humans do not exhibit these systematic shifts, indicating an alignment gap in current LLMs.

[Long] Do LLMs Really Memorize Personally Identifiable Information? Revisiting PII Leakage with a Cue-Controlled Memorization Framework ▶ *Underline Video*

Xiaoyu Luo, Yiyi Chen, Qiongxiu Li and Johannes Bjerva

Large Language Models (LLMs) have been reported to leak Personally Identifiable Information (PII), with successful PII reconstruction often interpreted as evidence of memorization. We propose a principled revision of memorization evaluation for LLMs, arguing that PII leakage should be evaluated under low lexical cue conditions, where target PII cannot be reconstructed through prompt-induced generalization or pattern completion. We formalize Cue-Resistant Memorization (CRM) as a cue-controlled evaluation framework and a necessary condition for valid memorization evaluation, explicitly conditioning on prompt-target overlap cues. Using CRM, we conduct a large-scale multilingual re-evaluation of PII leakage across 32 languages and multiple memorization paradigms. Revisiting reconstruction-based settings, including verbatim prefix-suffix completion and associative reconstruction, we find that their apparent effectiveness is driven primarily by direct surface-form cues rather than by true memorization. When such cues are controlled for, reconstruction success diminishes substantially. We further examine cue-free generation and membership inference, both of which exhibit extremely low true positive rates. Overall, our results suggest that previously reported PII leakage is better explained by cue-driven behavior than by genuine memorization, highlighting the importance of cue-controlled evaluation for reliably quantifying privacy-relevant memorization in LLMs.

[Short] Does Self-Consistency Improve the Recall of Encyclopedic Knowledge? ▶ *Underline Video*

Sho Hoshino, Ukyo Honda and Peinan Zhang

While self-consistency is known to improve performance on symbolic reasoning, its effect on the recall of encyclopedic knowledge is unclear due to a lack of targeted evaluation grounds. To address this, we establish such a knowledge recall split for the popular MMLU benchmark by applying a data-driven heuristic from prior work. We validate this split by showing that the performance patterns on the symbolic reasoning and knowledge recall subsets mirror those of GSM8K and MedMCQA, respectively. Using this solid ground, we find that self-consistency consistently improves performance across both symbolic reasoning and knowledge recall, even though its underlying CoT prompting is primarily effective for symbolic reasoning. As a result, we achieve an 89% accuracy on MMLU, the best performance to date with the use of GPT-4o.

[Findings] DualFact+: A Multimodal Fact Verification Framework for Procedural Video Captioning ▶ *Underline Video*

Cennet Oguz, Yasser HAMIDULLAH, Josef van Genabith and Simon Ostermann

Evaluating factual correctness in procedural video captions is challenging because captions must reflect both the abstract procedural roles (e.g., actions, ingredients, tools, locations) and their visual execution. Existing evaluation metrics, which rely on lexical overlap or holistic semantic similarity, often miss role-specific omissions and misclassify visually present but task-irrelevant content as hallucinations. We introduce DualFact+, a role-aware, fact-level evaluation framework that distinguishes conceptual facts, encoding ontology-based role typing of procedural steps (Action, Object or Ingredient, Tool, Location), from contextual facts, encoding video-grounded predicate-argument relations that specify how these roles are instantiated during execution. To enable complete and role-consistent evaluation, DualFact+ incorporates visually grounded implicit arguments and contrastive fact sets, and operates in two complementary modes: DualFact-C for text-based verification and DualFact-V for video-grounded verification. Experiments on YouCook3-Fact and CraftBench-Fact show that state-of-the-art captioning models produce fluent but often incomplete descriptions with systematic role-level errors. DualFact+ achieves stronger correlation with human factuality judgments than standard lexical and embedding-based metrics, highlighting the importance of role-aware evaluation for procedural video understanding.

[Findings] Easy Samples Are All You Need: Self-Evolving LLMs via Data-Efficient Reinforcement Learning ▶ *Underline Video*

Zhiyin Yu, Bo Zhang, Qibin Hou, Zhonghai Wu, Xiao Luo and Lei Bai

Previous LLMs-based RL studies typically follow either supervised learning with high annotation costs, or unsupervised paradigms using voting or entropy-based rewards. However, their performance remains far from satisfactory due to the substantial annotation cost and issues such as model collapse or reward hacking. To address these issues, we introduce a new perspective inspired by cognitive learning theory and propose a novel approach called EasyRL. The core of EasyRL is to simulate the human cognitive acquisition curve by integrating reliable knowledge transfer from easy labeled data with a progressive divide-and-conquer strategy that tackles increasingly difficult unlabeled data. Specifically, we initialize a warm-up model using supervised RL with few-shot labeled data. This is followed by a divide-and-conquer pseudo-labeling strategy on difficult unlabeled data, combining consistency-based selection for low-uncertainty cases and reflection-based resolution for medium-uncertainty cases. Finally, difficulty-progressive self-training with iterative pseudo-labeling and RL further strengthens the models reasoning capability. EasyRL provides a unified self-evolving framework that facilitates data-efficient post-training of LLMs. Experimental results on mathematical and scientific benchmarks demonstrate that EasyRL, using only 10% of easy labeled data, consistently outperforms state-of-the-art baselines.

[Long] Efficient Low-Resource Language Adaptation via Multi-Source Dynamic Logit Fusion ▶ *Underline Video*

Chen Zhang, Jiuheng Lin, Zhiyuan Liao and Yansong Feng

Adapting large language models (LLMs) to low-resource languages (LRLs) is constrained by the scarcity of task data and computational resources. Although Proxy Tuning offers a logit-level strategy for introducing scaling effects, it often fails in LRL settings because the

large models weak LRL competence might overwhelm the knowledge of specialized smaller models. We thus propose TriMix, a test-time logit fusion framework that dynamically balances capabilities from three different sources: LRL competence from a continually pretrained small model, task competence from high-resource language instruction tuning, and the scaling benefits of large models. It is data- and compute-efficient, requiring no LRL task annotations, and only continual pretraining on a small model. Experiments across four model families and eight LRLs show that TriMix consistently outperforms single-model baselines and Proxy Tuning. Our analysis reveals that prioritizing the small LRL-specialized model's logits is crucial for success, challenging the prevalent large-model-dominant assumption.

[Long] EIFFEL: a novel benchmark to measure bias of English heavy training on French idiomatic expressions ▶ *Underline Video*
Charlotte Noel, Nicholas Asher, Olivier Gouvert, Farah Benamara and Julie Hunter
Mainstream multilingual LLMs are generally trained on a much higher proportion of English than multilingual data, raising questions about their ability to capture linguistic features particular to non-English languages or to capture information important to non-anglophone cultures. We add to a growing effort to increase multilingual sensitivity in LLMs by developing a benchmark, EIFFEL, testing mastery of French idiomatic expressions in context. We fully explain the methodology, which exploits input from native French speakers, to make it reproducible for other languages. We compare mainstream multilingual LLMs with French-focused LLMs both on standard LLM benchmarks and EIFFEL; EIFFEL brings out the benefits of higher proportions of French data and shows limitations of standard benchmarks for measuring multilingual competence.

[Findings] Enhancing Multilingual RAG Systems with Debiased Language Preference-Guided Query Fusion ▶ *Underline Video*
Jeonghyun Park, Byeongjeong Kim, Seojin Hwang and Hwanhee Lee
Multilingual Retrieval-Augmented Generation (mRAG) systems often exhibit a perceived preference for high-resource languages, particularly English, resulting in the widespread adoption of English pivoting. While prior studies attribute this advantage to the superior English-centric capabilities of Large Language Models (LLMs), we find that such measurements are significantly distorted by structural priors inherent in evaluation benchmarks. Specifically, we identify exposure bias and a gold availability prior, both driven by the disproportionate concentration of resources in English, as well as cultural priors rooted in topic locality, as factors that hinder accurate assessment of genuine language preference. To address these biases, we propose DeLP (Debiased Language Preference), a calibrated metric designed to explicitly factor out these structural confounds. Our analysis using DeLP reveals that the previously reported English preference is largely a byproduct of evidence distribution rather than an inherent model bias. Instead, we find that retrievers fundamentally favor monolingual alignment between the query and the document language. Building on this insight, we introduce DELTA (DEbiased Language preference-guided Text Augmentation), a lightweight and efficient mRAG framework that strategically leverages monolingual alignment to optimize cross-lingual retrieval and generation. Experimental results demonstrate that DELTA consistently outperforms English pivoting and mRAG baselines across diverse languages.

[Findings] EULoInf: Efficient Hessian-Free Entropy Based Uncertainty-Aware Data Influence Approximation ▶ *Underline Video*
Runxin Cai, Jingtan Wang and Bryan Kian Hsiang Low
In Large Language Model post-training, high-quality data effectively enhances model performance with fine-tuning, highlighting the need to identify high-quality and beneficial fine-tuning data. However, one of the most popular data valuation paradigms, influence function and its variants, are computationally expensive due to their reliance on inverse Hessian-Vector Products (iHVP) computations that scale poorly with increasing model size. To examine whether influence values correlate with efficiently computable intrinsic features, we empirically investigate the distribution of top influential data for the model in fine-tuning, and observe that data with high influence tend to be those with high predictive uncertainty. Yet such highly uncertain samples exhibit a dual nature, which can be either beneficial or detrimental noisy data. Unlike traditional methods that treat uncertainty as a standalone criterion, we introduce a directional indicator to rigorously disentangle these opposing effects. Formally, we propose EULoInf (Entropy-based Uncertainty-aware Lookahead Influence), a computationally efficient valuation framework. By approximating influence via uncertainty and gradient based validation loss lookahead, EULoInf avoids iHVP computation, effectively reducing the iHVP-induced quadratic complexity in model parameters to linear time. We rigorously derive our framework from the influence function. Empirically, it matches or even outperforms prior methods across diverse data valuation tasks and LLM architectures, including mislabel detection and data selection, while reducing computational time and memory usage by over 50%.

[Findings] EVADE: LLM-Based Explanation Generation and Validation for Error Detection in NLI ▶ *Underline Video*
Longfei Zuo, Barbara Plank and Siyao Peng
High-quality datasets are critical for training and evaluating reliable NLP models. In tasks like natural language inference (NLI), human label variation (HLV) arises when multiple labels are valid for the same instance, making it difficult to separate annotation errors from plausible variation. An earlier framework, VariErr (Weber-Genzel et al., 2024), asks multiple annotators to explain their label decisions in the first round and flags errors through validity judgments in the second round. However, conducting two rounds of manual annotation is costly and may limit the coverage of plausible labels or explanations. Our study proposes a new framework, EVADE, for generating and validating explanations to detect errors using large language models (LLMs). We perform a comprehensive analysis comparing human- and LLM-detected errors for NLI across distribution comparison, validation overlap, and impact on model fine-tuning. Our experiments demonstrate that LLM validation refines generated explanation distributions to more closely align with human annotations, and that removing LLM-detected errors from training data yields improvements in fine-tuning performance than removing errors identified by human annotators. This highlights the potential to scale error detection, reducing human effort while improving dataset quality under label variation.

[Long] Evaluating Answer Leakage Robustness of LLM Tutors against Adversarial Student Attacks ▶ *Underline Video*

Jin Zhao, Marta Kneevi and Tanja Käser

Large Language Models (LLMs) are increasingly used in education, yet their default helpfulness often conflicts with pedagogical principles. Prior work evaluates pedagogical quality via answer leakage-the disclosure of complete solutions instead of scaffolding-but typically assumes well-intentioned learners, leaving tutor robustness under student misuse largely unexplored. In this paper, we study scenarios where students behave adversarially and aim to obtain the correct answer from the tutor. We evaluate a broad set of LLM-based tutor models, including different model families, pedagogically aligned models, and a multi-agent design, under a range of adversarial student attacks. We adapt six groups of adversarial and persuasive techniques to the educational setting and use them to probe how likely a tutor is to reveal the final answer. We evaluate answer leakage robustness using different types of in-context adversarial student agents, finding that they often fail to carry out effective attacks. We therefore introduce an adversarial student agent that we fine-tune to jailbreak LLM-based tutors, which we propose as the core of a standardized benchmark for evaluating tutor robustness. Finally, we present simple but effective defense strategies that reduce answer leakage and strengthen the robustness of LLM-based tutors in adversarial scenarios.

[Findings] Evaluating Reasoning Models for Queries with Presuppositions ▶ *Underline Video*

Rose Sathyanathan, Kinshuk Vasisht and Danish Pruthi

Millions of users turn to AI models for their information needs. It is conceivable that a large number of user queries contain assumptions that may be factually inaccurate. Prior work notes that large language models (LLMs) often fail to challenge such erroneous assumptions, and can reinforce users' misinformed opinions. However, given the recent advances, especially in model's reasoning capabilities, we revisit whether large reasoning models (LRMs) can reason about the underlying assumptions and respond to user queries appropriately. We construct queries with varying degrees of presuppositions spanning health, science, and general knowledge, and use it to evaluate several widely-deployed models When compared to non-reasoning models, we find that reasoning models achieve a slightly higher accuracy (2-11%), but they still fail to challenge a large fraction (26-42%) of false presuppositions. Further, reasoning models remain susceptible to how strongly the presupposition is expressed.

[Long] Evaluating Robustness of Large Language Models Against Multilingual Typographical Errors ▶ *Underline Video*

Raoyuan Zhao, Yihong Liu, Lena Altinger, Hinrich Schuetze and Michael A. Hedderich

Large language models (LLMs) are increasingly deployed in multilingual, real-world applications with user inputs – naturally introducing typographical errors (typos). Yet most benchmarks assume clean input, leaving the robustness of LLMs to typos across languages largely underexplored. To address this gap, we introduce MulTypo, a multilingual typo generation algorithm that simulates human-like errors based on language-specific keyboard layouts and typing behavior. We evaluate 18 open-source LLMs across three model families and five downstream tasks spanning language inference, multi-choice question answering, mathematical reasoning, and machine translation tasks. Our results show that typos consistently degrade performance, particularly in generative tasks and those requiring reasoning – while the natural language inference task is comparatively more robust. Instruction tuning improves clean-input performance but may increase brittleness under noise. We also observe language-dependent robustness: high-resource languages are generally more robust than low-resource ones, and translation from English is more robust than translation into English. Our findings underscore the need for noise-aware training and multilingual robustness evaluation. We release a Python package for MulTypo and make the source code publicly available at https://github.com/cisnlp/multypo.

[Findings] Evaluating the Impact of Reviewer Guideline Design on LLM-Based Automated Peer Review ▶ *Underline Video*

Haowen Li, Yoichi Ishibashi and Masafumi Oyamada

Peer review is an essential process in scientific research, yet the growing workload has made its automation increasingly necessary. In this study, we analyze how different types of reviewer guidelines, such as official conference guidelines and reviewer-imitating ones distilled from high-quality human reviews, affect automated peer review. Our experiments show that official conference guidelines produce review results most consistent with human judgments, suggesting that evaluation criteria refined through conference practice serve as effective guidance for automated reviewing as well. In contrast, reviewer-imitating guidelines, especially those enforcing strict rubric-style scoring, consistently degraded automated review performance, highlighting the importance of allowing subjective and holistic scoring.

[Long] Evaluation Pitfalls and Challenges in Multimedia Event Extraction ▶ *Underline Video*

Philipp Seeberger, Steffen Freisinger, Tobias Bocklet and Korbinian Riedhammer

Multimedia event extraction aims to jointly identify events and their arguments across multiple modalities, such as text and images, to support more comprehensive event understanding. While recent work reports steady and substantial progress, the reliability and comparability of these results critically depend on consistent and rigorous evaluation. In this work, we present the first systematic analysis of evaluation pitfalls in multimedia event extraction and identify three major sources of issues: inconsistent data processing, inconsistent task assumptions, and overly relaxed evaluation settings. We demonstrate, through a series of controlled experiments under a strict evaluation framework, that minor evaluation choices can cause large performance variations and lead to overestimation of a models ability to ground real-world events across modalities. Our findings highlight the need for comparable evaluation standards and encourage a shift toward more rigorous evaluation in multimedia event extraction.

[Findings] ExaGPT: Example-Based Machine-Generated Text Detection for Human Interpretability ▶ *Underline Video*

Ryuto Koike, Masahiro Kaneko, Ayana Niwa, Preslav Nakov and Naoaki Okazaki
Detecting texts generated by Large Language Models (LLMs) could cause grave mistakes due to incorrect decisions, such as undermining student's academic dignity. LLM text detection thus needs to ensure the interpretability of the decision, which can help users judge how reliably correct its prediction is. When humans verify whether a text is human-written or LLM-generated, they intuitively investigate with which of them it shares more similar spans. However, existing interpretable detectors are not aligned with the human decision-making process and fail to offer evidence that users easily understand. To bridge this gap, we introduce ExaGPT, an interpretable detection approach grounded in the human decision-making process for verifying the origin of a text. ExaGPT identifies a text by checking whether it shares more similar spans with human-written vs. with LLM-generated texts from a datastore. This approach can provide similar span examples that contribute to the decision for each span in the text as evidence. Our human evaluation demonstrates that providing similar span examples contributes more effectively to judging the correctness of the decision than existing interpretable methods. Moreover, extensive experiments in four domains and three generators show that ExaGPT massively outperforms prior interpretable detectors by up to +37.0 points of accuracy at a false positive rate of 1%.

[Short] Experiments or Outcomes? Probing Scientific Feasibility in Large Language Models ▶ *Underline Video*
Seyedali Mohammadi, Manas Gaur and Francis Ferraro
Scientific feasibility assessment asks whether a claim is consistent with established knowledge and whether experimental evidence could support or refute it. We frame feasibility assessment as a diagnostic reasoning task in which, given a hypothesis, a model predicts feasible or infeasible and justifies its decision. We evaluate large language models (LLMs) under controlled knowledge conditions (hypothesis-only, with experiments, with outcomes, or both) and probe robustness by progressively removing portions of the experimental and/or outcome context. Across multiple LLMs and two datasets, providing outcome evidence is generally more reliable than providing experiment descriptions. Outcomes tend to improve accuracy beyond what internal knowledge alone provides, whereas experimental text can be brittle and may degrade performance when the context is incomplete. These findings clarify when experimental evidence benefits LLM-based feasibility assessment and when it introduces fragility.

[Long] Exploring Concreteness Through a Figurative Lens ▶ *Underline Video*
Saptarshi Ghosh and Tianyu Jiang
Static concreteness ratings are widely used in NLP, yet a words concreteness can shift with context, especially in figurative language such as metaphor, where common concrete nouns can take abstract interpretations. While such shifts are evident from context, it remains unclear how LLMs understand concreteness internally. We conduct a layer-wise and geometric analysis of LLM hidden representations across four model families, examining how models distinguish literal vs. figurative uses of the same noun and how concreteness is organized in representation space. We find that LLMs separate literal and figurative usage in early layers, and that mid-to-late layers compress concreteness into a one-dimensional direction that is consistent across models. Finally, we show this geometric structure is practically useful: a single concreteness direction supports efficient figurative-language classification and enables training-free steering of generation toward more literal or more figurative rewrites.

[Findings] Extraction of Texters' Explicit Emotion Expressions in Crisis Conversations ▶ *Underline Video*
Greg Buda, Ignacio J. Tripodi, Kelly L. Zuromski, Margaret Meagher and Elizabeth A. Olson
Understanding the emotions that individuals in crisis express is a clinically relevant goal. Here, we introduce an automated method for extracting present and past personal emotion expressions from text-based crisis conversations, enabling nuanced analyses of how these emotional profiles vary by age. We develop a three-tier emotion taxonomy and leverage both real conversation data and synthetic sentences to train a transformer-based model that captures contextual distinctions between true personal emotion expressions and other mentions. Our RoBERTa-based classifier outperforms both a regex baseline and a model trained only on real conversation data, achieving an F1 score of 0.856. Subsequent analysis of 338,924 crisis conversations shows that age is correlated with distinct patterns in emotional expressions. These findings underscore the clinical value of age-sensitive emotion analysis and constitute an initial step toward characterizing lexical variations across demographic groups.

[Long] False Friends or Cognates? A Cross-lingual Semantic Ambiguity Evaluation for Galician, Portuguese and Spanish ▶ *Underline Video*
Marta Vázquez Abuín, Jose Camacho-Collados and Marcos Garcia
The linguistic proximity between Galician, Portuguese, and Spanish results in a lexical overlap that often conceals semantic interference. This is particularly evident in false friends, posing a challenge for NLP systems. In this work, we assess whether state-of-the-art language models can identify and process false friends among these languages. We introduce six cross-lingual datasets –created manually or using semi-automatic methods, with all instances being carefully verified– covering cognates and false friends. We evaluate a broad range of encoder and decoder models of varying sizes via zero-shot and few-shot settings. Our results highlight the challenging nature of the task, but also show the clear progress made by LLMs in recent years, particularly those of a larger size, with smaller language models struggling on the task. Notably, unlike other tasks where language distance poses additional challenges, we find that linguistic proximity itself introduces errors: closely related language pairs tend to perform worse, reflecting the challenge of semantic discrimination due to lexical overlap.

[Findings] Faster MoE LLM Inference for Extremely Large Models ▶ *Underline Video*

Haoqi Yang, Luohe Shi, Qiwei Li, Zuchao Li, Ping Wang, Hao Huang and Hai Zhao

In fine-grained sparse Mixture-of-Experts (MoE) models, a large pool of specialized experts replaces a small homogeneous set, shifting performance and throughput to be governed by inference-time expert activation. Yet most existing optimization recipes implicitly assume a fixed activation budget (e.g., a constant Top-k per layer), whose behavior in fine-grained MoEs is poorly understood. We first characterize runtime skipping strategies, quantifying the accuracy–efficiency trade-off of (i) uniform fixed activation and (ii) static layer-wise Top-k allocation found by search. Our analysis reveals that static skipping can already provide substantial throughput gains, but optimal static schedules vary significantly across models and routing mechanisms. We therefore introduce Adaptive Skipping with Entropy-Penalized Thresholding (ASET), a training-free policy that adapts token-level activation using router confidence and entropy while remaining within the model's original budget. Across the fine-grained MoEs we study, static skipping policies yield 10–78% throughput gains with minimal performance degradation, including 10% improvement on DeepSeek-V3 without measurable loss. On the OLMoE testbed, ASET yields a Pareto frontier between average activation and task quality. Overall, these results identify expert skipping as a practical lever for faster fine-grained MoE inference, with adaptive activation helping when fixed budgets are too rigid.

[Findings] FIND: Toward Multimodal Financial Reasoning and Question Answering for Indic Languages ▶ *Underline Video*

Sarmistha Das;Vaibhav Vishal;Syed Ibrahim Ahmad;Sriparna Saha;Manish Gupta

Financial decision-making in multilingual settings demands accurate numerical reasoning grounded in diverse modalities, yet existing benchmarks largely overlook this high-stakes, real-world challenge, especially for Indic languages. We introduce FinVQA, a benchmark for evaluating financial numerical and multimodal reasoning in multilingual Indic contexts. FinVQA spans English, Hindi, Bengali, Marathi, Gujarati, and Tamil, and comprises 18,900 samples across 14 financial domains. The dataset captures diverse reasoning paradigms under realistic constraints, and is structured across three difficulty levels (easy, moderate, hard) and four question formats: multiple choice, fill-in-the-blank, table matching, and true/false. To address these challenges, we propose FIND, a framework that combines supervised fine-tuning with constraint-aware decoding to promote faithful numerical reasoning, robust multimodal grounding, and structured decision-making. Together, FinVQA and FIND establish a rigorous evaluation and modeling paradigm for high-stakes multilingual multimodal financial reasoning.

[Short] Frame-Semantic Knowledge Injection for Event-Level Inference in LLMs ▶ *Underline Video*

Shahid Iqbal Rai, Danilo Croce and Roberto Basili

Large language models (LLMs) are fluent but often brittle when interpretation depends on external information (e.g., events or participant roles), as next-token prediction does not explicitly encode situation-level semantic constraints. FrameNet provides a structured account of semantics through its inventory of frames, roles, and relations. We present a scalable framework that injects frame-semantic knowledge into LLMs via LoRA, moving from fact-oriented prompting to principle-oriented supervision over the full FrameNet inventory. The supervision encodes semantic constraints through semantic types, sense-aware definitions, frame relations, and role-annotated examples. To test whether this knowledge generalizes beyond surface cues, we use Natural Language Inference (NLI) as a diagnostic task for event-level reasoning. Experiments on CONFER and SNLI show consistent gains over Meta-Llama-3.1-8B-Instruct in zero-shot and few-shot settings, especially for entailment and contradiction. Complementary semantic role labeling analyses further indicate improved sensitivity to frame, role, and span structure.

[Findings] From Interpretability to Performance: Optimizing Retrieval Heads for Long-Context Language Models ▶ *Underline Video*

Youmi Ma and Naoaki Okazaki

Advances in mechanistic interpretability have identified special attention heads, known as retrieval heads, that are responsible for retrieving information from the context. However, the role of these retrieval heads in improving model performance remains unexplored. This work investigates whether retrieval heads can be leveraged to enhance the long-context capabilities of LLMs. Specifically, we propose RetMask, a method that generates training signals by contrasting normal model outputs with those from an ablated variant in which the retrieval heads are masked. This mechanism-based approach achieves substantial improvements: +2.28 points on HELMET at 128K for Llama-3.1, with +70% gains on generation with citation and +32% on passage re-ranking, while preserving performance on general tasks. Experiments across three model families demonstrate that RetMask consistently improves long-context performance, with gains correlating with the sparsity of the retrieval score distribution: models with sparser distributions, where retrieval capabilities are concentrated in a small set of heads, respond more strongly, while those with less sparse distributions show more modest gains. These results validate the functional role of retrieval heads and show that mechanistic insights can be transformed into performance enhancements.

[Findings] From Signal Degradation to Computation Collapse: Uncovering the Two Failure Modes of LLM Quantization ▶ *Underline Video*

Chenxi Zhou, Pengfei Cao, Jiang Li, Bohan Yu, Jinyu Ye, Jun Zhao and Kang Liu

Post-Training Quantization (PTQ) is critical for the efficient deployment of Large Language Models (LLMs). While 4-bit quantization is widely regarded as an optimal trade-off, reducing the precision to 2-bit usually triggers a catastrophic "performance cliff." It remains unclear whether the underlying mechanisms differ fundamentally. Consequently, we conduct a systematic mechanistic analysis, revealing two qualitatively distinct failure modes: Signal Degradation, where the computational patterns remain intact but information precision is impaired by cumulative error; and Computation Collapse, where key components fail to function, preventing correct information processing and destroying the signal in the early layers. Guided by this diagnosis, we conduct mechanism-aware interventions, demonstrating that targeted, training-free repair can mitigate Signal Degradation, but remains ineffective for Computation Collapse. Our findings provide a

systematic diagnostic framework for PTQ failures and suggest that addressing Computation Collapse requires structural reconstruction rather than mere compensation.

[Findings] GALA: Geometric Data Selection with Strategic Prospecting for Large Language Model Self-training ► *Underline Video*
XIE XIE and Zhongwei Zhongwei
Self-training has emerged as a promising direction for autonomously improving large language models (LLMs). Existing approaches typically adopt a generate-and-filter paradigm based on rejection sampling, which could suffer from inefficiency and low-quality reasoning paths. Towards this end, this paper proposes a novel framework named Geometric Data Selection with Strategic Prospecting (GALA) for LLM self-training. The core of our GALA is to identify diverse and informative samples from redundant data and exploit them more strategically. In particular, our proposed GALA first conducts clustering on latent sentence embeddings and then selects an anchor sample from each cluster based on the geometric distance to reduce data redundancy. To further exploit these samples, we conduct strategic brainstorming and reflection for high-quality reasoning trajectory prospecting. In addition, we introduce a lightweight dynamic validation module to validate the reliability of mini-batches to ensure the overall quality of the data. Extensive experiments on various benchmarks validate the effectiveness of the proposed GALA against several competing baselines.

[Long] GenomeQA: Benchmarking General Large Language Models for Genome Sequence Understanding ► *Underline Video*
Weicai Long, Yusen Hou, Junning Feng, Houcheng Su, Shuo Yang, Donglin Xie and Yanlin Zhang
Large Language Models (LLMs) are increasingly adopted as conversational assistants in genomics, where they are mainly used to reason over biological knowledge, annotations, and analysis outputs through natural language interfaces. However, existing benchmarks either focus on specialized DNA models trained for sequence prediction or evaluate biological knowledge using text-only questions, leaving the behavior of general-purpose LLMs when directly exposed to raw genome sequences underexplored. We introduce GenomeQA, a benchmark designed to provide a controlled evaluation setting for general-purpose LLMs on sequence-based genome inference tasks. GenomeQA comprises 5,200 samples drawn from multiple biological databases, with sequence lengths ranging from 6 to 1,000 base pairs (bp), spanning six task families: Enhancer and Promoter Identification, Splice Site Identification, Taxonomic Classification, Histone Mark Prediction, Transcription Factor Binding Site Prediction, and TF Motif Prediction. Across six frontier LLMs, we find that models often outperform random baselines, particularly on tasks driven by local sequence cues such as GC content and short motifs, while performance degrades on tasks that require more indirect or multi-step inference over sequence patterns. GenomeQA establishes a diagnostic benchmark for studying and improving the use of general-purpose LLMs on raw genomic sequences.

[Findings] Grouped Adaptive Weight Sharing (GAWS): An Inference-Efficient Adaptation Method for Large Language Models ► *Underline Video*
Eman Alsuradi, Junhyun Lee, Kyenghun Lee, Hyeonmok Ko and Fahed Jubair
Although Low-Rank Adaptation (LoRA) revolutionized parameter-efficient fine-tuning, it often incurs an inference overhead due to the extra computation required by adapter layers. While most literature focuses on maximizing accuracy or minimizing parameter counts, this paper prioritizes single-request inference performance in the unmerged adapter setting, where adapters must remain decoupled from the base model at runtime. By analyzing LoRA adapters on GPUs, we identify segmented function calls as the primary source of this latency. To address this, we propose Grouped Adaptive Weight Sharing (GAWS), a novel adapter design based on structured Kronecker product decomposition. Experiments on T5-3B, GPT-2 Large, LLaMA3.2-3B, and RoBERTa-Large show that GAWS reduces latency to about 40% of the gap between the unmerged LoRA and the base model, while maintaining parameter efficiency and comparable accuracy. This positions GAWS as a Pareto-efficient solution for deploying adapted LLMs in latency-sensitive settings, balancing the high latency of compressed adapters with the accuracy of LoRA. The source code is available at: https://github.com/SamsungLabs/GAWS.

[Long] How do Role Models Shape Collective Morality? Exemplar-Driven Moral Learning in Multi-Agent Simulation ► *Underline Video*
Junjie Liao, Huacong Tang, Zhou Ziheng, Yizhou Wang and Fangwei Zhong
We investigate how role models shape collective morality. To explore this, we build a multi-agent simulation powered by a Large Language Models (LLMs), where agents with diverse intrinsic drives, ranging from cooperative to competitive, interact and adapt through a four-stage cognitive loop (plan-act-observe-reflect). We design four experimental games (Alignment, Collapse, Conflict, and Construction) and conduct motivational ablation studies to identify the key drivers of imitation. The results indicate that identity-driven conformity can substantially reshape the initial dispositions. Agents tend to adapt their values to align with a perceived successful exemplar, leading to rapid value convergence.

[Long] How Much Would a Clinician Edit This Draft? Evaluating LLM Alignment for Patient Message Response Drafting ► *Underline Video*
Parker Seegmiller, Joseph Gatto, Sarah E. Greer, Ganza Belise Isingizwe, Rohan Ray, Timothy E. Burdick and Sarah Masud Preum
Large language models (LLMs) show promise in drafting responses to patient portal messages, yet their integration into clinical workflows raises various concerns, including whether they would actually save clinicians time and effort in their portal workload. We investigate LLM alignment with individual clinicians through a comprehensive evaluation of the patient message response drafting task. We develop a novel taxonomy of thematic elements in clinician responses and propose a novel evaluation framework for assessing clinician editing load of

LLM-drafted responses at both content and theme levels. We release an expert-annotated dataset and conduct large-scale evaluations of local and commercial LLMs using various adaptation techniques including thematic prompting, retrieval-augmented generation, supervised fine-tuning, and direct preference optimization. Our results reveal substantial epistemic uncertainty in aligning LLM drafts with clinician responses. While LLMs demonstrate capability in drafting certain thematic elements, they struggle with clinician-aligned generation in other themes, particularly question asking to elicit further information from patients. Theme-driven adaptation strategies yield improvements across most themes. Our findings underscore the necessity of adapting LLMs to individual clinician preferences to enable reliable and responsible use in patient-clinician communication workflows.

[Findings] How Robust Are Large Language Models for Clinical Numeracy? An Empirical Study on Numerical Reasoning Abilities in Clinical Contexts ▶ *Underline Video*

Minh-Vuong Nguyen, Fatemeh Shiri, Zhuang Li and Karin Verspoor

Large Language Models (LLMs) are increasingly being explored for clinical question answering and decision support, yet safe deployment critically requires reliable handling of patient measurements in heterogeneous clinical notes. Existing evaluations of LLMs for clinical numerical reasoning provide limited operation-level coverage, restricted primarily to arithmetic computation, and rarely assess the robustness of numerical understanding across clinical note formats. We introduce ClinicNumRobBench, a benchmark of 1,624 context-question instances with ground-truth answers that evaluates four main types of clinical numeracy: value retrieval, arithmetic computation, relational comparison, and aggregation. To stress-test robustness, ClinicNumRobBench presents longitudinal MIMIC-IV vital-sign records in three semantically equivalent representations, including a real-world note-style variant derived from the Open Patients dataset, and instantiates queries using 42 question templates. Experiments on 17 LLMs show that value retrieval is generally strong, with most models exceeding 85% accuracy, while relational comparison and aggregation remain challenging, with some models scoring below 15%. Fine-tuning on medical data can reduce numeracy relative to base models by over 30%, and performance drops under note-style variation indicate LLM sensitivity to format. ClinicNumRobBench offers a rigorous testbed for clinically reliable numerical reasoning. Code and data URL are available on https://github.com/MinhVuong2000/ClinicNumRobBench.

[Findings] Hybrid-Vector Retrieval for Visually Rich Documents: Combining Single-Vector Efficiency and Multi-Vector Accuracy ▶ *Underline Video*

Juyeon KimGeon LeeDongwon ChoiTaeuk KimKijung Shin

Retrieval over visually rich documents is essential for tasks such as legal discovery, scientific search, and enterprise knowledge management. Existing approaches fall into two paradigms: single-vector retrieval, which is efficient but coarse, and multi-vector retrieval, which is accurate but computationally expensive. To address this trade-off, we propose HEAVEN, a plug-and-play two-stage hybrid-vector framework. In the first stage, HEAVEN efficiently retrieves candidate pages using a single-vector method over Visually-Summarized Pages (VS-Pages), which assemble representative visual layouts from multiple pages. In the second stage, it reranks candidates with a multi-vector method while filtering query tokens by linguistic importance to reduce redundant computations. To evaluate retrieval systems under realistic conditions, we also introduce ViMDoc, a benchmark for visually rich, multi-document, and long-document retrieval. Across four benchmarks, HEAVEN attains 99.87% of the Recall@1 performance of multi-vector models on average while reducing per-query computation by 99.82%, achieving efficiency and accuracy. Our code and datasets are available at: https://github.com/juyeonnn/HEAVEN

[Findings] ID10M-JAM: Stress-Testing Idiom Identification Under Challenging Context ▶ *Underline Video*

Kai Golan_Hashiloni, Lior Livyatan, Ofri Hefetz, Alon Mannor, Bar Cohen and Kfir Bar

Large language models (LLMs) achieve strong performance on idiom identification benchmarks, yet their robustness to misleading contextual signals remains largely untested. We introduce ID10M-JAM, an adversarial extension of the ID10M dataset designed to jam model understanding by injecting coherent but conflicting context before each target sentence. For every sentence containing a potential idiomatic expression (PIE), we construct variants that deliberately invert contextual expectations: placing literal cues before idiomatic uses and idiomatic cues before literal ones. All variants are validated by human annotators to ensure naturalness and unambiguous interpretation for human readers. ID10M-JAM exposes systematic vulnerabilities in LLMs contextual reasoning, pushing idiom identification to its breaking point.

[Findings] Idea First, Code Later: Disentangling Problem Solving from Code Generation in Evaluating LLMs for Competitive Programming ▶ *Underline Video*

Sama Hadhoud, Alaa Elsetohy, Frederikus Hudi, Jan Christian Blaise Cruz, Steven Halim and Alham Fikri Aji

Large Language Models (LLMs) increasingly succeed on competitive programming problems, yet existing evaluations conflate algorithmic reasoning with code-level implementation. We argue that competitive programming is fundamentally a problem-solving task and propose centering natural-language editorials in both solution generation and evaluation. Generating an editorial prior to code improves solve rates for some LLMs, with substantially larger gains when using expertly written gold editorials. However, even with gold editorials, models continue to struggle with implementation, while the gap between generated and gold editorials reveals a persistent problem-solving bottleneck in specifying correct and complete algorithms. Beyond pass/fail metrics, we diagnose reasoning errors by comparing model-generated editorials to gold standards using expert annotations and validate an LLM-as-a-judge protocol for scalable evaluation. We introduce a dataset of 83 ICPC-style problems with gold editorials and full test suites, and evaluate 19 LLMs, arguing that future benchmarks should explicitly separate problem solving from implementation.

[Long] Improving Autoformalization Using Direct Dependency Retrieval ▶ *Underline Video*

Shaoqi Wang, Lu Yu, Siwei Lou, Feng Yan, Chunjie Yang, Qing Cui and Jun Zhou

Statement autoformalization, a crucial first step in formal verification, aims to transform informal descriptions of math problems into machine-verifiable formal representations but remains a significant challenge. The core difficulty lies in the fact that existing language models hallucinate formal dependencies, including missing or incorrect definitions, lemmas, and theorems. Current dependency retrieval approaches exhibit poor precision and recall, and lack the scalability to leverage ever-growing public datasets. To bridge this gap, we propose a novel retrieval-augmented framework based on Direct Dependency Retrieval (DDR). DDR directly generates candidate formal dependencies from natural-language mathematical descriptions and verifies their existence in the formal library via an efficient Suffix Array Check (SAC). Built on a SAC-constructed dependency retrieval dataset of over 500,000 samples, a high-precision DDR model is fine-tuned and shown to significantly outperform state-of-the-art methods in both retrieval precision and recall, leading to superior advantage in the autoformalization tasks. SAC also contributes in assessing formalization difficulty and enabling explicit quantification of the hallucination in In-Context Learning (ICL).

[Findings] InferPilot: Autonomous Inference Attacks Against ML Services With LLM-Based Agents ▶ *Underline Video*

Yixin Wu, Rui Wen, Chi Cui, Michael Backes and Yang Zhang

Inference attacks have been widely studied and offer a systematic risk assessment of ML services; however, their implementation and the attack parameters for optimal estimation are challenging for non-experts. The emergence of advanced large language models presents a promising yet largely unexplored opportunity to develop autonomous agents as inference attack experts, helping address this challenge. In this paper, we propose InferPilot, an autonomous agent capable of independently conducting inference attacks without human intervention. We evaluate it on 20 target services. The evaluation shows that our agent, using GPT-4o, achieves a 100.0% task completion rate and near-expert attack performance, with an average token cost of only $0.627 per run. The agent can also be powered by many other representative LLMs and can adaptively optimize its strategy under service constraints. We further perform trace analysis, demonstrating that design choices, such as a multi-agent framework and task-specific action spaces, effectively mitigate errors such as bad plans, inability to follow instructions, task context loss, and hallucinations. We anticipate that such agents could empower non-expert ML service providers, auditors, or regulators to systematically assess the risks of ML services without requiring deep domain expertise.

[Findings] Information Representation Fairness in Long-Document Embeddings: The Peculiar Interaction of Positional and Language Bias ▶ *Underline Video*

Elias Schuhmacher, Andrianos Michail, Juri Opitz, Rico Sennrich and Simon Clematide

To be discoverable in an embedding-based search process, each part of a document should be reflected in its embedding representation. To quantify any potential reflection biases, we introduce a permutation-based evaluation framework. With this, we observe that state-of-the-art embedding models exhibit systematic positional and language biases when documents are longer and consist of multiple segments. Specifically, early segments and segments in higher-resource languages like English are over-represented, while later segments and segments in lower-resource languages are marginalized. In our further analysis, we find that the positional bias stems from front-loaded attention distributions in pooling-token embeddings, where early tokens receive more attention. To mitigate this issue, we introduce an inference-time attention calibration method that redistributes attention more evenly across document positions, increasing discoverabiltiy of later segments. Our evaluation framework and attention calibration is available at https://github.com/impresso/fair-sentence-transformers

[Long] Instant Personalized Large Language Model Adaptation via Hypernetwork ▶ *Underline Video*

Zhaoxuan Tan, Zixuan Zhang, Haoyang Wen, Zheng Li, Rongzhi Zhang, Pei Chen, Fengran Mo, Zheyuan Liu, Qingkai Zeng, Qingyu Yin and Meng Jiang

Personalized large language models (LLMs) tailor content to individual preferences using user profiles or histories. However, existing parameter-efficient fine-tuning (PEFT) methods, such as the "One-PEFT-Per-User" (OPPU) paradigm, require training a separate adapter for each user, making them computationally expensive and impractical for real-time updates. We introduce Profile-to-PEFT, a scalable framework that employs a hypernetwork, trained end-to-end, to map a user's encoded profile directly to a full set of adapter parameters (e.g., LoRA), eliminating per-user training at deployment. This design enables instant adaptation, generalization to unseen users, and privacy-preserving local deployment. Experimental results demonstrate that our method outperforms both prompt-based personalization and OPPU while using substantially fewer computational resources at deployment. The framework exhibits strong generalization to out-of-distribution users and maintains robustness across varying user activity levels and different embedding backbones. The proposed Profile-to-PEFT framework enables efficient, scalable, and adaptive LLM personalization suitable for large-scale applications.

[Long] Is this chart lying to me? Automating the detection of misleading visualizations ▶ *Underline Video*

Jonathan Tonglet, Jan Zimny, Tinne Tuytelaars and Iryna Gurevych

Misleading visualizations are a potent driver of misinformation on social media and the web. By violating chart design principles, they distort data and lead readers to draw inaccurate conclusions. Prior work has shown that both humans and multimodal large language models (MLLMs) are frequently deceived by such visualizations. Automatically detecting misleading visualizations and identifying the specific design rules they violate could help protect readers and reduce the spread of misinformation. However, the training and evaluation of AI models has been limited by the absence of large, diverse, and openly available datasets. In this work, we introduce Misviz, a benchmark of 2,604 real-world visualizations annotated with 12 types of misleaders. To support model training, we also release Misviz-synth, a synthetic

dataset of 81,814 visualizations generated using Matplotlib and based on real-world data tables. We perform a comprehensive evaluation on both datasets using state-of-the-art MLLMs, rule-based systems, and fine-tuned classifiers. Our results reveal that the task remains highly challenging. We release Misviz, Misviz-synth, and the accompanying code.

[Findings] JudgeMeNot: Personalizing Large Language Models to Emulate Judicial Reasoning in Hebrew ▶ *Underline Video*

Itay Razumenko, Arnon Sturm and Nir Grinberg

Despite significant advances in large language models, personalizing them for individual decision-makers remains an open problem. Here, we introduce a synthetic-organic supervision pipeline that transforms raw judicial decisions into instruction-tuning data, enabling parameter-efficient fine-tuning of personalized models for individual judges in low-resource settings. We compare our approach to state-of-the-art personalization techniques across three different tasks and settings. The results show that Causal Language Modeling followed by synthetically generated instruction-tuning significantly outperforms all other baselines, providing significant improvements across lexical, stylistic, and semantic similarity. Notably, our model-generated outputs are indistinguishable from the reasoning of human judges, highlighting the viability of efficient personalization, even in low-resource settings.

[Long] JurisBench: A Deep Benchmark for Assessing Large Language Models in Professional Legal Practice ▶ *Underline Video*

Ziang Chen, Guannan Li, Fanlin Ji, Yipeng Kang, Jiaqi Li, Muhan Zhang, Yangtao Zhang, Li Tianjiao, Jiannan Wang, Xin Guo, Song-Chun Zhu and BIN LING

Large Language Models (LLMs) have demonstrated strong cross-domain capabilities, yet their competence in specialized professional tasks remains underexamined. Existing legal benchmarks evaluate isolated tasks or exam-style questions, failing to capture the procedural interdependencies and adjudicative rigor inherent in professional practice. To bridge this gap, we construct JurisBench, a vertical, depth-oriented, domain-specific benchmark designed to evaluate LLMs across key stages of Chinese civil litigation. JurisBench introduces a Linear Depth Simulation track that mirrors the cognitive workflow of professional judges through four sequential, dependency-aware phases: Cause of Action prediction, Focus of Disputes identification, Rationale of the Judgment generation, and Result of the Judgment determination. Results reveal an illusion of competence: state-of-the-art models exhibit marked performance degradation in end-to-end pipelines due to cascading error propagation. We identify precise statutory grounding as a persistent bottleneck, highlighting a critical gap between fluent linguistic output and judicial reliability. JurisBench shifts evaluation from isolated legal knowledge to workflow-level task execution, providing a diagnostic framework for legal AI and a template for benchmark design in specialized domains.

[Long] Just Pass Twice: Efficient Token Classification with LLMs for Zero-Shot NER ▶ *Underline Video*

Ahmed Ewais, Ahmed Hashish and Amr Ali

Large language models encode extensive world knowledge valuable for zero-shot named entity recognition. However, their causal attention mechanism, where tokens attend only to preceding context, prevents effective token classification when disambiguation requires future context. Existing approaches use LLMs generatively, prompting them to list entities or produce structured outputs, but suffer from slow autoregressive decoding, hallucinated entities, and formatting errors. We propose Just Pass Twice (JPT), a simple yet effective method that enables causal LLMs to perform discriminative token classification with full bidirectional context. Our key insight is that concatenating the input to itself lets each token in the second pass attend to the complete sentence, requiring no architectural modifications. We combine these representations with definition-guided entity embeddings for flexible zero-shot generalization. Our approach achieves state-of-the-art results on zero-shot NER benchmarks, surpassing the previous best method by +7.9 F1 on average across CrossNER and MIT benchmarks, being over 20x faster than comparable generative methods.

[Findings] Just Use XML: Revisiting Joint Translation and Label Projection ▶ *Underline Video*

Thennal D K, Chris Biemann and Hans Ole Hatzel

Label projection is an effective technique for cross-lingual transfer, extending span-annotated datasets from a high-resource language to low-resource ones. Most approaches perform label projection as a separate step after machine translation, and prior work that combines the two reports degraded translation quality. We re-evaluate this claim with LabelPigeon, a novel framework that jointly performs translation and label projection via XML tags. We design a direct evaluation scheme for label projection, and find that LabelPigeon outperforms baselines and actively improves translation quality in 11 languages. We further assess translation quality across 203 languages and varying annotation complexity, finding consistent improvement attributed to additional fine-tuning. Finally, across 27 languages and three downstream tasks, we report substantial gains in cross-lingual transfer over comparable work, up to +40.2 F1 on NER. Overall, our results demonstrate that XML-tagged label projection provides effective and efficient label transfer without compromising translation quality.

[Findings] Kumatigi: Quality-Driven Data Augmentation for Low-Resource Machine Translation ▶ *Underline Video*

Cheick Tidiani Cissé

Neural machine translation for extremely low-resource languages faces compounding challenges: scarce parallel data, orthographic inconsistency, and absence of quality metadata for principled training. We present Kumatigi, a quality-annotated French-Bambara corpus combining systematic curation with data augmentation strategies tailored to Bambara. We provide 67k quality-scored pairs that enable targeted data filtering and address pervasive orthographic normalization issues in existing resources. Our dual-dataset generation framework strategically exploits round-trip translation, producing synthetic pairs for fluency reinforcement alongside back-translated pairs that preserve authentic vocabulary for coverage expansion. We further introduce linguistically-motivated augmentation techniques addressing Bambara's

orthographic variability, improving model robustness for real-world text. Experiments with LoRA-based fine-tuning demonstrate consistent improvements across automatic metrics, with our full system achieving up to +3-4 BLEU over strong baselines. Data generation and augmentation strategies contribute +1-2 BLEU beyond high-quality parallel data alone. Human evaluation by native speakers confirms these automatic improvements align with substantial gains in translation adequacy and fluency, with our best model approaching human reference translation quality. Our methodology provides a reproducible framework applicable to other under-resourced languages facing similar data challenges.

[Long] LAD-RAG: Layout-aware Dynamic RAG for Visually-Rich Document Understanding ▶ *Underline Video*

Zhivar Sourati, Zheng Wang, Marianne Menglin Liu, Yazhe Hu, Mengqing Guo, Sujeeth Bharadwaj, Kyu J. Han, Tao Sheng, Sujith Ravi, Morteza Dehghani and Dan Roth

Question answering over visually rich documents (VRDs) requires reasoning not only over isolated content but also over documents' structural organization and cross-page dependencies. However, conventional retrieval-augmented generation (RAG) methods encode content in isolated chunks during ingestion, losing structural and cross-page dependencies, and retrieve a fixed number of pages at inference, regardless of the specific demands of the question or context. This often results in incomplete evidence retrieval and degraded answer quality for multi-page reasoning tasks. To address these limitations, we propose LAD-RAG, a novel Layout-Aware Dynamic RAG framework. During ingestion, LAD-RAG constructs a symbolic document graph that captures layout structure and cross-page dependencies, adding it alongside standard neural embeddings to yield a more holistic representation of the document. During inference, an LLM agent dynamically interacts with the neural and symbolic indices to adaptively retrieve the necessary evidence based on the query. Experiments on MMLongBench-Doc, LongDocURL, DUDE, and MP-DocVQA demonstrate that LAD-RAG improves retrieval, achieving over 90% perfect recall on average without any topk tuning, and outperforming baseline retrievers by up to 20% in recall at comparable noise levels, yielding higher QA accuracy with minimal latency.

[Findings] Large Language Models Are Effective Human Annotation Assistants, But Not Good Independent Annotators ▶ *Underline Video*

Feng Gu, Zongxia Li, Carlos R. Colon, Benjamin Evans, Ishani Mondal and Jordan Lee Boyd-Graber

Event annotation is important for identifying, monitoring, and understanding sociological trends. Although expert annotators set the gold standard, they are expensive and inefficient. While state-of-the-art NLP models are an attractive alternative, they are often evaluated on standalone subtasks rather than entire workflows. Thus, we evaluate a holistic workflow that summarizes news with event coreference resolution and argument extraction in three modes: AI-only, AI assistance, and human only. Although AI's recall is seven times higher than the tf-idf baseline at coreference resolution, it is far from replacing experts. However, experts adopt AI-extracted arguments 60% of the time, reducing extraction time by 25%. Our code and data are in https://github.com/Obertura777/gtd-data.

[Short] Learning from Emptiness: De-biasing Listwise Rerankers with Content-Agnostic Probability Calibration ▶ *Underline Video*

Hang Lv, Hongchao Gu, Ruiqing Yang, Liangyue Li, Zulong Chen, Defu Lian, Hao Wang and Enhong Chen

Generative listwise reranking leverages global context for superior retrieval but is plagued by intrinsic position bias, where models exhibit structural sensitivity to input order independent of relevance. Existing mitigations present a dilemma: inference-time aggregation incurs prohibitive latency, while training-based methods often fail to eradicate ingrained priors, particularly in compact models. To resolve this dilemma, we propose CapCal (Content-Agnostic Probability Calibration), a training-free framework that mechanically decouples positional bias from ranking decisions. By estimating the bias distribution via content-free placeholders, CapCal rectifies output logits through an entropy-adaptive contrastive mechanism. Evaluations across 10 benchmarks confirm that CapCal achieves superior performance among training-free methods while preserving single-pass efficiency. Notably, it unlocks the latent potential of lightweight models (e.g., 0.6B), delivering absolute NDCG gains exceeding 10 points and outperforming computationally expensive data augmentation strategies.

[Findings] Learning How to Remember: A Meta-Cognitive Management Method for Structured and Transferable Agent Memory ▶ *Underline Video*

Sirui Liang, Pengfei Cao, Jian Zhao, Wenhao Teng, Xiangwen Liao, Jun Zhao and Kang Liu

Large language model (LLM) agents increasingly rely on accumulated memory to solve long-horizon decision-making tasks. However, most existing approaches store memory in fixed representations and reuse it at a single or implicit level of abstraction, which limits generalization and often leads to negative transfer when distribution shift. This paper proposes the Meta-Cognitive Memory Abstraction method (MCMA), which treats memory abstraction as a learnable cognitive skill rather than a fixed design choice. MCMA decouples task execution from memory management by combining a frozen task model with a learned memory copilot. The memory copilot is trained using direct preference optimization; it determines how experience should be structured, abstracted, and reused. Memories are further organized into a hierarchy of abstraction levels, enabling selective reuse based on task similarity. When no memory is transferable, MCMA transfers the ability to abstract and manage memory by transferring the memory copilot. Experiments on ALFWorld, ScienceWorld, and BabyAI demonstrate substantial improvements in performance, out-of-distribution generalization, and cross-task transfer over several baselines.

[Findings] Learning to Control Summaries with Score Ranking ▶ *Underline Video*

Hongye Liu, Liang Ding and Ricardo Henao

Recent advances in summarization research focus on improving summary quality across multiple criteria, such as completeness, conciseness,

and faithfulness, by jointly optimizing these dimensions. However, these efforts largely overlook the challenge of controlling summary generation with respect to individual criteria, especially in the presence of their inherent trade-offs. For example, enhancing conciseness can compromise completeness, and vice versa. In this work, we address this gap by proposing a loss function that aligns model outputs with fine-grained, model-based evaluation scores (e.g., from FineSurE), enabling both improvement in summary quality and dimension-specific control. Our approach improves the overall quality of summaries while maintaining the ability to selectively prioritize one criterion over others. Experiments on three pretrained models (LLaMA, Qwen, and Mistral) demonstrate that our method achieves performance comparable to state-of-the-art summarizers, while uniquely offering strong controllability over individual quality dimensions.

[Long] Leave My Images Alone: Preventing Multi-Modal Large Language Models from Analyzing Images via Visual Prompt Injection ▶ *Underline Video*

Zedian Shao, Hongbin Liu, Yuepeng Hu and Neil Zhenqiang Gong

Multi-modal large language models (MLLMs) have emerged as powerful tools for analyzing Internet-scale image data, offering significant benefits but also raising critical safety and societal concerns. In particular, these models may be misused to extract sensitive information from personal images, such as identifying individuals or revealing locations. In this work, we propose ImageProtector, a method designed to protect images from unauthorized analysis by MLLMs. Before an image is shared online, ImageProtector embeds a carefully crafted, nearly imperceptible perturbation that acts as a visual prompt injection attack on MLLMs. Consequently, when a malicious actor downloads and queries a protected image, the MLLM is consistently misled into generating a refusal response such as "Im sorry, I cant help with that request." We empirically demonstrate the effectiveness of ImageProtector across six MLLMs and four datasets. Additionally, we evaluate three potential countermeasures, Gaussian noise, DiffPure, and adversarial training, and show that while they partially mitigate the impact of ImageProtector, they simultaneously degrade model accuracy and/or efficiency.

[Long] Limited Linguistic Diversity in Embodied AI Datasets ▶ *Underline Video*

Selma Wanna, Agnes Luhtaru, Jonathan Salfity, Ryan Barron, Juston Moore, Cynthia Matuszek and Mitch Pryor

Language plays a critical role in Vision-Language-Action (VLA) models, yet the linguistic characteristics of the datasets used to train and evaluate these systems remain poorly documented. In this work, we present a systematic dataset audit of several widely used VLA corpora, aiming to characterize what kinds of instructions these datasets actually contain and how much linguistic variety they provide. We quantify instruction language along complementary dimensions, including lexical variety, duplication and overlap, semantic similarity, and syntactic complexity. Our analysis shows that many datasets rely on highly repetitive, template-like commands with limited structural variation, yielding a narrow distribution of instruction forms. We position these findings as descriptive documentation of the language signal available in current VLA training and evaluation data, intended to support more detailed dataset reporting, more principled dataset selection, and targeted curation or augmentation strategies that broaden language coverage.

[Long] Linear-Time and Constant-Memory Text Embeddings Based on Recurrent Language Models ▶ *Underline Video*

Tobias Grantner, Emanuel Sallinger and Martin Flechl

Transformer-based embedding models suffer from quadratic computational and linear memory complexity, limiting their utility for long sequences. We propose recurrent architectures as an efficient alternative, introducing a vertically chunked inference strategy that enables fast embedding generation with memory usage that becomes constant in the input length once it exceeds the vertical chunk size. By fine-tuning Mamba2 models, we demonstrate their viability as general-purpose text embedders, achieving competitive performance across a range of benchmarks while maintaining a substantially smaller memory footprint compared to transformer-based counterparts. We empirically validate the applicability of our inference strategy to Mamba2, RWKV, and xLSTM models, confirming consistent runtime-memory trade-offs across architectures and establishing recurrent models as a compelling alternative to transformers for efficient embedding generation.

[Long] LLMEval-Fair: A Large-Scale Longitudinal Study on Robust and Fair Evaluation of Large Language Models ▶ *Underline Video*

Ming Zhang, Yujiong Shen, Jingyi Deng, Yuhui Wang, Huayu Sha, Kexin Tan, Qiyuan Peng, Yue Zhang, Junzhe Wang, Shichun Liu, Yueyuan Huang, Jingqi Tong, Changhao Jiang, Yilong Wu, Zhihao Zhang, Mingqi Wu, Mingxu Chai, Zhiheng Xi, Shihan Dou, Tao Gui, Qi Zhang and Xuanjing Huang

Existing evaluation of Large Language Models (LLMs) on static benchmarks is vulnerable to data contamination and leaderboard overfitting, critical issues that obscure true model capabilities. To address this, we introduce LLMEval-Fair, a framework for dynamic evaluation of LLMs. LLMEval-Fair is built on a proprietary bank of 220k graduate-level questions, from which it dynamically samples unseen test sets for each evaluation run. Its automated pipeline ensures integrity via contamination-resistant data curation, a novel anti-cheating architecture, and a calibrated LLM-as-a-judge process achieving 90% agreement with human experts, complemented by a relative ranking system for fair comparison. An 30-month longitudinal study of nearly 60 leading models reveals a performance ceiling on knowledge memorization and exposes data contamination vulnerabilities undetectable by static benchmarks. The framework demonstrates exceptional robustness in ranking stability and consistency, providing strong empirical validation for the dynamic evaluation paradigm. LLMEval-Fair offers a robust and credible methodology for assessing the true capabilities of LLMs beyond leaderboard scores, promoting the development of more trustworthy evaluation standards.

[Findings] LLMs as Lab Engineers: A Benchmark for Analytical Method Lifecycle Management ▶ *Underline Video*

Xiaoyi Chen, Mahsa Monshizadeh, Chaoqi Zhang, Jianjun Lang, Yang Wu, Genevieve Mortensen, Xiaozhong Liu and Haixu Tang
We introduce ChemBench, a comprehensive benchmark for evaluating LLMs' capabilities in analytical chemistry scenarios. Unlike existing benchmarks focused on factual knowledge, ChemBench assesses model abilities to provide contextualized, practical guidance for complex analytical chemistry challenges, including instrument readiness checks, system suitability testing, method development, and troubleshooting for both liquid chromatography coupled mass spectrometry (LC-MS) and Gas Chromatography-Mass Spectrometry (GC-MS) platforms. We evaluate three enhancement approaches: chemistry-specialized models, human-guided Chain-of-Thought reasoning, and Retrieval-Augmented Generation (RAG). Our findings reveal that general-purpose commercial models often outperform domain-specialized ones, while RAG and reasoning significantly improve performance. The six-dimension evaluation framework (specificity, correctness, usefulness, feasibility, misinformation risk, and error handling) provides valuable insights into LLMs' real-world utility for chemistry researchers, establishing a foundation for developing more effective AI assistants for scientific research.

[Findings] Localized Cultural Knowledge is Conserved and Controllable in Large Language Models ▶ *Underline Video*
Veniamin Veselovsky, Berke Argn, Benedikt Stroebl, Chris Wendler, Robert West, James Evans, Thomas L. Griffiths and Arvind Narayanan
Large language models (LLMs), like human language learners, show patterns influenced by their dominant training language. Just as humans display language patterns influenced by their native tongue (semantic accents) when learning new languages, LLMs often default to English-centric responses even when generating in other languages. However, we observe that explicitly providing cultural context in prompts significantly improves the models' ability to generate culturally localized responses. We term this phenomenon the explicit-implicit localization gap, indicating that while cultural knowledge exists within LLMs, it may not naturally surface in multilingual interaction without explicitly including cultural context. In this paper, we (1) quantify this gap in multiple LLMs using a new cultural localization benchmark and find large (>10%) gaps in the majority of investigated models. (2) Demonstrate a fundamental trade-off between localization accuracy and output diversity. (3) Through mechanistic interpretability, we identify the underlying localization mechanisms within LLMs and show that these mechanisms are both language and task agnostic, with individual steering vectors effectively generalizing across different languages and culturally-relevant tasks.

[TACL] Localizing Factual Inconsistencies in Attributable Text Generation ▶ *Underline Video*
Arie Cattan, Paul Roit, Shiyue Zhang, David Wan, Roee Aharoni, Idan Szpektor, Mohit Bansal and Ido Dagan
There has been an increasing interest in detecting hallucinations in model-generated texts, both manually and automatically, at varying levels of granularity. However, most existing methods fail to precisely pinpoint the errors. In this work, we introduce QASemConsistency a new formalism for localizing factual inconsistencies in attributable text generation, at a fine-grained level. Drawing inspiration from Neo-Davidsonian formal semantics, we propose decomposing the generated text into minimal predicate-argument level propositions, expressed as simple question-answer (QA) pairs, and assess whether each individual QA pair is supported by a trusted reference text. As each QA pair corresponds to a single semantic relation between a predicate and an argument, QASemConsistency effectively localizes the unsupported information. We first demonstrate the effectiveness of the QASemConsistency methodology for human annotation, by collecting crowdsourced annotations of granular consistency errors, while achieving a substantial inter-annotator agreement. This benchmark includes more than 3K instances spanning various tasks of attributable text generation. We also implement several methods for automatically detecting localized factual inconsistencies, with both supervised entailment models and LLMs. Code, datasets and models will be made publicly available upon publication.

[Findings] Locate, Steer, and Improve: A Practical Survey of Actionable Mechanistic Interpretability in Large Language Models ▶ *Underline Video*
Hengyuan Zhang, Zhihao Zhang, Ercong Nie, Mingyang Wang, Zunhai Su, Yiwei Wang, Qianli Wang, Shuzhou Yuan, Xufeng Duan, Qibo Xue, Zeping Yu, Chenming Shang, Xiao Liang, Jing Xiong, Hui Shen, Chaofan Tao, Zhengwu Liu, Senjie Jin, Zhiheng Xi, Dongdong Zhang, Sophia Ananiadou, Tao Gui, Ruobing Xie, Hayden Kwok-Hay So, Hinrich Schütze, Xuanjing Huang, Qi Zhang and Ngai Wong
Mechanistic Interpretability (MI) has emerged as a vital approach to demystify the opaque decision-making of Large Language Models (LLMs). However, existing reviews primarily treat MI as an observational science, summarizing analytical insights while lacking a systematic framework for actionable intervention. To bridge this gap, we present a practical survey structured around the pipeline: "Locate, Steer, and Improve." We formally categorize Localizing (diagnosis) and Steering (intervention) methods based on specific Interpretable Objects to establish a rigorous intervention protocol. Furthermore, we demonstrate how this framework enables tangible improvements in Alignment, Capability, and Efficiency, effectively operationalizing MI as a practical engineering toolkit for model optimization. The curated paper list of this work is available at https://anonymous.4open.science/r/Act-MI-F068.

[Long] Locomo-Plus: Beyond-Factual Cognitive Memory Evaluation Framework for LLM Agents ▶ *Underline Video*
Yifei Li, Weidong Guo, Lingling Zhang, Rongman Xu, Muye Huang, Hui Liu, Lijiao Xu, Yu Xu and Jun Liu
Long-term conversational memory is a core capability for LLM-based dialogue systems, yet existing benchmarks and evaluation protocols primarily focus on surface-level factual recall. In realistic interactions, appropriate responses often depend on implicit constraints such as user state, goals, or values that are not explicitly queried later. To evaluate this setting, we introduce LoCoMo-Plus, a benchmark for assessing cognitive memory under cue–trigger semantic disconnect, where models must retain and apply latent constraints across long conversational contexts. We further show that conventional string-matching metrics and explicit task-type prompting are misaligned with such scenarios, and propose a unified evaluation framework based on constraint consistency. Experiments across diverse backbone models,

retrieval-based methods, and memory systems demonstrate that cognitive memory remains challenging and reveals failures not captured by existing benchmarks. Our code and evaluation framework are publicly available at https://github.com/xjtuleeyf/Locomo-Plus.

[Long] LogosKG: Hardware-Optimized Scalable and Interpretable Knowledge Graph Retrieval ▶ *Underline Video*

He Cheng, Yifu Wu, Saksham Khatwani, Maya Kruse, Dmitriy Dligach, Timothy A. Miller, Majid Afshar and Yanjun Gao

Knowledge graphs (KGs) are increasingly integrated with large language models (LLMs) to provide structured, verifiable reasoning. A core operation in this integration is multi-hop retrieval, yet existing systems struggle to balance efficiency, scalability, and interpretability. We introduce LogosKG, a novel, hardware-aligned framework that enables scalable and interpretable k-hop retrieval on large KGs by building on symbolic KG formulations and executing traversal as hardware-efficient operations over decomposed subject, object, and relation representations. To scale to billion-edge graphs, LogosKG integrates degree-aware partitioning, cross-graph routing, and on-demand caching. Experiments show substantial efficiency gains over CPU and GPU baselines without loss of retrieval fidelity. With proven performance in KG retrieval, a downstream two-round KG-LLM interaction demonstrates how LogosKG enables large-scale, evidence-grounded analysis of how KG topology, such as hop distribution and connectivity, shapes the alignment between structured biomedical knowledge and LLM diagnostic reasoning, thereby opening the door for next-generation KG-LLM integration. The source code is publicly available at https://github.com/LARK-NLP-Lab/LogosKG, and an online demo is available at https://lark-nlp-lab-logoskg.hf.space/.

[Long] Lost in the Mix: Evaluating LLM Understanding of Code-Switched Text ▶ *Underline Video*

Amr Mohamed, Yang Zhang, Michalis Vazirgiannis and Guokan Shang

Code-switching (CSW) is the act of alternating between two or more languages within a single discourse. This phenomenon is widespread in multilingual communities and increasingly prevalent online, exposing large language models (LLMs) to mixed-language inputs. We present a systematic evaluation of LLM comprehension under code-switching by generating linguistically grounded CSW variants of established benchmarks (Belebele, MMLU, XNLI) across five typologically diverse languages. Our contributions are: (i) a controlled pipeline for producing CSW test sets that respect linguistic constraints on code-switching; (ii) a multi-model, multi-language analysis showing that inserting non-English tokens into English consistently reduces accuracy on comprehension and reasoning benchmarks, whereas embedding English into non-English contexts often improves it; and (iii) a mitigation study contrasting in-context learning (ICL) with fine-tuning. Across model families, ICL cues yield inconsistent, and sometimes negative, effects, while fine-tuning on CSW data provides modest but reliable gains, partially recovering accuracy under CSW.

[Findings] Lost in Translation: Do LVLM Judges Generalize Across Languages? ▶ *Underline Video*

Md Tahmid Rahman Laskar, Mohammed Saidul Islam, Mir Tafseer Nayeem, Amran Bhuiyan, Mizanur Rahman, Shafiq Joty, Enamul Hoque and Jimmy Huang

Automatic evaluators such as reward models play a central role in the alignment and evaluation of large vision-language models (LVLMs). Despite their growing importance, these evaluators are almost exclusively assessed on English-centric benchmarks, leaving open the question of how well these evaluators generalize across languages. To answer this question, we introduce MM-JudgeBench, the first large-scale benchmark for multilingual and multimodal judge model evaluation, which includes over 60K pairwise preference instances spanning 25 typologically diverse languages. MM-JudgeBench integrates two complementary subsets: a general vision-language preference evaluation subset extending VL-RewardBench, and a chart-centric visual-text reasoning subset derived from OpenCQA, enabling systematic analysis of reward models (i.e., LVLM judges) across diverse settings. We additionally release a multilingual training set derived from MM-RewardBench, disjoint from our evaluation data, to support domain adaptation. By evaluating 22 LVLMs (15 open-source, 7 proprietary), we uncover substantial cross-lingual performance variance in our proposed benchmark. Our analysis further shows that model size and architecture are poor predictors of multilingual robustness, and that even state-of-the-art LVLM judges exhibit inconsistent behavior across languages. Together, these findings expose fundamental limitations of current reward modeling and underscore the necessity of multilingual, multimodal benchmarks for developing reliable automated evaluators.

[Long] MADE: A Living Benchmark for Multi-Label Text Classification with Uncertainty Quantification of Medical Device Adverse Events ▶ *Underline Video*

Raunak Agarwal, Markus Wenzel, Simon Baur, Jonas Zimmer, George Harvey and Jackie Ma

Machine learning in high-stakes domains such as healthcare requires not only strong predictive performance but also reliable uncertainty quantification (UQ) to support human oversight. Multi-label text classification (MLTC) is a central task in this domain, yet remains challenging due to label imbalances, dependencies, and combinatorial complexity. Existing MLTC benchmarks are increasingly saturated and may be affected by training data contamination, making it difficult to distinguish genuine reasoning capabilities from memorization. We introduce MADE, a living MLTC benchmark derived from medical device adverse event reports and continuously updated with newly published reports to prevent contamination. MADE features a long-tailed distribution of hierarchical labels and enables reproducible evaluation with strict temporal splits. We establish baselines across more than 20 encoder- and decoder-only models under fine-tuning and few-shot settings (instruction-tuned/reasoning variants, local/API-accessible). We systematically assess entropy-/consistency-based and self-verbalized UQ methods. Results show clear trade-offs: smaller discriminatively fine-tuned decoders achieve the strongest head-to-tail accuracy while maintaining competitive UQ; generative fine-tuning delivers the most reliable UQ; large reasoning models improve performance on rare labels yet exhibit surprisingly weak UQ; and self-verbalized confidence is not a reliable proxy for uncertainty. Our work is publicly available at https://hhi.fraunhofer.de/aml-demonstrator/made-benchmark.

[Findings] MAPLE: Multi-Aspect Panels of LLM Evaluators for Open-Ended Questions ► *Underline Video*

Michinori Jinji, Kyohei Atarashi, Koh Takeuchi and Hisashi Kashima

LLM-as-a-Judge, which uses LLMs to evaluate responses to open-ended questions, has seen significant growth in recent years. It has been adopted as a scalable alternative to manual human evaluation, such as crowdsourcing, which is often time-consuming and costly. However, the discrepancy between LLM-generated evaluations and human evaluations remains a critical problem in this field. To bridge this gap, we propose Multi-Aspect Panels of LLM Evaluators (MAPLE), a framework that orchestrates evaluations across multiple criteria using multiple LLMs. MAPLE integrates criterion-wise pairwise evaluations from multiple LLMs by estimating the importance of criteria and the reliability of individual evaluators. We conduct experiments with both open-source and closed-source models. Our results demonstrate that MAPLE achieves superior alignment with human evaluations compared to baselines, highlighting the importance of employing multi-agent and multi-criteria evaluation strategies.

[Findings] MaRF: Leveraging Representation-Level Fusion of Formula Semantics for Mathematical Information Retrieval ► *Underline Video*

Suyuan Wang, Hongbo Zheng and Nickvash Kani

Mathematical information retrieval (MIR) depends on jointly modeling natural-language context and mathematical expressions. While BERT-based dense retrievers are effective, they often dilute mathematical semantics because textual content dominates most training data and mathematical formulas differ fundamentally from natural language in structure and composition. Consequently, these models rely heavily on surrounding text, which reduces robustness in math-intensive scenarios with limited textual description. We propose MaRF, a dual-encoder representation-level fusion framework for MIR that explicitly integrates formula semantics into context-aware dense retrieval. By combining contextual and formula-specific representations, MaRF captures complementary information from both textual and symbolic views. Experiments on the ARQMath-3 benchmark demonstrate that MaRF substantially improves retrieval performance and robustness, outperforming strong baselines across MIR tasks. The source code and datasets are available at https://github.com/MLPgroup/MaRF.

[Long] Mask-to-Correct^+: Leveraging Retriever Diversity for Masking-guided Faithful Fact Correction ► *Underline Video*

Payel Santra, Lavisha Sharma, Madhusudan Ghosh and Partha Basuchowdhuri

The rapid spread of misinformation on social media highlights the need for robust, automated fact correction frameworks. However, existing works rely on supervised learning from manually annotated claim-evidence pairs, which are scarce and prone to biases, limiting their generalization across domains. Moreover, these methods overlook semantic faithfulness in their correction process. To address these challenges, we propose Mask-to-Correct (M_2C), a training-free, inference-only Retrieval Augmented Generation (RAG) based framework that leverages diversity-aware masking to identify erroneous spans of claims and evaluate the faithfulness of corrections using retrieved evidence. However, the effectiveness of RAG heavily depends on the choice of retriever, which may vary across queries. To mitigate this, we further introduce M_2C^+, an ensemble-based framework that combines corrections across multiple rankers to reduce retrieval bias and improve robustness. Extensive experiments on the benchmark datasets demonstrate that our proposed frameworks consistently outperform all baselines, achieving up to 14% improvement in SARI scores, without using gold evidence.

[Long] MED-COREASONER: Reducing Language Disparities in Medical Reasoning via Language-Informed Co-Reasoning ► *Underline Video*

Fan Gao, Sherry T. Tong, Jiwoong Sohn, Jiahao Huang, Junfeng Jiang, Ding Xia, Piyalitt Ittichaiwong, Kanyakorn Veerakanjana, Hyunjae Kim, Qingyu Chen, Edison Marrese-Taylor, Kazuma Kobayashi, Akiko Aizawa and Irene Li

While reasoning-enhanced large language models perform strongly on English medical tasks, a persistent multilingual gap remains, with substantially weaker reasoning in local languages, limiting equitable global medical deployment. To bridge this gap, we introduce Med-CoReasoner, a language-informed co-reasoning framework that elicits parallel English and local-language reasoning, abstracts them into structured concepts, and integrates local clinical knowledge into an English logical scaffold via concept-level alignment and retrieval. This design combines the structural robustness of English reasoning with the practice-grounded expertise encoded in local languages. To evaluate multilingual medical reasoning beyond multiple-choice settings, we construct MultiMed-X, a benchmark covering seven languages with expert-annotated long-form question answering and natural language inference tasks, comprising 350 instances per language. Experiments across three benchmarks show that Med-CoReasoner improves multilingual reasoning performance by an average of 5%, with particularly substantial gains in low-resource languages. Moreover, model distillation and expert evaluation analysis further confirm that Med-CoReasoner produces clinically sound and culturally grounded reasoning traces.

[Findings] MedLayBench-V: A Large-Scale Benchmark for Expert-Lay Semantic Alignment in Medical Vision Language Models ► *Underline Video*

Han Jang, Junhyeok Lee, Heeseong Eum and Kyu Sung Choi

Medical Vision-Language Models (Med-VLMs) have achieved expert-level proficiency in interpreting diagnostic imaging. However, current models are predominantly trained on professional literature, limiting their ability to communicate findings in the lay register required for patient-centered care. While text-centric research has actively developed resources for simplifying medical jargon, there is a critical absence of large-scale multimodal benchmarks designed to facilitate lay-accessible medical image understanding. To bridge this resource gap, we introduce MedLayBench-V, the first large-scale multimodal benchmark dedicated to expert-lay semantic alignment. Unlike naive

simplification approaches that risk hallucination, our dataset is constructed via a Structured Concept-Grounded Refinement (SCGR) pipeline. This method enforces strict semantic equivalence by integrating Unified Medical Language System (UMLS) Concept Unique Identifiers (CUIs) with micro-level entity constraints. MedLayBench-V provides a verified foundation for training and evaluating next-generation Med-VLMs capable of bridging the communication divide between clinical experts and patients.

[Findings] Memory Dial: A Training Framework for Controllable Memorization in Language Models ▶ *Underline Video*

Xiangbo Zhang and Ali Emami

Memorization in language models is widely studied but remains difficult to isolate and control. Understanding when and what models memorize is essential for explaining their predictions, yet existing approaches are post-hoc: they can detect memorization in trained models, but cannot disentangle its effects from architecture, data, or optimization. We introduce Memory Dial, a training framework that makes memorization an explicit, controllable variable. Memory Dial interpolates between standard cross-entropy and a temperature-sharpened objective via a single parameter, producing a family of models identical in architecture, data, and optimization, but varying in memorization pressure. Experiments across six architectures and five benchmarks demonstrate that: (1) reliably controls memorization, with seen-example accuracy increasing monotonically while unseen accuracy remains stable; (2) larger models are more responsive to memorization pressure; and (3) frequent sequences are easier to memorize than rare ones. Memory Dial provides a controlled experimental framework for studying how memorization behavior emerges and interacts with generalization in language models.

[Long] MetaBench: A Multi-task Benchmark for Assessing LLMs in Metabolomics ▶ *Underline Video*

Yuxing Lu, Xukai Zhao, J. Ben Tamo, Micky C. Nnamdi, Rui Peng, Shuang Zeng, Xingyu Hu, Jinzhuo Wang and May Dongmei Wang

Large Language Models (LLMs) have demonstrated remarkable capabilities on general text; however, their proficiency in specialized scientific domains that require deep, interconnected knowledge remains largely uncharacterized. Metabolomics presents unique challenges with its complex biochemical pathways, heterogeneous identifier systems, and fragmented databases. To systematically evaluate LLM capabilities in this domain, we introduce MetaBench, the first benchmark for metabolomics assessment. Curated from authoritative public resources, MetaBench evaluates five capabilities essential for metabolomics research: knowledge, understanding, grounding, reasoning, and research. Our evaluation of 25 open- and closed-source LLMs reveals distinct performance patterns across metabolomics tasks: while models perform well on text generation tasks, cross-database identifier grounding remains challenging even with retrieval augmentation. Model performance also decreases on long-tail metabolites with sparse annotations. With MetaBench, we provide essential infrastructure for developing and evaluating metabolomics AI systems, enabling systematic progress toward reliable computational tools for metabolomics research.

[Long] Mind the Gap in Cultural Alignment: Task-Aware Culture Management for Large Language Models ▶ *Underline Video*

Binchi Zhang, Xujiang Zhao, Jundong Li, Haifeng Chen and Zhengzhang Chen

Large language models (LLMs) are increasingly deployed in culturally sensitive real-world tasks. However, existing cultural alignment approaches fail to align LLMs' broad cultural values with the specific goals of downstream tasks and suffer from cross-culture interference. We propose CultureManager, a novel pipeline for task-specific cultural alignment. CultureManager synthesizes task-aware cultural data in line with target task formats, grounded in culturally relevant web search results. To prevent conflicts between cultural norms, it manages multi-culture knowledge learned in separate adapters with a culture router that selects the appropriate one to apply. Experiments across five national cultures and ten culture-sensitive tasks show consistent improvements over prompt-based and fine-tuning baselines. Our results demonstrate the necessity of task adaptation and modular culture management for effective cultural alignment.

[Long] MM-JudgeBias: A Benchmark for Evaluating Compositional Biases in MLLM-as-a-Judge ▶ *Underline Video*

Sua Lee, Sanghee Park and Jinbae Im

Multimodal Large Language Models (MLLMs) have been increasingly used as automatic evaluators, a paradigm known as MLLM-as-a-Judge. However, their reliability and vulnerabilities to biases remain underexplored. We find that many MLLM judges fail to reliably integrate key visual or textual cues, yielding unreliable evaluations when evidence is missing or mismatched, and exhibiting instability under semantically irrelevant perturbations. To address this, we systematically define Compositional Bias in MLLM-as-a-Judge systems and introduce MM-JudgeBias, a benchmark for evaluating it. MM-JudgeBias introduces controlled perturbations across Query, Image, and Response, and evaluates model behavior via two complementary metrics: Bias-Deviation (BD) for sensitivity and Bias-Conformity (BC) for stability. Our dataset of over 1,800 curated and refined multimodal samples, drawn from 29 source benchmarks, enables a fine-grained diagnosis of nine bias types across diverse tasks and domains. Experiments on 26 state-of-the-art MLLMs reveal systematic modality neglect and asymmetric evaluation tendencies, underscoring the need for more reliable judges.

[Long] Multi-Agent-as-Judge: Aligning LLM-Agent-Based Automated Evaluation with Multi-Dimensional Human Evaluation ▶ *Underline Video*

Jiaju Chen, Yuxuan Lu, Xiaojie Wang, Huimin Zeng, Jing Huang, Jiri Gesi, Ying Xu, Bingsheng Yao and Dakuo Wang

Nearly all human work is collaborative; thus, the evaluation of real-world NLP applications often requires multiple dimensions that align with diverse human perspectives. As real human evaluator resources are often scarce and costly, the emerging "LLM-as-a-judge" paradigm sheds light on a promising approach to leverage LLM agents to believably simulate human evaluators. Yet, to date, existing LLM-as-a-judge approaches face two limitations: persona descriptions of agents are often arbitrarily designed, and the frameworks are not generalizable to

other tasks. To address these challenges, we propose MAJ-EVAL, a Multi-Agent-as-Judge evaluation framework that can automatically construct multiple evaluator personas with distinct dimensions from relevant text documents (e.g., research papers), instantiate LLM agents with the personas, and engage in-group debates with multi-agents to generate multi-dimensional feedback. Our evaluation experiments in both the educational and medical domains demonstrate that MAJ-EVAL can generate evaluation results that better align with human experts ratings compared with conventional automated evaluation metrics and existing LLM-as-a-judge methods.

[Findings] Multi-Docker-Eval: A 'Shovel of the Gold Rush' Benchmark on Automatic Environment Building for Software Engineering ▶ *Underline Video*

Kelin Fu, Tianyu Liu, Zeyu Shang, Yingwei Ma, Jiaheng Liu, Jian Yang and Kaigui Bian

Automated environment configuration is a critical bottleneck in scaling software engineering (SWE) automation. To provide a reliable evaluation standard for this task, we present Multi-Docker-Eval benchmark. It includes 40 real-world repositories spanning 9 programming languages and measures both success in achieving executable states and efficiency under realistic constraints. Our extensive evaluation of state-of-the-art LLMs and agent frameworks reveals key insights: (1) the overall success rate of current models is low (F2P at most 37.7%), with environment construction being the primary bottleneck; (2) model size and reasoning length are not decisive factors, and open-source models like DeepSeek-V3.1 and Kimi-K2 are competitive in both efficiency and effectiveness; (3) agent framework and programming language also have significantly influence on success rate. These findings provide actionable guidelines for building scalable, fully automated SWE pipelines.

[Findings] Multimodal In-context Learning for ASR of Low-resource Languages ▶ *Underline Video*

Zhaolin Li and Jan Niehues

Automatic speech recognition (ASR) still covers only a small fraction of the worlds languages, mainly due to supervised data scarcity. In-context learning (ICL) with large language models (LLMs) addresses this problem, but prior work largely focuses on high-resource languages covered during training and text-only settings. This paper investigates whether speech LLMs can learn unseen languages with multimodal ICL (MICL), and how this learning can be used to improve ASR. We conduct experiments with two speech LLMs, Phi-4 and Qwen3-Omni, on three diverse endangered languages. Firstly, we find that MICL is effective for unseen languages, leveraging both speech and text modalities. We further show that cross-lingual transfer learning improves MICL efficiency on target languages without training on them. Moreover, we analyze attention patterns to interpret MICL mechanisms, and we observe layer-dependent preferences between audio and text context, with an overall bias towards text. Finally, we show that prompt-based ASR with speech LLMs performs poorly on unseen languages, motivating a simple ASR system that combines a stronger acoustic model with a speech LLM via MICL-based selection of acoustic hypotheses. Results show that MICL consistently improves ASR performance, and that cross-lingual transfer learning matches or outperforms corpus-trained language models without using target-language data. Our code is publicly available

[Long] NSF-SciFy: Mining the NSF Awards Database for Scientific Claims ▶ *Underline Video*

Delip Rao, Weiqiu You, Eric Wong and Chris Callison-Burch

We introduce NSF-SciFy, a comprehensive dataset of scientific claims and investigation proposals extracted from National Science Foundation award abstracts. While previous scientific claim verification datasets have been limited in size and scope, NSF-SciFy represents a significant advance with 2.8 million claims from 400,000 abstracts spanning all science and mathematics disciplines. We present two focused subsets: NSF-SciFy-MatSci with 114,000 claims from materials science awards, and NSF-SciFy-20K with 135,000 claims across five NSF directorates. Using zero-shot prompting, we develop a scalable approach for joint extraction of scientific claims and investigation proposals. We demonstrate the dataset's utility through three downstream tasks: non-technical abstract generation, claim extraction, and investigation proposal extraction. Fine-tuning language models on our dataset yields substantial improvements, with relative gains often exceeding 100%, particularly for claim and proposal extraction tasks. Our error analysis reveals that extracted claims exhibit high precision but lower recall, suggesting opportunities for further methodological refinement. NSF-SciFy enables new research directions in large-scale claim verification, scientific discovery tracking, and meta-scientific analysis.

[Findings] One Script Instead of Hundreds? On Pretraining Romanized Encoder Language Models ▶ *Underline Video*

Benedikt Ebing, Lennart Keller and Goran Glava

Exposing latent lexical overlap, script romanization has emerged as an effective strategy for improving cross-lingual transfer (XLT) in multilingual language models (mLMs). Most prior work, however, focused on setups that favor romanization the most: (1) transfer from high-resource Latin-script to low-resource non-Latin-script languages and/or (2) between genealogically closely related languages with different scripts. It thus remains unclear whether romanization is a good representation choice for pretraining general-purpose mLMs, or, more precisely, if information loss associated with romanization harms performance for high-resource languages. We address this gap by pretraining encoder LMs from scratch on both romanized and original texts for six typologically diverse high-resource languages, investigating two potential sources of degradation: (i) loss of script-specific information and (ii) dilution of language-specific representations from increased subword overlap. Using two romanizers with different fidelity profiles, we observe negligible performance loss for languages with segmental scripts, whereas languages with morphosyllabic scripts (Chinese and Japanese) suffer degradation that higher-fidelity romanization mitigates but cannot fully recover. Importantly, comparing monolingual LMs with their mLM counterpart, we find no evidence that increased subword overlap dilutes language-specific representations. We further show that romanization improves encoding efficiency (i.e., fertility) for segmental scripts at a negligible performance cost.

[Long] One Tokenizer To Rule Them All: Emergent Language Plasticity via Multilingual Tokenizers ▶ *Underline Video*

Diana Abagyan, Alejandro R. Salamanca, Andres Felipe Cruz-Salinas, Kris Cao, Hangyu Lin, Acyr Locatelli, Marzieh Fadaee, Ahmet Üstün and Sara Hooker

Pretraining massively multilingual Large Language Models (LLMs) for many languages at once is challenging due to limited model capacity, scarce high-quality data, and compute constraints. Moreover, the lack of language coverage in the tokenizer makes it harder to address the gap for new languages purely at the post-training stage. In this work, we study what relatively cheap interventions early on in training improve language plasticity, or adaptation capabilities of the model post-training to new languages. We focus on tokenizer design and propose using a universal tokenizer that is trained for more languages than the primary pretraining languages to enable efficient adaptation in expanding language coverage after pretraining. Our systematic experiments across diverse groups of languages and different training strategies show that a universal tokenizer enables significantly higher language adaptation, with up to 20.2% increase in win rates compared to tokenizers specific to pretraining languages. Furthermore, a universal tokenizer also leads to better plasticity towards languages that are completely unseen in the tokenizer and pretraining, by up to 5% win rate gain. We achieve this adaptation to an expanded set of languages with minimal compromise in performance on the majority of languages included in pretraining.

[Long] OPeRA: A Dataset of Observation, Persona, Rationale, and Action for Evaluating LLMs on Human Online Shopping Behavior Simulation ▶ *Underline Video*

Ziyi Wang, Yuxuan Lu, Wenbo Li, Amirali Amini, Bo Sun, Yakov Bart, Weimin Lyu, Jiri Gesi, Tian Wang, Jing Huang, Yu Su, Upol Ehsan, Malihe Alikhani, Toby Jia-Jun Li, Lydia Chilton and Dakuo Wang

Can Large Language models (LLMs) accurately simulate the next web action of a specific user? While LLMs have shown promising capabilities in generating believable human behaviors, evaluating their ability to mimic real user behaviors remains an open challenge, largely due to the lack of high-quality, publicly available datasets that capture both the observable actions and the internal reasoning of an actual human user. To address this gap, we introduce OPeRA, a novel dataset of Observation, Persona, Rationale, and Action collected from real human participants during online shopping sessions. OPeRA is the first public dataset that comprehensively captures: user personas, browser observations, fine-grained web actions, and self-reported just-in-time rationales. We developed both an online questionnaire and a custom browser plugin to gather this dataset with high fidelity. Using OPeRA, we establish the first benchmark to evaluate how well current LLMs can predict a specific users next action and rationale with a given persona and <observation, action, rationale> history. This dataset lays the groundwork for future research into LLM agents that aim to act as personalized digital twins for human.

[Findings] PAPERMIND: Benchmarking Agentic Reasoning and Critique over Scientific Papers in Multimodal LLMs ▶ *Underline Video*

Yanjun Zhao, Tianxin Wei, Jiaru Zou, Xuying Ning, Yuanchen Bei, Lingjie Chen, Simmi Rana, Wendy H. Yang, Hanghang Tong and Jingrui He

Understanding scientific papers requires more than answering isolated questions or summarizing content. It involves an integrated reasoning process that grounds textual and visual information, interprets experimental evidence, synthesizes information across sources, and critically evaluates scientific claims. However, existing benchmarks typically assess these abilities in isolation, making it difficult to evaluate scientific paper understanding as a unified set of interacting cognitive abilities. In this work, we introduce PaperMind, a benchmark designed to evaluate integrated and agent-oriented scientific reasoning over research papers. PaperMind is constructed from real scientific papers across seven domains, including agriculture, biology, chemistry, computer science, medicine, physics, and economics. It comprises four complementary task families that collectively operationalize distinct cognitive facets of scientific paper reasoning, including multimodal grounding, experimental interpretation, cross-source evidence reasoning, and critical assessment. By analyzing model behavior across multiple tasks, PaperMind enables a diagnostic evaluation of integrated scientific reasoning behaviors that are difficult to assess through isolated task evaluations. Extensive experiments on both open-source and closed-source multimodal LLMs reveal consistent performance gaps across tasks, highlighting persistent challenges in integrated scientific reasoning and critique. Our benchmark and dataset are available at https://github.com/Yanjun-Zhao/PaperMind.

[Long] Parallel Universes, Parallel Languages: A Comprehensive Study on LLM-based Multilingual Counterfactual Example Generation ▶ *Underline Video*

Qianli Wang, Van Bach Nguyen, Yihong Liu, Fedor Splitt, Nils Feldhus, Christin Seifert, Hinrich Schuetze, Sebastian Möller and Vera Schmitt

Counterfactuals refer to minimally edited inputs that cause a model's prediction to change, serving as a promising approach to explaining the model's behavior. Large language models (LLMs) excel at generating English counterfactuals and demonstrate multilingual proficiency. However, their effectiveness in generating multilingual counterfactuals remains unclear. To this end, we conduct a comprehensive study on multilingual counterfactuals. We first conduct automatic evaluations on both directly generated counterfactuals in the target languages and those derived via English translation across six languages. Although translation-based counterfactuals offer higher validity than their directly generated counterparts, they demand substantially more modifications and still fall short of matching the quality of the original English counterfactuals. Second, we find the patterns of edits applied to high-resource European-language counterfactuals to be remarkably similar, suggesting that cross-lingual perturbations follow common strategic principles. Third, we reveal that multilingual counterfactual data augmentation (CDA) yields larger model performance improvements than cross-lingual CDA, especially for lower-resource languages. Yet, the imperfections of the generated counterfactuals limit gains in model performance and robustness. Finally, we identify and categorize four

main types of errors that consistently appear in the generated counterfactuals across languages.

[Findings] PO-KGQA: Preference Optimization for Low-Resource Complex Knowledge Graph Question Answering ▶ *Underline Video*

Prerna Agarwal, Ayushman Kumar Singh and Srikanta Bedathur

Existing low-resource in-context learning-based knowledge graph question answering (KGQA) methods rely heavily on large language models (LLMs) to convert the natural language question into its corresponding logical form (LF), such as SPARQL, KoPL, etc. Recently, a few alignment techniques have been introduced that enable instruction-based fine-tuning of language models. They provide explicit negative signals and comparative objectives to learn how to avoid negative signals using preference optimization methods. Exploring such fine-tuning techniques with LLMs becomes very challenging due to the high computational resource requirements associated with them. Due to this, the focus has been shifted towards Small Language Models (SLMs), which offer advantages such as ease of (i) deployment for practical applications and (ii) instruction fine-tuning for specialized tasks. Motivated by this, in this work, we propose PO-KGQA: An SLM-based preference optimization framework for the complex KGQA task in a low-resource setting. Our extensive experiments demonstrate how PO-KGQA outperforms other fine-tuning alignment techniques on complex benchmarks such as KQA Pro by approximately 9% (avg).

[Long] PRISM: Probing Reasoning, Instruction, and Source Memory in LLM Hallucinations ▶ *Underline Video*

Yuhe Wu, Guangyu Wang, Yuran Chen, Jiatong Zhang, Yutong Zhang, Yujie Chen, Jiaming Shang, Guang Zhang and Zhuang Liu

As large language models (LLMs) evolve from conversational assistants into agents capable of handling complex tasks, they are increasingly deployed in high-risk domains. However, existing benchmarks largely rely on mixed queries and posterior evaluation, output-level scoring, which quantifies hallucination severity but offers limited insight into where and why hallucinations arise in the generation pipeline. We therefore reformulate hallucination evaluation as a diagnostic problem and propose PRISM, a controlled benchmark that disentangles hallucinations into four dimensions: knowledge missing, knowledge errors, reasoning errors, and instruction-following errors, grounded in three stages of generation (memory, instruction, and reasoning). PRISM contains 9,448 instances across 65 tasks and supports fine-grained, stage-aware diagnostic evaluation. Evaluating 24 mainstream open-source and proprietary LLMs, we uncover consistent trade-offs across instruction following, memory retrieval, and logical reasoning, showing that mitigation strategies often improve specific dimensions at the expense of others. We hope PRISM provides a framework for understanding the specific mechanisms behind LLMs hallucinations, ultimately accelerating the development of trustworthy large language models.

[Findings] PseudoSeer: a Search Engine for Pseudocode ▶ *Underline Video*

Levent Toksoz, Mukund Srinath, Gang Tan and C. Lee Giles

PseudoSeer is a novel search engine for academic pseudocode, enabling retrieval over 320,000 algorithm implementations extracted from the arXiv. Using the system's caption-reference pairs, we study asymmetric retrieval, matching short queries with a median length of five words against long documents of roughly 300 words composed primarily of natural language with limited LaTeX notation. Our evaluation reveals scaling limitations in embedding models: a 149M parameter encoder outperforms 1.5B parameter alternatives, while BM25 remains competitive with pretrained models. Analyzing attention patterns over 33,000 caption document pairs, we identify two factors driving these results: attention efficiency and attention concentration. Models that significantly attend to sinks or non-discriminative tokens leave less attention for discriminative content, while models with overly diffuse attention fail to form discriminative representations. Guided by these findings, PseudoSeers embedding model, trained via contrastive learning with efficient attention patterns, outperforms the best pretrained model by 8.7 points. A hybrid approach combining learned embeddings with BM25 reaches 66.5% R@10. PseudoSeer is deployed at pseudoseer.ist.psu.edu as both a practical search system and a benchmark for retrieval evaluation.

[Long] Putting HUMANS first: Efficient LAM Evaluation with Human Preference Alignment ▶ *Underline Video*

Woody Haosheng Gan, William Barr Held and Diyi Yang

The rapid proliferation of large audio models (LAMs) demands efficient approaches for model comparison, yet comprehensive benchmarks are costly. To fill this gap, we investigate whether minimal subsets can reliably evaluate LAMs while reducing costs and data redundancy. Analyzing 10 subset selection methods with 18 audio models across 40 tasks covering major LAM evaluation dimensions, we show that subsets of just 50 examples (0.3% of data) can achieve over 0.93 Pearson correlation with full benchmark scores. To understand how well these scores align with what practitioners ultimately care about, user satisfaction, we collect 776 human preference ratings from realistic voice assistant conversations, finding that both subsets and full benchmark achieve only 0.85 correlation with human. To better predict preferences, we trained regression models on these selected subsets, achieving 0.98 correlation, outperforming regression models trained on both random subsets and the full benchmark. This demonstrates that in regression modeling, well-curated subsets outpredict the full benchmark, showing quality over quantity. We open-source these regression-weighted subsets as the HUMANS benchmark, an efficient proxy for LAM evaluation that captures both benchmark performance and user preferences.

[Long] Quantifying and Understanding Uncertainty in Large Reasoning Models ▶ *Underline Video*

Yangyi Li, Chenxu Zhao and Mengdi Huai

Large Reasoning Models (LRMs) have recently demonstrated significant improvements in complex reasoning. While quantifying generation uncertainty in LRMs is crucial, traditional methods are often insufficient because they do not provide finite-sample guarantees for reasoning-answer generation. Conformal prediction (CP) stands out as a model-agnostic methodology that constructs statistically rigorous

uncertainty sets. However, existing CP methods ignore the logical connection between the reasoning trace and the final answer. Additionally, prior studies fail to interpret the origins of uncertainty coverage for LRMs as they typically overlook the specific training factors driving valid reasoning. Notably, it is challenging to disentangle reasoning quality from answer correctness, while simultaneously establishing theoretical guarantees for computationally efficient explanation methods. To address these challenges, we first propose a novel methodology that provides the uncertainty of the reasoning-answer structure with statistical guarantees. Subsequently, we develop a unified example-to-step explanation framework using Shapley values that identifies a provably sufficient subset of training data and their specific reasoning steps sufficient to achieve coverage. We also provide the theoretical analysis for our proposed methods. Extensive experiments on challenging reasoning datasets verify the effectiveness of the proposed methods.

[Long] QuDAR: Query-Wise Dual-Perspective Adaptive Retrieval ▶ *Underline Video*

Joeun Kim, Seunghyouk Yoon, Xuan-Bach Le, Youngeun Nam, Doyoung Kim, Hwanjun Song and Jae-Gil Lee

Retrieval-augmented generation(RAG) systems depend on retrieval modules to supply grounding evidence for large language models. While hybrid approaches combining sparse and dense retrievers improve performance, most rely on fixed weights that ignore query-specific and corpus-specific variation. Similarly, query expansion has long been used to enrich recall, but its integration with original queries is usually static and can introduce noise. We present QuDAR, a dual-perspective adaptive retrieval framework that adapts along two perspectives: retriever type (sparse vs. dense) and query format (original vs.expanded). Leveraging margin-derived confidence (e.g., top-1–top-2 score gaps) and blind LLM-based relevance scoring, QuDAR dynamically assigns query-specific weights, fusing lexical specificity with semantic breadth while mitigating noise. QuDAR is lightweight, retriever-agnostic, and broadly applicable. Experiments show consistent gains over static baselines, improving overall retrieval quality and yielding more stable performance across queries.

[Findings] Raw Pointer Rewriting with LLMs for Translating C to Safer Rust ▶ *Underline Video*

Yifei Gao, Chengpeng Wang, Pengxiang Huang, Xuwei Liu, Mingwei Zheng and Xiangyu Zhang

There has been a growing interest in translating C code to Rust due to Rust's robust memory and thread safety guarantees. Tools such as C2Rust enable syntax-guided transpilation from C to semantically equivalent Rust code. However, the resulting Rust programs often rely heavily on unsafe constructs, particularly raw pointers, which undermines Rusts safety guarantees. This paper aims to improve the memory safety of Rust programs generated by C2Rust by eliminating raw pointers. Specifically, we propose a raw pointer rewriting technique that lifts raw pointers in individual functions to appropriate Rust data structures. Technically, PR2 employs decision-tree-based prompting to guide the pointer lifting process. It also leverages code change analysis to guide the repair of errors introduced during rewriting, effectively addressing errors encountered during compilation and test case execution.We implement PR2 and evaluate it using gpt-4o-mini on 28 real-world C projects. It is shown that PR2 successfully eliminates 18.57% of local raw pointers across these projects, significantly enhancing the safety of the translated Rust code. On average, PR2 completes the transformation of a project in 5.02 hours, at a cost of $1.13. Our code is available at https://github.com/bhcsayx/PR2.

[Long] REaR: Retrieve, Expand and Refine for Effective Multitable Retrieval ▶ *Underline Video*

Rishita Agarwal, Himanshu Singhal, Peter Baile Chen, Manan Roy Choudhury, Dan Roth and Vivek Gupta

Answering natural language queries over relational data often requires retrieving and reasoning over multiple tables, yet most retrievers optimize only for query-table relevance and ignore table-table compatibility. We introduce REaR (Retrieve, Expand and Refine), a three-stage, LLM-free framework that separates semantic relevance from structural joinability for efficient, high-fidelity multi-table retrieval. REaR (i) retrieves query-aligned tables, (ii) expands these with structurally joinable tables via fast, precomputed column-embedding comparisons, and (iii) refines them by pruning noisy or weakly related candidates. Empirically, REaR is retriever-agnostic and consistently improves dense/ sparse retrievers on complex table QA datasets (BIRD, MMQA, and Spider) by improving both multi-table retrieval quality and downstream SQL execution. Despite being LLM-free, it delivers performance competitive with state-of-the-art LLM-augmented retrieval systems (e.g., ARM) while achieving much lower latency and cost. Ablations confirm complementary gains from expansion and refinement, underscoring REaR as a practical, scalable building block for table-based downstream tasks (e.g., Text-to-SQL).

[Long] ReFEree: Reference-Free and Fine-Grained Method for Evaluating Factual Consistency in Real-World Code Summarization ▶ *Underline Video*

Suyoung Bae, CheolWon Na, Jaehoon Lee, Yumin Lee, YunSeok Choi and Jee-Hyong Lee

As Large Language Models (LLMs) have become capable of generating long and descriptive code summaries, accurate and reliable evaluation of factual consistency has become a critical challenge. However, previous evaluation methods are primarily designed for short summaries of isolated code snippets. Consequently, they struggle to provide fine-grained evaluation of multi-sentence functionalities and fail to accurately assess dependency context commonly found in real-world code summaries. To address this, we propose ReFEree, a reference-free and fine-grained method for evaluating factual consistency in real-world code summaries. We define factual inconsistency criteria specific to code summaries and evaluate them at the segment level using these criteria along with dependency information. These segment-level results are then aggregated into a fine-grained score. We construct a code summarization benchmark with human-annotated factual consistency labels. The evaluation results demonstrate that ReFEree achieves the highest correlation with human judgment among 13 baselines, improving 15-18% over the previous state-of-the-art. Our code and data are available at https://github.com/bsy99615/ReFEree.git.

[Findings] Reference-Free Evaluation of Taxonomies ▶ *Underline Video*

Pascal Wullschleger, Majid Zarharan, Donnacha Daly, Marc Pouly and Jennifer Foster
We introduce two reference-free metrics for quality evaluation of taxonomies in the absence of labels. The first metric evaluates robustness by calculating the correlation between semantic and taxonomic similarity, addressing error types not considered by existing metrics. The second uses Natural Language Inference to assess logical adequacy. Both metrics are tested on five taxonomies and are shown to correlate well with F1 against ground truth taxonomies. We further demonstrate that our metrics can predict downstream performance in hierarchical classification when used with label hierarchies.

[TACL] ResearchQA: Evaluating Scholarly Question Answering at Scale Across 75 Fields with Survey-Mined Questions and Rubrics ▶ *Underline Video*
Li S. Yifei, Allen Chang, Chaitanya Malaviya and Mark Yatskar
Evaluating long-form responses to research queries heavily relies on expert annotators, restricting attention to areas like AI where researchers can conveniently enlist colleagues. Yet, research expertise is abundant: survey articles consolidate knowledge spread across the literature. We introduce ResearchQA, a resource for evaluating LLM systems by distilling survey articles from 75 research fields into 21K queries and 160K rubric items. Queries and rubrics are jointly derived from survey sections, where rubric items list query-specific answer evaluation criteria, i.e., citing papers, making explanations, and describing limitations. 31 Ph.D. annotators in 8 fields judge that 90% of queries reflect Ph.D. information needs and 87% of rubric items warrant emphasis of a sentence or longer. We leverage ResearchQA to evaluate 18 systems in 7.6K head-to-heads. No parametric or retrieval-augmented system we evaluate exceeds 70% on covering rubric items, and the highest-ranking system shows 75% coverage. Error analysis reveals that the highest-ranking system fully addresses less than 11% of citation rubric items, 48% of limitation items, and 49% of comparison items. We release our data to facilitate more comprehensive multi-field evaluations.

[Long] REST: Stress Testing Large Reasoning Models by Asking Multiple Problems at Once ▶ *Underline Video*
Zhuoshi Pan, Qizhi Pei, Yu Li, Zinan Tang, Qiyao Sun, H. Vicky Zhao, Conghui He and Lijun Wu
Recent Large Reasoning Models (LRMs) have achieved remarkable progress, yet their evaluation still relies on a narrow paradigm: evaluating one question at a time. This single-question setup suffers from two major limitations: (1) vulnerability to data contamination and diminishing difficulty, forcing costly creation of new questions with significant human effort, (2) failure to evaluate models under multi-context pressure, a key requirement for real-world deployment. To bridge this gap, we present REST (Reasoning Evaluation through Simultaneous Testing), a stress-testing framework that exposes LRMs to multiple problems simultaneously. Beyond basic reasoning, REST evaluates two under-tested capabilities: contextual priority allocation and robustness against contextual interference. Our evaluation of more than 30 advanced reasoning models on 9 reasoning benchmarks reveals several striking findings: Even state-of-the-art (SOTA) models such as DeepSeek-R1 exhibit substantial performance degradation under stress testing, challenging the prevailing assumption that "LLMs are multi-problem solvers". Crucially, REST demonstrates stronger discriminative power than existing benchmarks, revealing performance gaps among models that exhibit similar, near-ceiling performance under traditional evaluation. Some key insights emerge from our analysis: (1) the "overthinking trap" is a critical factor contributing to the performance degradation; (2) models trained with the "Long2Short" technique preserve more of their single-problem accuracy under REST, outperforming their standard-trained counterparts. These results establish REST as a cost-efficient, future-proof evaluation paradigm while reducing reliance on continuous human annotation. Code is available at https://github.com/opendatalab/REST.

[Long] Rethinking Evaluation for LLM Hallucination Detection: A Desiderata, A New RAG-based Benchmark, New Insights ▶ *Underline Video*
Wenbo Chen, Veena Padmanabhan, Tootiya Giyahchi, Elaine Wong and Leman Akoglu
Hallucination, broadly referring to unfaithful, fabricated, or inconsistent content generated by LLMs, has wide-ranging implications. Therefore, a large body of effort has been devoted to detecting LLM hallucinations, as well as designing benchmark datasets for evaluating these detectors. In this work, we first establish a desiderata of properties for hallucination detection benchmarks (HDBs) to exhibit for effective evaluation. A critical look at existing HDBs through the lens of our desiderata reveals that none of them exhibits all the properties. We identify two largest gaps: (1) RAG-based grounded benchmarks with long context are severely lacking (partly because length impedes human annotation); and (2) Existing benchmarks do not make available realistic label noise for stress-testing detectors although real-world use-cases often grapple with label noise due to human or automated/weak annotation. To close these gaps, we build and open-source a new RAG-based HDB called TRIVIA+ that underwent a rigorous human annotation process. Notably, our benchmark exhibits all desirable properties including (1) TRIVIA+ contains samples with the longest context in the literature; and (2) we design and share three sets of noisy labels with different, sample-dependent noise schemes. Finally, we perform experiments on RAG-based HDBs, including our TRIVIA+, using popular SOTA detectors that reveal new insights: (i) ample room remains for current detectors to reach the performance ceiling on RAG-based HDBs, (ii) the basic LLM-as-a-Judge baseline performs competitively, and (iii) label noise hinders detection performance. We expect that our findings, along with our proposed benchmark, will motivate and foster needed research on hallucination detection for RAG-based tasks.

[Long] Rethinking Meeting Effectiveness: A Benchmark and Framework for Temporal Fine-grained Automatic Meeting Effectiveness Evaluation ▶ *Underline Video*
Yihang Li and Chenhui Chu

Evaluating meeting effectiveness is crucial for improving organizational productivity. Current approaches rely on post-hoc surveys that yield a single coarse-grained score for an entire meeting. The reliance on manual assessment is inherently limited in scalability, cost, and reproducibility. Moreover, a single score fails to capture the dynamic nature of collaborative discussions. We propose a new paradigm for evaluating meeting effectiveness centered on novel criteria and temporal fine-grained approach. We define effectiveness as the rate of objective achievement over time and assess it for individual topical segments within a meeting. To support this task, we introduce the AMI Meeting Effectiveness (AMI-ME) dataset, a new meta-evaluation dataset containing 2,459 human-annotated segments from 130 AMI Corpus meetings. We also develop an automatic effectiveness evaluation framework that uses a Large Language Model (LLM) as a judge to score each segment's effectiveness relative to the overall meeting objectives. Through substantial experiments, we establish a comprehensive benchmark for this new task and evaluate the framework's generalizability across distinct meeting types, ranging from business scenarios to unstructured discussions. Furthermore, we benchmark end-to-end performance starting from raw speech to measure the capabilities of a complete system. Our results validate the framework's effectiveness and provide strong baselines to facilitate future research in meeting analysis and multi-party dialogue. Our dataset and code will be publicly available.

[Long] ReTRE: Benchmarking LLM Transfer Robustness with Structure-Preserving Variants ► *Underline Video*

ZhongDong Li, Weijie Shi, Yue Cui, Haolun Ma, Yuanjun Liu, Jiawei Li, An Liu, Jia Zhu and Jiajie Xu

Large language models (LLMs) have achieved strong performance on standard benchmarks, yet their performance is not robust across different task manifestations. It remains unclear how performance changes under controlled task rewrites that preserve the original solution structure, while varying the rewrite type and level. To address this question, we introduce ReTRE (Rewrite-based Transfer Robustness Evaluation), an evaluation benchmark inspired by learning transfer theory that probes transfer robustness along two rewrite levels: Near Transfer and Far Transfer. ReTRE employs a multi-agent system to construct textual and visual variants while preserving the structure of the original solution. Evaluations on mathematical and science tasks across state-of-the-art multimodal LLMs reveal a consistent transfer gap: performance exhibits a general declining trend as transfer similarity drops and strong text performance can face performance decline under cross-modal transfer. Crucially, we identify a divergence between post-training paradigms: reinforcement learning preserves transfer robustness, whereas supervised fine-tuning tends to overfit the training distribution, leading to severe degradation in far-transfer performance despite strong in-distribution accuracy.

[Long] REZE: Representation Regularization for Domain-adaptive Text Embedding Pre-finetuning ► *Underline Video*

Seungmin Lee, Jeonghwan Lee, Hyunkuk Lim, Sejoon Kim and Mingi Sung

Recent text embedding models are often adapted to specialized domains via contrastive pre-finetuning (PFT) on a naive collection of scattered, heterogeneous tasks. However, this approach often introduces task-induced bias alongside domain knowledge, leading to uncontrolled representation shifts that distort the pretrained embedding geometry and cause substantial performance degradation. To address this issue, we propose REZE, a representation regularization framework that explicitly controls representation shift during embedding pre-finetuning. REZE operates on the relations of anchor-positive pairs and decomposes them in an eigenspace. It then measures task-wise dispersion along each eigencomponent to identify task-variant directions and applies adaptive soft-shrinkage to suppress task-induced noise while preserving task-invariant semantic structure, without inference-time overhead. Experiments across multiple embedding backbones and specialized benchmarks show that REZE outperforms standard pre-finetuning and isotropy-oriented post-hoc regularization in most settings, remaining stable where existing PFT variants collapse. Embedding space analyses further confirm that REZE induces controlled shifts aligned with the original embedding manifold, underscoring representation shift control as a key principle for robust embedding pre-finetuning under heterogeneous supervision.

[Findings] RIGOURATE: Quantifying Scientific Exaggeration with Evidence-Aligned Claim Evaluation ► *Underline Video*

Joseph James, Chenghao Xiao, Yucheng Li, Nafise Sadat Moosavi and Chenghua Lin

Scientific rigour tends to be sidelined in favour of bold statements, leading authors to overstate claims beyond what their results support. We present RIGOURATE, a two-stage multimodal framework that retrieves supporting evidence from a papers body and assigns each claim an overstatement score. The framework consists of a dataset of over 10K claim-evidence sets from ICLR and NeurIPS papers, annotated using eight LLMs, with overstatement scores calibrated using peer-review comments and validated through human evaluation. It employes a fine-tuned reranker for evidence retrieval and a fine-tuned model to predict overstatement scores with justification. Compared to strong baselines, RIGOURATE enables improved evidence retrieval and overstatement detection. Overall, our work operationalises evidential proportionality and supports clearer, more transparent scientific communication.

[Findings] Robustness via Referencing: Defending against Prompt Injection Attacks by Referencing the Executed Instruction ► *Underline Video*

Yulin Chen, Haoran Li, Yuan Sui, Yue Liu, Yufei He, XiaolingBai XiaolingBai, Chi Fei, Li Yabo, Haozhe Ma, Yangqiu Song and Bryan Hooi

Prompt injection attacks manipulate large language models (LLMs) by misleading them to deviate from the original input instructions and execute maliciously injected instructions, because of their instruction-following capabilities and inability to distinguish between the original input instructions and maliciously injected instructions. Currently, various prompt injection defense methods have been proposed, including prompt-engineering-based approaches and fine-tuning methods. Most of these methods instruct the model to follow the original input instructions, suppressing its inherent tendencies to follow the injected instructions. However, experimental results reveal that suppressing the model's instruction-following tendencies is challenging. After analyzing successful attack cases, we

find that the LLMs can correctly reference the instructions they are executing in some cases. Motivated by this finding, we propose a defense method that leverages LLMs instruction-following abilities rather than suppressing them. Our approach prompts LLMs to generate responses that include both the answers and their corresponding instruction references. Based on these references, we filter out answers whose references are not to the original input instructions. We conduct comprehensive experiments to evaluate the effectiveness of our proposed method. The results show that our approach outperforms prompt-engineering-based baselines and is comparable to fine-tuning methods, reducing the ASR to nearly 0% in some scenarios. Moreover, our approach has minimal impact on overall utility.

[Findings] RoleConflictBench: A Benchmark of Role Conflict Scenarios for Evaluating LLMs' Contextual Sensitivity ▶ *Underline Video*

Jisu Shin, Hoyun Song, Juhyun Oh, Changgeon Ko, Eunsu Kim, Chani Jung and Alice Oh

People often encounter role conflicts—social dilemmas where the expectations of multiple roles clash and cannot be simultaneously fulfilled. As large language models (LLMs) increasingly navigate these social dynamics, a critical research question emerges. When faced with such dilemmas, do LLMs prioritize dynamic contextual cues or the learned preferences? To address this, we introduce RoleConflictBench, a novel benchmark designed to measure the contextual sensitivity of LLMs in role conflict scenarios. To enable objective evaluation within this subjective domain, we employ situational urgency as a constraint for decision-making. We construct the dataset through a three-stage pipeline that generates over 13,000 realistic scenarios across 65 roles in five social domains by systematically varying the urgency of competing situations. This controlled setup enables us to quantitatively measure contextual sensitivity, determining whether model decisions align with the situational contexts or are overridden by the learned role preferences. Our analysis of 10 LLMs reveals that models substantially deviate from this objective baseline. Instead of responding to dynamic contextual cues, their decisions are predominantly governed by the preferences toward specific social roles.

[Findings] ROSCO-Omni: Multimodal LLM-Based Communication Understanding for Non- and Minimally-Speaking Autistic Individuals ▶ *Underline Video*

Siddhant Shah and Kristina Johnson

Approximately 30% of autistic individuals remain non- or minimally-speaking throughout their lives, yet communicate richly through gestures, vocalizations, facial expressions, and augmentative devices. Interpreting this communication is an inherently multimodal task: caregivers rely on the simultaneous integration of visual cues, auditory signals, and contextual understanding to infer intent. Despite this natural alignment with multimodal large language models (MLLMs), research in this intersection remains narrowly focused on diagnosis rather than communication understanding. We address this gap by reframing the problem around two complementary dimensions: communicative actions (the physical modality) and communicative functions (the pragmatic intent). We analyze the ROSCO dataset, containing 2,903 caregiver-annotated video samples from 27 non- and minimally-speaking individuals, with multi-label annotations capturing up to three concurrent actions and two functions per sample across 6 action and 6 function classes. We further propose ROSCO-Omni, a teacher-student distillation framework that generates label-guided instruction data from a high-capability teacher MLLM and uses it to finetune a student MLLM for domain-specialized inference. ROSCO-Omni achieves performance comparable to closed-source models, demonstrating that open-source MLLMs can be adapted to understand communication in this underserved population.

[Long] SACTOR: LLM-Driven Correct and Idiomatic C to Rust Translation with Static Analysis and FFI-Based Verification ▶ *Underline Video*

Tianyang Zhou, Ziyi Zhang, Haowen Lin, Somesh Jha, Mihai Christodorescu, Kirill Levchenko and Varun Chandrasekaran

Translating software written in C to Rust has significant benefits in improving memory safety. However, manual translation is cumbersome, error-prone, and often produces unidiomatic code. Large language models (LLMs) have demonstrated promise in producing idiomatic translations, but offer no correctness guarantees. We propose SACTOR, an LLM-driven C-to-Rust translation tool that employs a two-step process: an initial "unidiomatic" translation to preserve semantics, followed by an "idiomatic" refinement to align with Rust standards. SACTOR leverages static analysis of the C source to handle pointer semantics and dependency resolution. To validate correctness of our function-wise incremental translation that mixes C and Rust, we use end-to-end testing via the foreign function interface. We evaluate SACTOR on 200 programs from two public datasets and on two more complex scenarios (a 50-sample subset of CRust-Bench and the libogg library), comparing multiple LLMs. Across datasets, SACTOR delivers high end-to-end correctness and produces safe, idiomatic Rust with up to 7x fewer Clippy warnings; On CRust-Bench, SACTOR achieves an average (across samples) of 85% unidiomatic and 52% idiomatic success, and on libogg it attains full unidiomatic and up to 78% idiomatic coverage on GPT-5.

[Findings] SDC-LoRA: Singular-Subspace Drift Controlled LoRA to Mitigate Knowledge Forgetting ▶ *Underline Video*

Geyuan Zhang, Xiaofei Zhou, Shihao Liu, Jingyuan Tian and Jizheng Ma

Knowledge forgetting is a central challenge when adapting LLMs to new tasks. Prior studies indicate that pretrained knowledge is concentrated in the principal singular subspace of pretrained weight W_0; so recent Low-Rank Adaptation (LoRA) variants initialize LoRA in the minor subspace to steer early updates away from principal directions and mitigate forgetting. However, we observe that during fine-tuning, the update direction progressively shifts from the minor to the principal subspace, which is called as Singular-subspace Drift (SD), thereby allocating more energy to the directions that carry pretrained knowledge and leaving a persistent risk of forgetting. To address this issue, we propose Singular-subspace Drift Controlled LoRA (SDC-LoRA), which constrains the growth of update energy in the principal singular subspace of W_0 and thus mitigate SD. SDC-LoRA proposes Principal Subspace Energy-Controlled Learning, using Spectral

Calibration factor _sc to selectively downscale gradients along the principal singular subspace of W_0 while keeping minor-subspace updates unchanged. Across extensive experiments with LLaMA-3.1-8B-Instruct and Qwen2.5-7B-Chat on MetaMathQA and CodeFeedback, SDC-LoRA mitigates forgetting on MMLU, TruthfulQA, and HellaSwag while matching or improving GSM8K and HumanEval, offering a practical route to adapt LLMs without sacrificing prior knowledge.

[Long] SDiaReward: Modeling and Benchmarking Spoken Dialogue Rewards with Modality and Colloquialness ▶ *Underline Video*

Jingyu Lu, Yuhan Wang, Fan Zhuo, Xize Cheng, Changhao Pan, Xueyi Pu, Yifu Chen, Chenyuhao Wen, Tianle Liang and Zhou Zhao

The rapid evolution of end-to-end spoken dialogue systems demands transcending mere textual semantics to incorporate paralinguistic nuances and the spontaneous nature of human conversation. However, current methods struggle with two critical gaps: the modality gap, involving prosody and emotion, and the colloquialness gap, distinguishing written scripts from natural speech. To address these challenges, we introduce SDiaReward, an end-to-end multi-turn reward model trained on SDiaReward-Dataset, a novel collection of episode-level preference pairs explicitly targeting these gaps. It operates directly on full multi-turn speech episodes and is optimized with pairwise preference supervision, enabling joint assessment of modality and colloquialness in a single evaluator. We further establish ESDR-Bench, a stratified benchmark for robust episode-level evaluation. Experiments demonstrate that SDiaReward achieves state-of-the-art pairwise preference accuracy, significantly outperforming general-purpose audio LLMs. Further analysis suggests that SDiaReward captures relative conversational expressiveness beyond superficial synthesis cues, improving generalization across domains and recording conditions.

[Findings] Similarity-Distance-Magnitude Activations ▶ *Underline Video*

Allen Schmaltz

We introduce the Similarity-Distance-Magnitude (SDM) activation function, a more robust and interpretable formulation of the standard softmax activation function, adding Similarity (i.e., correctly predicted depth-matches into training) awareness and Distance-to-training-distribution awareness to the existing output Magnitude (i.e., decision-boundary) awareness, and enabling interpretability-by-exemplar via dense matching. We further introduce the SDM estimator, based on a data-driven partitioning of the class-wise empirical CDFs via the SDM activation, to control the class- and prediction-conditional accuracy among selective classifications. When used as the final-layer activation over pre-trained language models for selective classification, the SDM estimator is more robust to covariate shifts and out-of-distribution inputs than existing calibration methods using softmax activations, while remaining informative over in-distribution data.

[Findings] SimpleOCR: Rendering Visual Questions to Teach MLLMs to Read ▶ *Underline Video*

Yibo Peng, Peng Xia, Ding Zhong, Kaide Zeng, Siwei Han, Yiyang Zhou, Jiaqi Liu, Ruiyi Zhang and Huaxiu Yao

Despite the rapid advancements in Multimodal Large Language Models (MLLMs), a critical question regarding their visual grounding mechanism remains unanswered: do these models genuinely read text embedded in images, or do they merely rely on parametric shortcuts in the text prompt? In this work, we diagnose this issue by introducing the Visualized-Question (VQ) setting, where text queries are rendered directly onto images to structurally mandate visual engagement. Our diagnostic experiments on Qwen2.5-VL reveal a startling capability-utilization gap: despite possessing strong OCR capabilities, models suffer a performance degradation of up to 12.7% in the VQ setting, exposing a deep-seated modality laziness. To bridge this gap, we propose SimpleOCR, a plug-and-play training strategy that imposes a structural constraint on the learning process. By transforming training samples into the VQ format with randomized styles, SimpleOCR effectively invalidates text-based shortcuts, compelling the model to activate and optimize its visual text extraction pathways. Empirically, SimpleOCR yields robust gains without architectural modifications. On four representative OOD benchmarks, it surpasses the base model by 5.4% and GRPO based on original images by 2.7%, while exhibiting extreme data efficiency, achieving superior performance with 30x fewer samples (8.5K) than recent RL-based methods. Furthermore, its plug-and-play nature allows seamless integration with advanced RL strategies like NoisyRollout to yield complementary improvements. Code is available at https://github.com/aiming-lab/SimpleOCR.

[Findings] Simul-COMET: A Quality Metric for Simultaneous Interpretation in Distant Language Pair Considering Word Order Difference ▶ *Underline Video*

Kosuke Doi, Mana Makinae, Yusuke Sakai, Hidetaka Kamigaito and Taro Watanabe

In simultaneous interpretation (SI), interpreters perform real-time translation by segmenting the source speech into chunks and translating them in the order they appear. Since surface-matching metrics such as BLEU correlate poorly with human evaluations, translation quality is often evaluated using neural metrics that measure semantic similarity, such as COMET. However, while SI translation ideally exhibits high monotonicity, COMET tends to assign higher scores to offline translations with long-distance reordering, because it is trained on such offline translation data. To address this gap, we propose Simul-COMET, a variation of COMET adapted for SI evaluation specifically designed for monotonicity. We train Simul-COMET on the SI-style translation data, which was converted from the offline translation of the COMET training data by leveraging large language models. In English–Japanese translation experiments, we demonstrate that Simul-COMET assigns higher scores to SI-style translations than to offline ones. Moreover, Simul-COMET shows stronger alignment with evaluation scores provided by professional interpreters than the original COMET. Simul-COMET is available at https://github.com/kosuked/simul-comet.

[Long] Simulated Students in Tutoring Dialogues: Substance or Illusion? ▶ *Underline Video*

Alexander Scarlatos, Jaewook Lee, Simon Woodhead and Andrew Lan

Advances in large language models (LLMs) enable many new innovations in education. However, evaluating the effectiveness of new technology requires real students, which is time-consuming and hard to scale up. Therefore, many recent works on LLM-powered tutoring

solutions have used simulated students for both training and evaluation, often via simple prompting. Surprisingly, little work has been done to ensure or even measure the quality of simulated students. In this work, we formally define the student simulation task, propose a set of evaluation metrics that span linguistic, behavioral, and cognitive aspects, and benchmark a wide range of student simulation methods on these metrics. We experiment on a real-world math tutoring dialogue dataset, where both automated and human evaluation results show that prompting strategies for student simulation perform poorly; supervised fine-tuning and preference optimization yield much better but still limited performance, motivating future work on this challenging task.

[Long] SMARTER: A Data-efficient Framework to Improve Toxicity Detection with Explanation via Self-augmenting Large Language Models ▶ *Underline Video*

Huy Nghiem, Advik Sachdeva and Hal Daumé III

To address toxic content on social media, we introduce SMARTER, a data-efficient 2-stage framework for explainable content moderation using Large Language Models (LLMs). In Stage 1, we leverage LLMs' own outputs to generate synthetic explanations for correct and incorrect labels, enabling preference optimization with minimal supervision. In Stage 2, we refine explanation quality through cross-model training, allowing weaker models to align with stronger ones. Experiments on 3 benchmarks (HateXplain, Latent Hate, Implicit Hate) show SMARTER achieves up to 13% macro-F1 improvement over few-shot baselines using only 6-57% of training data. Our framework offers a scalable strategy for low-data settings by harnessing LLMs' self-improvement for explainable moderation.

[Long] Social Dynamics as Critical Vulnerabilities that Undermine Objective Decision-Making in LLM Collectives ▶ *Underline Video*

Changgeon Ko, Jisu Shin, Hoyun Song, Huije Lee, Eui Jun Hwang and Jong C. Park

Large language model (LLM) agents are increasingly acting as human delegates in multi-agent environments, where a representative agent integrates diverse peer perspectives to make a final decision. Drawing inspiration from social psychology, we investigate how the reliability of this representative agent is undermined by the social context of its network. We define four key phenomena—social conformity, perceived expertise, dominant speaker effect, and rhetorical persuasion—and systematically manipulate the number of adversaries, relative intelligence, argument length, and argumentative styles. Our experiments demonstrate that the representative agents accuracy consistently declines as social pressure increases: larger adversarial groups, more capable peers, and longer arguments all lead to significant performance degradation. Furthermore, rhetorical strategies emphasizing credibility or logic can further sway the agent's judgment, depending on the context. These findings reveal that multi-agent systems are sensitive not only to individual reasoning but also to the social dynamics of their configuration, highlighting critical vulnerabilities in AI delegates that mirror the psychological biases observed in human group decision-making.

[Findings] SocraticKG: Knowledge Graph Construction via QA-Driven Fact Extraction ▶ *Underline Video*

Sanghyeok Choi, Woosang Jeon, Kyuseok Yang and Taehyeong Kim

Constructing Knowledge Graphs (KGs) from unstructured text provides a structured framework for knowledge representation and reasoning, yet current LLM-based approaches struggle with a fundamental trade-off: factual coverage often leads to relational fragmentation, while premature consolidation causes information loss. To address this, we propose SocraticKG, an automated KG construction method that introduces question-answer pairs as a structured intermediate representation to systematically unfold document-level semantics prior to triple extraction. By employing 5W1H-guided QA expansion, SocraticKG captures contextual dependencies and implicit relational links typically lost in direct KG extraction pipelines, providing explicit grounding in the source document that helps mitigate implicit reasoning errors. Evaluation on the MINE benchmark demonstrates that our approach effectively addresses the coverage-connectivity trade-off, achieving superior factual retention while maintaining high structural cohesion even as extracted knowledge volume substantially expands. These results highlight that QA-mediated semantic scaffolding plays a critical role in structuring semantics prior to KG extraction, enabling more coherent and reliable graph construction in subsequent stages.

[Findings] SParK-Eval: Evaluating Structure-Aware Knowledge Acquisition in LLMs for Domain Adaptation to Industrial Records ▶ *Underline Video*

Ekant Muljibhai Amin, Yuta Koreeda and Yasuhiro Sogawa

Large Language Models (LLMs) often underperform in domain adaptation for industrial settings, where available corpora are limited and structurally diverse. These corpora frequently include non-natural formats such as tables, entity lists, or bullet-point instructions that hinder effective learning. To understand and improve domain-adaptive pretraining on such data, we introduce SParK-Eval (Structure-aware Parametric Knowledge Evaluation), a framework that constructs question-answer pairs from pretraining data and annotates each with its input structure (e.g., natural sentence, table, list). This enables fine-grained analysis of how input structure affects parametric knowledge acquisition during DAPT. Additionally, we propose a prompt-based input normalization method that converts diverse inputs into coherent natural sentences, providing a reference for isolating structural effects. Our experiments show that LLMs acquire substantially more knowledge from natural sentences than from their structurally non-standard counterparts. These findings underscore the importance of structure-aware evaluation in diagnosing learning challenges and guiding effective domain adaptation strategies.

[Long] Specializing Large Models for Oracle Bone Script Interpretation via Component-Grounded Multimodal Knowledge Augmentation ▶ *Underline Video*

Jianing Zhang, Runan Li, Honglin Pang, Ding Xia, Zhou Zhu, Qian Zhang, Chuntao Li and Xi Yang

Deciphering ancient Chinese Oracle Bone Script (OBS) is a challenging task that offers insights into the beliefs, systems, and culture of the

ancient era. Existing approaches treat decipherment as a closed-set image recognition problem, which fails to bridge the "interpretation gap": while individual characters are often unique and rare, they are composed of a limited set of recurring, pictographic components that carry transferable semantic meanings. To leverage this structural logic, we propose an agent-driven Vision-Language Model (VLM) framework that integrates a VLM for precise visual grounding with an LLM-based agent to automate a reasoning chain of component identification, graph-based knowledge retrieval, and relationship inference for linguistically accurate interpretation. To support this, we also introduce OB-Radix, an expert-annotated dataset providing structural and semantic data absent from prior corpora, comprising 1,022 character images (934 unique characters) and 1,853 fine-grained component images across 478 distinct components with verified explanations. By evaluating our system across three benchmarks of different tasks, we demonstrate that our framework yields more detailed and precise decipherments compared to baseline methods.

[Findings] Stable and Explainable Personality Trait Evaluation in Large Language Models with Internal Activations ► *Underline Video*

Xiaoxu Ma, Xiangbo Zhang and Zhenyu Weng

Evaluating personality-related tendencies in Large Language Models (LLMs) helps characterize model behavior, compare models beyond task accuracy, and support responsible deployment in socially interactive settings. However, existing questionnaire-based evaluation methods exhibit limited stability and offer little explainability, as their results are highly sensitive to minor variations in prompt phrasing or role-play configurations. To address these limitations, we propose an internal-activation-based approach, termed Persona-Vector Neutrality Interpolation (PVNI), for stable and explainable personality trait evaluation in LLMs. PVNI extracts a persona vector associated with a target personality trait from the model's internal activations using contrastive prompts. It then estimates the corresponding neutral score by interpolating along the persona vector as an anchor axis, enabling an interpretable comparison between the neutral prompt representation and the persona direction. We provide a theoretical analysis of the effectiveness and generalization properties of PVNI. Extensive experiments across diverse LLMs demonstrate that PVNI yields substantially more stable personality trait evaluations than existing methods, even under questionnaire and role-play variants.

[TACL] Start Making Sense(s): A Developmental Probe of Attention Specialization Using Lexical Ambiguity ► *Underline Video*

Pamela Riviere and Sean Trott

Transformer language model (LM) contextualization mechanisms remain poorly understood. Here, we leverage lexical ambiguity, where a single word has multiple meanings, to isolate attention mechanisms that contribute to word sense disambiguation. We take a "'developmental' approach: first, using publicly available checkpoints for the Pythia models, we identify inflection points in disambiguation performance across each model in the suite; in 14M and 410M specifically, we identify attention heads whose attention to disambiguating words covaries with overall disambiguation performance across development. We then stress-test the robustness of these heads to stimulus perturbations: in 14M, we find limited robustness, but in 410M, we identify multiple heads with surprisingly generalizable behavior, including the part-of-speech of the disambiguating cue. Then, in a causal analysis, we find that ablating the target heads demonstrably impairs disambiguation performance, particularly in 14M. The developmental analyses of 14M are also reproduced across all random seeds of the model. Together, these results suggest: first, that disambiguation benefits from a constellation of mechanisms, some of which (especially in 14M) are highly sensitive to the position and part-of-speech of the disambiguating cue; and second, that larger models (410M) may contain heads with more robust disambiguation behavior. They also join a growing body of work that highlights the value of adopting a developmental perspective when probing LM mechanisms.

[Long] Subject-level Inference for Realistic Text Anonymization Evaluation ► *Underline Video*

Myeong Seok Oh, Dong-Yun Kim, Hanseok Oh, Chaean Kang, Joeun Kang, Xiaonan Wang, Hyunjung Park, Young Cheol Jung and Hansaem Kim

Current text anonymization evaluation relies on span-based metrics that fail to capture what an adversary could actually infer, and assumes a single data subject, ignoring multi-subject scenarios. To address these limitations, we present SPIA (Subject-level PII Inference Assessment), the first benchmark that shifts the unit of evaluation from text spans to individuals, comprising 675 documents across legal and online domains with novel subject-level protection metrics. Extensive experiments show that even when over 90% of PII spans are masked, subject-level inference protection drops as low as 33%, leaving the majority of personal information recoverable through contextual inference. Furthermore, target-subject-focused anonymization leaves non-target subjects substantially more exposed than the target subject. We show that subject-level inference-based evaluation is essential for ensuring safe text anonymization in real-world settings.

[Long] SURE or Not? Investigating Semantic Understanding in Dense Retrieval Models ► *Underline Video*

Lingdi Kong, Xuanang Chen, Ben He and Le Sun

Dense retrieval has become a core technique in applications like web search and retrieval-augmented generation. Despite their empirical success, it remains unclear whether these models truly understand semantics. To address this gap, this paper conducts a systematic investigation by introducing SURE, a benchmark for Semantic Understanding in dense REtrieval built upon the MSMARCO, NQ, and FiQA datasets. SURE characterizes semantic understanding in dense retrieval along three dimensions: semantic precision, semantic abstraction, and semantic equivalence. We evaluate ten representative models ranging from 110M to 8B parameters, including both general-purpose and domain-specific models. Results show that current dense retrievers struggle to distinguish fine-grained semantic differences across texts with varying information density, and to recognize semantic consistency under lexical paraphrasing. Moreover, larger models do not necessarily

exhibit stronger semantic understanding, and diverse training data generally enhances semantic understanding on challenging retrieval tasks.

[Long] Survey Response Generation: Generating Closed-Ended Survey Responses In-Silico with Large Language Models ▶ *Underline Video*

Georg Ahnert, Anna-Carolina Haensch, Barbara Plank and Markus Strohmaier

Many _in-silico_ simulations of human survey responses with large language models (LLMs) focus on generating closed-ended survey responses, whereas LLMs are typically trained to generate open-ended text instead. Previous research has used a diverse range of methods for generating closed-ended survey responses with LLMs and a standard practice remains to be identified. In this paper, we systematically investigate the impact that various Survey Response Generation Methods have on predicted survey responses. We present the results of 32 mio. simulated survey responses across 8 Survey Response Generation Methods, 4 political attitude surveys, and 10 open-weight language models. We find significant differences between the Survey Response Generation Methods in both individual-level and subpopulation-level alignment. Our results show that Restricted Generation Methods perform best overall, and that reasoning output does not consistently improve alignment. Our work underlines the significant impact that Survey Response Generation Methods have on simulated survey responses, and we develop practical recommendations on the application of Survey Response Generation Methods.

[Long] Sycophants in the Courtroom: Are LLMs Fragile to Juridical Authority and Evolving Legal Standards? ▶ *Underline Video*

Lorenzo Molfetta, Alessio Cocchieri, Luca Ragazzi, Ilaria Bartolini, Marco Patella and Gianluca Moro

In medicine, claims remain valid when supported by empirical evidence grounded in stable biological reality. In law, by contrast, truth is contingent, defined by jurisdiction, temporal validity, and the hierarchy of authoritative sources. The recent success of large language models (LLMs) on medical licensing examinations has encouraged an expectation of comparable legal competence. This analogy, however, obscures a critical distinction between domains. Unlike in medicine, legal performance often depends less on inference than on determining when external authority is applicable, valid, and non-contradictory. We introduce a comparative diagnostic framework evaluating legal reasoning against medical baselines along four axes (knowledge recall, grounding, confidence, and robustness), uncovering a sharp domain asymmetry when applied to a new benchmark that encodes temporal validity and normative relationships. While medical LLMs reliably benefit from verified sources, legal LLMs struggle to assess when retrieved citations are useful or misleading, exhibiting overconfidence in perturbed contexts and sensitivity to superficial formatting cues. Increased model scale amplifies this tendency, revealing that stronger instruction following can coincide with weaker resistance to authoritative perturbations. These findings show that LLMs treat law as unstructured text rather than binding precedent, while revealing a tendency to over-trust authoritative but false information when external references conflict with a model's internal knowledge.

[Long] Tackling the Root of Misinformation by Teaching Laypeople about Logical Fallacies via Socratic Questioning and Critical Argumentation ▶ *Underline Video*

Minjing Shi, Junling Wang, Jingwei Ni, Sankalan Pal Chowdhury and Mrinmaya Sachan

Identifying logical fallacies (LFs) in everyday discourse is challenging for many people. This challenge is amplified in the era of Large Language Models (LLMs), where malicious agents can deploy fallacious arguments to disseminate misinformation at scale. In this work, we explore the potential of LLMs as part of the solution. We introduce LFTutor, an intelligent tutoring system which uses LLMs to tutor humans and help them learn about logical fallacies. LFTutor integrates intent-driven Socratic questioning and critical argumentation principles to actively engage learners to reflect on their reasoning. Through both automatic and human evaluations, we demonstrate that LFTutor significantly outperforms baseline LLMs lacking such pedagogical strategies. This work highlights the promise of combining LLMs with pedagogical scaffolding to foster critical thinking and argument literacy in the age of AI.

[Long] Text-to-Distribution Prediction with Quantile Tokens and Neighbor Context ▶ *Underline Video*

Yilun Zhu, Yuan Zhuang, Nikhita Vedula, Dushyanta Dhyani, Shaoyuan Xu, Moyan Li, Mohsen Bayati, Bryan Wang and Shervin Malmasi

Many applications of LLM-based text regression require predicting a full conditional distribution rather than a single point value. We study distributional regression under empirical-quantile supervision, where each input is paired with multiple observed quantile outcomes, and the target distribution is represented by a dense grid of quantiles. We address two key limitations of current approaches: the lack of local grounding for distribution estimates, and the reliance on shared representations that create an indirect bottleneck between inputs and quantile outputs. In this paper, we introduce Quantile Token Regression, which, to our knowledge, is the first work to insert dedicated quantile tokens into the input sequence, enabling direct input-output pathways for each quantile through self-attention. We further augment these quantile tokens with retrieval, incorporating semantically similar neighbor instances and their empirical distributions to ground predictions with local evidence from similar instances. We also provide the first theoretical analysis of loss functions for quantile regression, clarifying which distributional objectives each optimizes. Experiments on the Inside Airbnb and StackSample benchmark datasets with LLMs ranging from 1.7B to 14B parameters show that quantile tokens with neighbors consistently outperform baselines (4 points lower MAPE and 2x narrower prediction intervals), with especially large gains on smaller and more challenging datasets where quantile tokens produce substantially sharper and more accurate distributions.

[Findings] The Overlooked Role of Graded Relevance Thresholds in Multilingual Dense Retrieval ▶ *Underline Video*

Tomer Wullach, Ori Shapira and Amir David Nissan Cohen

Dense retrieval models are typically fine-tuned with contrastive learning objectives that require binary relevance judgments, even though

relevance is inherently graded. We analyze how graded relevance scores and the threshold used to convert them into binary labels affect multilingual dense retrieval. Using a multilingual dataset with LLM-annotated relevance scores, we examine monolingual, multilingual mixture, and cross-lingual retrieval scenarios. Our findings show that the optimal threshold varies systematically across languages and tasks, often reflecting differences in resource level. A well-chosen threshold can improve effectiveness, reduce the amount of fine-tuning data required, and mitigate annotation noise, whereas a poorly chosen one can degrade performance. We argue that graded relevance is a valuable but underutilized signal for dense retrieval, and that threshold calibration should be treated as a principled component of the fine-tuning pipeline.

[Long] Tiny Scales, Great Challenges: The Limits of Multimodal LLMs in Scale Recognition ▶ *Underline Video*

Jihang Jin, Ronghao Chen, Hao Zhang, Ziyan Liu, Huacan Wang, Qi Ye and Jingping Liu

Visual scale recognition is a fundamental aspect for humans to perceive physical quantities in the real world, and it is crucial for enabling human-like intelligence in multimodal large language models (MLLMs). However, existing benchmarks typically focus on a single type of quantity (e.g., time) or a specific format (e.g., dials), lacking a comprehensive evaluation of scale recognition capabilities. To address these problems, we propose ScaleBench, a visual scale recognition benchmark built using images from COCO, Open Images, and Flickr, designed to comprehensively evaluate the scale recognition capabilities of MLLMs. To ensure high data quality, we develop detailed annotation guidelines and procedures, resulting in a total of 6,574 annotated samples. Based on this benchmark, we evaluate multiple closed-source and open-source MLLMs. Experimental results reveal that the best-performing model achieves only 42.60% accuracy, far lower than the 97.40% of humans. Furthermore, we conduct in-depth experimental analyses and provide future research directions. Our benchmark and implementation codes are available at https://github.com/Sonder-hang/ScaleBench.

[Findings] Topology Matters: Measuring Memory Leakage in Multi-Agent LLMs ▶ *Underline Video*

Jinbo Liu, Defu Cao, Yifei Wei, Tianyao Su, Yuan Liang, Yushun Dong, Yan Liu, Yue Zhao and Xiyang Hu

Graph topology is a fundamental determinant of memory leakage in multi-agent LLM systems, yet its effects remain poorly quantified. We introduce MAMA (Multi-Agent Memory Attack), a controlled evaluation framework for comparing topology-conditioned memory leakage in multi-agent LLM systems. MAMA operates on synthetic documents containing labeled Personally Identifiable Information (PII) entities, from which we generate sanitized task instructions. We execute a two-phase protocol: Engram (seeding private information into a target agent's memory) and Resonance (multi-round interaction where an attacker attempts extraction). Over 10 rounds, we measure leakage using a two-stage recovery criterion that combines exact-match extraction with LLM-based inference over the attacker's final output. We evaluate six canonical topologies (complete, circle, chain, tree, star, star-ring) across n4,5,6, attacker-target placements, and base models. Results are consistent: denser connectivity, shorter attacker-target distance, and higher target centrality increase leakage; most leakage occurs in early rounds and then plateaus; model choice shifts absolute rates but preserves broad structural trends; spatiotemporal/location attributes leak more readily than identity credentials or regulated identifiers. We distill practical guidance for system design: favor sparse or hierarchical connectivity, maximize attacker-target separation, and restrict hub/shortcut pathways via topology-aware access control. Our code is available at https://github.com/llll1121/mama-eval.

[Long] Towards Trustworthy Smart Contract Synthesis: A Multi-Agent Framework with Lean-Based Verification ▶ *Underline Video*

Bowei Zhang, Hanbing Liu, Qixin Tian, Siyu Chen, Ziyuan Wang and Qi Qi

Smart Contracts are the foundation of Decentralized Finance (DeFi), executing financial logic without trusted intermediaries. Recent advances in large language models (LLMs) have substantially lowered the barrier to smart contract development by enabling code generation from natural language. However, because smart contracts are immutable and directly manage financial assets, this accessibility introduces a critical trust gap: generated contracts are easy to produce but hard to trust. To bridge this gap, we present LeVer, the first trustworthy smart contract synthesis framework that integrates LLM-based generation with Lean-based auto-formalization and Verification. LeVer employs a closed-loop multi-agent architecture to iteratively generate, verify, attack, and repair contracts, providing both formal guarantees and empirical robustness. To facilitate the adoption of automated formal verification in smart contract generation and audition, we open-source our framework and datasets at: https://github.com/gl-bowei/LeVer

[Findings] Towards Understanding the Robustness of Sparse Autoencoders ▶ *Underline Video*

Ahson Saiyed, Sabrina Sadiekh and Chirag Agarwal

Large Language Models (LLMs) remain vulnerable to optimization-based jailbreak attacks that exploit internal gradient structure. While Sparse Autoencoders (SAEs) are widely used for interpretability, their robustness implications remain underexplored. We present a study of integrating pretrained SAEs into transformer residual streams at inference time, without modifying model weights or blocking gradients. Across four model families (Gemma, LLaMA, Mistral, Qwen) and two strong white-box attacks (GCG, BEAST) plus three black-box benchmarks, SAE-augmented models achieve up to a 5x reduction in jailbreak success rate relative to the undefended baseline and reduce cross-model attack transferability. Parametric ablations reveal (i) a monotonic dose-response relationship between L0 sparsity and attack success rate, and (ii) a layer-dependent defense-utility tradeoff, where intermediate layers balance robustness and clean performance. These findings are consistent with a representational bottleneck hypothesis: sparse projection reshapes the optimization geometry exploited by jailbreak attacks.

[Findings] UbuntuGuard: A Culturally-Grounded Policy Benchmark for Equitable AI Safety in African Languages.

► *Underline Video*

Tassallah Abdullahi, Macton Mgonzo, Mardiyyah Oduwole, Paul Okewunmi, Abraham Owodunni, Ritambhara Singh and Carsten Eickhoff

Current guardian models are predominantly Western-centric and optimized for high-resource languages, leaving low-resource African languages vulnerable to evolving harms, cross-lingual failures, and cultural misalignment. Moreover, most guardian models rely on rigid, predefined safety categories that fail to generalize across diverse linguistic and sociocultural contexts. Achieving robust safety requires flexible, runtime-enforceable policies and benchmarks that reflect local norms, harm scenarios, and cultural expectations. We introduce UbuntuGuard, the first policy-based safety benchmark for African languages built from adversarial queries authored by 155 domain experts across sensitive fields, including healthcare. From these expert-crafted queries, we derive context-specific safety policies and reference responses that capture culturally grounded risk signals, enabling policy-aligned evaluation of guardian models. We evaluate 15 models, comprising seven general-purpose LLMs and eight guardian models across three distinct variants: static, dynamic, and multilingual. Our findings reveal that existing English-centric benchmarks overestimate real-world multilingual safety, cross-lingual transfer provides partial but insufficient coverage, and dynamic models, while better equipped to leverage policies at inference time, still struggle to fully localize African-language contexts. These findings highlight the urgent need for multilingual, culturally grounded safety benchmarks to enable the development of reliable and equitable guardian models for low-resource languages.

[Findings] Uncertainty-Aware Contrastive Sentence Embedding With Local Context Representation for Text Classification ► *Underline Video*

Han Liu, Jiaqing Zhan, Zhichao Chen and Qin Zhang

In real-world applications of natural language processing, it is essential to effectively adapt a pre-trained model to a downstream task. While text classification is undertaken as a downstream task, it is crucial to produce meaningful sentence embedding that is adaptive to the task. In this paper, we explore how to effectively adapt a pre-trained model for extracting meaningful context representations from sentences, and propose an uncertainty-aware contrastive sentence embedding approach that involves addressing language ambiguity and inter-class separability for a text classification task. Specifically, we design an end-to-end strategy for driving the process of learning to transform a word embedding matrix into a contextualized sentence vector and to quantify the representation uncertainty of the sentence, while the word embedding matrix is produced by a pre-trained model without fine-tuning, and a label-wise contrastive learning strategy is designed to enhance intra-class compactness and inter-class separability. The results on public data sets show that a considerable improvement of text classification accuracy is achieved by adopting the proposed approach in comparison with using those state-of-the-art methods.

[Long] Uncertainty-Aware Test-Time Search for Optimization Problem Solving ► *Underline Video*

Linlin Yu, Xujiang Zhao, Dong Li, Yanchi Liu, Wei Cheng, Zhengzhang Chen, Chen Zhao, Feng Chen and Haifeng Chen

Automatically solving optimization problems from natural language descriptions with both efficiency and reliability is highly desirable but remains challenging. Language model hallucinations and the limited availability of labeled datasets often result in misaligned formulations, code errors, and feasibility failures We propose UMCTS, an Uncertainty-aware Monte Carlo Tree Search framework that combines the language understanding capability of large language models with the reliability of well-established solvers. UMCTS structures the solution process into four stages: global instruction, assumptions, mathematical formulation, and solver code generation. It employs Monte Carlo Tree Search with semantic-equivalence pruning, prior-guided exploration, and solver-based feasibility checks. An LLM judge provides numerical reward signals, qualitative error information, and uncertainty estimates. These signals are backpropagated to guide the search and flag unreliable outputs. Across six public benchmarks, UMCTS achieves state-of-the-art solution accuracy, improves efficiency by reducing token usage.

[Long] UPDESH: Synthesizing Grounded Instruction Tuning Data for 13 Indic Languages ► *Underline Video*

Pranjal Chitale, Varun Gumma, Sanchit Ahuja, Prashant Kodali, Manan Uppadhyay, Deepthi Sudharsan and Sunayana Sitaram

Developing culturally grounded multilingual AI systems remains challenging, particularly for low-resource languages. While synthetic data offers promise, its effectiveness in multilingual and multicultural contexts is underexplored. We investigate bottom-up synthetic data generation using large open-source LLMs (>= 235B parameters) grounded in language-specific Wikipedia content, complementing dominant top-down translation-based approaches from English. We introduce Updesh, a high-quality large-scale synthetic instruction-following dataset comprising 9.5M data points across 13 Indian languages and English, encompassing diverse reasoning and generative tasks emphasizing on enhancing long-context and multi-turn capabilities while improving alignment with Indian cultural contexts. Comprehensive evaluation using automated metrics and 10K human assessments confirms high data quality. Downstream evaluations performed by fine-tuning models on various datasets and assessing performance across 13 diverse multilingual datasets and model comparative evaluations, demonstrate that models trained on Updesh consistently obtain significant improvements on NLG tasks and remain competitive on NLU tasks. Improvements are most pronounced for low and medium-resource languages, effectively narrowing performance gaps with high-resource languages. Our findings provide empirical evidence that effective multilingual AI development requires multi-faceted, culturally grounded data curation strategies beyond translation-based approaches.

[Findings] UrBLiMP: A Benchmark for Evaluating the Linguistic Competence of Large Language Models in Urdu ► *Underline Video*

Farah Adeeba, Brian Dillon, Hassan Sajjad and Rajesh Bhatt

Evaluating how large language models (LLMs) capture the grammatical structure of low-resource languages remains underexplored. This paper presents the Urdu Benchmark of Linguistic Minimal Pairs (UrBLiMP), a diagnostic suite of 5,696 minimal pairs that contrast grammatical acceptability across ten core syntactic and morpho-syntactic phenomena in Urdu. The dataset is constructed from the Urdu

Treebank and diverse text corpora, and human validation achieves a 96.1% inter-annotator agreement, confirming its reliability. We evaluate twenty one multilingual LLMs, including LLaMA-3-70B and Gemma-3-27B-PT, and additionally assess the proprietary GPT-4o model using grammar-prompting techniques. GPT-4o (grammar-prompted) attains the highest average accuracy (97.4%), reaching near-human performance on regular phenomena such as aspect agreement and ergativity. However, all models continue to struggle with flexible syntactic patterns like word-order variation and long-distance subject-verb agreement. UrBLiMP provides the first controlled evaluation framework for probing morpho-syntactic competence in Urdu and highlights both the progress and remaining challenges of multilingual and proprietary LLMs in low-resource settings.

[Long] Valid Survey Simulations with Limited Human Data: The Roles of Prompting, Fine-Tuning, and Rectification ► *Underline Video*

Stefan Krsteski, Giuseppe Russo, Serina Chang, Robert West and Kristina Gligori

Surveys provide valuable insights into public opinion and behavior, but their execution is costly and slow. Large language models (LLMs) have been proposed as a scalable, low-cost substitute for human respondents, but their outputs are often biased and yield invalid estimates. We study the interplay between synthesis methods that use LLMs to generate survey responses and rectification methods that debias population estimates, and explore how human responses are best allocated between them. Using two panel surveys with questions on nutrition, politics, and economics, we find that synthesis alone introduces substantial bias (24-86%), whereas combining it with rectification reduces bias below 5% and increases effective sample size by up to 14%. Overall, we challenge the common practice of using all human responses for fine-tuning, showing that under a fixed budget, allocating most to rectification results in far more effective estimation.

[Findings] Verified Critical Step Optimization for LLM Agents ► *Underline Video*

Mukai Li, Qingcheng Zeng, Tianqing Fang, Zhenwen Liang, Linfeng Song, Qi Liu, Haitao Mi and Dong Yu

As large language model agents tackle increasingly complex long-horizon tasks, effective post-training becomes critical. Prior work faces fundamental challenges: outcome-only rewards fail to precisely attribute credit to intermediate steps, estimated step-level rewards introduce systematic noise, and Monte Carlo sampling approaches for step reward estimation incur prohibitive computational cost. Inspired by findings that only a small fraction of high-entropy tokens drive effective RL for reasoning, we propose Critical Step Optimization (CSO), which focuses preference learning on verified critical steps, decision points where alternate actions demonstrably flip task outcomes from failure to success. Crucially, our method starts from failed policy trajectories rather than expert demonstrations, directly targeting the policy models weaknesses. We use a process reward model (PRM) to identify candidate critical steps, leverage expert models to propose high-quality alternatives, then continue execution from these alternatives using the policy model itself until task completion. Only alternatives that the policy successfully executes to correct outcomes are verified and used as DPO training data, ensuring both quality and policy reachability. This yields fine-grained, verifiable supervision at critical decisions while avoiding trajectory-level coarseness and step-level noise. Experiments on GAIA-Text-103 and XBench-DeepSearch show that CSO achieves 37% and 26% relative improvement over the SFT baseline and substantially outperforms other post-training methods, while requiring supervision at only 16% of trajectory steps. This demonstrates the effectiveness of selective verification-based learning for agent post-training.

[Findings] VLURes: Benchmarking Long-Text Grounding and Cross-Lingual Robustness in Vision Language Models ► *Underline Video*

Jesse Atuhurra, Iqra Ali, Tomoya Iwakura, Hidetaka Kamigaito and Tatsuya Hiraoka

We introduce VLURes, a multilingual benchmark for evaluating Vision-Language Models (VLMs) under long-text grounding: selecting and reasoning over the image-relevant subset of article-length text that contains distractors and ungrounded claims. VLURes contains 4,000 web-curated image + long-text pairs across English (En), Japanese (Ja), Swahili (Sw), and Urdu (Ur) and 10 topical categories, and defines eight tasks spanning image-only perception (OR, SU, RU, SS, IC) and image+text grounding (ITM, Unrelatedness, VQA). To construct web-realistic pairs, we apply language-adapted CLIP alignment to select representative images and filter weakly grounded pages. Across 10 proprietary and open VLMs evaluated under zero-shot and one-shot prompting, with and without rationales, the best model (GPT-4o) reaches 90.8% overall accuracy but remains 6.7 points below human performance (97.5%) on Object Recognition, and cross-lingual sensitivity persists, while open models are substantially weaker and often lack reliable multilingual VL support. VLURes provides a practical testbed for long-text grounding and multilingual robustness in web-realistic agent settings.

[Long] VRPO: Rethinking Value Modeling for Robust RL under Noisy Supervision in LLM Post-Training ► *Underline Video*

Dingwei Zhu, Shihan Dou, Zhiheng Xi, Senjie Jin, Guoqiang Zhang, Jiazheng Zhang, Junjie Ye, Mingxu Chai, Enyu Zhou, Ming Zhang, Yuhui Wang, Caishuang Huang, Chenhao Huang, Yunke Zhang, Yuran Wang, Tao Gui, Qi Zhang, Xipeng Qiu and Xuanjing Huang

Reinforcement Learning (RL) in real-world environments often suffers from ambiguous or incomplete reward supervision, which undermines policy stability and generalization. Such noise may cause models to ignore key information or even collapse in advantage estimation. We find that a strong value model is essential for absorbing unstable signals and producing reliable advantages, offering denser and more robust supervision than the reward model. To better optimize noisy supervision, we propose VRPO, a framework that enhances value modeling for robust RL in LLM post-training. VRPO integrates (1) auxiliary losses guided by entropy and perplexity from a frozen language model, and (2) a variational information bottleneck, enabling the value model to filter noise and capture key words. This design allows the value model to correct noise rewards and generate more reliable advantage estimates, transforming it from a passive predictor into an active noise regulator. Experiments on multi-turn dialogue, math reasoning, and science QA with both rule-based and model-based rewards show that

VRPO consistently outperforms baselines such as PPO and GRPO. Our work highlight the central role of the value model in Robust RL and provide a principled and practical approach to policy optimization under noisy supervision.

[Long] What About the Scene With the Hitler Reference? HAUNT: A Framework to Probe LLMs Self-consistency in Closed Domains Via Adversarial Nudge ▶ *Underline Video*

Arka Dutta, Sujan Dutta, Rijul Magu, Soumyajit Datta, Munmun De Choudhury and Ashiqur R. KhudaBukhsh

Hallucinations pose a critical challenge to the real-world deployment of large language models (LLMs) in high-stakes domains. In this paper, we present a framework for stress testing factual fidelity in LLMs in the presence of adversarial nudge. Our framework consists of three steps. First, we instruct the LLM to produce sets of truths and lies consistent with the closed domain in question. Next, we instruct the LLM to verify the same set of assertions as truths and lies consistent with the same closed domain. Finally, we test the robustness of the LLM against the lies generated (and verified) by itself. Our extensive evaluation, conducted using five widely known proprietary and six open LLMs across two closed domains of popular movies and novels, reveals a wide range of susceptibility to adversarial nudges: even among the strongest proprietary LLMs, Claude exhibits strong resilience, GPT and Grok demonstrate moderate resilience, while Gemini and DeepSeek show weak resilience and open models fall short significantly.

[Long] Which Reasoning Trajectories Teach Students to Reason Better? A Simple Metric of Informative Alignment ▶ *Underline Video*

Yuming Yang, Mingyoung Lai, Wanxu Zhao, Xiaoran Fan, Zhiheng Xi, Mingqi Wu, Chiyue Huang, Jun Zhao, Haijun Lv, Tong Jian, Yunhua Zhou, Yicheng Zou, Qipeng Guo, Tao Gui, Qi Zhang and Xuanjing Huang

Long chain-of-thought (CoT) trajectories provide rich supervision signals for distilling reasoning from teacher to student LLMs. However, both prior work and our experiments show that trajectories from stronger teachers do not necessarily yield better students, highlighting the importance of data-student suitability in distillation. Existing methods assess suitability primarily through student likelihood, favoring trajectories that align closely with the student model's current behavior but overlooking more informative ones. Addressing this, we propose Rank-Surprisal Ratio (RSR), a simple metric that captures both alignment and informativeness to assess the suitability of a reasoning trajectory. RSR is motivated by the observation that effective trajectories typically balance learning signal strength and behavioral alignment by combining low absolute probability with relatively high-ranked tokens under the student model. Concretely, RSR is defined as the ratio of a trajectory's average token-wise rank to its average negative log-likelihood, and is straightforward to compute and interpret. Across five student models and reasoning trajectories from 11 diverse teachers, RSR strongly correlates with post-training reasoning performance (average Spearman 0.86), consistently outperforming existing metrics. We further demonstrate its practical utility in both trajectory selection and teacher selection.

[Long] Why Mean Pooling Works: Quantifying Second-Order Collapse in Text Embeddings ▶ *Underline Video*

Tomomasa Hara, Hiroto Kurita, Masaaki Imaizumi, Kentaro Inui and Sho Yokoi

For constructing text embeddings, mean pooling, which averages token embeddings, is the standard approach. This paper examines whether mean pooling actually works well in real models. First, we note that mean pooling can collapse information beyond the first-order statistics of the token embeddings, such as second-order statistics that capture their spatial structure, potentially mapping distinct token embedding distributions to similar text embeddings. Motivated by this concern, we propose a simple metric to quantify such a collapse induced by mean pooling. Then, using this metric, we empirically measure how often this collapse occurs in actual models and texts, and find that modern text encoders are robust to this collapse. In particular, contrastive fine-tuned text encoders tend to be less prone to the collapse than their pretrained backbone models. We also find that the robustness of these text encoders lies in the concentration of token embeddings within each text. In addition, we find that robustness to the collapse, as quantified by our proposed metric, correlates with downstream task performance. Overall, our findings offer a new perspective on why modern text encoders remain effective despite relying on seemingly coarse mean pooling.

[Long] YIELD: A Large-Scale Dataset and Evaluation Framework for Information Elicitation Agents ▶ *Underline Video*

Victor De Lima and Grace Hui Yang

Most conversational agents (CAs) are designed to satisfy user needs through user-driven interactions. However, many real-world settings, such as academic interviewing, judicial proceedings, and journalistic investigations, involve broader institutional decision-making processes and require agents that can elicit information from users. In this paper, we introduce Information Elicitation Agents (IEAs) in which the agents goal is to elicit information from users to support the agent's institutional or task-oriented objectives. To enable systematic research on this setting, we present YIELD, a 26M-token dataset of 2,281 ethically sourced, human-to-human dialogues. Moreover, we formalize information elicitation as a finite-horizon POMDP and propose novel metrics tailored to IEAs. Pilot experiments on multiple foundation LLMs show that training on YIELD improves their alignment with real elicitation behavior and findings are corroborated by human evaluation. We release YIELD under CC BY 4.0. The dataset, project code, evaluation tools, and fine-tuned model adapters are available at: https://github.com/infosenselab/yield.

Session 17: Oral/Poster/Demo G - 11:00-12:30

Demo Session G

11:00-12:30 (Grand Hall)

[Demo] A Dynamic Self-Evolving Extraction System ▶ *Underline Video*

Moin Aminnaseri, Hannah Kim and Estevam Hruschka

The extraction of structured information from raw text is a fundamental component of many NLP applications, including document retrieval, ranking, and relevance estimation. High-quality extractions often require domain-specific accuracy, up-to-date understanding of specialized taxonomies, and the ability to incorporate emerging jargon and rare outliers. In many domains–such as medical, legal, and HR–the extraction model must also adapt to shifting terminology and benefit from explicit reasoning over structured knowledge. We propose DySECT, a Dynamic Self-Evolving Extraction and Curation Toolkit, which continually improves as it is used. The system incrementally populates a versatile, self-expanding knowledge base (KB) with triples extracted by the LLM. The KB further enriches itself through the integration of probabilistic knowledge and graph-based reasoning, gradually accumulating domain concepts and relationships. The enriched KB then feeds back into the LLM extractor via prompt tuning, sampling of relevant few-shot examples, or fine-tuning using KB-derived synthetic data. As a result, the system forms a symbiotic closed-loop cycle in which extraction continuously improves knowledge, and knowledge continuously improves extraction.

[Demo] AgentFactory: A Self-Evolving Framework Through Executable Subagent Accumulation and Reuse ▶ *Underline Video*

Zhang Zhang; Shuqi Lu; Hongjin Qian; Di He; Zheng Liu;

Building LLM-based agents has become increasingly important. Recent works on LLM-based agent self-evolution primarily record successful experiences as textual prompts or reflections, which cannot reliably guarantee efficient task re-execution in complex scenarios. We propose AgentFactory, a new self-evolution paradigm that preserves successful task solutions as executable subagent code rather than textual experience. Crucially, these subagents are continuously refined based on execution feedback, becoming increasingly robust and efficient as more tasks are encountered. Saved subagents are pure Python code with standardized documentation, enabling portability across any Python-capable system. We demonstrate that AgentFactory enables continuous capability accumulation: its library of executable subagents grows and improves over time, progressively reducing the effort required for similar tasks without manual intervention. Our implementation is open-sourced at https://github.com/zzatpku/AgentFactory, and our demonstration video is available at https://youtu.be/iKSsuAXJHW0.

[Demo] AI Steerability 360: A Toolkit for Steering Large Language Models ▶ *Underline Video*

Pierre Dognin

The AI Steerability 360 toolkit is an extensible, opensource Python library for steering LLMs. Steering abstractions are designed around four model control surfaces: input (modification of the prompt), structural (modification of the model's weights or architecture), state (modification of the model's activations and attentions), and output (modification of the decoding or generation process). Steering methods exert control on the model through a common interface, termed a steering pipeline, which additionally allows for the composition of multiple steering methods. Comprehensive evaluation and comparison of steering methods/pipelines is facilitated by use case classes (for defining tasks) and a benchmark class (for performance comparison on a given task). The functionality provided by the toolkit significantly lowers the barrier to developing and comprehensively evaluating steering methods. The toolkit is Hugging Face native and is released under an Apache 2.0 license at https://github.com/IBM/AISteer360.

[Demo] AutoFigure-Edit: Generating Editable Scientific Illustrations via Reference-Guided Styling ▶ *Underline Video*

Zhen Lin, Qiujie Xie, Minjun Zhu, Shichen Li, Qiyao Sun, Enhao Gu, Yiran Ding, Ke Sun, Fang Guo, Panzhong Lu, Zhiyuan Ning, Yixuan Weng and Yue Zhang

High-quality scientific illustrations are essential for communicating complex scientific and technical concepts, yet existing automated systems remain limited in editability, stylistic controllability, and efficiency. We present AutoFigure-Edit, an end-to-end system that generates fully editable scientific illustrations from long-form scientific text while enabling flexible style adaptation through user-provided reference images. By combining long-context understanding, reference-guided styling, and native SVG editing, it enables efficient creation and refinement of high-quality scientific illustrations. To facilitate further progress in this field, we release the video at https://youtu.be/10IH8SyJjAQ, the full codebase at https://github.com/ResearAI/AutoFigure-Edit and provide a live demo for easy access and interactive use at https://autofigure.cc/.

[Demo] Expert Calibration Lens for Pruning Mixture of Experts ▶ *Underline Video*

Luis Frentzen Salim, Chia-Chun Wu, Tran Van Nhiem, Lun-Wei Ku and Yung-Hui Li

Expert pruning is a practical deployment technique for Mixture-of-Experts (MoE) models. It reduces resource usage and mitigates expert redundancy, but its success depends strongly on the calibration set used for pruning. In domain-general settings, it is unclear which properties of the calibration data drive good pruning outcomes, and the effects of calibration perturbations are often unintuitive. We observe, for example, that calibration sets in different languages can lead to very similar pruning results despite appearing dissimilar on the surface. To address this, we propose Expert Calibration Lens, a lightweight analysis tool that compares expert activation patterns across datasets to predict the impact of calibration perturbations without repeatedly running expensive pruning procedures. We use activations that are quick to compute and evaluate the resulting analysis for downstream task performance.

[Demo] GAMED.AI: A Hierarchical Multi-Agent Framework for Automated Educational Game Generation ► *Underline Video*
Shiven Agarwal, Yash Shah, Ashish Raj Shekhar, Priyanuj Bordoloi and Vivek Gupta
We introduce GameDAI, a hierarchical multi-agent framework that transforms instructor-provided questions into fully playable, pedagogically grounded educational games validated through formal mechanic contracts. Built on phase-based LangGraph sub-graphs, deterministic Quality Gates, and structured Pydantic schemas, GameDAI supports two template families encompassing 15 interaction mechanics across spatial reasoning, procedural execution, and higher-order Bloom's Taxonomy objectives. Evaluated on 200 questions spanning five subject domains, the system achieves a 90% validation pass rate, 98.3% schema compliance, and 73% token reduction over ReAct agents 73,500 to 19,900 tokens/game) at $0.46 per game. Within this model configuration, these results suggest that phase-bounded architectural structure correlates more strongly with alignment quality than prompting strategy alone. Our demonstration lets attendees generate Bloom's-aligned games from natural language in under 60 seconds, inspect Quality Gate outputs at each pipeline phase, and browse a curated library of 50 games spanning all 15 mechanic types.

[Demo] Media-to-Insights: A Multi-Agent AI System for Continuous Media Monitoring, Analysis, and Reporting ► *Underline Video*
Ashraf Hatim Elneima, Ozan Yilmaz, Hadi Nasrallah, Sanjika Hewavitharana, Mohamed Al-Badrashiny and Hassan Sawaf
Continuous monitoring of high-volume media streams requires systems that go beyond keyword alerts to deliver structured, actionable intelligence. We present a multi-agent media monitoring system that processes streaming articles through three stages: (1) a Matching Agent that uses a hybrid keyword-then-semantic matching approach, reducing agent invocations by \{}~20% (2) a batched multi-agent feature extraction, reducing core feature-extraction calls from 7 to 2 per article - a 71% reduction - with bounded quality tradeoffs; and (3) a Report Generation Agent that uses deterministic deduplication and density-based clustering. Four autonomous life-cycle agents manage the evolution of watchers.

Oral Session G: Hierarchical Structure Prediction, Syntax, and Parsing & Phonology, Morphology and Word Segmentation

11:00-12:30 (Promenade)

[Long] Communicating in Emergent Language with an Induced Morphological Phrasebook ► *Underline Video*
Brendon Boldt and David Mortensen
We build rule-based emergent language (EL) agents using form-meaning mappings induced from ELs ("morphological phrasebooks") and test their communicative performance in the EL environment with its neural network agents. This contributes three things: First, it assesses the quality of the morphemes discovered by the induction algorithm in situ, which we find to be effective for communicating in the EL. Second, it allows us to uncover morphosyntactic properties of EL through ablating the algorithms which induce and utilize morphemes, showing that the ELs rely on repetition as well as morpheme ordering to convey meaning. Third, we find that the normalized pointwise mutual information of forms and meanings in the morphemes serves as a metric of compositionality that is more closely correlated with the ability of the phrasebook-agents to "speak" and "hear" an EL than existing metrics such as topographic similarity.

[Long] Corpus-Dependent Subcharacter Encoding via HMM-Guided Code Assignment ► *Underline Video*
Tatsuya Hiraoka
We propose a corpus-dependent alternative to byte encoding that learns fixed-length atomic codes for characters directly from text, which we refer to as Latom (Learned Atom-based Encoding). We instantiate this framework by training an HMM on N-repeated character sequences to estimate "atom" posteriors, followed by a Hungarian assignment yielding a globally optimal one-to-one character-code mapping. Across 14 languages, the encodings improve intrinsic metrics, including token counts after subword tokenization and bigram perplexity, with appropriate code lengths. On Amazon Reviews in six languages, Latom improves text classification accuracy and reduces decoding errors in language model generation. Overall, these results demonstrate that character encodings can be learned from corpus statistics while remaining reversible and compatible with standard tokenization pipelines.

[Long] GiLT: Augmenting Transformer Language Models with Dependency Graphs ► *Underline Video*
Tianyu Huang, Yida Zhao, Chuyan Zhou and Kewei Tu
Augmenting Transformers with linguistic structures effectively enhances the syntactic generalization performance of language models. Previous work in this direction focuses on syntactic tree structures of languages, in particular constituency tree structures. We propose Graph-Infused Layers Transformer Language Model (GiLT) which leverages dependency graphs for augmenting Transformer language models. Unlike most previous work, GiLT does not insert extra structural tokens in language modeling; instead, it injects structural information into language modeling by modulating attention weights in the Transformer with features extracted from the dependency graph that is incrementally constructed along with token prediction. In our experiments, GiLT with semantic dependency graphs achieves better syntactic generalization while maintaining competitive perplexity in comparison with Transformer language model baselines. In addition, GiLT can be finetuned from a pretrained language model to achieve improved downstream task performance. Our code is released at

https://github.com/cookie-pie-oops/GiLT-LM.

[Long] HiChunk: Evaluating and Enhancing Retrieval Augmented Generation with Hierarchical Chunking ▶ *Underline Video*
Wensheng Lu, Keyu Chen, Zhifeng Shen, Ruizhi Qiao and Xing Sun
Retrieval-Augmented Generation (RAG) enhances the response capabilities of language models by integrating external knowledge sources. However, document chunking as an important part of RAG system often lacks effective evaluation tools. This paper first analyzes why existing RAG evaluation benchmarks are inadequate for assessing document chunking quality, specifically due to evidence sparsity. Based on this conclusion, we propose HiCBench, which includes manually annotated multi-level document chunking points, synthesized evidence-dense question answer(QA) pairs, and their corresponding evidence sources. We also propose HiChunk, a hierarchical document structuring framework using fine-tuned LLMs and the Auto-Merge retrieval algorithm to enhance retrieval quality. Experiments demonstrate that HiCBench effectively evaluates the impact of different chunking methods across the entire RAG pipeline. Moreover, HiChunk achieves better chunking quality within reasonable time consumption, thereby enhancing the overall performance of RAG systems. Source code is available at https://github.com/TencentCloudADP/hichunk.

[Long] HiKEY: Hierarchical Multimodal Retrieval for Open-Domain Document Question Answering ▶ *Underline Video*
Joongmin Shin, Gyuho Shim, Jeongbae Park, Jaehyung Seo and Heuiseok Lim
Retrieval-augmented generation (RAG) for document-based Open-domain Question Answering (ODQA) on large-scale industrial corpora faces two critical bottlenecks: routing failure in locating the correct document and evidence fragmentation in integrating scattered information. Existing approaches relying on flat text chunks or page-level images inherently struggle to (i) precisely pinpoint the target document among thousands of candidates and (ii) organically connect multimodal evidence, such as tables and figures, within a limited token budget. To address these challenges, we propose HiKEY, a hierarchical tree-based multimodal retrieval framework that elevates document hierarchy to a first-class retrieval signal. Instead of simple chunking, HiKEY reconstructs a logical heterogeneous graph via Document Hierarchical Parsing (DHP), explicitly encoding parent-child relationships. Adopting a hierarchical coarse-to-fine strategy, the framework (1) performs global routing to rapidly prune the search space using hierarchical indexing, and (2) conducts fine-grained retrieval to rank sections by employing a multimodal fusion strategy that captures the most discriminative evidence. Finally, HiKEY assembles a token-efficient evidence subgraph via a hybrid structural-semantic packing strategy. Experiments on ODQA benchmarks demonstrate that HiKEY significantly outperforms page- and chunk-based baselines, improving retrieval recall by up to 12.9% and end-to-end QA performance by up to 6.8%.

[Long] How Tokenization Limits Phonological Knowledge Representation in Language Models and How to Improve Them ▶ *Underline Video*
Disen Liao and Freda Shi
Tokenization is the first step in every language model (LM), yet it never takes the sounds of words into account. We investigate how tokenization influences text-only LMs' ability to represent phonological knowledge. Through a series of probing experiments, we show that subword-based tokenization systematically weakens the encoding of both local (e.g., rhyme) and global (e.g., syllabification) phonological features. To quantify this effect, we introduce the syllabification-tokenization alignment distance (STAD), a metric that measures the misalignment between a model's tokenization and the natural syllable boundaries of words, and find that higher misalignment correlates with poorer phonological representations, providing a simple diagnostic for phonology-aware tokenization. To address these limitations, we propose a lightweight IPA-based fine-tuning method that infuses phonological awareness into LMs, leading to consistent improvements across three phonology-related tasks while largely preserving math and general reasoning ability, with 1.1% and 0.9% drops on GSM8K and MMLU, respectively.

[Long] Modal Dependency Parsing as Structured Prediction over Source-Cue Scope ▶ *Underline Video*
Jayeol Chun and Nianwen Xue
Modal dependency parsing-the task of identifying a semantic graph that represents who is responsible for an event-centered claim and with what degree of certainty-relies on recognizing source-introducing cues and correctly linking them to their associated content. However, prior work has largely focused on identifying sources only, treating cue expressions and their modal coverage as auxiliary signals. In this work, we propose a structured prediction framework that leverages large language models (LLMs) to explicitly identify source-cue pairs as well as their respective scope, which together define the modal contexts governing downstream source attribution for events. By concentrating learning at the source-cue level and constraining event-level decisions to a small, scope-defined candidate set, our top-down approach enables more efficient inference in long, event-rich documents. Experiments show this approach surpasses prior state-of-the-art results by 3 and 4% for English and Chinese datasets, respectively.

[Short] Prefix Parsing is Just Parsing ▶ *Underline Video*
Clemente Pasti, Andreas Opedal, Timothy J. O' Donnel, Ryan Cotterell and Tim Vieira
Prefix parsing asks whether an input prefix can be extended to a complete string generated by a given grammar. In the weighted setting, it also provides prefix probabilities, which are central to context-free language modeling, psycholinguistic analysis, and syntactically constrained generation from large language models. We introduce the prefix grammar transformation, an efficient reduction of prefix parsing to ordinary parsing. Given a grammar, our method constructs another grammar that generates exactly the prefixes of its original strings. Prefix parsing is then solved by applying any ordinary parsing algorithm on the transformed grammar without

modification. The reduction is both elegant and practical: the transformed grammar is only a small factor larger than the input, and any optimized implementation can be used directly, eliminating the need for bespoke prefix-parsing algorithms. We also present a strategy, based on algorithmic differentiation, for computing the next-token weight vector, i.e., the prefix weights of all one-token extensions, enabling efficient prediction of the next token. Together, these contributions yield a simple, general, and efficient framework for prefix parsing.

[Long] SPEAK: Spiking Neurons as an Entropy-Aware Tokenizer for Large Language Models ▶ *Underline Video*

Ming Chen, Wenyao Li, Chao Liang, Shi Gu, Peng Lin, De Ma, Huajin Tang, Qian Zheng and Gang Pan

Tokenizers play a critical role in large language model studies. Despite recent advances, existing tokenizers fail to explicitly leverage historical tokenization results when making subsequent token decisions, nor do they selectively utilize such history based on contextual relevance. We propose SPEAK, a tokenizer that integrates spiking neurons to explicitly leverage historical tokenization results. Furthermore, we introduce an entropy-aware reset mechanism that selectively leverages history based on contextual relevance, which is determined by token-level entropy. High-entropy tokens are treated as contextual boundaries, whereas low-entropy tokens between consecutive such boundaries exhibit strong contextual relevance. Accordingly, we induce hard reset at high-entropy tokens to discard irrelevant historical tokenization results, and soft reset at low-entropy tokens to preserve and leverage relevant history. Experiments on 2 language models and 5 datasets spanning 16 languages demonstrate superior cross-lingual adaptability, with competitive performance and efficiency. Our code is publicly available at https://github.com/zju-bmi-lab/SPEAK.

Oral Session G: Human-AI Interaction/Cooperation and Human-Centric NLP 2

11:00-12:30 (Harbor B-C)

[Long] "Excuse me, may I say something..." CoLabScience, A Proactive AI Assistant for Biomedical Discovery and LLM-Expert Collaborations ▶ *Underline Video*

Yang Wu, Jinhong Yu, Jingwei Xiong, Zhimin Tao and Xiaozhong Liu

The integration of Large Language Models (LLMs) into scientific workflows presents exciting opportunities to accelerate biomedical discovery. However, the reactive nature of LLMs, which respond only when prompted, limits their effectiveness in collaborative settings that demand foresight and autonomous engagement. In this study, we introduce CoLabScience, a proactive LLM assistant designed to enhance biomedical collaboration between AI systems and human experts through timely, context-aware interventions. At the core of our method is PULI (Positive-Unlabeled Learning-to-Intervene), a novel framework trained with a reinforcement learning objective to determine when and how to intervene in streaming scientific discussions, by leveraging the team's project proposal and long- and short-term conversational memory. To support this work, we introduce BSDD (Biomedical Streaming Dialogue Dataset), a new benchmark of simulated research discussion dialogues with intervention points derived from PubMed articles. Experimental results show that PULI significantly outperforms existing baselines in both intervention precision and collaborative task utility, highlighting the potential of proactive LLMs as intelligent scientific assistants.

[Long] Believing without Seeing: Quality Scores for Contextualizing Vision-Language Model Explanations ▶ *Underline Video*

Keyu He, Tejas Srinivasan, Brihi Joshi, Xiang Ren, Jesse Thomason and Swabha Swayamdipta

When people query Vision-Language Models (VLMs) but cannot see the accompanying visual context (e.g. for blind and low-vision users), augmenting VLM predictions with natural language explanations can signal which model predictions are reliable. However, prior work has found that explanations can easily convince users that inaccurate VLM predictions are correct. To remedy undesirable overreliance on VLM predictions, we propose evaluating two complementary qualities of VLM-generated explanations via two quality scoring functions. We propose Visual Fidelity, which captures how faithful an explanation is to the visual context, and Contrastiveness, which captures how well the explanation identifies visual details that distinguish the model's prediction from plausible alternatives. On the A-OKVQA, VizWiz, and MMMU-Pro tasks, these quality scoring functions are better calibrated with model correctness than existing explanation qualities. We conduct a user study in which participants have to decide whether a VLM prediction is accurate without viewing its visual context. We observe that showing our quality scores alongside VLM explanations improves participants' accuracy at predicting VLM correctness by 11.1%, including a 15.4% reduction in the rate of falsely believing incorrect predictions. These findings highlight the utility of explanation quality scores in fostering appropriate reliance on VLM predictions.

[Long] Effects of Collaboration on the Performance of Interactive Theme Discovery Systems ▶ *Underline Video*

Alvin Chen, Rohan Das, Dananjay Srinivas, Alexandra Barry, Maksim Seniw and Maria Pacheco

NLP-assisted solutions to support qualitative data analysis have gained considerable traction. However, no unified evaluation framework exists which can account for the many different settings in which qualitative researchers may employ them. In this paper, we propose a framework to evaluate the way collaboration settings may produce different research outcomes across a variety of interactive systems. Specifically, we study the impact of synchronous vs. asynchronous collaboration using three different NLP-assisted qualitative research tools and present a comprehensive analysis of the differences in the consistency, cohesiveness, and correctness of their outcomes.

[Long] False Friends in the Shell: Unveiling the Emoticon Semantic Confusion in Large Language Models ▶ *Underline Video*

Weipeng Jiang, Xiaoyu Zhang, Juan Zhai, Shiqing Ma, Chao Shen and Yang Liu

Emoticons are widely used in digital communication to convey affective intent, yet their safety implications for Large Language Models (LLMs) remain largely unexplored. In this paper, we identify emoticon semantic confusion, a vulnerability where LLMs misinterpret ASCII-based emoticons to perform unintended and even destructive actions. To systematically study this phenomenon, we develop an automated data generation pipeline and construct a dataset containing 3,757 code-oriented test cases spanning 21 meta-scenarios, four programming languages, and varying contextual complexities. Our study on six LLMs reveals that emoticon semantic confusion is pervasive, with an average confusion ratio exceeding 38%. More critically, over 90% of confused responses yield 'silent failures', which are syntactically valid outputs but deviate from user intent, potentially leading to destructive security consequences.Furthermore, we observe that this vulnerability readily transfers to popular agent frameworks, while existing prompt-based mitigations remain largely ineffective. We call on the community to recognize this emerging vulnerability and develop effective mitigation methods to uphold the safety and reliability of human-LLM interactions.

[Long] MeepleLM: A Virtual Playtester Simulating Diverse Subjective Experiences ▶ *Underline Video*

Zizhen Li, Chuanhao Li, Yibin Wang, Jianwen Sun, Yukang Feng, Jiaxin Ai, Fanrui Zhang, Mingzhu Sun, Yifei Huang and Kaipeng Zhang

Recent advancements have expanded the role of Large Language Models (LLMs) in board games from playing agents to creative co-designers. However, a critical gap remains: current systems lack the capacity to offer constructive critique grounded in the emergent user experience. Bridging this gap is fundamental for harmonizing Human-AI collaboration, as it empowers designers to refine their creations via external perspectives while steering models away from biased or unpredictable outcomes. Automating this evaluation presents two challenges: inferring the latent dynamics connecting static rules to gameplay without an explicit engine, and modeling the subjective heterogeneity of diverse player groups. To address these, we curate a comprehensive dataset of 1,727 structurally corrected rulebooks and 150K reviews selected via rigorous quality scoring and facet-aware sampling. We augment this data with Mechanics-Dynamics-Aesthetics (MDA) reasoning to explicitly bridge the causal gap between written rules and player experience. We further distill distinct player personas and introduce MeepleLM, a specialized model that internalizes persona-specific reasoning patterns to accurately simulate the subjective feedback of diverse player archetypes. Extensive experiments demonstrate that MeepleLM significantly outperforms latest commercial models (e.g., GPT-5.1, Gemini3-Pro) in community alignment and critique quality, achieving a 70% preference rate in user studies assessing practical utility. MeepleLM serves as a reliable virtual playtester that provides experience-grounded feedback, offering a practical step towards audience-aligned Human-AI collaboration.

[Long] Modeling Annotator Disagreement with Demographic-Aware Experts and Synthetic Perspectives ▶ *Underline Video*

Yinuo Xu, Veronica Derricks, Allison Earl and David Jurgens

We present an approach to modeling annotator disagreement in subjective NLP tasks through both architectural and data-centric innovations. Our model, DEM-MoE (Demographic-Aware Mixture of Experts), routes inputs to expert subnetworks based on annotator demographics, enabling it to better represent structured, group-level variation compared to prior models. DEM-MoE consistently performs competitively across demographic groups, and shows especially strong results on datasets with high annotator disagreement. To address sparse demographic coverage, we test whether LLM-generated synthetic annotations via zero-shot persona prompting can be used for data imputation. We show these synthetic judgments align moderately well with human annotations on our data and offer a scalable way to potentially enrich training data. We then propose and evaluate approaches for blending real and synthetic data using strategies tailored to dataset structure. We find that the optimal strategies depend on dataset structure. Together, these contributions improve the representation of diverse perspectives.

[Long] SMILE-Next: Teaching Large Language Models to Detect, Classify, and Reason about Laughter ▶ *Underline Video*

Lee Jung-Mok, Kim Sung-Bin, Joohyun Chang, Lee Hyun and Tae-Hyun Oh

Laughter is a complex social signal that conveys communicative intent beyond amusement. While prior work has focused on isolated laughter analysis tasks, a comprehensive understanding of laughter in real-world scenarios remains underexplored. We introduce SMILE-Next, a dataset for real-world laughter understanding with multimodal textual representations and question-answer annotations across three tasks: laughter detection, laughter type classification, and laughter reasoning. Building on this dataset, we propose a laughter expert LLM that leverages disentangled multimodal textual cues, together with a Mixture-of-Laugh-Experts framework and laughter-specific self-instruction for task-adaptive specialization. Experimental results show that the combination of our proposed components substantially outperforms multimodal LLM baselines, advancing robust real-world laughter understanding.

[Long] User Perceptions vs. Proxy LLM Judges: Privacy and Helpfulness in LLM Responses to Privacy-Sensitive Scenarios ▶ *Underline Video*

Xiaoyuan Wu, Roshni Kaushik, Wenkai Li, Lujo Bauer and Koichi Onoue

Large language models (LLMs) are rapidly being adopted for tasks like drafting emails, summarizing meetings, and answering health questions. In these settings, users may need to share private information (e.g., contact details, health records). To evaluate LLMs' ability to identify and redact such information, prior work introduced real-life, scenario-based benchmarks (e.g., ConfAIde, PrivacyLens) and found that LLMs can leak private information in complex scenarios. However, these evaluations relied on proxy LLMs to judge the helpfulness and privacy-preservation quality of LLM responses, rather than directly measuring users' perceptions. To understand how users perceive the helpfulness and privacy-preservation quality of LLM responses to privacy-sensitive scenarios, we conducted a user study (n=94) using

90 PrivacyLens scenarios. We found that users had low agreement with each other when evaluating identical LLM responses. In contrast, five proxy LLMs reached high agreement, yet each proxy LLM had low correlation with users' evaluations. These results indicate that proxy LLMs cannot accurately estimate users' wide range of perceptions of utility and privacy in privacy-sensitive scenarios. We discuss the need for more user-centered studies to measure LLMs' ability to help users while preserving privacy, and for improving alignment between LLMs and users in estimating perceived privacy and utility.

[Long] Value of Information: A Framework for Human-Agent Communication ▶ *Underline Video*
Yijiang River Dong, Tiancheng Hu, Zheng Hui, Caiqi Zhang, Ivan Vuli, Andreea Bobu and Nigel Collier
Large Language Model (LLM) agents deployed for real-world tasks face a fundamental dilemma: user requests are underspecified, yet agents must decide whether to act on incomplete information or interrupt users for clarification. Existing approaches either rely on brittle confidence thresholds that require task-specific tuning, or fail to account for the varying stakes of different decisions. We introduce a decision-theoretic framework that resolves this trade-off through the Value of Information (VoI), enabling agents to dynamically weigh the expected utility gain from asking questions against the cognitive cost imposed on users. Our inference-time method requires no hyperparameter tuning and adapts seamlessly across contexts, from casual games to medical diagnosis. Experiments across four diverse domains (20 Questions, medical diagnosis, flight booking, and e-commerce) show that VoI consistently matches or exceeds the best manually-tuned baselines, achieving up to 1.36 utility points higher in high-cost settings. This work provides a parameter-free framework for adaptive agent communication that explicitly balances task risk, query ambiguity, and user effort.

Oral Session G: Industry 6

11:00-12:30 (Balboa)

[Industry] Are Large Language Models Economically Viable for Industry Deployment? ▶ *Underline Video*
Abdullah Mohammad, Sushant Kumar Ray, Pushkar Arora, Rafiq Ali, Ebad Shabbir, Gautam Siddharth Kashyap, Jiechao Gao and Usman Naseem
Generative AI, powered by Large Language Models (LLMs), is increasingly deployed in industry across healthcare decision support, financial analytics, enterprise retrieval, and conversational automation, where reliability, efficiency, and cost control are critical. In such settings, models must satisfy strict constraints on energy, latency, and hardware utilization, not accuracy alone. Yet prevailing evaluation pipelines remain accuracy-centric, creating a Deployment-Evaluation Gap, the absence of operational and economic criteria in model assessment. To address this gap, we present EDGE-EVAL, a industry-oriented benchmarking framework that evaluates LLMs across their full lifecycle on legacy NVIDIA Tesla T4 GPUs. Benchmarking LLaMA and Qwen variants across three industrial tasks, we introduce five deployment metrics, Economic Break-Even (Nbreak), Intelligence-Per-Watt (IP W), System Density (rhosys), Cold-Start Tax (Ctax), and Quantization Fidelity (Qret), capturing profitability, energy efficiency, hardware scaling, serverless feasibility, and compression safety. Our results reveal a clear efficiency frontier, models in the < 2B parameter class dominate larger baselines across economic and ecological dimensions. LLaMA-3.2-1B (INT4) achieves ROI break-even in 14 requests (median), delivers 3x higher energy-normalized intelligence than 7B models, and exceeds 6,900 tokens/s/GB under 4-bit quantization. We further uncover an efficiency anomaly, while QLoRA reduces memory footprint, it increases adaptation energy by up to 7x for small models, challenging prevailing assumptions about quantization-aware training in edge deployment.

[Industry] Detecting AI-Generated Content on Social Media with Multi-modal Language Models ▶ *Underline Video*
Chenyang Yang, Shen Yan, Yibo Yang, Litao Hu, Yuchen Liu, Yuan Zeng, Hanchao Yu, Yinan Zhu, Sumedha Singla, Brian Vanover, Huijun Qian, Zihao Wang, Fujun Liu, Aashu Singh, Jianyu Wang and Xuewen Zhang
Generative AI has enabled the creation of photorealistic images and videos that are increasingly disseminated on social media, often used for spam, misinformation, manipulation, and fraud. Existing AI-generated content (AIGC) detection methods face challenges including poor generalization to new generation models, reliance on single modalities, and lack of interpretable explanations. We present our pipeline at Meta that addresses these issues by continuously curating diverse multi-modal social media data and training a compact vision-language model for detection and explanation. Our model achieves state-of-the-art detection performance on public benchmarks and demonstrates robust detection and explanation capabilities on internal social media datasets across multiple platforms. This work highlights the feasibility of effective, interpretable AIGC detection in dynamic social media environments.

[Industry] From Pixels to Policies: Reinforcing Spatial Reasoning in Language Models for Content-Aware Layout Design ▶ *Underline Video*
Sha Li, Stefano Petrangeli, Yu Shen and Xiang Chen
We introduce LaySPA, a reinforcement learning framework that equips large language models (LLMs) with explicit and interpretable spatial reasoning for content-aware graphic layout design. LaySPA addresses two key challenges: LLMs limited spatial reasoning and the lack of opacity in design decision making. Instead of operating at the pixel level, we reformulate layout design as a policy learning problem over a structured textual spatial environment that explicitly encodes canvas geometry, element attributes, and inter-element relationships.

LaySPA produces dual-level outputs comprising interpretable reasoning traces and structured layout specifications, enabling transparent and controllable design decision making. Layout design policy is optimized via a multi-objective spatial critique that decomposes layout quality into geometric validity, relational coherence, and aesthetic consistency, and is trained using relative group optimization to stabilize learning in open-ended design spaces. Experiments demonstrate that LaySPA improves structural validity and visual quality, outperforming larger proprietary LLMs and achieving performance comparable to specialized SOTA layout generators while requiring fewer annotated samples and reduced latency.

[Industry] How to Train a Real-World Silicon Concierge? Internalizing Complex Business Workflow to Only OneModel ► *Underline Video*

Yongqi Tong, Xiaoyun Feng, Lyuxin Xue, JIANSHE LI, Xin Zhang and Jiang-Ming Yang

Traditional industrial agents rely on modular pipelines, including Router, Retriever, Planner, Executor, Responder, Reviewer and so on, which inevitably fracture into a labyrinth of ad-hoc patches, leading to cascading errors and high latency. We propose OneModel, an applicable paradigm shift from external workflows to internalized knowledge representation. Unlike modular systems that slice fluid user intents into static steps, OneModel consolidates complex business logic and SOPs directly into the models parameters. Through Continual Pre-training (CPT) and logic-compilation SFT, we transform fragmented business rules into the models intuitive reasoning within a unified attention space. Deployed in our global financial service system, OneModel effectively breaks the impossible triangle of latency, accuracy, and complexity. Online A/B testing demonstrates end-to-end latency reduction of more than 50% (18.7s -> 8s) while the Intelligent Resolution Rate (IRR) jumps from 64.3% to 83.3%. The results demonstrate our paradigm OneModel effectively replaces brittle engineering logic with internalized cognitive intuition, offering a scalable and future-proof blueprint for transitioning industrial agents from complex, error-prone workflow to unified model architectures.

[Industry] Know What You See: Grounded localization of product components ► *Underline Video*

Manan Soni, Abinesh Kanagarajan and Shyam Mohan

Many real-world decisions about products (e.g. how they function, how they should be used) depend on their components rather than the object as a whole. Accurately knowing product component has applications like automated defect detection, visual spare-parts search, and verified assembly. However, existing object detectors treat components as isolated objects, ignoring their inherent structure. We propose Know What You See (KWYS), where we localize components by grounding them using a textual knowledge base (e.g., manuals or web descriptions). KWYS converts raw text into a hierarchical component taxonomy, which then guides an open-vocabulary object detector using a hierarchical verification algorithm. We evaluate on 1,000 product images across 5 diverse categories, improving component localization accuracy by 11% along with reducing component hallucinations by 25%.

[Industry] MTIVE: Multi-Task Image Verification Engine Using Vision-Language Models for E-commerce ► *Underline Video*

Yu-Tong Cao, Vishnu Prabhakaran, Arunita Das, Purav Aggarwal and Anoop Saladi;

Vision-language models show promise for e-commerce automation but struggle with noisy real-world images and multi-task requirements. We introduce MTIVE, a curriculum learning framework that progressively adapts base models through three stages: continued pre-training on large-scale e-commerce datasets with contrastive learning and diverse dialogue templates, instruction tuning on synthetic data, and modular task-specific expert training. Our architecture uses frozen base weights with stacked LoRA adapters, shared modules for domain knowledge and lightweight task-specific experts, enabling continual learning without catastrophic forgetting. MTIVE outperforms open-source and proprietary baselines in both standard and continual learning settings.

[Industry] Multi-Task Reinforcement Learning for Enhanced Multimodal LLM-as-a-Judge ► *Underline Video*

Junjie Wu, Xuan Kan, Zihao He, Shunwen Tan, Bo Pan and Kaitai Zhang

Multimodal Large Language Models (MLLMs) have been widely adopted as MLLM-as-aJudges due to their strong alignment with human judgment across various visual tasks. However, most existing judge models are optimized for single-task scenarios and struggle to generalize to diverse contexts, which is a critical requirement for reliable evaluation. To address this limitation, we propose Multi-Task Reinforcement Learning for MLLM-as-a-Judge (MT-RL-Judge), a framework that jointly optimizes the judge model across multiple tasks, leveraging the generalization capabilities of RL. Experimental results against several strong baselines demonstrate that MT-RL-Judge outperforms strong baselines in both judgment consistency and correlation with human preferences. Furthermore, our approach exhibits robust generalization on out-of-distribution tasks, further validating its effectiveness.

[Industry] Pre-Deployment Advertisement Ranking under Data Scarcity via Context-Aware Criteria Generation with VLMs ► *Underline Video*

Kyungho Kim, Yeonje Choi, Gyurim Hwang, Sejin Chung, Hongseok Lee, Myeong Ho Song, Yeongho Kim, Sunwoo Kim, Jongha Lee, Juyeon Kim and Kijung Shin

Vision-Language Models (VLMs) perform well on general multimodal tasks, yet applying them to real-world advertisement (ad) evaluation is challenging due to strong brand specificity and limited labeled data. We introduce a new practical task, brand-specific ad ranking, which aims to rank ads for a target brand prior to deployment by modeling brand-specific effectiveness. To this end, we propose ADvisor, which derives explicit brand-aware decision criteria using VLMs, augments limited brand context with ads from similar brands, and applies reflection-based scoring for ranking. Experiments on real-world advertising data from 10 brands, collected through actual ad campaigns, show that ADvisor

outperforms strong baselines by up to 7.2%. Further analyses show the generated criteria capture meaningful brand specificity, and ADvisor also performs strongly in online A/B testing. Our code is available at https://anonymous.4open.science/r/ADvisor-2C0C/.

[Industry] SSR-A: Spatial- and Semantic-Aware Instructions and Curriculum Reinforcement for Advertisement Compliant Rectification ▶ *Underline Video*

Te Cao, Mengge Xue, Zhenyu Hu, Yuan Chen, Liqun Liu, Peng Shu, Huan Yu and Jie Jiang

While advertising is a cornerstone of commercial growth, it is constrained by online violation detection systems that reject non-compliant content at a million-scale daily. Advertisers urgently require automated solutions to rectify these advertisements, especially visual ads, as manual fixing is unscalable. Although recent safety-driven methods can achieve compliance, they typically suffer from over-editing, destroying the original commercial intent and perceptual similarity. To address this, we present SSR-A, a framework tailored for the minimalist rectification of non-compliant image ads. Instead of fine-tuning image editing models directly, SSR-A focuses on translating violation policies into targeted editing instructions. We first introduce a Spatial- and Semantic-Aware Instruction Synthesis Pipeline, where MLLMs synthesize candidate instructions, incorporating spatial grounding and semantic guidance, and select the optimal instruction via multi-dimensional evaluation. Furthermore, we align the model using Curriculum Reinforcement Learning, employing GRPO with multi-faceted rewards to progressively navigate the trade-off between compliance and visual preservation. Extensive experiments and online A/B tests show that SSR-A significantly outperforms state-of-the-art baselines in both compliance and preservation of visual and commercial consistency.

[Industry] UNIVID: Unified Vision-Language Model for Video Moderation ▶ *Underline Video*

Kejuan Yang, Yizhuo Zhang, Mingyuan Du, Yue Zhang, Dixin Zheng, Kaili Zhao, Yang Xiao, Hanzhong Liang and Kenan Xiao

Global-scale video moderation faces a dual challenge: the need for fine-grained multimodal reasoning and the demand for interpretable outputs to support downstream enforcement. Traditional moderation systems often rely on fragmented black-box classifiers that are difficult to maintain and lack transparency. In this paper, we present UNIVID, a Unified Vision-Language model for Video Moderation. Unlike standard classification models, UNIVID generates policy-aware captions that serve as an interpretable intermediate representation, enabling human-verifiable decisions and multi-task reusability. While existing open-source and commercial VLMs often suffer from safety-guardrail refusals and lack fine-grained policy alignment, we develop a specialized training data recipe that combines expert human-refined labels with synthetic data to align the model with our safety guidelines. By integrating UNIVID as the core captioner, we design a novel end-to-end video moderation system that reduces violation leakage by 42.7% and overkill rate by 37.0% relatively. Meanwhile, by replacing over 1,000 policy-specific models with a single UNIVID backbone, we recycle extensive computational resources while significantly reducing engineering maintenance overhead. To our knowledge, this is one of the first reports of a high-efficiency captioning VLM successfully supporting industrial-scale moderation and cross-functional business.

Oral Session G: Mathematical, Symbolic, Neurosymbolic, and Logical Reasoning

11:00-12:30 (Harbor G)

[Long] AgentCoMa: A Compositional Benchmark Mixing Commonsense and Mathematical Reasoning in Real-World Scenarios ▶ *Underline Video*

Lisa Alazraki, Lihu Chen, Ana Brassard, Joe Stacey, Hossein A. Rahmani and Marek Rei

Large Language Models (LLMs) have achieved high accuracy on complex commonsense and mathematical problems that involve the composition of multiple reasoning steps. However, current compositional benchmarks testing these skills tend to focus on either commonsense or math reasoning, whereas LLM agents solving real-world tasks would require a combination of both. In this work, we introduce an Agentic Commonsense and Math benchmark (AgentCoMa), where each compositional task requires a commonsense reasoning step and a math reasoning step. We test it on 61 LLMs of different sizes, model families, and training strategies. We find that LLMs can usually solve both steps in isolation, yet their accuracy drops by nearly 30% on average when the two are combined. This is a substantially greater performance gap than the one we observe in prior compositional benchmarks that combine multiple steps of the same reasoning type. In contrast, non-expert human annotators can solve the compositional questions and the individual steps in AgentCoMa with similarly high accuracy. Furthermore, we conduct a series of interpretability studies to better understand the performance gap, examining neuron patterns, attention maps and membership inference. Our work underscores a substantial degree of model brittleness in the context of mixed-type compositional reasoning and offers a test bed for future improvement.

[Long] Challenging the Boundaries of Reasoning: An Olympiad-Level Math Benchmark for Large Language Models ▶ *Underline Video*

Haoxiang Sun, Yingqian Min, Zhipeng Chen, Wayne Xin Zhao and Ji-Rong Wen

The rapid advancement of large reasoning models has saturated existing math benchmarks, underscoring the urgent need for more challenging evaluation frameworks. To address this, we introduce OlymMATH, a rigorously curated, Olympiad-level math benchmark comprising 350 problems, each with parallel English and Chinese versions. OlymMATH is the first benchmark to unify dual evaluation paradigms within a

single suite: (1) natural language evaluation through OlymMATH-EASY and OlymMATH-HARD, comprising 200 computational problems with numerical answers for objective rule-based assessment, and (2) formal verification through OlymMATH-LEAN, offering 150 problems formalized in Lean 4 for rigorous process-level evaluation. All problems are manually sourced from printed publications to minimize data contamination, verified by experts, and span four core domains. Extensive experiments reveal the benchmark's significant challenge, and our analysis also uncovers consistent performance gaps between languages and identifies cases where models employ heuristic "guessing" rather than rigorous reasoning. To further support community research, we release 582k+ reasoning trajectories, a visualization tool, and expert solutions at https://github.com/RUCAIBox/OlymMATH.

[Long] Characterizing the Expressivity of Local Attention in Transformers ▶ ***Underline Video***

Jiaoda Li and Ryan Cotterell

The transformer is the most popular neural architecture for language modeling. The cornerstone of the transformer is its global attention mechanism, which lets the model aggregate information from all preceding tokens before generating the next token. One common variant of attention is called local attention, which restricts each token to aggregating information from a bounded window of predecessors, reducing the quadratic cost of global attention to linear. Although this restriction is usually motivated by efficiency, it has also been found to improve model quality, a phenomenon that has so far lacked a satisfactory explanation. We provide a formal account of this phenomenon in terms of recognizer expressivity. It has been shown that fixed-precision transformers with global attention correspond to a fragment of linear temporal logic containing a single past operator. We additionally prove that adding local attention introduces a second temporal operator, strictly enlarging the class of recognizable regular languages. Moreover, global and local attention are expressively complementary: neither subsumes the other, and combining them yields the richest fragment. Experiments on formal language recognition and natural language modeling corroborate the theory, showing that hybrid global–local transformers outperform their global-only counterparts.

[Long] Discovery and Reinforcement of Tool-Integrated Reasoning Chains via Rollout Trees ▶ ***Underline Video***

Kun Li, Zenan Xu, Junan Li, Zengrui Jin, Jinghao Deng, Zexuan Qiu and Bo Zhou

Tool-Integrated Reasoning has emerged as a key paradigm to augment Large Language Models (LLMs) with computational capabilities, yet integrating tool-use into long Chain-of-Thought (long CoT) remains underexplored, largely due to the scarcity of training data and the challenge of integrating tool-use without compromising the model's intrinsic long-chain reasoning. In this paper, we introduce DART (Discovery And Reinforcement of Tool-Integrated Reasoning Chains via Rollout Trees), a reinforcement learning framework that enables spontaneous tool-use during long CoT reasoning without additional human annotation. DART operates by constructing dynamic rollout trees during training to discover valid tool-use opportunities, branching out at promising positions to explore tool-integrated trajectories. Subsequently, a tree-based process advantage estimation identifies and credits specific sub-trajectories where tool invocation positively contributes to the solution, effectively reinforcing these beneficial behaviors during training. Extensive experiments on challenging benchmarks like AIME and GPQA-Diamond demonstrate that DART significantly outperforms existing methods, successfully harmonizing tool execution with long CoT reasoning.

[Long] Immediate Inference: The Missing Foundation in Large Language Model Logical Reasoning ▶ ***Underline Video***

Sihang Jiang, Zhiyu Lu, Keyi Wang, Jiaqing Liang, Yanghua Xiao, Xiaojun Meng and Jiansheng Wei

While extensive research has evaluated LLMs on complex reasoning tasks, the foundational building blocks of logical reasoning remain underexplored. We introduce IIBench, a benchmark evaluating immediate inference (elementary operations over categorical propositions). Our evaluation reveals that even SoTA models exhibit systematic deficiencies in immediate inference, and establishes immediate inference as foundational: it mediates approximately 40% of the effect on syllogistic reasoning, with near-perfect correlation (= 0.98) across reasoning benchmarks. Our analysis reveals that models lack robust operator grounding, oscillating between structural reasoning and surface pattern matching with inconsistent handling of quantifiers and negation.

[Long] Learning from Contrasts: Synthesizing Reasoning Paths from Diverse Search Trajectories ▶ ***Underline Video***

Peiyang Liu, Zhirui Chen, Xi Wang, Di Liang, Youru Li, Zhi Cai and Wei Ye

Monte Carlo Tree Search (MCTS) has been widely used for automated reasoning data exploration, but current supervision extraction methods remain inefficient. Standard approaches retain only the single highest-reward trajectory, discarding the comparative signals present in the many explored paths. Here we introduce Contrastive Reasoning Path Synthesis (CRPS), a framework that transforms supervision extraction from a filtering process into a synthesis procedure. CRPS uses a structured reflective process to analyze the differences between high- and low-quality search trajectories, extracting explicit information about strategic pivots and local failure modes. These insights guide the synthesis of reasoning chains that incorporate success patterns while avoiding identified pitfalls. We show empirically that models fine-tuned on just 60K CRPS-synthesized examples match or exceed the performance of baselines trained on 590K examples derived from standard rejection sampling, a 20x reduction in dataset size. Furthermore, CRPS improves generalization on out-of-domain benchmarks, demonstrating that learning from the contrast between success and failure produces more transferable reasoning capabilities than learning from success alone.

[Long] Rethinking Sample Polarity in Reinforcement Learning with Verifiable Rewards ▶ ***Underline Video***

Xinyu Tang, Yuliang Zhan, Zhixun Li, Xin Zhao, Zhenduo Zhang, Zujie Wen, Zhiqiang Zhang and Jun Zhou

Large reasoning models (LRMs) are typically trained using reinforcement learning with verifiable reward (RLVR) to enhance their reasoning abilities. In this paradigm, policies are updated using both positive and negative self-generated rollouts, which correspond

to distinct sample polarities. In this paper, we provide a systematic investigation into how these sample polarities affect RLVR training dynamics and behaviors. We find that positive samples sharpen existing correct reasoning patterns, while negative samples encourage exploration of new reasoning paths. We further explore how adjusting the advantage values of positive and negative samples at both the polarity level and the token level affects RLVR training. Based on these insights, we propose an Adaptive and Asymmetric token-level Advantage shaping method for Policy Optimization, namely A3PO, that more precisely allocates advantage signals to key tokens across different polarities. Experiments across five reasoning benchmarks demonstrate the effectiveness of our approach.

[Long] Semantic-Aware Logical Reasoning via a Semiotic Framework ► *Underline Video*
Yunyao Zhang, Xinglang Zhang, Junxi Sheng, Wenbing Li, Junqing Yu, Yi-Ping Phoebe Chen, Wei Yang and Zikai Song
Logical reasoning is a fundamental capability of large language models (LLMs). However, existing studies largely overlook the interplay between logical complexity and semantic complexity, limiting their robustness under abstract propositions, ambiguous contexts, and conflicting stances, which are central to human reasoning. We propose LogicAgent, a semiotic-square-guided framework that jointly addresses these two axes of difficulty. The semiotic square provides a principled structure for multi-perspective semantic analysis, and LogicAgent integrates automated deduction with reflective verification to manage logical complexity across deeper reasoning chains. To evaluate reasoning under coupled semantic and logical complexity, we introduce RepublicQA, a benchmark that contains abstract propositions with systematically constructed contrary and contradictory forms, providing a semantically rich setting for assessing logical reasoning in LLMs. Experiments show that LogicAgent achieves state-of-the-art performance on RepublicQA with a 6.25% average gain, and generalizes well to four mainstream logical reasoning benchmarks with an additional 7.05% improvement, highlighting the effectiveness of our semiotic-grounded multi-perspective reasoning in boosting LLMs logical performance.

[Long] Student Guides Teacher: Weak-to-Strong Inference via Spectral Orthogonal Exploration ► *Underline Video*
Dayu Wang, Jiaye Yang, Weikang Li, Jiahui Liang, Yang Li, Deguo Xia and Jizhou Huang
Large Language Models (LLMs) often suffer from "Reasoning Collapse" on challenging mathematical reasoning tasks, where stochastic sampling produces lexical variations of the same erroneous logic rather than genuine semantic exploration. We observe that failed reasoning traces are often associated with a low-rank bias manifold in the model's hidden-state geometry, which reduces exploration toward corrective solution directions. To address this, we propose Spectral Orthogonal Exploration (SOE), a geometric inference framework under a "Student Guides Teacher" paradigm. Instead of using a weak auxiliary agent for imitation, SOE uses it as an orthogonal probe to introduce semantically heterogeneous reasoning signals into the teacher's orthogonal complement of its dominant subspace. This intervention steers the teacher toward more diverse reasoning trajectories and improves exploration beyond standard sampling. Experiments on mathematical benchmarks show that SOE improves average accuracy by 62.4% and average sampling efficiency by 113.7% over baseline methods, suggesting that geometric interventions can be effective for mitigating reasoning collapse in mathematical reasoning. We further provide preliminary evidence that SOE is also effective on logic and code generation benchmarks. Code is available at https://github.com/dayuwang401/spectral-orthogonal-exploration.

Oral Session G: Natural Language Generation 2

11:00-12:30 (Harbor H-I)

[Long] Act-Adaptive Margin: Dynamically Calibrating Reward Models for Subjective Ambiguity ► *Underline Video*
Feiteng Fang, Dingwei Chen, Xiang Huang, Ting-En Lin, Yuchuan Wu, Xiong Liu, Xinge Ye, Ziqiang Liu, Haonan Zhang, Liang Zhu, Hamid Alinejad-Rokny, Min Yang and Yongbin Li
Currently, most reinforcement learning tasks focus on domains like mathematics and programming, where verification is relatively straightforward. However, in subjective tasks such as role-playing, alignment techniques struggle to make progress, primarily because subjective reward modeling using the Bradley-Terry model faces significant challenges when dealing with ambiguous preferences. To improve reward modeling in subjective tasks, this paper proposes AAM (Act-Adaptive Margin), which enhances reward modeling by dynamically calibrating preference margins using the model's internal parameter knowledge. We design two versions of AAM that efficiently generate contextually-appropriate preference gaps without additional human annotation. This approach fundamentally improves how reward models handle subjective rewards by better integrating generative understanding with preference scoring. To validate AAM's effectiveness in subjective reward modeling, we conduct evaluations on RewardBench, JudgeBench, and challenging role-playing tasks. Results show that AAM significantly improves subjective reward modeling performance, enhancing Bradley-Terry reward models by 2.95% in general tasks and 4.85% in subjective role-playing tasks. Furthermore, reward models trained with AAM can help downstream alignment tasks achieve better results. Our test results show that applying rewards generated by AAM-Augmented RM to preference learning techniques (e.g., GRPO) achieves state-of-the-art results on CharacterEval and Charm. The code and dataset will be released upon acceptance.

[Long] Attention Basin: Why Contextual Position Matters in Large Language Models ► *Underline Video*
Zihao Yi, Zhenqing Ling, Delong Zeng, Haohao Luo, Zhe Xu, Wei Liu, Jian Luan, Wanxia Cao and Ying Shen
The performance of Large Language Models (LLMs) is significantly sensitive to the contextual position of information in the input. To

investigate the mechanism behind this positional bias, our extensive experiments reveal a consistent phenomenon we term the attention basin: when presented with a sequence of structured items (e.g., retrieved documents or few-shot examples), models systematically assign higher attention to the items at the beginning and end of the sequence, while neglecting those in the middle. Crucially, our analysis further reveals that allocating higher attention to critical information is key to enhancing model performance. Based on these insights, we introduce Attention-Driven Reranking (AttnRank), a two-stage framework that (i) estimates a model's intrinsic positional attention preferences using a small calibration set, and (ii) reorders retrieved documents or few-shot examples to align the most salient content with these high-attention positions. AttnRank is a model-agnostic, training-free, and plug-and-play method with minimal computational overhead. Experiments on multi-hop QA and few-shot in-context learning tasks demonstrate that AttnRank achieves substantial improvements across 10 large language models of varying architectures and scales, without modifying model parameters or training procedures.

[Long] Attribution, Citation, and Quotation: A Survey of Evidence-based Text Generation with Large Language Models ▶ *Underline Video*

Tobias Schreieder, Tim Schopf and Michael Färber

The increasing adoption of large language models (LLMs) has raised serious concerns about their reliability and trustworthiness. As a result, a growing body of research focuses on evidence-based text generation with LLMs, aiming to link model outputs to supporting evidence to ensure traceability and verifiability. However, the field is fragmented due to inconsistent terminology, isolated evaluation practices, and a lack of unified benchmarks. To bridge this gap, we systematically analyze 134 papers, introduce a unified taxonomy of evidence-based text generation with LLMs, and investigate 300 evaluation metrics across seven key dimensions. Thereby, we focus on approaches that use citations, attribution, or quotations for evidence-based text generation. Building on this, we examine the distinctive characteristics and representative methods in the field. Finally, we highlight open challenges and outline promising directions for future work.

[Long] CSRP: Chain-of-Thought Reasoning for Chinese Text Correction via Reinforcement Learning with Efficiency-Aware Rewards ▶ *Underline Video*

Wei Tian, Yuhao Zhou and Man Lan

Large Language Model (LLM) based Chinese Grammatical Error Correction (CGEC) systems face two critical challenges: general-purpose models lack specialized linguistic priors for subtle grammatical distinctions, and Supervised Fine-Tuning (SFT) with Maximum Likelihood Estimation fails to optimize for precision-focused metrics, leading to systematic over-correction. We propose CSRP, a three-stage framework that progressively builds correction capability through Continual Pre-training (CPT) on 5.9M balanced samples to internalize domain knowledge, Chain-of-Thought SFT with explicit error reasoning for diagnostic transparency, and Group Relative Policy Optimization with a novel Efficiency-Aware Reward that explicitly penalizes unnecessary edits. On the NACGEC benchmark, CSRP achieves state-of-the-art performance with 50.99 F_0.5 and 57.17 precision, substantially outperforming previous best results while effectively mitigating the over-correction bias inherent in MLE-trained models. Our method also advances CSCD spelling correction to 59.61 F1, surpassing GPT-4 by 5.20 points. Comprehensive ablation studies demonstrate that the RL alignment stage contributes a 8% relative gain over the SFT baseline, and that this gain is orthogonal to the contribution of large-scale CPT, validating that explicit optimization for edit efficiency is essential for high-quality grammatical error correction. Our code is available at https://github.com/TW-NLP/ChineseErrorCorrector.

[Long] Decoupling Task-Solving and Output Formatting in LLM Generation ▶ *Underline Video*

Haikang Deng, Po-Nien Kung and Nanyun Peng

Large language models (LLMs) are increasingly adept at solving complex problems, such as mathematical reasoning and automatic evaluation. However, performance often degrades when prompts intertwine task instructions with rigid formatting requirements. This entanglement creates competing goals for the model, hindering its reasoning capabilities. To address this, we introduce Deco-G, a decoding framework that explicitly decouples format adherence from problem solving. Deco-G delegates format adherence to a separate Format Estimation Module (FEM), which performs probabilistic lookahead to estimate future format compliance rate and reweighs token probabilities, allowing the LLM to focus solely on task resolution. To make this approach both practical and efficient, we introduce three key innovations: instruction-aware distillation, a flexible trie-building algorithm, and HMM state pruning. Experiments across mathematical reasoning, event argument extraction, and LLM-as-a-judge demonstrate that Deco-G constantly gains over prompting or structured generation baselines, with guaranteed format compliance.

[Long] DFAMS: Dynamic-flow guided Federated Alignment based Multi-prototype Search ▶ *Underline Video*

Zhibang Yang, Xinke Jiang, Rihong Qiu, Ruiqing Li, Yihang Zhang, Yue Fang, Yongxin Xu, Hongxin Ding, Xu Chu, Junfeng Zhao and Yasha Wang

Federated Retrieval (FR) routes queries across multiple external knowledge sources, to mitigate hallucinations of LLMs, when necessary external knowledge is distributed. However, existing methods struggle to retrieve high-quality and relevant documents for ambiguous queries, especially in cross-domain scenarios, which significantly limits their effectiveness in supporting downstream generation tasks. Inspired by Dynamic Information Flow (DIF), we propose DFAMS, a novel framework that leverages DIF to identify latent query intents and construct semantically aligned knowledge partitions for accurate retrieval across heterogeneous sources. Specifically, DFAMS probes the DIF in LLMs by leveraging gradient signals from a few annotated queries and employing Shapley value-based attribution to trace neuron activation paths associated with intent recognition and subdomain boundary detection. Then, DFAMS leverages DIF to train an alignment module via multi-prototype contrastive learning, enabling fine-grained intra-source modeling and

inter-source semantic alignment across knowledge bases. Experimental results across five benchmarks show that DFAMS outperforms advanced FR methods by up to 14.37% in knowledge classification accuracy, 5.38% in retrieval recall, and 6.45% in downstream QA accuracy, demonstrating its effectiveness in complex FR scenarios. Our code is publicly available at https://github.com/Artessay/DFAMS.

[Long] IF-CRITIC: Towards a Fine-Grained LLM Critic for Instruction-Following Evaluation ▶ *Underline Video*

Bosi Wen, Yilin Niu, Cunxiang Wang, Pei Ke, Xiaoying Ling, Ying Zhang, Aohan Zeng, Hongning Wang and Minlie Huang

Instruction-following is a fundamental ability of Large Language Models (LLMs), requiring their generated outputs to follow multiple constraints imposed in input instructions. Numerous studies have attempted to enhance this ability through preference optimization or reinforcement learning based on reward signals from LLM-as-a-Judge. However, existing evaluation models for instruction-following still possess many deficiencies, such as substantial costs and unreliable assessments. To this end, we propose IF-CRITIC, an LLM critic for fine-grained, efficient, and reliable instruction-following evaluation. We first develop a checklist generator to decompose instructions and generate constraint checklists. With the assistance of the checklists, we collect high-quality critique training data through a multi-stage critique filtering mechanism and employ a constraint-level preference optimization method to train IF-CRITIC. Extensive experiments show that the evaluation performance of IF-CRITIC can beat strong LLM-as-a-Judge baselines, including o4-mini and Gemini-3-Pro. With the reward signals provided by IF-CRITIC, LLMs can achieve substantial performance gains in instruction-following optimization under lower computational overhead compared to strong LLM critic baselines. Our code and model are available at https://github.com/thu-coai/IF-CRITIC.

[Long] MUR: Momentum Uncertainty guided Reasoning for Large Language Models ▶ *Underline Video*

Hang Yan, Fangzhi Xu, Rongman Xu, Yifei Li, Jian Zhang, Haoran Luo, Xiaobao Wu, Anh Tuan Luu, Haiteng Zhao, Qika Lin and Jun Liu

Large Language Models (LLMs) have achieved impressive performance on reasoning-intensive tasks, yet optimizing their reasoning efficiency remains an open challenge. While Test-Time Scaling (TTS) improves reasoning quality, it often leads to overthinking, wasting tokens on redundant computations. This work investigates how to efficiently and adaptively guide LLM TTS without additional training. Inspired by the concept of momentum in physics, we propose Momentum Uncertainty-guided Reasoning (MUR), which dynamically allocates thinking budgets to critical reasoning steps by tracking and aggregating step-wise uncertainty over time. To support flexible inference-time control, we introduce -control, a simple mechanism that tunes the reasoning budget via a single hyperparameter. We provide in-depth theoretical proof to support the superiority of MUR in terms of stability and biases. MUR is comprehensively evaluated against various TTS methods across four challenging benchmarks (MATH-500, AIME24, AIME25, and GPQA-diamond) using different sizes of recent Qwen3 models (1.7B, 4B, and 8B). Results demonstrate that MUR reduces computation by over 45% on average while improving accuracy by 0.33-3.46%.

[Long] SWAN: Semantic Watermarking with Abstract Meaning Representation ▶ *Underline Video*

Ziping Ye, Gourab Dey, Christos Christodoulopoulos, Charith Peris, Anil Ramakrishna, Weitong Ruan, Aram Galstyan, Kai-Wei Chang, Rahul Gupta and Ninareh Mehrabi

We introduce SWAN (Semantic Watermarking with Abstract Meaning Representation), a novel framework that embeds watermark signatures into the semantic structure of a sentence using Abstract Meaning Representation (AMR). In contrast to existing watermarking methods, which typically encode signatures by adjusting token selection preferences during text generation, SWAN embeds the signature directly in the sentences semantic representation. As the signature is encoded at the semantic structure level, any paraphrase that preserves meaning, automatically preserves the signature. SWAN is training-free: watermark injection is achieved by prompting an LLM to generate sentences guided by a selected AMR template while maintaining contextual coherence, and detection uses an off-the-shelf AMR parser followed by a simple one-proportion z-test. Empirical evaluation on the RealNews benchmark shows SWAN matches state-of-the-art detection performance on unaltered watermarked text, while significantly improving robustness against paraphrasing, increasing detection AUC by up to 13.9 percentage points compared to prior methods. These results demonstrate that SWAN's approach of anchoring watermarks in AMR semantic structures provides a simple, effective, and prompt-based method for robust text provenance verification under paraphrasing, opening new avenues for semantic-level watermarking research.

Oral Session G: Question Answering 2

11:00-12:30 (Harbor D)

[Long] CypherSmith: Transforming Text-to-Cypher Generation for LLMs with Synthetic Data ▶ *Underline Video*

Zeyu Zhang, Kexuan Sun, Zheng Tang, Jens-S. Vöckler, Thien Huu Nguyen and Thuy Vu

Knowledge Graph (KG) retrieval is a promising augmentation to address knowledge gaps and hallucinations in LLMs. As KGs in practice are stored in graph databases (e.g., Wikidata, Freebase), accurate retrieval requires translating natural language questions into structured queries (query generation). A key challenge of query generation is Text-to-Cypher, which generates Cypher queries for property graphs (e.g., Neo4j), a paradigm increasingly adopted in industry for their scalable architectures and expressive schemas. However, compared to other query generation tasks such as Text-to-SQL or Text-to-SPARQL, Text-to-Cypher remains underexplored due to scarce public KGs and datasets. Existing datasets are small, domain-limited, and lack diversity, constraining LLM progress. To address this, we introduce

CypherSmith, an instruction-tuning dataset over 12x larger than prior public Text-to-Cypher datasets, spanning diverse domains to better support LLM fine-tuning. Our key distinction lies in fully leveraging open-source LLMs for large-scale synthetic data generation and introducing a novel likelihood-based filtering technique to ensure high-quality Text-to-Cypher data. Extensive experiments demonstrate the effectiveness of CypherSmith, achieving state-of-the-art LLM performance.

[Long] From Answers to Arguments: Toward Trustworthy Clinical Diagnostic Reasoning with Toulmin-Guided Curriculum Goal-Conditioned Learning ▶ *Underline Video*

Chen Zhan, Xiaoyu Tan, Gengchen Ma, Yu-Jie Xiong, Xiaoyan Jiang and Xihe Qiu

The integration of Large Language Models (LLMs) into clinical decision support is critically obstructed by their opaque and often unreliable reasoning. In the high-stakes domain of healthcare, correct answers alone are insufficient; clinical practice demands full transparency to ensure patient safety and enable professional accountability. A pervasive and dangerous weakness of current LLMs is their tendency to produce "correct answers through flawed reasoning." This issue is far more than a minor academic flaw; such process errors signal a fundamental lack of robust understanding, making the model prone to broader hallucinations and unpredictable failures when faced with real-world clinical complexity. In this paper, we establish a framework for trustworthy clinical argumentation by adapting the Toulmin model to the diagnostic process. We propose a novel training pipeline: Curriculum Goal-Conditioned Learning (CGCL), designed to progressively train LLM to generate diagnostic arguments that explicitly follow this Toulmin structure. CGCL's progressive three-stage curriculum systematically builds a solid clinical argument: (1) extracting facts and generating differential diagnoses; (2) justifying a core hypothesis while rebutting alternatives; and (3) synthesizing the analysis into a final, qualified conclusion. We validate CGCL using T-Eval, a quantitative framework measuring the integrity of the diagnosis reasoning. Experiments show that our method achieves diagnostic accuracy and reasoning quality comparable to resource-intensive Reinforcement Learning (RL) methods, while offering a more stable and efficient training pipeline.

[Long] From Query to Counsel: Structured Reasoning with a Multi-Agent Framework and Dataset for Legal Consultation ▶ *Underline Video*

Mingfei Lu, Yi Zhang, Mengjia Wu and Yue Feng

Legal consultation question answering (Legal CQA) presents unique challenges compared to traditional legal QA tasks, including the scarcity of high-quality training data, complex task composition, and strong contextual dependencies. To address these, we construct JurisCQAD, a large-scale dataset of over 43,000 real-world Chinese legal queries annotated with expert-validated positive and negative responses, and design a structured task decomposition that converts each query into a legal element graph integrating entities, events, intents, and legal issues. We further propose JurisMA, a modular multi-agent framework supporting dynamic routing, statutory grounding, and stylistic optimization. Combined with the element graph, the framework enables strong context-aware reasoning, effectively capturing dependencies across legal facts, norms, and procedural logic. Trained on JurisCQAD and evaluated on a refined LawBench, our system significantly outperforms both general-purpose and legal-domain LLMs across multiple lexical and semantic metrics, demonstrating the benefits of interpretable decomposition and modular collaboration in Legal CQA.

[Long] HopWeaver: Cross-Document Synthesis of High-Quality and Authentic Multi-Hop Questions ▶ *Underline Video*

Zhiyu Shen, Jiyuan Liu, Yunhe Pang, Yanghui Rao, Fu Lee Wang and Jianxing Yu

Multi-Hop Question Answering (MHQA) is crucial for evaluating the model's capability to integrate information from diverse sources. However, creating extensive and high-quality MHQA datasets is challenging: (i) manual annotation is expensive, and (ii) current synthesis methods often produce simplistic questions or require extensive manual guidance. This paper introduces HopWeaver, the first cross-document framework synthesizing authentic multi-hop questions without human intervention. HopWeaver synthesizes bridge and comparison questions through an innovative pipeline that identifies complementary documents and constructs authentic reasoning paths to ensure true multi-hop reasoning. We further present a comprehensive system for evaluating the synthesized multi-hop questions. Empirical evaluations demonstrate that the synthesized questions achieve comparable or superior quality to human-annotated datasets at a lower cost. Our framework provides a valuable tool for the research community: it can automatically generate challenging benchmarks from any raw corpus, which opens new avenues for both evaluation and targeted training to improve the reasoning capabilities of advanced question answering models, especially in domains with scarce resources.

[Long] MTR-Bench: A Comprehensive Benchmark for Multi-Turn Reasoning Evaluation ▶ *Underline Video*

Xiaoyuan Li, Keqin Bao, Yubo Ma, Moxin Li, Wenjie Wang, Rui Men, Yichang Zhang, Fuli Feng and Dayiheng Liu

Recent advances in Large Language Models (LLMs) have shown promising results in complex reasoning tasks. However, current evaluations predominantly focus on single-turn reasoning scenarios, leaving interactive tasks largely unexplored. We attribute it to the absence of comprehensive datasets and scalable automatic evaluation protocols. To fill these gaps, we present MTR-Bench for LLMs' Multi-Turn Reasoning evaluation. Comprising 4 classes, 40 tasks and 3600 instances, MTR-Bench covers diverse reasoning capabilities, fine-grained difficulty granularity, and necessitates multi-turn interactions with the environments. Moreover, MTR-Bench features fully-automated framework spanning both dataset constructions and model evaluations, which enables scalable assessment without human interventions. Extensive experiments reveal that even the cutting-edge reasoning models fall short of multi-turn, interactive reasoning tasks. And the further analysis upon these results brings valuable insights for future research in interactive AI systems.

[Long] Nested Browser-Use Learning for Agentic Information Seeking ▶ *Underline Video*

Baixuan Li, Jialong Wu, Wenbiao Yin, Kuan Li, Zhongwang Zhang, Huifeng Yin, Zhengwei Tao, Liwen Zhang, Pengjun Xie, Jingren Zhou, Yong Jiang, Wentao Zhang and Zhiqiang Gao

Information-seeking (IS) agents have achieved strong performance across a range of wide and deep search tasks, yet their tool use remains largely restricted to API-level snippet retrieval and URL-based page fetching, limiting access to the richer information available through real browsing. While full browser interaction could unlock deeper capabilities, its fine-grained control and verbose page content returns introduce substantial complexity for ReAct-style function-calling agents. To bridge this gap, we propose Nested Browser-Use Learning (NestBrowse), which introduces a minimal and complete browser-action framework that decouples interaction control from page exploration through a nested structure. This design simplifies agentic reasoning while enabling effective deep-web information acquisition. Empirical results on challenging deep IS benchmarks demonstrate that NestBrowse offers clear benefits in practice. Further in-depth analyses underscore its efficiency.

[Long] Reinforcing Agentic Search Via Reward Density Optimization ▶ *Underline Video*

Kun Luo, Hongjin Qian, Zheng Liu, Ziyi Xia, Shitao Xiao, Zhao Cao, Siqi Bao, Jun Zhao and Kang Liu

Reinforcement Learning with Verifiable Rewards (RLVR) is a promising approach for enhancing agentic search. However, its performance is often hindered by reward sparsity, whereby agents receive very limited positive feedback despite incurring significant exploration costs. In this paper, we formalize this challenge as a new research problem termed Reward Density Optimization, which aims to improve the reward obtained per unit of exploration cost. To address this problem, we introduce InfoFlow, a systematic framework that operates along three complementary dimensions: 1) Sub-goal Scaffolding: which decomposes long-horizon tasks into intermediate objectives and assigns process-level rewards to provide denser learning signals; 2) Pathfinding Hints: which injects corrective guidance into stalled trajectories to increase the ratio of successful trials; and 3) Dual-agent Refinement: which employs a dual-agent architecture to offload the cognitive burden of deep exploration. We evaluate InfoFlow on several popular agentic search benchmarks, where it significantly outperforms strong baselines and enables lightweight LLMs to achieve performance comparable to that of advanced proprietary models.

[Long] Rethinking Table Pruning in TableQA: From Sequential Revisions to Gold Trajectory-Supervised Parallel Search ▶ *Underline Video*

Yu Guo, Shenghao Ye, Shuangwu Chen, Zijian Wen, Tao Zhang, Qirui Bai, Dong Jin, Yunpeng Hou, Huasen He, Jian Yang and Xiaobin Tan

Table Question Answering (TableQA) benefits significantly from table pruning, which extracts compact sub-tables by eliminating redundant cells to streamline downstream reasoning. However, existing pruning methods typically rely on sequential revisions driven by unreliable critique signals, often failing to detect the loss of answer-critical data. To address this limitation, we propose TabTrim, a novel table pruning framework which transforms table pruning from sequential revisions to gold trajectory-supervised parallel search. TabTrim derives a gold pruning trajectory using the intermediate sub-tables in the execution process of gold SQL queries, and trains a pruner and a verifier to make the step-wise pruning result align with the gold pruning trajectory. During inference, TabTrim performs parallel search to explore multiple candidate pruning trajectories and identify the optimal sub-table. Extensive experiments demonstrate that TabTrim achieves state-of-the-art performance across diverse tabular reasoning tasks: TabTrim-8B reaches 73.5% average accuracy, outperforming the strongest baseline by 3.2%, including 79.4% on WikiTQ and 61.2% on TableBench.

[Long] Routing with Generated Data: Annotation-Free LLM Skill Estimation and Expert Selection ▶ *Underline Video*

Tianyi Niu, Justin Chen, Genta Indra Winata, Shi-Xiong Zhang, Supriyo Chakraborty, Sambit Sahu, Yue Zhang, Elias Stengel-Eskin and Mohit Bansal

Large Language Model (LLM) routers dynamically select optimal models for given inputs. Existing approaches typically assume access to ground-truth labeled data, which is often unavailable in practice, especially when user request distributions are heterogeneous and unknown. We introduce Routing with Generated Data (RGD), a challenging setting in which routers are trained exclusively on generated queries and answers produced from high-level task descriptions by generator LLMs. We evaluate query-answer routers (using both queries and labels) and query-only routers across four diverse benchmarks and 12 models, finding that query-answer routers degrade faster than query-only routers as generator quality decreases. Our analysis reveals two crucial characteristics of effective generators: they must accurately respond to their own questions, and their questions must produce sufficient performance differentiation among the model pool. We then show how filtering for these characteristics can improve the quality of generated data. We further propose CASCAL, a novel query-only router that estimates model correctness through consensus voting and identifies model-specific skill niches via hierarchical clustering. CASCAL is substantially more robust to generator quality, outperforming the best query-answer router by 4.6% absolute accuracy when trained on weak generator data.

Oral Session G: Retrieval-Augmented Language Models

11:00-12:30 (Harbor A)

[Long] Beyond Single-shot Writing: Deep Research Agents are Unreliable at Multi-turn Report Revision ▶ *Underline Video*

Bingsen Chen, Boyan Li, Ping Nie, Yuyu Zhang, Xi Ye and Chen Zhao

Existing benchmarks for Deep Research Agents (DRAs) treat report generation as a single-shot writing task, which fundamentally diverges from how human researchers iteratively draft and revise reports via self-reflection or peer feedback. Whether DRAs can reliably revise reports with user feedback remains unexplored. We introduce Mr Dre, an evaluation suite that establishes multi-turn report revision as a new axis. Mr Dre consists of (1) a unified long-form report evaluation protocol spanning comprehensiveness, factuality, and presentation, and (2) a human-verified feedback simulation pipeline for systematic multi-turn revision evaluation. Our analysis of five diverse DRAs reveals a critical limitation: while agents can address most user feedback, they also regress on 16–27% of previously covered content and citation quality. Over multiple revision turns, even the best-performing agents leave significant headroom, as they continue to disrupt content outside the feedback's scope and fail to preserve earlier edits. We further show that these issues are not easily resolvable through inference-time fixes such as prompt engineering and a dedicated sub-agent for revision.

[Long] Culinary Crossroads: A RAG Framework for Enhancing Diversity in Cross-Cultural Recipe Adaptation ▶ *Underline Video*
Tianyi Hu, Andrea Morales-Garzón, Jingyi Zheng, Maria Maistro and Daniel Hershcovich
In cross-cultural recipe adaptation, the goal is not only to ensure cultural appropriateness and retain the original dishs essence, but also to provide diverse options for various dietary needs and preferences. Retrieval-Augmented Generation (RAG) is a promising approach, combining the retrieval of real recipes from the target cuisine for cultural adaptability with large language models (LLMs) for relevance. However, it remains unclear whether RAG can generate diverse adaptation results. Our analysis shows that RAG tends to overly rely on a limited portion of the context across generations, failing to produce diverse outputs even when provided with varied contextual inputs. This reveals a key limitation of RAG in creative tasks with multiple valid answers: it fails to leverage contextual diversity for generating varied responses. To address this issue, we propose CARRIAGE, A plug-and-play RAG framework for cross-cultural recipe adaptation that enhances diversity in both retrieval and context organization. To our knowledge, this is the first RAG framework that explicitly aims to generate highly diverse outputs to accommodate multiple user preferences. Our experiments show that CARRIAGE achieves Pareto efficiency in terms of diversity and quality of recipe adaptation compared to closed-book LLMs.

[TACL] Dissecting GraphRAG: A Modular Analysis of Knowledge Structuring for Factoid Question Answering ▶ *Underline Video*
Noriki Nishida, Rumana Ferdous Munne, Shanshan Liu, Narumi Tokunaga, Yuki Yamagata, Fei Cheng, Koiji Kozaki and Yuji Matsumoto
We present a systematic analysis of module-level design choices in GraphRAG, a retrieval-augmented generation framework that integrates structured knowledge graphs into question answering. Focusing on triple extraction, community clustering, and report generation, we evaluate multiple strategies across two knowledge-intensive benchmarks. Our results show that high-quality triple extraction is critical, as the accuracy and coverage of the resulting knowledge graph can become a bottleneck for downstream reasoning. We also find that the granularity of fundamental knowledge units, as determined by community clustering, has a significant impact on downstream performance: achieving a balance between factual detail and topical coherence within each unit is important to enable precise and comprehensive retrieval and to facilitate effective multi-hop reasoning. In addition, simple template-based reporting outperforms LLM-based summarization in both accuracy and efficiency. These findings provide practical guidance for the structure-aware design of retrieval-augmented systems.

[Long] LogicPoison: Logical Attacks on Graph Retrieval-Augmented Generation ▶ *Underline Video*
Yilin Xiao, Jin Chen, Qinggang Zhang, Yujing Zhang, Chuang Zhou, Longhao Yang, Lingfei Ren, Xin Yang and Xiao Huang
Graph-based Retrieval-Augmented Generation (GraphRAG) enhances the reasoning capabilities of Large Language Models (LLMs) by grounding their responses in structured knowledge graphs. Leveraging community detection and relation filtering techniques, GraphRAG systems demonstrate inherent resistance to traditional RAG attacks, such as text poisoning and prompt injection. However, in this paper, we find that the security of GraphRAG systems fundamentally relies on the topological integrity of the underlying graph, which can be undermined by implicitly corrupting the logical connections, without altering surface-level text semantics. To exploit this vulnerability, we propose LogicPoison, a novel attack framework that targets logical reasoning rather than injecting false contents. Specifically, LogicPoison employs a type-preserving entity swapping mechanism to perturb both global logic hubs for disrupting overall graph connectivity and query-specific reasoning bridges for severing essential multi-hop inference paths. This approach effectively reroutes valid reasoning into dead ends while maintaining surface-level textual plausibility. Comprehensive experiments across multiple benchmarks demonstrate that LogicPoison successfully bypasses GraphRAG's defenses, significantly degrading performance and outperforming state-of-the-art baselines in both effectiveness and stealth. Our code is available at https://github.com/Jord8061/logicPoison.

[Long] Long Context Modeling with Ranked Memory-Augmented Retrieval ▶ *Underline Video*
Ghadir Alselwi, Hao Xue, Shoaib Jameel, Basem Suleiman, Flora Salim and Imran Razzak
Effective long-term memory management is crucial for language models handling extended contexts. We introduce the Enhanced Ranked Memory Augmented Retrieval ERMAR framework, which dynamically ranks memory entries based on relevance. Unlike prior models, ERMAR employs a novel relevance scoring mechanism and a pointwise re-ranking model for key-value embeddings, inspired by learning-to-rank techniques in information retrieval. By integrating historical usage patterns and adaptive retrieval, ERMAR achieves state-of-the-art results on standard benchmarks, demonstrating superior scalability and performance in long-context tasks.

[Long] Query-Efficient Agentic Graph Extraction Attacks on GraphRAG Systems ▶ *Underline Video*
Shuhua Yang, Jiahao Zhang, Yilong Wang, Dongwon Lee and Suhang Wang
Graph-based retrieval-augmented generation (GraphRAG) systems construct knowledge graphs over document collections to support

multi-hop reasoning. While prior work shows that GraphRAG responses may leak retrieved subgraphs, the feasibility of query-efficient reconstruction of the hidden graph structure remains unexplored under realistic query budgets. We study a budget-constrained black-box setting where an adversary adaptively queries the system to steal its latent entity-relation graph. We propose AGEA (Agentic Graph Extraction Attack), a framework that leverages a novelty-guided exploration-exploitation strategy, external graph memory modules, and a two-stage graph extraction pipeline combining lightweight discovery with LLM-based filtering. We evaluate AGEA on medical, agriculture, and literary datasets across Microsoft-GraphRAG and LightRAG systems. Under identical query budgets, AGEA significantly outperforms prior attack baselines, recovering up to 90% of entities and relationships while maintaining high precision. These results demonstrate that modern GraphRAG systems are highly vulnerable to structured, agentic extraction attacks, even under strict query limits. The code is available at https://github.com/shuashua0608/AGEA.

[Long] STEM: Structure-Tracing Evidence Mining for Knowledge Graphs-Driven Retrieval-Augmented Generation ► *Underline Video*

Peng Yu, En Xu, Bin Chen, Haibiao Chen and Yinfei Xu

Knowledge Graph-based Question Answering (KGQA) plays a pivotal role in complex reasoning tasks but remains constrained by two persistent challenges: the structural heterogeneity of Knowledge Graphs(KGs) often leads to semantic mismatch during retrieval, while existing reasoning path retrieval methods lack a global structural perspective. To address these issues, we propose Structure-Tracing Evidence Mining (STEM), a novel framework that reframes multi-hop reasoning as a schema-guided graph search task. First, we design a Semantic-to-Structural Projection pipeline that leverages KG structural priors to decompose queries into atomic relational assertions and construct an adaptive query schema graph. Subsequently, we execute globally-aware node anchoring and subgraph retrieval to obtain the final evidence reasoning graph from KG. To more effectively integrate global structural information during the graph construction process, we design a Triple-Dependent GNN (Triple-GNN) to generate a Global Guidance Subgraph (Guidance Graph) that guides the construction. STEM significantly improves both the accuracy and evidence completeness of multi-hop reasoning graph retrieval, and achieves State-of-the-Art performance on multiple multi-hop benchmarks.

[Long] UniversalRAG: Retrieval-Augmented Generation over Corpora of Diverse Modalities and Granularities ► *Underline Video*

Woongyeong Yeo, Kangsan Kim, Soyeong Jeong, Jinheon Baek and Sung Ju Hwang

Retrieval-Augmented Generation (RAG) has shown substantial promise in improving factual accuracy by grounding model responses with external knowledge relevant to queries. However, most existing approaches are limited to a text-only corpus, and while recent efforts have extended RAG to other modalities such as images and videos, they typically operate over a single modality-specific corpus. In contrast, real-world queries vary widely in the type of knowledge they require, which a single type of knowledge source cannot address. To address this, we introduce UniversalRAG, an any-to-any RAG framework designed to retrieve and integrate knowledge from heterogeneous sources with diverse modalities and granularities. Specifically, motivated by the observation that forcing all modalities into a unified representation space derived from a single aggregated corpus causes a modality gap, where the retrieval tends to favor items from the same modality as the query, we propose modality-aware routing, which dynamically identifies the most appropriate modality-specific corpus and performs targeted retrieval within it, and further justify its effectiveness with a theoretical analysis. Moreover, beyond modality, we organize each modality into multiple granularity levels, enabling fine-tuned retrieval tailored to the complexity and scope of the query. We validate UniversalRAG on 10 benchmarks of multiple modalities, showing its superiority over various modality-specific and unified baselines.

[Long] UR^2: Unify RAG and Reasoning through Reinforcement Learning ► *Underline Video*

Weitao Li, Boran Xiang, Xiaolong Wang, Jingyi Ren, Ante Wang, Zhinan Gou, Weizhi Ma and Yang Liu

Large Language Models (LLMs) have shown strong capabilities through two complementary paradigms: Retrieval-Augmented Generation (RAG) for knowledge grounding and Reinforcement Learning from Verifiable Rewards (RLVR) for complex reasoning. However, existing attempts to unify these paradigms remain narrow in scope, typically limited to open-domain QA with fixed retrieval settings, which constrains generalization to broader domains. To address this limitation, we propose UR^2 (Unified RAG and Reasoning), a general reinforcement learning framework that dynamically coordinates retrieval and reasoning. UR^2 introduces two key designs: a difficulty-aware curriculum that selectively invokes retrieval only for challenging instances, and a hybrid knowledge access strategy that combines domain-specific offline corpora with on-the-fly LLM-generated summaries. Together, these components mitigate the imbalance between retrieval and reasoning and improve robustness to noisy information. Experiments on open-domain QA, MMLU-Pro, medical, and mathematical reasoning tasks show that UR^2, built on Qwen-2.5-3/7B and LLaMA-3.1-8B, consistently outperforms existing RAG and RL baselines, and achieves performance comparable to GPT-4o-mini and GPT-4.1-mini on several benchmarks. We will release all code, models, and data.

Oral Session G: Sentiment Analysis, Stylistic Analysis, and Argument Mining

11:00-12:30 (Regatta)

[Long] Are Emotion and Rhetoric Neurons in LLM? Neuron Recognition and Adaptive Masking for Emotion-Rhetoric Prediction Steering ► *Underline Video*

Li Zheng, Xin Zhang, Shuyi He, Fei Li, Chong Teng, Jiang-Ming Yang, Donghong Ji and Zhuang Li

Accurate comprehension and controllable generation of emotion and rhetoric are pivotal for enhancing the reasoning capabilities of large language models (LLMs). Existing studies mostly rely on external optimizations, lacking in-depth exploration of internal representation mechanisms, thus failing to achieve fine-grained steering at the neuron level. A handful of works on neurons are confined to emotions, neglecting rhetoric neurons and their intrinsic connections. Traditional neuron masking also exhibits counterintuitive phenomena, making reliable verification of neuron functionality infeasible. To address these issues, we systematically investigate the neurons representation mechanisms and inherent associations of 6 emotion categories and 4 core rhetorical devices. We propose a neuron identification framework that integrates multi-dimensional screening, and design an adaptive masking method incorporating dynamic filtering, attenuation masking, and feedback optimization, enabling reliable functional validation of neuron functionality. Through neuron regulation, we achieve directed induction of non-target sentences and enhancement of emotion tasks via rhetoric neurons. Experiments on 5 commonly used datasets validate the effectiveness of our method, providing a novel paradigm for the fine-grained steering of emotion and rhetoric expressions in LLMs.

[Long] Arg-LLaDA: Argument Summarization via Large Language Diffusion Models and Sufficiency-Aware Refinement ▶ *Underline Video*

Hao Li, Yizheng Sun, Viktor Schlegel, Kailai Yang, Riza Batista-Navarro and Goran Nenadic

Argument summarization aims to generate concise, structured representations of complex, multi-perspective debates. While recent work has advanced the identification and clustering of argumentative components, the generation stage remains underexplored. Existing approaches typically rely on single-pass generation, offering limited support for factual correction or structural refinement. To address this gap, we introduce ArgLLaDA, a novel large language diffusion framework that iteratively improves summaries via sufficiency-guided remasking and regeneration. Our method combines a flexible masking controller with a sufficiency-checking module to identify and revise unsupported, redundant, or incomplete spans, yielding more faithful, concise, and coherent outputs. Empirical results on two benchmark datasets demonstrate that ArgLLaDA surpasses state-of-the-art baselines in 7 out of 10 automatic evaluation metrics. In addition, human evaluations reveal substantial improvements across core dimensions, coverage, faithfulness, and conciseness,validating the effectiveness of our iterative, sufficiency-aware generation strategy.

[Long] Dynamic Emotion and Personality Profiling for Multimodal Deception Detection ▶ *Underline Video*

Li Zheng, Yanyi Luo, Hao Fei, Yuzhe Ding, Yujie Huang, Fei Li, Chong Teng and Donghong Ji

Deception detection is of great significance for ensuring information security and conducting public opinion analysis, with personality factors and emotion cues playing a critical role. However, existing methods lack sample-level dynamic annotations for emotions and personality. In this paper, we propose an innovative multi-model multi-prompt annotation scheme and a strict label quality evaluation standard, and establish a multimodal joint detection dataset DDEP for deception, emotion, and personality. Meanwhile, we propose Rel-DDEP, an adaptive reliability-weighted fusion framework. Our framework quantifies uncertainty by mapping modal features to a high-dimensional Gaussian distribution space. It then performs reliability-weighted fusion and incorporates an alignment module and a sorting constraint module to achieve joint detection of deception, emotion, and personality. Experimental results on the MDPE and DDEP datasets show that our Rel-DDEP significantly outperforms the existing state-of-the-art baseline models in three tasks. The F1 score of the deception detection increases by 2.53%, that of the emotion detection increases by 2.66%, and that of the personality detection increases by 9.30%. The experiments fully verify the necessity of annotating dynamic emotion and personality labels for each sample and the effectiveness of reliability-weighted fusion.

[Long] LLM-MC-Affect: LLM-Based Monte Carlo Modeling of Affective Trajectories and Latent Ambiguity for Interpersonal Dynamic Insight ▶ *Underline Video*

Yu-Zheng Lin, Bono Po-Jen Shih, John Paul Martin Encinas, Elizabeth Victoria Achom, Karan Patel, Jesus Horacio Pacheco, Sicong Shao, Jyotikrishna Dass, Soheil Salehi and Pratik Satam

Emotional coordination is a core property of human interaction that shapes how relational meaning is constructed in real time. While text-based affect inference has become increasingly feasible, prior approaches often treat sentiment as a deterministic point estimate for individual speakers, failing to capture the inherent subjectivity, latent ambiguity, and sequential coupling found in mutual exchanges. We introduce LLM-MC-Affect, a probabilistic framework that characterizes emotion not as a static label, but as a continuous latent probability distribution defined over an affective space. By leveraging stochastic LLM decoding and Monte Carlo estimation, the methodology approximates these distributions to derive high-fidelity sentiment trajectories that explicitly quantify both central affective tendencies and perceptual ambiguity. These trajectories enable a structured analysis of interpersonal coupling through sequential cross-correlation and slope-based indicators, identifying leading or lagging influences between interlocutors. To validate the interpretive capacity of this approach, we utilize teacher-student instructional dialogues as a representative case study, where our quantitative indicators successfully distill high-level interaction insights such as effective scaffolding. This work establishes a scalable and deployable pathway for understanding interpersonal dynamics, offering a generalizable solution that extends beyond education to broader social and behavioral research.

[Long] Persona-Ẽ: A Human-Grounded Dataset for Personality-Shaped Emotional Responses to Textual Events ▶ *Underline Video*

Yuqin Yang, Haowu Zhou, Haoran Tu, Zhiwen Hui, Shiqi Yan, HaoYang Li, Dong She, Xianrong Yao, Yang Gao and Zhanpeng Jin

Most affective computing research treats emotion as a static property of text, focusing on the writer's sentiment while overlooking the reader's perspective. This approach ignores how individual personalities lead to diverse emotional appraisals of the same event. Although role-playing Large Language Models (LLMs) attempt to simulate such nuanced reactions, they often suffer from "personality illusion"—relying

on surface-level stereotypes rather than authentic cognitive logic. A critical bottleneck is the absence of ground-truth human data to link personality traits to emotional shifts. To bridge the gap, we introduce Persona-Eš (Persona-Event2Emotion), a large-scale dataset grounded in annotated MBTI and Big Five traits to capture reader-based emotional variations across news, social media, and life narratives. Extensive experiments reveal that state-of-the-art LLMs struggle to capture precise appraisal shifts, particularly in social media domains. Crucially, we find that personality information significantly improves comprehension, with the Big Five traits alleviating "personality illusion.'

[Long] QA-MoE: Towards a Continuous Reliability Spectrum with Quality-Aware Mixture of Experts for Robust Multimodal Sentiment Analysis ▶ *Underline Video*

*Yitong Zhu, Yuxuan Jiang, Guanxuan Jiang, Bojing Hou, Peng Yuan Zhou, Ge lin and Yuyang Wang**

Multimodal Sentiment Analysis (MSA) aims to infer human sentiment from textual, acoustic, and visual signals. In real-world scenarios, however, multimodal inputs are often compromised by dynamic noise or modality missingness. Existing methods typically treat these imperfections as discrete cases or assume fixed corruption ratios, which limits their adaptability to continuously varying reliability conditions. To address this, we first introduce a Continuous Reliability Spectrum to unify missingness and quality degradation into a single framework. Building on this, we propose QA-MoE, a Quality-Aware Mixture-of-Experts framework that quantifies modality reliability via self-supervised aleatoric uncertainty. This mechanism explicitly guides expert routing, enabling the model to suppress error propagation from unreliable signals while preserving task-relevant information. Extensive experiments indicate that QA-MoE achieves competitive or state-of-the-art performance across diverse degradation scenarios and exhibits a promising One-Checkpoint-for-All property in practice.

[Long] Single-Pass, Depth-Selective Reading for Multi-Aspect Sentiment Analysis ▶ *Underline Video*

Yan Xia, Zhuangzhuang Pan, Amirrudin Kamsin and Chee Seng Chan

Aspect-Term Sentiment Analysis (ATSA) in multi-aspect sentences faces a fundamental tradeoff between efficiency and expressiveness. Existing models either re-encode the sentence for each aspect or rely on static use of deep representations, leading to redundant computation and limited adaptivity. We argue that Transformer depth is a costly, queryable resource, and propose DABS, a single-pass inference framework that encodes each sentence once to construct a reusable, depth-ordered substrate. Each aspect then queries this shared representation to selectively read relevant tokens and abstraction levels, without re-encoding. This decouples shared sentence encoding from lightweight, aspect-conditioned readout. Experiments on four ATSA benchmarks show that DABS achieves competitive performance while reducing end-to-end computation by up to 60% in multi-aspect settings (M 2). Further analyses indicate that adaptive depth querying is most beneficial for linguistically complex cases such as negation and contrast. Code is publicly available at https://github.com/panzhzh/acl-dabs.

[Long] Teaching LLMs Human-Like Editing of Inappropriate Argumentation via Reinforcement Learning ▶ *Underline Video*

Timon Ziegenbein, Maja Stahl and Henning Wachsmuth

Editing human-written text has become a standard use case of large language models (LLMs), for example, to make one's arguments more appropriate for a discussion. Comparing human to LLM-generated edits, however, we observe a mismatch in editing strategies: While LLMs often perform multiple scattered edits and tend to change meaning notably, humans rather encapsulate dependent changes in self-contained, meaning-preserving edits. In this paper, we present a reinforcement learning approach that teaches LLMs human-like editing to improve the appropriateness of arguments. Our approach produces self-contained sentence-level edit suggestions that can be accepted or rejected independently. We train the approach using group relative policy optimization with a multi-component reward function that jointly optimizes edit-level semantic similarity, fluency, and pattern conformity as well as argument-level appropriateness. In automatic and human evaluation, it outperforms competitive baselines and the state of the art in human-like editing, with multi-round editing achieving appropriateness close to full rewriting.

[Long] Uncovering Sentiment Analysis Circuit in Large Language Model ▶ *Underline Video*

Shichen Li, Zhouyang Wang, Zhongqing Wang and Peifeng Li

Large Language Models (LLMs) can perform sentiment analysis via natural language instructions, yet their predictions are highly sensitive to prompt phrasing. Prior work has shown that sentiment is encoded linearly in LLM representations, but the models ability to utilize this information remains surprisingly fragile to prompt variations. We leverage Sparse Autoencoders (SAEs) and circuit-level analysis to uncover causal mechanisms underlying sentiment prediction. We identify a sentiment analysis circuit and find that prompt sensitivity may stem from task activation failure. The model encodes the sentiment feature consistently, but different prompts trigger varying degrees of circuit activation. Based on this insight, we propose a simple inference-time intervention method that amplifies circuit features to compensate for insufficient activation. Experiments across diverse datasets, templates, and languages show consistent improvements, offering an interpretable and training-free alternative to manual prompt engineering.

Oral Session G: Special Theme: Explainability of NLP Models 2

11:00-12:30 (Harbor E-F)

[Long] Controllable LLM Reasoning via Sparse AutoencoderBased Steering ▶ *Underline Video*

Yi Fang, Wenjie Wang, Mingfeng Xue, Boyi Deng, Fengli Xu, Dayiheng Liu and Fuli Feng
Large Reasoning Models (LRMs) exhibit human-like cognitive reasoning strategies (e.g. backtracking, cross-verification) during reasoning process, which improves their performance on complex tasks. Currently, reasoning strategies are autonomously selected by LRMs themselves. However, such autonomous selection often produces inefficient or even erroneous reasoning paths. To make reasoning more reliable and flexible, it is important to develop methods for controlling reasoning strategies. Existing methods struggle to control fine-grained reasoning strategies due to conceptual entanglement in LRMs' hidden states. To address this, we leverage Sparse Autoencoders (SAEs) to decompose strategy-entangled hidden states into a disentangled feature space. To identify the few strategy-specific features from the vast pool of SAE features, we propose SAE-Steering, an efficient two-stage feature identification pipeline. SAE-Steering first recalls features that amplify the logits of strategy-specific keywords, filtering out over 99% of features, and then ranks the remaining features by their control effectiveness. Using the identified strategy-specific features as control vectors, SAE-Steering outperforms existing methods by over 15% in control effectiveness. Furthermore, controlling reasoning strategies can redirect LRMs from erroneous paths to correct ones, achieving a 7% absolute accuracy improvement.

[Long] DEFT: Demystifying VLN Failures via a Unified Dual-View Explainability Framework for LLM-based Agents ▶ *Underline Video*
Yawen Wang, Yihan Dai, Jianming Chen, Junjie Wang and Qing Wang
Large Language Models (LLMs) have emerged as central planners in Vision-and-Language Navigation (VLN), yet their complexity increasingly obscures their internal decision-making. Existing interpretability methods typically isolate temporal criticality from feature salience, creating an alignment gap and failing to account for the behavioral instability of black-box agents. To address this, we propose DEFT, a unified dual-view framework that demystifies agent behavior by jointly analyzing when a decision is pivotal and what visual evidence grounds it. Featuring a dual-head architecture with a shared latent representation, DEFT employs a Mask Head for counterfactual-based criticality detection and an Action Head that leverages an ensemble of surrogates to recover robust visual cues. Extensive experiments on MatterPort3D across three LLM-based agents demonstrate that DEFT outperforms baselines in both temporal and feature fidelity. User studies further validate its utility, showing 78% alignment with human intuition.

[Long] From Weights to Activations: Is Steering the Next Frontier of Adaptation? ▶ *Underline Video*
Simon Ostermann, Daniil Gurgurov, Tanja Baeumel, Michael A. Hedderich, Sebastian Lapuschkin, Wojciech Samek and Vera Schmitt
Post-training adaptation of large language models is commonly achieved through parameter updates or input based methods such as fine-tuning, parameter-efficient adaptation, and prompting. In parallel, a growing body of work modifies internal activations at inference time to influence model behavior, an approach known as steering. Despite increasing use, steering is rarely analyzed within the same conceptual framework as established adaptation methods. In this work, we argue that steering should be regarded as a form of model adaptation. We introduce a set of functional criteria for adaptation methods and use them to compare steering approaches with classical alternatives. This analysis positions steering as a distinct adaptation paradigm based on targeted interventions in activation space, enabling local and reversible behavioral change without parameter updates. The resulting framing clarifies how steering relates to existing methods, motivating a unified taxonomy for model adaptation.

[Long] LangSAE Editing: Improving Multilingual Information Retrieval via Post-hoc Language Identity Removal ▶ *Underline Video*
Dongjun Kim, Jeongho Yoon, Chanjun Park and Heuiseok Lim
Dense retrieval in multilingual settings often searches over mixed-language collections, yet multilingual embeddings encode language identity alongside semantics. This language signal can inflate similarity for same-language pairs and crowd out relevant evidence written in other languages. We propose LANGSAE EDITING, a post-hoc sparse autoencoder trained on pooled embeddings that enables controllable removal of language-identity signal directly in vector space. The method identifies language-associated latent units using cross-language activation statistics, suppresses these units at inference time, and reconstructs embeddings in the original dimensionality, making it compatible with existing vector databases without retraining the base encoder or re-encoding raw text. Experiments across multiple languages show consistent improvements in ranking quality and cross-language coverage, with especially strong gains for script-distinct languages.

[Long] Latent Agents: A Post-Training Procedure for Internalized Multi-Agent Debate ▶ *Underline Video*
John Seon Keun Yi, Aaron Mueller and Dokyun Lee
Multi-agent debate has been shown to improve reasoning in large language models (LLMs). However, it is compute-intensive, requiring generation of long transcripts before answering questions. To address this inefficiency, we develop a framework that distills multi-agent debate into a single LLM through a two-stage fine-tuning pipeline combining debate structure learning with internalization via dynamic reward scheduling and length clipping. Across multiple models and benchmarks, our internalized models match or exceed explicit multi-agent debate performance using up to 93% fewer tokens. We then investigate the mechanistic basis of this capability through activation steering, finding that internalization creates agent-specific subspaces: interpretable directions in activation space corresponding to different agent perspectives. We further demonstrate a practical application: by instilling malicious agents into the LLM through internalized debate, then applying negative steering to suppress them, we show that distillation makes harmful behaviors easier to localize and control with smaller reductions in general performance compared to steering base models. Our findings offer a new perspective for understanding multi-agent capabilities in distilled models and provide practical guidelines for controlling internalized reasoning behaviors.

[Long] Towards a Mechanistic Understanding of Large Reasoning Models: A Survey of Training, Inference, and Failures ► *Underline Video*

Yi Hu, Jiaqi Gu, Ruxin Wang, Zijun Yao, Hao Peng, Xiaobao Wu, Jianhui Chen, Muhan Zhang and Liangming Pan

Reinforcement learning (RL) has catalyzed the emergence of Large Reasoning Models (LRMs) that have pushed reasoning capabilities to new heights. While their performance has garnered significant excitement, exploring the internal mechanisms driving these behaviors has become an equally critical research frontier. This paper provides a comprehensive survey of the mechanistic understanding of LRMs, organizing recent findings into three core dimensions: 1) training dynamics, 2) reasoning mechanisms, and 3) unintended behaviors. By synthesizing these insights, we aim to bridge the gap between black-box performance and mechanistic transparency. Finally, we discuss under-explored challenges to outline a roadmap for future mechanistic studies, including the need for applied interpretability, improved methodologies, and a unified theoretical framework.

[Long] Towards Intrinsic Interpretability of Large Language Models: A Survey of Design Principles and Architectures ► *Underline Video*

Yutong Gao, Qinglin Meng, Yuan Zhou and Liangming Pan

While Large Language Models (LLMs) have achieved strong performance across many NLP tasks, their opaque internal mechanisms hinder trustworthiness and safe deployment. Existing surveys in explainable AI largely focus on post-hoc explanation methods that interpret trained models through external approximations. In contrast, intrinsic interpretability, which builds transparency directly into model architectures and computations, has recently emerged as a promising alternative. This paper presents a systematic review of the recent advances in intrinsic interpretability for LLMs, categorizing existing approaches into five design paradigms: functional transparency, concept alignment, representational decomposability, explicit modularization, and latent sparsity induction. We further discuss open challenges and outline future research directions in this emerging field. The paper list is available at: \{ } Survey-Intrinsic-Interpretability-of-LLMs

[Long] Tracing Logit Trajectories Across Layer Depth: Dataset-Level Explainability for Language Models ► *Underline Video*

Jeesu Jung and Sangkeun Jung

Sentence-level explanations can miss the bigger picture of how a black-box model behaves across data, which matters most for complex criteria like safety that cannot be defined by a single rule. We trace Logit-Trajectory, which tracks adjacent-layer logit updates as vectors and aggregates them into a reproducible dataset-level trajectory pattern, enabling depth-wise explainability through signals such as coherence and angular rotation. Across 6 languages and 5 NLP tasks, we show these trajectory summaries reveal consistent depth-wise patterns that divergence- and similarity-based baselines often wash out due to scalarization. As a case study where dataset-level intermediate decision structure matters, we evaluate safety classification, reporting both trajectory-level visual separability and classification performance.

[Long] Tracing the Roots: A Multi-Agent Framework for Uncovering Data Lineage in Post-Training LLMs ► *Underline Video*

Yu Li, Xiaoran shang, Qizhi Pei, Yun Zhu, Xin Gao, Honglin Lin, Zhanping Zhong, Zhuoshi Pan, Zheng Liu, Xiaoyang Wang, Conghui He, Dahua Lin, Feng Zhao and Lijun Wu

Post-training data plays a pivotal role in shaping the capabilities of Large Language Models (LLMs), yet datasets are often treated as isolated artifacts, overlooking the systemic connections that underlie their evolution. To disentangle these complex relationships, we introduce the concept of data lineage to the LLM ecosystem and propose an automated multi-agent framework to reconstruct the evolutionary graph of dataset development. Through large-scale lineage analysis, we characterize domain-specific structural patterns, such as vertical refinement in Math-oriented datasets and horizontal aggregation in General-domain corpora. Moreover, we uncover pervasive systemic issues, including structural redundancy induced by implicit dataset intersections and the propagation of benchmark contamination along lineage paths. To demonstrate the practical value of lineage analysis for data construction, we leverage the reconstructed lineage graph to create a lineage-aware diversity-oriented dataset. By anchoring instruction sampling at upstream leaf sources, this approach mitigates downstream homogenization and hidden redundancy, yielding a more diverse post-training corpus. We further highlight lineage-centric analysis as an efficient and robust topological alternative to sample-level dataset comparison for large-scale data ecosystems. By grounding data construction in explicit lineage structures, our work advances post-training data curation toward a more systematic and controllable paradigm.

Poster Session G (Coronado)

11:00-12:30 (Coronado)

[Long] AdabNER: Arabic Digital Archive Books with Nested Entity Recognition ► *Underline Video*

Aya Mourad and Mustafa Jarrar

Most studies on Arabic Named Entity Recognition (NER) have focused on news texts and social media posts, while the large and rich corpus of literary Arabic books has been underrepresented. We introduce AdabNER, the first large-scale nested NER dataset for Modern Standard Arabic (MSA) literary texts, comprising the first 6,000 words annotated from each of 138 books spanning ten literary genres, including history, biography, literary criticism, and travel literature, and covering works from the 1880s to the 2020s. The corpus comprises about 876K tokens, manually annotated using a nested 21 entity tag annotation scheme, yielding 78,530 entity mentions, 18.96% of which are

nested. We fine-tuned five pre-trained Arabic BERT encoders in two settings: stratified and leave-book-out, achieving F1 scores of 0.86 and 0.83 with AraBERTv2, respectively. We also evaluated five large language models through few-shot in-context learning, including open-source models and the closed-source Gemini 3 Pro, with Gemini 3 Pro achieving the highest LLM F1 score of 0.59. Supervised results degraded under out-of-domain evaluation; however, joint multi-domain training reduced this gap to less than a 1% F1 loss, demonstrating that domain-diverse training data is key to robust Arabic NER, though broader validation beyond the experiments reported is needed. AdabNER and its annotation guidelines are publicly available at https://doi.org/10.5281/zenodo.19468385.

[Findings] Addressing Overthinking in Large Vision-Language Models via Gated Perception-Reasoning Optimization ▶ *Underline Video*

Xingjian Diao, Zheyuan Liu, Chunhui Zhang, Weiyi Wu, Keyi Kong, Lin Shi, Kaize Ding, Soroush Vosoughi and Jiang Gui

Large Vision-Language Models (LVLMs) have exhibited strong reasoning capabilities through chain-of-thought mechanisms that generate step-by-step rationales. However, such slow-thinking approaches often lead to overthinking, where models produce excessively verbose responses even for simple queries, resulting in test-time inefficiency and even degraded accuracy. Prior work has attempted to mitigate this issue via adaptive reasoning strategies, but these methods largely overlook a fundamental bottleneck: visual perception failures. We argue that stable reasoning critically depends on low-level visual grounding, and that reasoning errors often originate from imperfect perception rather than insufficient deliberation. To address this limitation, we propose Gated Perception-Reasoning Optimization (GPRO), a meta-reasoning controller that dynamically routes computation among three decision paths at each generation step: a lightweight fast path, a slow perception path for re-examining visual inputs, and a slow reasoning path for internal self-reflection. To learn this distinction, we derive large-scale failure attribution supervision from approximately 790k samples, using teacher models to distinguish perceptual hallucinations from reasoning errors. We then train the controller with multi-objective reinforcement learning to optimize the trade-off between task accuracy and computational cost under uncertainty. Experiments on five benchmarks demonstrate that GPRO substantially improves both accuracy and efficiency, outperforming recent slow-thinking methods while generating significantly shorter responses.

[Long] Alexandria: A Multi-Domain Dialectal Arabic Machine Translation Dataset for Culturally Inclusive and Linguistically Diverse LLMs ▶ *Underline Video*

Abdellah El Mekki, Samar M. Magdy, Houdaifa Atou, Ruwa AbuHweidi, Baraah Qawasmeh, Omer Nacar, Thikra Al-hibiri, Razan Saadie, Hamzah Alsayadi, Nadia Ghezaiel Hammouda, Alshima Alkhazimi, Aya Hamod, Al-Yas Al-Ghafri, Wesam El-Sayed, Asila Al sharji, Mohamad Ballout, Anas Belfathi, Karim Ghaddar, Serry Sibaee, Alaa Aoun, Areej Asiri, Lina Abureesh, Ahlam Bashiti, Majdal Yousef, Abdulaziz Hafiz, Yehdih Mohamed, Emira Hamedtou, Brakehe Brahim, Rahaf Alhamouri, Youssef Nafea, Aya El Aatar, Walid Al-Dhabyani, Emhemed Hamed, Sara Shatnawi, Fakhraddin Alwajih, Khalid Elkhidir, Ashwaq Alasmari, Abdurrahman Gerrio, Omar Alshahri, AbdelRahim A. Elmadany, Ismail Berrada, Amir Azad Adli Alkathiri, Fadi A Zaraket, Mustafa Jarrar, Yahya Mohamed El Hadj, Hassan Alhuzali and Muhammad Abdul-Mageed

Arabic is a highly diglossic language where most daily communication occurs in regional dialects rather than Modern Standard Arabic (MSA). Despite this, machine translation (MT) systems often generalize poorly to dialectal input, limiting their utility for millions of speakers. We introduce Alexandria, a large-scale, community-driven, human-translated dataset designed to bridge this gap. Alexandria covers 13 Arab countries and 11 high-impact domains, including health, education, and agriculture. Unlike previous resources, Alexandria provides unprecedented granularity by associating contributions with city-of-origin metadata, capturing authentic local varieties beyond coarse regional labels. The dataset consists of parallel English-Dialectal Arabic multi-turn conversational scenarios annotated with speaker-addressee gender configurations, enabling the study of gender-conditioned variation in dialectal use. Comprising 107K total turns, Alexandria serves as both a training resource and as a rigorous benchmark for evaluating MT and Large Language Models (LLMs). Our automatic and human evaluation benchmarks the current capabilities of Arabic-aware LLMs in translating across diverse Arabic dialects and sub-dialects while exposing significant persistent challenges. The Alexandria dataset, the creation prompts, the translation and revision guidelines, and the evaluation code are publicly available in the following repository: https://github.com/UBC-NLP/Alexandria

[Findings] All Prompts Are Created Equal? Evaluating Robustness of LLM Judges Against Non-Adversarial Prompt Variations ▶ *Underline Video*

Savita Bhat and Vasudeva Varma

LLM-based evaluation systems (LLM judges) have emerged as a scalable alternative to expensive human evaluations. Although LLM judges demonstrate 70-80% agreement with human evaluators, their robustness under semantically equivalent prompt variations remains underexplored. Through systematic evaluation of 8 models across 4 NLG tasks using 10 semantically equivalent paraphrases per prompt (~115000 evaluations), we identify a critical accuracy-robustness gap: attribute verifiability affects the robustness more than model choice, with factually verifiable attributes achieving 0.71 accuracy versus 0.19 for subjective attributes. Our investigations discover three key insights: 1) Task structure characteristics influence the robustness and in turn accuracy, 2) Attribute verifiability as the strongest predictor-factually verifiable attribute achieve 0.71 accuracy versus 0.19 for subjective attributes, 3) No single winning model-smallest model (Llama-3.1-8B) exhibits second-best performance, while the strongest model (Llama-4) from the same family significantly lag behind, thus demonstrating that general capability improvements do not necessarily result in evaluation robustness. With these findings, we propose a diagnostic framework grounded in attribute verifiability that enables principled decisions about evaluation automation. Our work establishes new standards for assessing LLM judge reliability beyond simple accuracy metrics.

[Long] Beyond the Leaderboard: Rethinking Medical Benchmarks for Large Language Models ▶ *Underline Video*

Wenxuan Wang, Zizhan Ma, Guo Yu, Yiu-Fai Cheung, Meidan Ding, Jie Liu, Wenting Chen and Linlin Shen

Large language models (LLMs) show significant potential in healthcare, prompting numerous benchmarks to evaluate their capabilities. However, concerns persist regarding the reliability of these benchmarks, which often lack clinical fidelity, robust data management, and safety-oriented evaluation metrics. To address these shortcomings, we introduce MedCheck, the first lifecycle-oriented assessment framework specifically designed for medical benchmarks. Our framework deconstructs benchmark development into five stages from design to governance, and provides a comprehensive checklist of 46 medically-tailored criteria. Using MedCheck, we conducted an in-depth empirical evaluation of 56 medical LLM benchmarks. Our analysis uncovers widespread, systemic issues, including a profound disconnect from clinical practice, a crisis of data integrity due to unmitigated contamination risks, and a systematic neglect of safety-critical evaluation dimensions like model robustness and uncertainty awareness. Based on these findings, MedCheck is both a diagnostic tool for existing benchmarks and an actionable guideline for a more standardized, reliable, and transparent approach to evaluating AI in healthcare.

[Long] Can LLM Agents Simulate Multi-Turn Human Behavior? Evidence from Real Online Customer Behavior Data ▶ *Underline Video*

Yuxuan Lu, Jing Huang, Yan Han, Bingsheng Yao, Sisong Bei, Yaochen Xie, Yisi Sang, Qi He and Dakuo Wang

Recent research shows that LLM Agents can generate "believable" human behaviors via prompt-only methods, and such agents have been increasingly adopted in downstream applications. However, existing evaluation of these agents only focuses on qualitative believability (whether human raters think they are accurate), leaving open questions of whether LLM agents can accurately generate step-by-step actions mimicking a particular human's behavior in a multi-turn interaction task. In this work, we take shopping as a case study and present the first large-scale quantitative evaluation of state-of-the-art LLMs' ability to accurately simulate human behavior. Using real-world data from 31,865 online shopping sessions containing 230,965 user actions, our evaluation reveals that prompt-based LLMs (DeepSeek-R1, Llama, Claude) achieve only 11.86% accuracy in generating human actions, highlighting a substantial gap in actual behavioral accuracy. Through experiments, we also showcase that strategies as simple as fine-tuning LLMs on real human click-through data augmented with synthesized reasoning traces can greatly enhance models' performance. The fine-tuned Qwen2.5-7B achieves 17.26% action generation accuracy and 33.86% F1 score on final purchase prediction, representing substantial improvements of 5.4% and 13.85% over prompt-only baselines. This work establishes the first rigorous benchmark and dataset for human behavior simulation and provides actionable insights for developing more accurate LLM agents for future downstream applications.

[Findings] Can LLMs Estimate Student Struggles? Human-AI Difficulty Alignment with Proficiency Simulation for Item Difficulty Prediction ▶ *Underline Video*

Ming Li

Accurate estimation of item (question or task) difficulty is critical for educational assessment but suffers from the cold start problem. While Large Language Models demonstrate superhuman problem-solving capabilities, it remains an open question whether they can perceive the cognitive struggles of human learners. In this work, we present a large-scale empirical analysis of Human-AI Difficulty Alignment for over 20 models across diverse domains such as medical knowledge and mathematical reasoning. Our findings reveal a systematic misalignment where scaling up model size is not reliably helpful; instead of aligning with humans, models converge toward a shared machine consensus. We observe that high performance often impedes accurate difficulty estimation, as models struggle to simulate the capability limitations of students even when being explicitly prompted to adopt specific proficiency levels. Furthermore, we identify a critical lack of introspection, as models fail to predict their own limitations. These results suggest that general problem-solving capability does not imply an understanding of human cognitive struggles, highlighting the challenge of using current models for automated difficulty prediction.

[Long] CaseFacts: A Benchmark for Legal Fact-Checking and Precedent Retrieval ▶ *Underline Video*

Akshith Reddy Putta, Jacob Devasier and Chengkai Li

Automated Fact-Checking has largely focused on verifying general knowledge against static corpora, overlooking high-stakes domains like law where truth is evolving and technically complex. We introduce CaseFacts, a benchmark for verifying colloquial legal claims against U.S. Supreme Court precedents. Unlike existing resources that map formal texts to formal texts, CaseFacts challenges systems to bridge the semantic gap between layperson assertions and technical jurisprudence while accounting for temporal validity. The dataset consists of 6,294 claims categorized as Supported, Refuted, or Overruled. We construct this benchmark using a multi-stage pipeline that leverages Large Language Models (LLMs) to synthesize claims from expert case summaries, employing a novel semantic similarity heuristic to efficiently identify and verify complex legal overrulings. Experiments with state-of-the-art LLMs reveal that the task remains challenging; notably, augmenting models with unrestricted web search degrades performance compared to closed-book baselines due to the retrieval of noisy, non-authoritative precedents. We release CaseFacts to spur research into legal fact verification systems.

[Findings] ChangJuan: A Comprehensive Benchmark for Book-Length Chinese Story Evaluation ▶ *Underline Video*

Dingyi Yang, Mingshuo Wang and Qin Jin

Automatic evaluation of book-length stories remains underexplored, particularly for non-English literature. We introduce ChangJuan, the first benchmark for book-length Chinese story evaluation, comprising 300 novels with metadata, human ratings, and large-scale user reviews. To mitigate the subjectivity of raw reviews, we propose a distillation method to aggregate them into generally agreed viewpoints (pros and cons) across key evaluation aspects such as plot and character. We conduct systematic experiments to benchmark current LLMs, analyze

aspect importance, and examine genre differences. For book-length story evaluation, we propose an enhanced summary-based method that leverages length-detail balanced summaries and representative excerpts, generates aspect-specific reviews, and considers genre-aware aspect weighting to assign a final score. Using this framework and our distilled viewpoints, we fine-tune an 8B model, CLEM, which outperforms open-source baselines and raises Qwen3's Kendall's tau correlation with human judgments from 24.8 to 34.1. Our datasets and codes are available at https://github.com/DingyiYang/ChangJuan.

[Long] Clozing the Gap: Exploring Why Language Model Surprisal Outperforms Cloze Surprisal ► *Underline Video*

Sathvik Nair and Byung-Doh Oh

How predictable a word is can be generally quantified in two ways: using human responses to the cloze task or using probabilities from language models (LMs). When used as predictors of processing effort, LM probabilities outperform probabilities derived from cloze data. However, it is important to establish that LM probabilities do so for the right reasons, since different predictors can lead to different scientific conclusions about the role of prediction in language comprehension. We present evidence for three hypotheses about the apparent advantage of LM probabilities: not suffering from low resolution, distinguishing semantically similar words, and accurately assigning probabilities to low-frequency words. These results call for efforts to improve the resolution of cloze studies, coupled with experiments on whether human-like prediction is also as sensitive to the fine-grained distinctions made by LM probabilities.

[Findings] CMTD: Cognitive Modeling with Traits and Distortions for Multimodal Emotion Recognition in Conversations ► *Underline Video*

Minh-Tien Nguyen, Huu-Loi Le, Manh-Cuong Phan and Hajime Hotta

This paper introduces a new multi-agent framework, CMTD (Cognitive Modeling with Traits and Distortions), for multimodal emotion recognition in conversations (MERC). Instead of relying on shallow analysis of emotions, CMTD reconstructs a cognitive model by taking advantage of stable personality traits, dynamic cognitive distortions, visual and acoustic features of interlocutors to enhance the emotional intelligence of LLMs. CMTD includes trait, distortion detection, vision, and speech agents that provide psychological and multimodal indicators for the fusion agent to make the final prediction. Experimental results on MELD and IEMOCAP show that traits temper negativity bias from distortions, and cognitive modeling with psychological, visual, and acoustic information can improve the performance of MERC. CMTD is flexible and easy to adapt to advanced emotional AI systems (Github link: https://github.com/Shaun-le/CMTD.git).

[Long] CommonLID: Re-evaluating State-of-the-Art Language Identification Performance on Web Data ► *Underline Video*

Pedro Ortiz Suarez, Laurie Burchell, Catherine Arnett, Rafael Mosquera-Gómez, Sara Hincapie-Monsalve, Thom Vaughan, Damian Stewart, Malte Ostendorff, Idris Abdulmumin, Vukosi Marivate, Shamsuddeen Hassan Muhammad, Atnafu Lambebo Tonja, Hend Al-Khalifa, Nadia Ghezaiel Hammouda, Verrah Otiende, Tack Hwa Wong, Jakhongir Saydaliev, Melika Nobakhtian, Muhammad Ravi Shulthan Habibi, Chalamalasetti Kranti, Carol Muchemi, Khang Nguyen, Faisal Muhammad Adam, Luis Frentzen Salim, Reem Alqifari, Cynthia Amol, Joseph Marvin Imperial, Ilker Kesen, Ahmad Mustafid, Pavel Stepachev, Leshem Choshen, David Anugraha, Hamada Nayel, Seid Muhie Yimam, Vallerie Alexandra Putra, My Chiffon Nguyen, Azmine Toushik Wasi, Gouthami Vadithya, Rob van der Goot, Lanwenn ar C'horr, Karan Dua, Andrew Yates, Mithil Bangera, Yeshil Bangera, Hitesh Laxmichand Patel, Shu Okabe, Fenal Ashokbhai Ilasariya, Dmitry Gaynullin, Genta Indra Winata, Yiyuan Li, Juan Pablo Martínez, Amit Agarwal, Ikhlasul Akmal Hanif, Raia Abu Ahmad, Esther Adenuga, Filbert Aurelian Tjiaranata, Weerayut Buaphet, Michael Anugraha, Sowmya Vajjala, Benjamin Rice, Azril Hafizi Amirudin, Jesujoba O. Alabi, Srikant Panda, Yassine Toughrai, Bruhan Kyomuhendo, Daniel Ruffinelli, Akshata A, Manuel Goulão, Ej Zhou, Ingrid Gabriela Franco Ramirez, Cristina Aggazzotti, Konstantin Dobler, Jun Kevin, Quentin Pagès, Nicholas Andrews, Nuhu Ibrahim, Mattes Ruckdeschel, Amr Keleg, Mike Zhang, Casper Muziri, Saron Samuel, Sotaro Takeshita, Kun Kerdthaisong, Luca Foppiano, Rasul Dent, Tommaso Green, Ahmad Mustapha Wali, Kamohelo Makaaka, Vicky Feliren, Inshirah Idris, Hande Celikkanat, Abdulhamid Abubakar, Jean Maillard, Benoît Sagot, Thibault Clérice, Kenton Murray and Sarah Luger

Language identification (LID) is a fundamental step in curating multilingual corpora. However, LID models still perform poorly for many languages, especially on the noisy and heterogeneous web data often used to train multilingual language models. In this paper, we introduce CommonLID, a community-driven, human-annotated LID benchmark for the web domain, covering 109 languages. Many of the included languages have been previously under-served, making CommonLID a key resource for developing more representative high-quality text corpora. We show CommonLID's value by using it, alongside five other common evaluation sets, to test eight popular LID models. We analyse our results to situate our contribution and to provide an overview of the state of the art. In particular, we highlight that existing evaluations overestimate LID accuracy for many languages in the web domain. We make CommonLID and the code used to create it available under an open, permissive license.

[Long] Comparing human and language models sentence processing difficulties on complex structures ► *Underline Video*

Samuel Joseph Amouyal, Aya Meltzer-Asscher and Jonathan Berant

Large language models (LLMs) that fluently converse with humans are a reality - but do LLMs experience human-like processing difficulties? We systematically compare human and LLM sentence comprehension across seven challenging linguistic structures. We collect sentence comprehension data from humans and five families of state-of-the-art LLMs, varying in size and training procedure in a unified experimental framework. Our results show LLMs overall struggle on the target structures, but especially on garden path (GP) sentences. Indeed, while the strongest models achieve near perfect accuracy on non-GP structures (93.7% for GPT-5), they struggle on GP structures (46.8% for GPT-5). Additionally, when ranking structures based on average performance, rank correlation between humans and models increases with parameter

count. For each target structure, we also collect data for their matched baseline without the difficult structure. Comparing performance on the target vs. baseline sentences, the performance gap observed in humans holds for LLMs, with two exceptions: for models that are too weak performance is uniformly low across both sentence types, and for models that are too strong the performance is uniformly high. Together, these reveal convergence and divergence in human and LLM sentence comprehension, offering new insights into the similarity of humans and LLMs.

[Findings] Comprehensiveness Metrics for Automatic Evaluation of Factual Recall in Text Generation ▶ *Underline Video*

Adam Dejl, James Barry, Alessandra Pascale and Javier Carnerero-Cano

Despite demonstrating remarkable performance across a wide range of tasks, large language models (LLMs) have also been found to frequently produce outputs that are incomplete or selectively omit key information. In sensitive domains, such omissions can result in significant harm comparable to that posed by factual inaccuracies, including hallucinations. In this study, we address the challenge of evaluating the comprehensiveness of LLM-generated texts, focusing on the detection of missing information or underrepresented viewpoints. We investigate three automated evaluation metrics: (1) an NLI-based method that decomposes texts into atomic statements and uses natural language inference (NLI) to identify missing facts, (2) a Q&A-based metric that extracts question-answer pairs and compares responses across sources, and (3) an end-to-end approach that directly identifies missing content using LLMs. Our experiments demonstrate the surprising effectiveness of the simple end-to-end metric compared to more complex metrics, though at the cost of reduced robustness, interpretability and result granularity. We further assess the comprehensiveness of responses from several popular open-weight LLMs when answering user queries based on multiple sources.

[Findings] Context-Driven and Reference-Guided Data Augmentation for Subtitle Translation ▶ *Underline Video*

Hitoshi Ito, Naoto Shirai, Kazutaka Kinugawa, Hideya Mino, Rei Endo and Yoshihiko Kawai

Large language models (LLMs) have demonstrated strong performance in translation tasks. Subtitle translation presents unique challenges, such as preserving the original works worldview and the distinctive speaking styles of its characters. Achieving high-quality translations that reflect these stylistic nuances typically requires bilingual data for a specific movie, which is often scarce or unavailable. Thus, we propose a data augmentation method that uses LLMs to improve translation performance for specific movies, even when only a few hundred bilingual sentence pairs are available. The method expands source-side data by rewriting original subtitles using information that can be extracted from the context, such as character profiles and scene descriptions, to maintain the tone and thematic consistency of the movie. For translation, the augmented sentences are aligned with manually translated originals using structural similarity, which enables style-preserving bilingual data generation via one-shot learning. Experimental results show that data augmented using the proposed method effectively improves BLEU scores for film subtitle translation, and achieves superior stylistic quality in human evaluation.

[Long] CoreCodeBench: Decoupling Code Intelligence via Fine-Grained Repository-Level Tasks ▶ *Underline Video*

Lingyue Fu, Hao Guan, Bolun Zhang, Haowei Yuan, Yaoming Zhu, Lin Qiu, ZongYu Wang, Xuezhi Cao, Xunliang Cai, Weiwen Liu, Weinan Zhang and Yong Yu

The evaluation of Large Language Models (LLMs) for software engineering has shifted towards complex, repository-level tasks. However, existing benchmarks predominantly rely on coarse-grained pass rates that treat programming proficiency as a monolithic capability, obscuring specific cognitive bottlenecks. Furthermore, the static nature of these benchmarks renders them vulnerable to data contamination and performance saturation. To address these limitations, we introduce CoreCodeBench, a configurable repository-level benchmark designed to dissect coding capabilities through atomized tasks. Leveraging our automated framework, CorePipe, we extract and transform Python repositories into a comprehensive suite of tasks that isolate distinct cognitive demands within identical code contexts. Unlike static evaluations, CoreCodeBench supports controllable difficulty scaling to prevent saturation and ensures superior data quality. It achieves a 78.55% validity yield, significantly surpassing the 31.7% retention rate of SWE-bench-Verified. Extensive experiments with state-of-the-art LLMs reveal a significant capability misalignment, evidenced by distinct ranking shifts across cognitive dimensions. This indicates that coding proficiency is non-monolithic, as strength in one aspect does not necessarily translate to others. These findings underscore the necessity of our fine-grained taxonomy in diagnosing model deficiencies and offer a sustainable, rigorous framework for evolving code intelligence. Code of CorePipe framework and data of CoreCodeBench are available in https://github.com/AGI-Eval-Official/CoreCodeBench and https://huggingface.co/collections/tubehhh/corecodebench.

[Long] Cross-Modal Taxonomic Generalization in (Vision-) Language Models ▶ *Underline Video*

Tianyang Xu, Marcelo Sandoval-Castaneda, Karen Livescu, Greg Shakhnarovich and Kanishka Misra

What is the interplay between semantic representations learned by language models (LM) from surface form alone to those learned from more grounded evidence? We study this question for a scenario where part of the input comes from a different modality—in our case, in a vision-language model (VLM), where a pretrained LM is aligned with a pretrained image encoder. As a case study we focus on the task of predicting hypernyms of objects represented in images. We do so in a VLM setup where the image encoder and LM are kept frozen, and only the intermediate mappings are learned. We progressively deprive the VLM of explicit evidence for hypernyms, and test whether the knowledge is recoverable from the LM. We find that the LMs we study can recover this knowledge and generalize even in the most extreme version of this experiment (when the model receives no evidence of a hypernym during training). Additional experiments suggest that this cross-modal taxonomic generalization persists under counterfactual image-label mappings only when the counterfactual data have high visual similarity within each category. Taken together, these findings suggest that cross-modal generalization in LMs arises from an

interaction between linguistic structure and the information present in the input.

[Long] Cultural Benchmarking of LLMs in Standard and Dialectal Arabic Dialogues ▶ *Underline Video*

Muhammad Dehan Al Kautsar, Saeed Almheiri, Momina Ahsan, Bilal Elbouardi, Younes Samih, Sarfraz Ahmad, Amr Keleg, Omar El Herraoui, Kareem Elzeky, Abed Alhakim Freihat, Mohamed Anwar, Zhuohan Xie, Junhong Liang, Mohammad Rustom Al Nasar, Preslav Nakov and Fajri Koto

There is a significant gap in evaluating cultural reasoning in LLMs using conversational datasets that capture culturally rich and dialectal contexts. Most Arabic benchmarks focus on short text snippets in Modern Standard Arabic (MSA), overlooking the cultural nuances that naturally arise in dialogues. To address this gap, we introduce ArabCulture-Dialogue, a culturally grounded conversational dataset covering 13 Arabic-speaking countries, in both MSA and each country's respective dialect, spanning 12 daily-life topics and 54 fine-grained subtopics. We utilize the dataset to form three benchmarking tasks: (i) multiple-choice cultural reasoning, (ii) machine translation between MSA and dialects, and (iii) dialect-steering generation. Our experiments indicate that the performance gap between MSA and Arabic dialects still exists, whereby the models perform worse on all three tasks in the dialectal setup, compared to the MSA one.

[Findings] Demystifying Multi-Agent Debate: The Role of Confidence and Diversity ▶ *Underline Video*

Xiaochen Zhu, Caiqi Zhang, Yizhou Chi, Tom Stafford, Nigel Collier and Andreas Vlachos

Multi-agent debate (MAD) is widely used to improve large language model (LLM) performance through test-time scaling, yet recent work shows that vanilla MAD often underperforms simple majority vote despite higher computational cost. Studies show that, under homogeneous agents and uniform belief updates, debate preserves expected correctness and therefore cannot reliably improve outcomes. Drawing on findings from human deliberation and collective decision-making, we identify two key mechanisms missing from vanilla MAD: (i) diversity of initial viewpoints and (ii) explicit, calibrated confidence communication. We propose two lightweight interventions. First, a diversity-aware initialisation that selects a more diverse pool of candidate answers, increasing the likelihood that a correct hypothesis is present at the start of debate. Second, a confidence-modulated debate protocol in which agents express calibrated confidence and condition their updates on others confidence. We show theoretically that diversity-aware initialisation improves the prior probability of MAD success without changing the underlying update dynamics, while confidence-modulated updates enable debate to systematically drift to the correct hypothesis. Empirically, across six reasoning-oriented QA benchmarks, our methods consistently outperform vanilla MAD and majority vote. Our results connect human deliberation with LLM-based debate and demonstrate that simple, principled modifications can substantially enhance debate effectiveness.

[Long] Evaluating the Impact of Verbal Multiword Expressions on Machine Translation ▶ *Underline Video*

Linfeng Liu, Saptarshi Ghosh and Tianyu Jiang

Verbal multiword expressions (VMWEs) remain difficult for machine translation because their meanings are often not recoverable from their component words. In this study, we analyze the impact of three VMWE categories—verbal idioms, verb-particle constructions, and light verb constructions—on machine translation quality from English to multiple languages. Using both established multiword expression datasets and standard machine translation datasets, we evaluate how state-of-the-art translation systems handle these expressions. Our experimental results consistently show that VMWEs negatively affect translation quality, with deeper analysis indicating that this degradation is primarily attributable to the VMWE itself rather than general sentence-level difficulty. We release our code and evaluation framework to test new MT systems for the community.

[Findings] Evaluation Pitfalls and Sparsity Limitations in LLM-based Confidence Estimates for Classification ▶ *Underline Video*

Elena Merdjanovska, Omar Zaidan and Andreas Rücklé

Confidence estimation is essential when LLMs are used for classification, indicating when predictions can be trusted. However, common approaches such as verbalization produce extremely sparse outputs. For instance, Qwen3-32B verbalizes only eight unique confidence values on SST-2, with over half being exactly 95%, a pattern we observe consistently across four datasets and two LLMs. Besides limiting practical utility, we show that this sparsity critically affects evaluation: the choice of interpolation in area under the accuracy-rejection curve (AUARC) dramatically alters rankings, with consistency sampling dropping from best to worst under stepwise versus linear interpolation. We advocate for standardizing stepwise interpolation for a fairer comparison. Under such a fair evaluation, we find that weighting verbalized digits by token probabilities, a method we term verbalization logprobs, addresses sparsity and achieves the best AUARC (+2.3 points over vanilla verbalization) without incurring additional inference cost.

[Long] EVM-QuestBench: An Execution-Grounded Benchmark for Natural-Language Transaction Code Generation ▶ *Underline Video*

Pei Yang, wanyi Chen, Ke Wang, Lynn Ai, Eric Yang and TIANYU SHI

Large language models are increasingly applied to various development scenarios. However, in on-chain transaction scenarios, even a minor error can cause irreversible loss for users. Existing evaluations often overlook execution accuracy and safety. We introduce EVM-QuestBench, an execution-grounded benchmark for natural-language transaction-script generation on EVM-compatible chains. The benchmark employs dynamic evaluation: instructions are sampled from template pools, numeric parameters are drawn from predefined intervals, and validators verify outcomes against these instantiated values. EVM-QuestBench contains 107 tasks (62 atomic, 45 composite). Its modular architecture enables rapid task development. The runner executes scripts on a forked EVM chain with snapshot isolation;

composite tasks apply step-efficiency decay. We evaluate 20 models with 5 independent rounds each and find large performance gaps, with split scores revealing persistent asymmetry between single-action precision and multi-step workflow completion. Code: https://github.com/OpenEdgeHQ/EVM-quest-bench.

[Long] FairQE: Multi-Agent Framework for Mitigating Gender Bias in Translation Quality Estimation ▶ *Underline Video*

Jinhee Jang, Juhwan Choi, Dongjin Lee, Seunguk Yu and YoungBin Kim

Quality Estimation (QE) aims to assess machine translation quality without reference translations, but recent studies have shown that existing QE models exhibit systematic gender bias. In particular, they tend to favor masculine realizations in gender-ambiguous contexts and may assign higher scores to gender-misaligned translations even when gender is explicitly specified. To address these issues, we propose FairQE, a multi-agent-based, fairness-aware QE framework that mitigates gender bias in both gender-ambiguous and gender-explicit scenarios. FairQE detects gender cues, generates gender-flipped translation variants, and combines conventional QE scores with LLM-based unbiased reasoning through a dynamic bias-aware aggregation mechanism. This design preserves the strengths of existing QE models while calibrating their gender-related biases in a plug-and-play manner. Extensive experiments across multiple gender bias evaluation settings demonstrate that FairQE consistently improves gender fairness over strong QE baselines. Moreover, under MQM-based meta-evaluation following the WMT 2023 Metrics Shared Task, FairQE achieves competitive or improved general QE performance. These results show that gender bias in QE can be effectively mitigated without sacrificing evaluation accuracy, enabling fairer and more reliable translation evaluation.

[Long] From 1,000,000 Users to Every User: Scaling Up Personalized Preference for User-level Alignment ▶ *Underline Video*

Jia-Nan Li, Jian Guan, Songhao Wu, Wei Wu and Rui Yan

Large language models (LLMs) have traditionally been aligned through one-size-fits-all approaches that assume uniform human preferences, fundamentally overlooking the diversity in user values and needs. This paper introduces a comprehensive framework for scalable personalized alignment of LLMs. We establish a systematic preference space characterizing psychological and behavioral dimensions, alongside diverse persona representations for robust preference inference in real-world scenarios. Building upon this foundation, we introduce AlignX, a large-scale dataset of over 1.3 million personalized preference examples, and develop two complementary alignment approaches: in-context alignment directly conditioning on persona representations and preference-bridged alignment modeling intermediate preference distributions. Extensive experiments demonstrate substantial improvements over existing methods, with an average 17.06% accuracy gain across four benchmarks while exhibiting a strong adaptation capability to novel preferences, robustness to limited user data, and precise preference controllability. These results validate our approach toward user-adaptive AI systems.

[Findings] Generics are not quantificational: A new path from language models to semantic theory ▶ *Underline Video*

Gustavo Cilleruelo Calderón, Mahrad Almotahari, Emily Allaway, Barry Haddow and Alexandra Birch

Generic sentences express generalizations that tolerate exceptions without explicitly communicating information about quantities. For example, the sentence Ravens are black is true even though there are albino ravens. The sentence doesnt explicitly communicate the number or frequency of black ravens. Whether generics semantically encode information about quantities implicitly is controversial. This work takes a large-scale distributional approach to the semantic debate. It compares thousands of naturally occurring generics and quantificational sentences using language-model probabilities. It shows that language models recover many semantic facts about quantifiers. It also shows that they recover semantic facts about surface distributional differences between generics and their quantificational counterparts. Accordingly, and contrary to dominant views in other fields, we formulate an empirical argument to the effect that generics are not quantificational.

[Long] GenesisFunc: Multi-Agent Data Generation for Accurate and Generalizable Function-Calling ▶ *Underline Video*

Hao-Xiang Xu, Chong Deng, Jiaqing Liu, Wen Wang, Qian Chen, Lujia Bao, Xiangang Li and Zhen-Hua Ling

Large Language Models (LLMs) extend their capabilities through function-calling (FC), which relies on training data with high quality, diversity, and broad coverage of scenario. However, obtaining and annotating real function-calling data is challenging, while synthetic data from existing pipelines often suffers from unreliable APIs, limited tool scalability, insufficient diversity, and weak quality control. To address these, we present GenesisFunc, an automated pipeline for generating FC training data. Starting from reliable tools in widely used public benchmarks, our GenesisFunc employs a multi-agent framework to support a dialogue generation system that produces conversations spanning diverse scenarios, while maintaining both diversity and quality throughout the process. The accuracy of the data is further reinforced through a multi-stage evaluation system. We fine-tune an 8B LLM on the synthetic dataset and show through extensive experiments that it outperforms similarly sized open-source models in in-domain FC performance and out-of-domain generalization, while reaching FC capabilities comparable to some of the latest API-based models. In addition, our method demonstrates strong potential to scale effectively across downstream tools, underscoring its real-world applicability.

[Findings] GreekMMLU: A Native-Sourced Multitask Benchmark for Evaluating Language Models in Greek ▶ *Underline Video*

Yang ZHANG, Mersin Konomi, Christos Xypolopoulos, Konstantinos Divriotis, Konstantinos Skianis, Giannis Nikolentzos, Giorgos Stamou, Guokan Shang and Michalis Vazirgiannis

Large Language Models (LLMs) are commonly trained on multilingual corpora that include Greek, yet reliable evaluation benchmarks for Greek, particularly those based on authentic, native-sourced content, remain limited. Existing datasets are often machine-translated from English, failing to capture Greek linguistic and cultural characteristics. We introduce GreekMMLU, a native-sourced benchmark for massive multitask language understanding in Greek, comprising 21,805 multiple-choice questions across 45 subject areas, organized under a newly

defined subject taxonomy and annotated with educational difficulty levels spanning primary to professional examinations. All questions are sourced or authored in Greek from academic, professional, and governmental exams. We publicly release 16,857 samples and reserve 4,948 samples for a private leaderboard to enable robust and contamination-resistant evaluation. Evaluations of over 80 open- and closed-source LLMs reveal substantial performance gaps between frontier and open-weight models, as well as between Greek-adapted models and general multilingual ones. Finally, we provide a systematic analysis of factors influencing performance, including model scale, adaptation, and prompting, and derive insights for improving LLM capabilities in Greek.

[Findings] Grounded Concreteness: Human-Like Concreteness Sensitivity in Vision-Language Models ▶ *Underline Video*

Aryan Roy, Zekun Wang and Christopher J. MacLellan

Do vision-language models (VLMs) develop more human-like sensitivity to linguistic concreteness than text-only large language models (LLMs) when both are evaluated with text-only prompts? We study this question with a controlled comparison between matched Llama text backbones and their Llama Vision counterparts across multiple model scales, treating multimodal pretraining as an ablation on perceptual grounding rather than access to images at inference. We measure concreteness effects at three complementary levels: (i) output behavior, by relating question-level concreteness to QA accuracy; (ii) embedding geometry, by testing whether representations organize along a concreteness axis; and (iii) attention dynamics, by quantifying context reliance via attention-entropy measures. In addition, we elicit token-level concreteness ratings from models and evaluate alignment to human norm distributions, testing whether multimodal training yields more human-consistent judgments. Across benchmarks and scales, VLMs show larger gains on more concrete inputs, exhibit clearer concreteness-structured representations, produce ratings that better match human norms, and display systematically different attention patterns consistent with increased grounding.

[Long] Hard2Verify: A Step-Level Verification Benchmark for Open-Ended Frontier Math ▶ *Underline Video*

Shrey Pandit, Austin Xu, Xuan-Phi Nguyen, Yifei Ming, Caiming Xiong and Shafiq Joty

Large language model (LLM)-based reasoning systems have recently achieved gold medal-level performance in the IMO 2025 competition, writing mathematical proofs where, to receive full credit, each step must be not only correct but also sufficiently supported. To train LLM-based reasoners in such challenging, open-ended settings, strong verifiers capable of catching step-level mistakes are necessary prerequisites. We introduce Hard2Verify, a human-annotated, step-level verification benchmark produced with over 500 hours of human labor. Hard2Verify is designed to rigorously assess step-level verifiers at the frontier: Verifiers must provide step-level annotations or identify the first error in responses generated by frontier LLMs for very recent, challenging, and open-ended math questions. We evaluate 29 generative critics and process reward models, demonstrating that, beyond a few standouts, open-source verifiers lag closed source models. We subsequently analyze what drives poor performance in step-level verification, the impacts of scaling verifier compute, as well as fundamental questions such as self-verification and verification-generation dynamics.

[Long] HAT: Hallucination Annotation for Translation ▶ *Underline Video*

Rajen Chatterjee, Xintong Li, Paisarn Charoenpornsawat and Allen Lee

Hallucinations in machine translation (MT), outputs that may be fluent yet unfaithful to the source content, remain a critical obstacle. They hinder the reliable deployment of MT systems in real-world applications. Despite growing attention to this phenomenon, progress has been constrained by the lack of large-scale, high-quality benchmarks dedicated to hallucination detection. We introduce HAT (Hallucination Annotation for Translation), a novel dataset designed to advance research on this problem. HAT comprises 350,959 span-level annotated samples across 38 language pairs, with approximately 8,000-10,000 samples per pair partitioned into training, development, and test sets. Annotations were produced by professional translators under rigorous quality control protocols to ensure reliability. We provide a detailed analysis of hallucination distributions and establish benchmark performance using a diverse set of baselines, including automatic MT evaluation metrics as well as large language models. By providing the first large-scale, systematically annotated resource for hallucination detection in MT, HAT enables the development of more faithful translation models and lays the groundwork for future research on building trustworthy machine translation systems.

[Findings] HORIZON: A Benchmark for In-the-wild User Behaviour Modeling ▶ *Underline Video*

Arnav Goel, Pranjal A Chitale, Bhawna Paliwal, Bishal Santra and Amit Sharma

User behavior in the real world is diverse, cross-domain, and spans long time horizons. Existing user modeling benchmarks however remain narrow, focusing mainly on short sessions and next-item prediction within a single domain. Such limitations hinder progress toward robust and generalizable user models. We present HORIZON, a new benchmark that reformulates user modeling along three axes i.e. dataset, task, and evaluation. Built from a large-scale, cross-domain reformulation of Amazon Reviews, HORIZON covers 54M users and 35M items, enabling both pretraining and realistic evaluation of models in heterogeneous environments. Unlike prior benchmarks, it challenges models to generalize across domains, users, and time, moving beyond standard missing-positive prediction in the same domain. We propose new tasks and evaluation setups that better reflect real-world deployment scenarios. These include temporal generalization, sequence-length variation, and modeling unseen users, with metrics designed to assess general user behavior understanding rather than isolated next-item prediction. We benchmark popular sequential recommendation architectures alongside LLM-based baselines that leverage long-term interaction histories. Our results highlight the gap between current methods and the demands of real-world user modeling, while establishing HORIZON as a foundation for research on temporally robust, cross-domain, and general-purpose user models.

[Long] Identifying the Periodicity of Information in Natural Language ► *Underline Video*

Yulin OU, Yu Wang, Yang Xu and Hendrik Buschmeier

Recent theoretical advancement of information density in natural language has brought the following question on desk: To what degree does natural language exhibit periodicity pattern in its encoded information? We address this question by introducing a new method called AutoPeriod of Surprisal (APS). APS adopts a canonical periodicity detection algorithm and is able to identify any significant periods that exist in the surprisal sequence of a single document. By applying the algorithm to a set of corpora, we have obtained the following interesting results: Firstly, a considerable proportion of human language demonstrates a strong pattern of periodicity in information; Secondly, new periods that are outside the distributions of typical structural units in text (e.g., sentence boundaries, elementary discourse units, etc.) are found and further confirmed via harmonic regression modeling. We conclude that the periodicity of information in language is a joint outcome from both structured factors and other driving factors that take effect at longer distances. The advantages of our periodicity detection method and its potentials in LLM-generation detection are further discussed.

[Long] Improving Long-Context Translation via Self-Supervised Dual Learning ► *Underline Video*

Shanbo Cheng, Shuaijie She, Yu Bao, Jianbing Zhang, Jiajun Chen and Shujian Huang

Large language models (LLMs) with long context windows offer the potential to translate entire documents in a single pass, yet they frequently suffer from catastrophic information distortion, undermining the strict faithfulness required for translation. This challenge is compounded by the scarcity of document-level parallel data, which makes both supervised fine-tuning and reliable evaluation prohibitively expensive. We propose LongDu, a self-supervised post-training framework that improves long-document translation reliability via round-trip consistency. Given monolingual documents, LongDu samples multiple candidate translations, back-translates each candidate, and optimizes the model to prefer translations that best reconstruct the source. To make this signal robust for long-form generation, we design a reward that filters trivial failure modes (e.g., copying and local language drift) before applying a reconstruction and fluency score, enabling stable reinforcement learning without human annotations. We additionally introduce Long-CIRT, an automatic evaluation protocol that quantifies information distortion by measuring how much a LLMs performance degrades after a translation cycle. Across multiple base models, LongDu substantially improves information retention and translation quality, with gains that generalize beyond the training length range and to unseen target languages.

[Long] Inflated Excellence or True Performance? Rethinking Medical Diagnostic Benchmarks with Dynamic Evaluation ► *Underline Video*

Xiangxu Zhang, Lei Li, Yanyun Zhou, Xiao Zhou, Yingying Zhang and Xian Wu

Medical diagnostics is a high-stakes and complex domain that is critical to patient care. However, current evaluations of large language models (LLMs) remain limited in capturing key challenges of clinical diagnostic scenarios. Most rely on benchmarks derived from public exams, raising contamination bias that can inflate performance, and they overlook the confounded nature of real consultations beyond textbook cases. Recent dynamic evaluations offer a promising alternative, but often remain insufficient for diagnosis-oriented benchmarking, with limited coverage of clinically grounded confounders and trustworthiness beyond accuracy. To address these gaps, we propose DyReMe, a dynamic benchmark for medical diagnostics that provides a controlled and scalable stress test of diagnostic robustness. Unlike static exam-style questions, DyReMe generates fresh, consultation-style cases that incorporate clinically grounded confounders, such as differential diagnoses and common misdiagnosis factors. It also varies expression styles to capture heterogeneous patient-style descriptions. Beyond accuracy, DyReMe evaluates LLMs on three additional clinically relevant dimensions: veracity, helpfulness, and consistency. Our experiments show that this dynamic approach yields more challenging assessments and exposes substantial weaknesses of stateof-the-art LLMs under clinically confounded diagnostic settings. These findings highlight the urgent need for evaluation frameworks that better assess trustworthy medical diagnostics under clinically grounded confounders.

[Findings] Inventive Problem Solving with LLMs: A Benchmark for TRIZ Reasoning ► *Underline Video*

Zhu Wang and Brian Uzzi

Large language models are increasingly used in inventive problem-solving, but effective support requires more than open-ended idea generation. Inventive problem-solving requires improving one aspect of a technical system without unintentionally worsening another. TRIZ (Theory of Inventive Problem Solving) provides a unique and structured framework for this setting by representing engineering trade-offs as contradictions and linking them to standardized inventive principles. However, prior TRIZ–LLM evaluations are typically small-scale, case studies in focused areas of technology, and rarely grounded in patent text, which makes it difficult to assess structured reasoning at scale. We introduce TRIZBench, a dataset and benchmark for TRIZ reasoning grounded in open technical sources and U.S.\{} patents. TRIZBench evaluates the core TRIZ workflow through three tasks: contradiction prediction, inventive principle prediction, and grounded TRIZ reasoning. Experiments with multiple LLM baselines show that detecting contradictions is easier than recovering correct trade-off pairs, while principle prediction benefits from explicitly exploiting TRIZ structure. Our findings further underscore the importance of grounding. We show that semantic retrieval enables evidence-based justifications and helps explain why LLMs fail. Dataset and code are available at https://github.com/ellenzhuwang/trizbench.

[Long] Label Effects: Shared Heuristic Reliance in Trust Assessment by Humans and LLM-as-a-Judge ► *Underline Video*

Xin Sun, Di Wu, Sijing Qin, Isao Echizen, Abdallah El Ali and Saku Sugawara

Large language models (LLMs) are increasingly used as automated evaluators (LLM-as-a-Judge). This work challenges its reliability by

showing that trust judgments by LLMs are biased by disclosed source labels. Using a counterfactual design, we find that both humans and LLM judges assign higher trust to information labeled as human-authored than to the same content labeled as AI-generated. Eye-tracking data reveal that humans rely heavily on source labels as heuristic cues for judgments. We analyze LLM internal states during judgment. Across label conditions, models allocate denser attention to the label region than the content region, and this label dominance is stronger under Human labels than AI labels, consistent with the human gaze patterns. Besides, decision uncertainty measured by logits is higher under AI labels than Human labels. These results indicate that the source label is a salient heuristic cue for both humans and LLMs. It raises validity concerns for label-sensitive LLM-as-a-Judge evaluation, and we cautiously raise that aligning models with human preferences may propagate human heuristic reliance into models, motivating debiased evaluation and alignment.

[Long] Large Language Models Are Bad Dice Players: LLMs Struggle to Generate Random Numbers from Statistical Distributions ▶ *Underline Video*

Minda Zhao, Yilun Du and Mengyu Wang

As large language models (LLMs) transition from chat interfaces to integral components of stochastic pipelines and systems approaching general intelligence, the ability to faithfully sample from specified probability distributions has become a functional requirement rather than a theoretical curiosity. We present the first large-scale, statistically powered audit of native probabilistic sampling in frontier LLMs, benchmarking 11 models across 15 distributions. To disentangle failure modes, we employ a dual-protocol design: Batch Generation, where a model produces N=1000 samples within one response, and Independent Requests, comprising N=1000 stateless calls. We observe a sharp protocol asymmetry: batch generation achieves only modest statistical validity, with a 7% median pass rate, while independent requests collapse almost entirely, with 10 of 11 models passing none of the distributions. Beyond this asymmetry, we reveal that sampling fidelity degrades monotonically with distributional complexity and aggravates as the sampling horizon N increases. Finally, we demonstrate how the propagation of these failures into downstream real-world application tasks introduces systematic biases: models fail to enforce uniform answer-position constraints in Multiple Choice Question generation and systematically violate demographic targets in attribute-constrained text-to-image prompt synthesis. These findings indicate that current LLMs lack a functional internal sampler, necessitating external tools for applications requiring statistical guarantees.

[Findings] LifeSim: Long-Horizon User Life Simulator for Personalized Assistant Evaluation ▶ *Underline Video*

Feiyu Duan, Xuanjing Huang and zhongyu wei

The rapid advancement of large language models (LLMs) has accelerated progress toward universal AI assistants. However, existing benchmarks for personalized assistants remain misaligned with real-world user-assistant interactions, failing to capture the complexity of external contexts and users' cognitive states. To bridge this gap, we propose LifeSim, a user simulator that models user cognition through the Belief-Desire-Intention (BDI) model within physical environments for coherent life trajectories generation, and simulates intention-driven user interactive behaviors. Based on LifeSim, we introduce LifeSim-Eval, a comprehensive benchmark for multi-scenario, long-horizon personalized assistance. LifeSim-Eval covers 8 life domains and 1,200 diverse scenarios, and adopts a multi-turn interactive method to assess models' abilities to complete explicit and implicit intentions, recover user profiles, and produce high-quality responses. Under both single-scenario and long-horizon settings, our experiments reveal that current LLMs face significant limitations in handling implicit intention and long-term user preference modeling.

[Findings] LRBench and Judge-R1: Principled Evaluation and Training of LLM-Based Judges for Long-Context Reasoning ▶ *Underline Video*

Xinyi Zhao, Haoqi Hu, Ziyu Wang and Jinfeng Xiao

Large language models (LLMs) are increasingly used as judges to evaluate, rank, and supervise other models, yet their reliability in judging LLMs' reasoning process under long-context settings remains underexplored. Existing benchmarks either overly rely on human annotators, who may miss subtle flaws in lengthy reasoning chains, or focus solely on final responses while ignoring the underlying context and reasoning process. We introduce Long-Reason Bench (LRBench), a large-scale benchmark for evaluating LLM-based judges. LRBench comprises over 100K annotated instances spanning medical, legal, and academic-review scenarios, with fine-grained labels indicating violations of six core principles: Logical Correctness, Factual Consistency, Bias and Fairness, Groundedness, Helpfulness, and Harmlessness. Experimental results reveal that state-of-the-art LLM judges struggle to identify nuanced reasoning errors in long contexts. To improve judge reliability, we further present Judge-R1, which combines reinforcement learning with multi-turn search to enable grounded and principle-aware evaluation. Across domains and principles, Judge-R1 consistently outperforms single-turn baselines, enabling scalable and trustworthy evaluation of LLM reasoning. Our dataset and code are available at https://github.com/Xinyi-0724/Judge-R1.

[Findings] ltzGLUE: Luxembourgish General Language Understanding Evaluation ▶ *Underline Video*

Alistair Plum, Felicia Körner, Anne-Marie Lutgen, Laura Bernardy, Fred Philippy, Emilia Milano, Nils Rehlinger, Cedric Lothritz, Tharindu Ranasinghe, Barbara Plank and Christoph Purschke

This paper presents ltzGLUE, the first Natural Language Understanding (NLU) benchmark for Luxembourgish (LTZ) based on the popular GLUE benchmark for English. Although NLU tasks are available for many european languages nowadays, LTZ is one of the official national languages that is often overlooked. We introduce new tasks and reuse existing ones to introduce the first official NLU benchmark and accompanying evaluation of encoder models for the language. Our tasks include common natural language processing tasks in binary and multi-class classification settings, including named entity recognition, topic classification, and intent classification. We evaluate various

pre-trained language models for LTZ to present an overview of the current capabilities of these models on the LTZ language.

[Long] Macaron: Controlled, Human-Written Benchmark for Multilingual and Multicultural Reasoning via Template-Filling ▶ *Underline Video*

Alaa Elsetohy, Sama Hadhoud, Haryo Wibowo, Chenxi Whitehouse, Genta Winata, Fajri Koto and Alham Aji

Multilingual benchmarks rarely test reasoning over culturally grounded premises: translated datasets keep English-centric scenarios, while culture-first datasets often lack control over the reasoning required. We propose Macaron, a template-first benchmark that factorizes reasoning type and cultural aspect across question languages. Using 100 language-agnostic templates that cover 7 reasoning types, 22 cultural aspects, native annotators create scenario-aligned English and local-language multiple-choice questions, and systematically derived True/False questions. Macaron contains 11,862 instances spanning 20 countries/cultural contexts, 10 scripts, and 20 languages and dialects (including low-resource ones like Amharic, Yoruba, Zulu, Kyrgyz, and some Arabic dialects). In zero-shot evaluation of 21 multilingual LLMs, reasoning-mode models achieve the strongest performance (80.8% overall) and near-parity between English and local languages (DeltaMC = -1.3%), while open-weight models degrade substantially in local languages (DeltaMC = -6.8%) and often approach chance on T/F tasks. Culture-grounded mathematical and counting templates are consistently the hardest. The data can be accessed here https://huggingface.co/datasets/AlaaAhmed2444/Macaron.

[Long] Measuring User's Mental Models of Speech Translation in Human-AI Collaboration ▶ *Underline Video*

HyoJung Han, Nishant Balepur, Jordan Boyd-Graber and Marine Carpuat

Millions of people use machine translation (MT) tools daily, yet little is known about their perception of what systems can and cannot do. This paper studies users' mental models of speech translation systems through a new framework based on cross-lingual question answering, where users either accept MT output or request professional re-translation to answer questions based on the information presented in a foreign language. By analyzing user behavior and accuracy trends across varying translation qualities, we examine to what extent they can predict where the system is likely to be wrong, and how this mental model evolves. Users develop stronger mental models with practice, especially when they have some knowledge of the source language, primarily by relying on surface-level error cues. Moreover, providing speech transcriptions can help users develop better mental models. Our results show the promise of cross-lingual question answering as a downstream task for studying MT mental models and advancing our understanding of human-AI collaboration.

[Findings] Mending the Holes: Mitigating Reward Hacking in Reinforcement Learning for Multilingual Translation ▶ *Underline Video*

Yifeng Liu, Siqi Ouyang, Yatish Hosmane Revanasiddappa and Lei Li

Large Language Models (LLMs) have demonstrated remarkable capability in machine translation on high-resource language pairs, yet their performance on low-resource translation still lags behind. Existing post-training methods rely heavily on high-quality parallel data, which are often scarce or unavailable for low-resource languages. In this paper, we introduce WALAR, a reinforcement training method using only monolingual text to elevate LLMs' translation capabilities on massive low-resource languages while retaining their performance on high-resource languages. Our key insight is based on the observation of failure modes (or "holes") in existing source-based multilingual quality estimation (QE) models. Reinforcement learning (RL) using these QE models tends to amplify such holes, resulting in poorer multilingual LLMs. We develop techniques including word alignment and language alignment to mitigate such holes in WALAR's reward for RL training. We continually trained an LLM supporting translation of 101 languages using WALAR. The experiments show that our new model outperforms LLaMAX, one of the strongest open-source multilingual LLMs by a large margin on 1,414 language directions on Flores-101 dataset.

[Long] Mitigating Tokenization-Induced Distance Distortion in Long-Context Multilingual Machine Translation ▶ *Underline Video*

Khotso Selialia, Antoine Nzeyimana and Fatima M. Anwar

Multilingual neural machine translation (MNMT) models degrade in performance as input context length increases, causing positional encoding schemes to misinterpret token distances. Existing absolute and relative positional encodings rely on fixed token indices and implicitly assume uniform semantic density, which breaks down for long-context inputs. We introduce DCARPE, a tokenization-aware adaptive positional encoding that conditions relative positional bias on input-level sequence length and fragmentation statistics, allowing the model to reinterpret positional distance when tokenization-induced inflation arises rather than semantic factors. Evaluations on JW300 and out-of-distribution FLORES-200 demonstrate consistent improvements in long-context robustness, achieving gains of up to +10.81 ChrF++ and +8.00 BLEU over baselines.

[Findings] Multi-lingual Functional Evaluation for Large Language Models ▶ *Underline Video*

Victor Ojewale, Inioluwa Deborah Raji and Suresh Venkatasubramanian

Multi-lingual competence in large language models is often evaluated via static data benchmarks such as Belebele, M-MMLU and M-GSM. However, these evaluations often fail to provide an adequate understanding of the practical performance and robustness of models across multi-lingual settings. In response, we create multi-lingual functional benchmarks – Cross-Lingual Grade School Math Symbolic (CL-GSM Symbolic) and Cross-Lingual Instruction-Following Eval (CL-IFEval)– by translating existing functional benchmark templates from English to five additional languages that span the range of resources available for NLP: French, Spanish, Hindi, Arabic and Yoruba. Our results show that the gap between static and functional evaluations is highly uneven: across models, performance drops from M-GSM to CL-GSM

Symbolic by 24%, 17%, and 18% in English, French, and Spanish, while the drop from Belebele to CL-IFEval ranges from 15% to 24% across languages, and the drop from M-MMLU to CL-IFEval is much smaller (0.5% to 3%).Similarly, we find that model robustness across languages varies significantly, with certain languages (eg. Arabic, English) being the most consistently well performing across evaluation iterations.

[Long] Multi-View Attention Multiple-Instance Learning Enhanced by LLM Reasoning for Cognitive Distortion Detection ► *Underline Video*

Jun Seo Kim, Hyemi Kim, Woo Joo Oh, Hongjin Cho, Hochul Lee and Hye Hyeon Kim

Cognitive distortions have been closely linked to mental health disorders, yet their automatic detection remains challenging due to contextual ambiguity, co-occurrence, and semantic overlap. We propose a novel framework that combines Large Language Models (LLMs) with a Multiple-Instance Learning (MIL) architecture to enhance interpretability and expression-level reasoning. Each utterance is decomposed into Emotion, Logic, and Behavior (ELB) components, which are processed by LLMs to infer multiple distortion instances, each with a predicted type, expression, and model-assigned salience score. These instances are integrated via a Multi-View Gated Attention mechanism for final classification. Experiments on Korean (KoACD) and English (Therapist QA) datasets demonstrate that incorporating ELB and LLM-inferred salience scores improves classification performance, especially for distortions with high interpretive ambiguity. Our results suggest a psychologically grounded and generalizable approach for fine-grained reasoning in mental health NLP. The dataset and implementation details are publicly accessible.

[Findings] NASH: Numerically Aware Scoring Heuristic for Robust Semantic Similarity ► *Underline Video*

Yu-Shiang Huang, Yun-Yu Lee, Tzu-Hsin Chou, Che Lin and Chuan-Ju Wang

Numerical precision is critical in financial NLP, yet embedding-based semantic similarity metrics exhibit numerical blindness—failing to distinguish contradictory values within similar contexts. We introduce NASH (Numerically Aware Scoring Hueristic), a model-agnostic metric that decouples numerical verification from textual semantic evaluation through a three-stage pipeline: (1) modal separation via numeric masking, (2) dual-channel similarity estimation through masked-text similarity and context-aware numeric alignment, and (3) IDF-weighted aggregation. NASH functions as a drop-in enhancement to existing embedding-based metrics. Validated on our proposed NumFinE financial numerical evaluation benchmark and established semantic similarity datasets (STS-B, Financial-STS), NASH achieves substantial improvements in numerical sensitivity (up to +159.6% on listwise ranking) while preserving general semantic performance, establishing a reliable standard for numeracy-aware evaluation.

[Long] Neo-Classic: A Benchmark for Evaluating Linguistic-Aesthetic Reasoning in Classical Chinese Poetry ► *Underline Video*

Han Zhang, Zihan Gu, Zhiyuan Wang, Tianyi Ma, Jiacheng Lu, Xinyan Zhang, Yuhao Wei and Cheng Hua

While Large Language Models (LLMs) achieve high accuracy on established Classical Chinese Poetry benchmarks, it remains challenging to distinguish transferable Linguistic-Aesthetic Reasoning from reliance on familiar pre-training patterns. To address this issue, we introduce Neo-Classic, an evaluation benchmark that combines a constructionist Out-of-Sample (OOS) dataset with a suite of reverse understanding probes. Unlike traditional benchmarks that rely on verification or generation over historical corpora, Neo-Classic comprises strictly metrical poetry authored by contemporary experts, reducing the possibility of direct retrieval. We evaluate state-of-the-art models, including Qwen3-Max, Gemini-3-Pro, and DeepSeek-V3.2, across five behavioral probes designed to test hierarchical constraint satisfaction. Our results reveal two primary limitations. First, a performance gap of 20%-50% emerges when models transition from historical to contemporary texts. Second, models exhibit substantial difficulties in discourse-level ordering tasks, with standard accuracy remaining low (0-13%). Although expert-level guidance improves the performance of reasoning-enhanced models to 36%, a notable gap with human experts persists. These findings suggest that while current LLMs capture local formal patterns, they struggle with global hierarchical planning required for robust Linguistic-Aesthetic Reasoning.

[Long] NoisyCausal: A Benchmark for Evaluating Causal Reasoning Under Structured Noise ► *Underline Video*

Zhi Xu and Yun Fu

Causal reasoning in natural language requires identifying relevant variables, understanding their interactions, and reasoning about effects and interventions, often under noisy or ambiguous conditions. While large language models (LLMs) exhibit strong general reasoning abilities, they struggle to disentangle correlation from causation, particularly when observations are partially incorrect or irrelevant information is present. In this work, we introduce NoisyCausal, a new benchmark designed to evaluate causal reasoning under structured noise. Each instance is generated from a ground-truth causal graph and contextualized with a natural language scenario by injecting controllable forms of noise, such as irrelevant distractors, value perturbations, confounding, and partial observability. Moreover, we propose a modular reasoning framework that combines LLMs with explicit causal structure to address these challenges. Our method prompts the LLM to extract variables, construct a causal graph from context, and then reformulates the reasoning task as a structured prompt grounded in this graph. Rather than relying on statistical patterns alone, the LLM is guided by symbolic structure, enabling more interpretable and robust inference. Experimental results show that our method significantly outperforms standard prompting and reasoning baselines on NoisyCausal. Furthermore, it generalizes well to external benchmarks such as Cladder without task-specific tuning. Our findings highlight the importance of combining causal abstractions with language-driven reasoning to achieve faithful and robust causal understanding in LLMs.

[Findings] NovBench: Evaluating Large Language Models on Academic Paper Novelty Assessment ► *Underline Video*

Wenqing Wu, Yi Zhao, Yuzhuo Wang, Siyou Li, Juexi Shao, Yunfei Long and Chengzhi Zhang
Novelty is a core requirement in academic publishing and a central focus of peer review, yet the growing volume of submissions has placed increasing pressure on human reviewers. While large language models (LLMs), including those fine-tuned on peer review data, have shown promising results in generating review comments, the absence of dedicated benchmarks has limited systematic evaluation of their ability to assess research novelty. To address this gap, we introduce NovBench, the first large-scale benchmark designed to evaluate LLMs' capability to generate novelty evaluations in support of human peer review. NovBench comprises 1,684 paper-review pairs from a leading NLP conference, including novelty descriptions extracted from paper introductions and corresponding expert-written novelty evaluations. We focus on both sources because the introduction provides a standardized and explicit articulation of novelty claims, while expert-written novelty evaluations constitute one of the current gold standards of human judgment. Furthermore, we propose a four-dimensional evaluation framework (including Relevance, Correctness, Coverage, and Clarity) to assess the quality of LLM-generated novelty evaluations. Extensive experiments on both general and specialized LLMs under different prompting strategies reveal that current models exhibit limited understanding of scientific novelty, and that fine-tuned models often suffer from instruction-following deficiencies. Our findings underscore the need for targeted fine-tuning strategies that jointly improve novelty comprehension and instruction adherence.

[Findings] On Finding Inconsistencies in Documents ▶ *Underline Video*
Charles Lovering, Seth Ebner, Brandon Smock, Michael Krumdick, Saad Rabbani, Ahmed Muhammad, Varshini Reddy and Chris Tanner
Professionals in academia, law, and finance audit their documents because inconsistencies can result in monetary, reputational, and scientific costs. Language models (LMs) have the potential to dramatically speed up this auditing process. To understand their abilities, we introduce a benchmark, FIND (Finding INconsistencies in Documents), where each example is a document with an inconsistency inserted manually by a domain expert. Despite the documents being long, technical, and complex, the best-performing model ('gpt-5') recovered 64% of the inserted inconsistencies. Surprisingly, 'gpt-5' also found inconsistencies already present in the original documents. For example, on 50 arXiv papers, we judged 136 out of 196 of the model's suggestions to be legitimate inconsistencies missed by the original authors. However, despite these findings, even the best models miss almost half of the inconsistencies in FIND, demonstrating that inconsistency detection is still a challenging task.

[Findings] OptiVerse: A Comprehensive Benchmark towards Optimization Problem Solving ▶ *Underline Video*
Xinyu Zhang, Boxuan Zhang, Yuchen Wan, Lingling Zhang, YiXing Yao, Bifan Wei, Yaqiang Wu and Jun Liu
While Large Language Models (LLMs) demonstrate remarkable reasoning, complex optimization tasks remain challenging, requiring domain knowledge and robust implementation. However, existing benchmarks focus narrowly on Mathematical Programming and Combinatorial Optimization, hindering comprehensive evaluation. To address this, we introduce OptiVerse, a comprehensive benchmark of 1,000 curated problems spanning neglected domains, including Stochastic Optimization, Dynamic Optimization, Game Optimization, and Optimal Control, across three difficulty levels: Easy, Medium, and Hard. The experiments with 22 LLMs of different sizes reveal sharp performance degradation on hard problems, where even advanced models like GPT-5.2 and Gemini-3 struggle to exceed 27% accuracy. Through error analysis, we identify that modeling & logic errors remain the primary bottleneck. Consequently, we propose a Dual-View Auditor Agent that improves the accuracy of the LLM modeling process without introducing significant time overhead. OptiVerse will serve as a foundational platform for advancing LLMs in solving complex optimization challenges.

[Long] PEAR: Pairwise Evaluation for Automatic Relative Scoring in Machine Translation ▶ *Underline Video*
Lorenzo Proietti, Roman Grundkiewicz and Matt Post
We present PEAR (Pairwise Evaluation for Automatic Relative Scoring), a supervised quality estimation (QE) metric family that reframes reference-free machine translation (MT) evaluation as a graded pairwise comparison. Given a source segment and two candidate translations, PEAR predicts the direction and magnitude of their quality difference. The metrics are trained using pairwise supervision derived from differences in human judgments, with an additional regularization term that encourages sign inversion under candidate order reversal. On the WMT24 meta-evaluation benchmark, PEAR outperforms strictly matched single-candidate QE baselines trained with the same data and backbones, isolating the benefit of the proposed pairwise formulation. Despite using substantially fewer parameters than recent large metrics, PEAR surpasses far larger QE models and reference-based metrics. Our analysis further indicates that PEAR yields a less redundant evaluation signal relative to other top metrics. Finally, we show that PEAR is an effective utility function for minimum Bayes risk (MBR) decoding, reducing pairwise scoring cost at negligible impact.

[Long] PICTURE: Enhancing Theory-of-Mind in Large Language Models by Revealing, Not Hiding, Characters' Lack of Knowledge ▶ *Underline Video*
Eojin Jeon and SangKeun Lee
Simulating human-like Theory of Mind (ToM) has been a longstanding problem in natural language processing (NLP). To address this, existing works introduce a reasoning step of event hiding (a.k.a. perspective-taking), where events unknown to a character are removed before question answering. However, resorting to event hiding for ToM reasoning presents a performance degradation issue due to the strict output format constraints involved in event hiding. To mitigate this issue, we propose generating perspective-taking outputs as free-form explanations without event hiding, but this poses a notable yet underexplored challenge: LLMs need to inhibit responses to events unknown to characters, because the absence of event hiding exposes LLMs to these events throughout reasoning. To address this challenge, we hypothesize and empirically verify that LLMs can achieve such inhibition if a characters lack of knowledge about events is made explicit

during reasoning. Based on this finding, we introduce PICTURE, a new prompting method that enables LLMs to generate a character's lack of knowledge within free-form Chain-of-Thought (CoT). Experimental results show that PICTURE outperforms existing prompting methods by an average of 7.3% on false-belief tasks.

[Long] PodBench: A Comprehensive Benchmark for Instruction-Aware Audio-Oriented Podcast Script Generation ▶ *Underline Video*

Chenning Xu, Mao Zheng, Mingyu Zheng and Mingyang Song

Podcast script generation requires LLMs to synthesize structured, context-grounded dialogue from diverse inputs, yet systematic evaluation resources for this task remain limited. To bridge this gap, we introduce PodBench, a benchmark comprising 800 samples with inputs up to 21K tokens and complex multi-speaker instructions. We propose a multifaceted evaluation framework that integrates quantitative constraints with LLM-based quality assessment. Extensive experiments reveal that while proprietary models generally excel, open-source models equipped with explicit reasoning demonstrate superior robustness in handling long contexts and multi-speaker coordination compared to standard baselines. However, our analysis uncovers a persistent divergence where high instruction following does not guarantee high content substance. PodBench offers a reproducible testbed to address these challenges in long-form, audio-centric script generation.

[Long] PolitNuggets: Benchmarking Agentic Discovery of Long-Tail Political Facts ▶ *Underline Video*

Yifei Zhu

Large Reasoning Models (LRMs) embedded in agentic frameworks have transformed information retrieval from static, long-context question answering into open-ended exploration. Yet real-world use requires models to discover and synthesize long-tail facts from dispersed sources, a capability that remains under-evaluated. We introduce PolitNuggets, a multilingual benchmark for agentic information synthesis via constructing political biographies for 400 global elites, covering over 10000 political facts. We standardize evaluation with an optimized Supervisor–Searcher multi-agent system and propose FactNet, an evidence-conditional protocol that scores discovery, fine-grained accuracy, and efficiency. Across models and settings, we find that current systems often struggle with fine-grained details, and vary substantially in efficiency. Finally, using benchmark diagnostics, we relate agent performance to underlying model capabilities, highlighting the importance of short-context extraction, multilingual robustness, and reliable tool use.

[Findings] ProMCP: Profiling Token Flows and Latency Costs in Model Context Protocol-Based LLM Agents ▶ *Underline Video*

Sumera Anjum, Weijian Zheng, Rajkumar Kettimuthu, Heng Fan and Yunhe Feng

The Model Context Protocol (MCP) aims to standardize the integration of Large Language Models (LLMs) with external tools, yet existing research primarily evaluates functional capabilities while treating the underlying protocol as an opaque black box. This oversight obscures critical inefficiencies in token flows and latency distributed across MCPs decoupled Host-Client-Server architecture. In this paper, we introduce ProMCP, an end-to-end profiling and instrumentation framework that decomposes the MCP workflow into a six-stage communication pipeline, enabling granular attribution of computational costs. We evaluate widely varying deployment topologies, from air-gapped local models to commercial off-the-shelf (OTS) clients, across 20 servers and 169 tools from MCP-Bench and MCP-Universe. Our analysis reveals a distinct inversion in performance bottlenecks: topologies with customized clients devote 56-72% of total tokens and 60-67% of latency to planning and schema injection, whereas OTS clients concentrate over 85% of latency in final answer synthesis. Crucially, actual tool execution constitutes a negligible fraction of the total cost across all configurations. These findings establish a quantitative baseline for protocol overhead and demonstrate that future optimization must target schema orchestration and transport efficiency rather than tool execution speed. The code is available at: https://github.com/ResponsibleAILab/ProMCP.

[Long] Quantifying and Mitigating Socially Desirable Responding in LLMs: A Desirability-Matched Graded Forced-Choice Psychometric Study ▶ *Underline Video*

Kensuke Okada, Yui Furukawa and Kyosuke Bunji

Human self-report questionnaires are increasingly used in NLP to benchmark and audit large language models (LLMs), from persona consistency to safety and bias assessments. Yet these instruments presume honest responding; in evaluative contexts, LLMs can instead gravitate toward socially preferred answers—a form of socially desirable responding (SDR)—biasing questionnaire-derived scores and downstream conclusions. We propose a psychometric framework to quantify and mitigate SDR in questionnaire-based evaluation of LLMs. To quantify SDR, the same inventory is administered under HONEST versus FAKE-GOOD instructions, and SDR is computed as a direction-corrected standardized effect size from item response theory (IRT)-estimated latent scores. This enables comparisons across constructs and response formats, as well as against human instructed-faking benchmarks. For mitigation, we construct a graded forced-choice (GFC) Big Five inventory by selecting 30 cross-domain pairs from an item pool via constrained optimization to match desirability. Across nine instruction-following LLMs evaluated on synthetic personas with known target profiles, Likert-style questionnaires show consistently large SDR, whereas desirability-matched GFC substantially attenuates SDR while largely preserving the recovery of the intended persona profiles. These results highlight a model-dependent SDR-recovery trade-off and motivate SDR-aware reporting practices for questionnaire-based benchmarking and auditing of LLMs.

[Long] Reasoning Gets Harder for LLMs Inside A Dialogue ▶ *Underline Video*

Ivan Kartá, Mateusz Lango and Ondej Duek

Large Language Models (LLMs) achieve strong performance on many reasoning benchmarks, yet these evaluations typically focus on

isolated tasks that differ from real-world usage in task-oriented dialogue (TOD). In this setting, LLMs must perform reasoning inherently while generating text and adhering to instructions on role, format, and style. This mismatch raises concerns about whether benchmark performance accurately reflects models reasoning robustness in TOD setting. We investigate how framing reasoning tasks within TOD affects LLM performance by introducing a new dynamic benchmark covering eight travel-related tasks that require arithmetic, spatial, temporal, and commonsense reasoning. Each problem is presented in both isolated and dialogue-based variants, enabling controlled comparison while mitigating data contamination. Experiments on nine LLMs reveal a substantial and consistent performance gap between isolated and dialogue settings. Through ablations and qualitative analysis, we show that this gap is largely driven by the multi-turn nature of dialogue, with additional effects from role conditioning and tool-use requirements. Our results highlight the need to evaluate LLM reasoning in realistic interactive scenarios.

[Findings] RECOR: Reasoning-focused Multi-turn Conversational Retrieval Benchmark ▶ *Underline Video*

Mohammed Ali, Abdelrahman Abdallah, Amit Agarwal, Hitesh Laxmichand Patel and Adam Jatowt

Existing benchmarks treat multi-turn conversation and reasoning-intensive retrieval separately, yet real-world information seeking requires both. To bridge this gap, we present a benchmark for reasoning-based conversational information retrieval comprising 707 conversations (2,971 turns) across eleven domains. To ensure quality, our Decomposition-and-Verification framework transforms complex queries into fact-grounded multi-turn dialogues through multi-level validation, where atomic facts are verified against sources and explicit retrieval reasoning is generated for each turn. Comprehensive evaluation reveals that combining conversation history with reasoning doubles retrieval performance (Baseline.236 -> History+Reasoning.479 nDCG@10), while reasoning-specialized models substantially outperform dense encoders. Despite these gains, further analysis highlights that implicit reasoning remains challenging, particularly when logical connections are not explicitly stated in the text.

[Findings] Redefining Machine Simultaneous Interpretation: From Incremental Translation to Human-Like Strategies ▶ *Underline Video*

Qianen Zhang, Zeyu Yang and Satoshi Nakamura

Simultaneous Machine Translation (SiMT) requires high-quality translations under strict real-time constraints, which traditional policies with only READ/WRITE actions cannot fully address. We extend the action space of SiMT with four adaptive actions: Sentence_Cut, Drop, Partial_Summarization and Pronominalization, which enable real-time restructuring, omission, and simplification while preserving semantic fidelity. We adapt these actions in a large language model (LLM) framework and construct training references through action-aware prompting. To evaluate both quality and word-level monotonicity, we further develop a latency-aware TTS pipeline that maps textual outputs to speech with realistic timing. Experiments on the ACL60/60 English-Chinese, English-German and English-Japanese benchmarks show that our framework consistently improves semantic metrics and achieves lower delay compared to reference translations and salami-based baselines. Notably, combining Drop and Sentence_Cut leads to consistent improvements in the balance between fluency and latency. These results demonstrate that enriching the action space of LLM-based SiMT provides a promising direction for bridging the gap between human and machine interpretation.

[Findings] Remedy-R: Generative Reasoning for Machine Translation Evaluation without Error Annotations ▶ *Underline Video*

Shaomu Tan, Ryosuke Mitani, Ritvik Choudhary, Qiyu Wu, Toshiyuki Sekiya and Christof Monz

Over the years, scalar MT metrics have advanced rapidly on benchmarks. Yet they remain black boxes, offering little insight into their decisions and sometimes degrading under out-of-distribution inputs. We introduce Remedy-R, a reasoning-driven generative MT metric trained with reinforcement learning from pairwise translation preferences, without requiring error-span annotations or distillation from closed LLMs. Unlike scalar MT metrics that only outputs translation quality scores, Remedy-R produces step-by-step analyses of accuracy, fluency, and completeness, enabling more interpretable assessments. With only 60K pairwise training samples across two language pairs, Remedy-R remains competitive with top scalar metrics and GPT-4-based judges on WMT22-24 metric benchmarks, generalizes to other languages, and shows strong robustness on OOD stress tests. Moreover, Remedy-R generates self-reflective feedback that can be reused for translation refinement. We validate the faithfulness of such feedback with GPT-4 and show that a simple evaluate-revise pipeline leveraging Remedy-Rs analyses consistently improves translation quality across diverse models without any task-specific tuning.

[Long] ReTraceQA: Evaluating Reasoning Traces of Small Language Models in Commonsense Question Answering ▶ *Underline Video*

Francesco Maria Molfese, Luca Moroni, Ciro Porcaro, Simone Conia and Roberto Navigli

While Small Language Models (SLMs) have demonstrated promising performance on an increasingly wide array of commonsense reasoning benchmarks, current evaluation practices rely almost exclusively on the accuracy of their final answers, neglecting the validity of the reasoning processes that lead to those answers. To address this issue, we present ReTraceQA, a novel benchmark that introduces process-level evaluation for commonsense reasoning tasks. Our expert-annotated dataset reveals that in a substantial portion of instances (14-24%), SLMs provide correct final answers despite flawed reasoning processes, suggesting that the capabilities of SLMs are often overestimated by evaluation metrics that focus only on comparing the final answer with the ground truth. Indeed, we show that, when employing strong Large Language Models (LLMs) as automated judges for reasoning-aware evaluation rather than answer-only metrics, SLM performance drops significantly across all models and datasets, with scores decreasing by up to 25%.

[Long] SciCoQA: Quality Assurance for Scientific Paper–Code Alignment ▶ *Underline Video*

Tim Baumgärter and Iryna Gurevych

Discrepancies between scientific papers and their code undermine reproducibility, a concern that grows as automated research agents scale scientific output beyond human review capacity. Whether LLMs can reliably detect such discrepancies has not been systematically measured. To this end, we present SciCoQA, a dataset of 635 paper-code discrepancies (92 real, 543 synthetic) for this cross-modal verification task. Across 22 evaluated models, even the best-performing LLMs, Gemini 3.1 Pro and GPT-5 Mini, detect only 46.7% of real-world discrepancies, revealing a critical gap in automated scientific quality assurance. We construct SciCoQA from GitHub issues and reproducibility papers, and propose a synthetic generation pipeline to scale beyond AI to Physics, Quantitative Biology, and other computational sciences. We further introduce a taxonomy of discrepancy types and categories to characterize the occurring mismatches. Our analysis shows that models particularly struggle with omitted paper details, long-context inputs, and papers outside their pre-training corpus.

[Long] SMART: Evaluating LLMs' Mathematical Reasoning via a Human Cognitive Process-Inspired Benchmark ▶ *Underline Video*

Yujie Hou, Mei Wang, Yaoyao Zhong, Ting Zhang, Xuetao Ma and Hua Huang

Large Language Models (LLMs) have achieved remarkable performance across a wide range of mathematical benchmarks. However, concerns remain as to whether these successes reflect genuine reasoning or superficial pattern recognition. Existing evaluation methods, which typically focus either on the final answer or on the intermediate reasoning steps, reduce mathematical reasoning to a shallow input-output mapping, overlooking its inherently multi-stage and multi-dimensional cognitive nature. Inspired by P'olyas problem-solving theory, we propose SMART, a benchmark that decomposes mathematical problem-solving into four cognitive dimensions: Semantic Understanding, Mathematical Reasoning, Arithmetic Computation, and Reflection & RefinemenT, and introduces dimension-specific tasks to measure the corresponding cognitive processes of LLMs. We apply SMART to 22 state-of-the-art open- and closed-source LLMs and uncover substantial discrepancies in their capabilities across dimensions. Our findings reveal genuine weaknesses in current models and motivate a new metric, the All-Pass Score, designed to better capture true problem-solving capability.

[Findings] SPD-Faith Bench: Diagnosing and Improving Faithfulness in Chain-of-Thought for Multimodal Large Language Models ▶ *Underline Video*

Weijiang Lv, Yaoxuan Feng, Xiaobo Xia, Jiayu Wang, Yan Jing, Wenchao Chen and Bo Chen

Chain-of-Thought reasoning is widely used to improve the interpretability of multimodal large language models (MLLMs), yet the faithfulness of the generated reasoning traces remains unclear. Prior work has mainly focused on perceptual hallucinations, leaving reasoning level unfaithfulness underexplored. To isolate faithfulness from linguistic priors, we introduce SPD-Faith Bench, a diagnostic benchmark based on fine-grained image difference reasoning that enforces explicit visual comparison. Evaluations on state-of-the-art MLLMs reveal two systematic failure modes, perceptual blindness and perception-reasoning dissociation. We trace these failures to decaying visual attention and representation shifts in the residual stream. Guided by this analysis, we propose SAGE, a train-free visual evidence-calibrated framework that improves visual routing and aligns reasoning with perception. Our results highlight the importance of explicitly evaluating faithfulness beyond response correctness. Our benchmark and codes are available at https://anonymous.4open.science/r/SPD-Faith/.

[Long] Stereotype Bias in a Bilingual Setting: A Culturally Grounded Evaluation in Kazakhstan ▶ *Underline Video*

Nurkhan Laiyk, Daniil Orel, Ayana Mussabayeva, Maiya Goloburda, Kamila Kuishibekova, Liya Goloburda, Diana Turmakhan, Preslav Nakov, Yuxia Wang and Fajri Koto

Stereotype bias in language models has been widely examined in English, but remains largely understudied in bilingual contexts where multiple linguistic and cultural systems interact. This gap is especially important in regions where language use reflects complex historical and sociopolitical influences. In this work, we focus on Kazakhstan, a bilingual society where Kazakh, a low-resource Turkic language, and Russian, a high-resource Slavic language, are both actively used and frequently code-mixed in everyday communication. We introduce Aqbileq, a high-quality, human-verified dataset consisting of 5,634 stereotype-bearing statements in Kazakh, Russian, and code-mixed forms, covering six culturally salient domains. We evaluate both multilingual and Kazakh-specific language models using perplexity-based scoring and pretraining simulations, and find that stereotype bias is most pronounced in code-mixed inputs. Our results highlight the limitations of existing evaluation frameworks and emphasize the need for culturally grounded, linguistically inclusive benchmarks to better assess and mitigate bias in language models.

[Findings] StruNRAG: Evaluation of OCR-Induced Structural Noise on RAG Robustness ▶ *Underline Video*

Mengna Gao, Dapeng Yin, Shuyue Zhu, Bingxuan Hou, Zhanpeng Ni and Junli Wang

Retrieval-Augmented Generation (RAG) systems rely on Optical Character Recognition (OCR) to ingest knowledge from unstructured documents. However, OCR engines often struggle with complex layouts, introducing Structural Noise, such as line insertion and paragraph interleaving, which disrupts the semantic flow of the text. Existing evaluations largely overlook this dimension, operating on the assumption of structurally perfect input. To bridge this gap, we introduce StruNRAG, a dedicated benchmark for evaluating RAG robustness against OCR-induced structural perturbations. We construct a bilingual dataset of 2,132 question-answer pairs derived from complex Chinese and English documents and systematically inject three categories of real-world structural noise: line insertion, paragraph interleaving, and line interleaving. Our evaluation of mainstream retrievers and Large Language Models (LLMs) reveals a nuanced interaction between noise and pipeline stages: while structural distortions consistently degrade retrieval performance, the generation stage exhibits unexpected robustness. Advanced LLMs demonstrate robustness against local noise (e.g., line insertion), but struggle to maintain reasoning capabilities

under severe structural disruption that fragments global context. These findings indicate that while LLMs are capable of compensating for minor parsing errors, future RAG optimizations must take into account the effects of structural noise. Our code and datasets are available at https://github.com/GaoMengnana/StruNRAG.

[Findings] Syntax as a Rosetta Stone: Universal Dependencies for In-Context Coptic Translation ▶ *Underline Video*

Abhishek Purushothama, Emma Thronson, Alexia Guo and Amir Zeldes

This paper proposes a novel in-context learning approach to support low resource machine translation for the Coptic language, using prompts based on Universal Dependencies parses of input sentences. Building on existing work using bilingual dictionaries to support inference for vocabulary items, we add several representations of syntactic analyses to our inputs, specifically exploring the inclusion of raw parser outputs, verbalizations of parses in plain English, and explanations of specific difficult constructions identified in input subgraphs and how they can be translated. Our results show that while syntactic information alone is not as useful as dictionary-based glosses, combining retrieved dictionary items with syntactic information achieves significant gains across model sizes, achieving new state-of-the-art results for the language.

[Long] The Prosody of Emojis ▶ *Underline Video*

Giulio Zhou, Tsz Kin Lam, Alexandra Birch and Barry Haddow

Prosodic features such as pitch, timing, and intonation are central to spoken communication, conveying emotion, intent, and discourse structure. In text-based settings, where these cues are absent, emojis act as visual surrogates that add affective and pragmatic nuance. This study examines how emojis influence prosodic realisation in speech and how listeners interpret prosodic cues to recover emoji meanings. Unlike previous work, we directly link prosody and emojis by analysing human speech data collected through a controlled elicited production task. Using Bayesian multilevel modelling, we show that speakers systematically adapt their prosody based on emoji cues, and that listeners can recover intended meanings significantly above chance. Furthermore, our results reveal a clear hierarchy in prosodic shifts: greater semantic differences between emojis correspond to increased prosodic divergence. These findings suggest that emojis are meaningful carriers of prosodic intent that bridge the gap between digital text and spoken production.

[Findings] Too Fast, Too Shallow – LLMs, Including Reasoning LLMs, Are Unreliable Constitutional Reasoners ▶ *Underline Video*

Reto Gubelmann and Peter Hongler

We assess LLMs' constitutional reasoning abilities using three different, newly developed datasets on three different constitutional questions in three different constitutional frameworks, comprising two different languages; the structure and content of the datasets is informed by legal expertise and grounded in the state of the art in philosophy of language. Our results indicate that the 19 LLMs tested, including the reasoning LLMs, while not being uniformly subject to political bias, are still not reliable constitutional reasoners, as they are heavily influenced by logically irrelevant aspects of the reasoning. Of the 196k evaluations run in our main experiment, the LLMs label less than 70% correctly, and open-weight reasoning LLMs as well as gpt-4o are outperformed by moderately sized open-weight non-reasoning LLMs. None of the LLMs tested consistently show slow, systematic, rule-based system 2 thinking.

[Long] Trajectory2Task: Training Robust Tool-Calling Agents with Synthesized Yet Verifiable Data for Complex User Intents ▶ *Underline Video*

Ziyi Wang, Yuxuan Lu, Yimeng Zhang, Pei Chen, Ziwei Dong, Jing Huang, Jiri Gesi, Xianfeng Tang, Chen Luo, Qun Liu, Yisi Sang, Hanqing Lu, Manling Li, Jin Lai and Dakuo Wang

Tool-calling agents are increasingly deployed in real-world customer-facing workflows. Yet most studies on tool-calling agents focus on idealized settings with general, fixed, and well-specified tasks. In real-world applications, user requests are often (1) ambiguous, (2) changing over time, or (3) infeasible due to policy constraints, and training and evaluation data that cover these diverse, complex interaction patterns remain under-represented. To bridge the gap, we present Trajectory2Task a verifiable data generation pipeline for studying tool use at scale under three realistic user scenarios: ambiguous intent, changing intent, and infeasible intents. The pipeline first conducts multi-turn exploration to produce valid tool-call trajectories. It then converts these trajectories into user-facing tasks with controlled intent adaptations. This process yields verifiable task that support closed-loop evaluation and training. We benchmark several state-of-the-art LLMs on the generated complex user scenario tasks and observe frequent failures. Finally, using successful trajectories obtained from task rollouts, we fine-tune lightweight LLMs and find consistent improvements across all three conditions, along with better generalization to unseen tool-use domains, indicating stronger tool-calling ability.

[Findings] TripTide: A Benchmark for Adaptive Travel Planning under Disruptions ▶ *Underline Video*

Priyanshu Karmakar, Soumyabrata Chaudhuri, Shubhojit Mallick, Manish Gupta, Abhik Jana and Shreya Ghosh

Recent work, such as TripCraft and TravelPlanner, has shown the promise of Large Language Models (LLMs) for personalized, constraint-aware travel itinerary generation. However, real-world travel often involves disruptions such as transit cancellations, weather-related closures, or overbooked attractions. To address this gap, we introduce TripTide, the first benchmark designed to evaluate the ability of LLMs to revise travel itineraries under realistic disruptions. TripTide models both disruption severity and traveler tolerance, enabling systematic evaluation of how LLMs respond to unexpected travel events. The benchmark simulates scenarios where existing itineraries must be revised while preserving the travelers original intent and respecting practical constraints. We conduct a three-fold evaluation of itinerary revision quality: (i) Automatic metrics measuring Preservation of Intent, Responsiveness, and Adaptability (semantic, spatial, and sequential),

(ii) LLM-as-a-Judge evaluation assessing the quality and plausibility of revised itineraries and (iii) Human evaluation examining overall revision quality and user satisfaction. Our findings show that LLMs generally preserve semantic intent and sequential structure, while spatial deviations are more pronounced in shorter itineraries and diminish for longer ones. However, the ability to handle disruptions degrades as itinerary length increases, highlighting limitations in long-horizon itinerary revision. The TripTide benchmark provides a foundation for systematically evaluating robustness and adaptability in LLM-based travel planning systems.

[Long] What Do Prosody and Text Convey? Characterizing How Meaningful Information is Distributed Across Multiple Channels ▶ *Underline Video*

Aditya Yadavalli, Tiago Pimentel, Tamar Regev, Ethan Wilcox and Alex Warstadt

Prosody—the melody of speech—conveys critical information often not captured by the words or text of a message. In this paper, we propose an information-theoretic approach to quantify how much is conveyed by prosody that is not recoverable from text alone, and, crucially, what prosody conveys. Our approach applies large speech and language models to estimate the mutual information between a particular dimension of an utterance's meaning (e.g., its emotion) and any of its communication channels (e.g., audio or text). We then use this approach to quantify the information conveyed by audio and text about sarcasm, emotion, and questionhood, using speech from television and podcasts. We find that for sarcasm and emotion, the audio channel, and by implication the prosodic channel, transmits over an order of magnitude more information about these features than the text channel alone, at least when long-term context beyond the current sentence is unavailable. For questionhood, prosody provides comparatively less additional information. We conclude by outlining a program applying our approach to more dimensions of meaning, communication channels, and languages.

[Long] When Personalization Tricks Detectors: The Feature-Inversion Trap in Machine-Generated Text Detection ▶ *Underline Video*

Lang Gao, Xuhui Li, Chenxi Wang, Mingzhe Li, Wei Liu, Zirui Song, Jinghui Zhang, Rui Yan, Preslav Nakov and Xiuying Chen

As large language models (LLMs) increasingly imitate personal writing styles, personalization has become a key challenge for machine-generated text (MGT) detection. Yet personalized MGT detection remains largely underexplored. In this work, we introduce StyloBench, the first benchmark for evaluating detector robustness under personalization, built from literary and blog texts paired with their LLM-generated imitations. Experiments across diverse detectors show pronounced performance instability under personalization, with frequent inversions relative to general-domain behavior. To better understand this limitation, we conduct an in-depth analysis and attribute it to a feature-inversion trap, i.e., features that are effective for separating human-written text (HWT) from MGT in general flip their effect in personalized contexts, ultimately misleading detectors. Motivated by this, we propose StyloCheck, a diagnostic framework for predicting detector robustness under personalization. StyloCheck identifies the inverted features and quantifies detector dependence using perturbed texts pronounced in the features. In our experiments, StyloCheck predicts both the direction and magnitude of cross-domain performance shifts with an 85% correlation to actual outcomes. We hope this work will raise awareness of the structural risks introduced by personalization and motivate more robust approaches to personalized MGT detection. The code is available at: https://github.com/mbzuai-nlp/Personalized_MGT_Detect

[Findings] Who Watches the Watchmen? Humans Disagree With Translation Metrics on Unseen Domains ▶ *Underline Video*

Finn Schmidt, Jan Philip Wahle, Terry Ruas and Bela Gipp

Automatic evaluation metrics are central to the development of machine translation systems, yet their robustness under domain shift remains unclear. Most metrics are developed on the Workshop on Machine Translation (WMT) benchmarks, raising concerns about their robustness to unseen domains. Prior studies that analyze unseen domains vary translation systems, annotators, or evaluation conditions, confounding domain effects with human annotation noise. To address these biases, we introduce a systematic multi-annotator Cross-Domain Error-Span-Annotation dataset (CD-ESA), comprising 18.8k human error span annotations across three language pairs, where we fix annotators within each language pair and evaluate translations of the same six translation systems across one seen news domain and two unseen technical domains. Using this dataset, we first find that automatic metrics appear surprisingly robust to domain-shifts at the segment level (up to 0.69 agreement), but this robustness largely disappears once we account for human label variation. Averaging annotations increases inter-annotator agreement by up to +0.11. Metrics struggle on the unseen chemical domain compared to humans (inter-annotator agreement of 0.78-0.83 vs. 0.96). We recommend comparing metric-human agreement against inter-annotator agreement, rather than comparing raw metric-human agreement alone, when evaluating across different domains.

[Long] Zero-shot Large Language Models for Automatic Readability Assessment ▶ *Underline Video*

Riley Grossman and Yi Chen

Unsupervised automatic readability assessment (ARA) methods have important practical and research applications (e.g., ensuring medical or educational materials are suitable for their target audiences). In this paper, we propose a new zero-shot prompting methodology for ARA and present the first comprehensive evaluation of using large language models (LLMs) as an unsupervised ARA method by testing 10 diverse open-source LLMs (e.g., different sizes and developers) on 14 diverse datasets (e.g., different text lengths and languages). Our findings show that our proposed prompting methodology outperforms prior methods on 13 of the 14 datasets. Furthermore, we propose LAURAE, which combines LLM and readability formula scores to improve robustness by capturing both contextual and shallow (e.g., sentence length) features of readability. Our evaluation demonstrates that LAURAE robustly outperforms prior methods across languages, text lengths, and amounts of technical language.

[Long] I See What You Did There: Can Large Vision-Language Models Understand Multimodal Puns? ▶ *Underline Video*

Naen Xu, Jiayi Sheng, Changjiang Li, Chunyi Zhou, Yuyuan Li, Tianyu Du, Jun Wang, Zhihui Fu, Jinbao Li and Shouling Ji
Puns are a common form of rhetorical wordplay that exploits polysemy and phonetic similarity to create humor. In multimodal puns, visual and textual elements synergize to ground the literal sense and evoke the figurative meaning simultaneously. Although Vision-Language Models (VLMs) are widely used in multimodal understanding and generation, their ability to understand puns has not been systematically studied due to a scarcity of rigorous benchmarks. To address this, we first propose a multimodal pun generation pipeline. We then introduce MultiPun, a dataset comprising diverse types of puns alongside adversarial non-pun distractors. Our evaluation reveals that most models struggle to distinguish genuine puns from these distractors. Moreover, we propose both prompt-level and model-level strategies to enhance pun comprehension, with an average improvement of 16.5% in F1 scores. Our findings provide valuable insights for developing future VLMs that master the subtleties of human-like humor via cross-modal reasoning.

Poster Session G (Grand Hall)

11:00-12:30 (Grand Hall)

[Long] A Survey of Inductive Reasoning for Large Language Models ▶ *Underline Video*
Kedi Chen, Dezhao Ruan, Yuhao Dan, Yaoting Wang, Siyu Yan, Xuecheng Wu, Yinqi Zhang, Qin Chen, Jie Zhou, Liang He, Biqing Qi, Linyang Li, Qipeng Guo, Xiaoming Shi and Wei Zhang
Reasoning is an important task for large language models (LLMs). Among all the reasoning paradigms, inductive reasoning is one of the basic types, which is characterized by its particular-to-general thinking process and the non-uniqueness of its answers. The inductive mode is crucial for knowledge generalization and aligns better with human cognition, so it is a fundamental mode of learning, hence attracting increasing interest. Despite the importance of inductive reasoning, there is no systematic summary of it. Therefore, this paper presents the first comprehensive survey of inductive reasoning for LLMs. First, methods for improving inductive reasoning are categorized into three main areas: post-training enhancement, test-time exploration, and data augmentation. Then, current benchmarks of inductive reasoning are summarized, and a unified sandbox-based evaluation approach with the observation coverage metric is derived. Finally, we offer some analyses regarding the source of inductive ability and how simple model architectures and data help with inductive tasks, providing a solid foundation for future research.

[Long] A Survey of Large Language Model-Based Search Agents ▶ *Underline Video*
Yunjia Xi, Jianghao Lin, Yongzhao Xiao, Zheli Zhou, Rong Shan, Te Gao, Jiachen Zhu, Weiwen Liu, Yong Yu and Weinan Zhang
The advent of Large Language Models (LLMs) has significantly revolutionized web search. The emergence of LLM-based Search Agents marks a pivotal shift towards deeper, dynamic, autonomous information seeking. These agents can comprehend user intentions and environment context and execute multi-turn retrieval with dynamic planning, extending search capabilities far beyond the web. Leading examples like OpenAI's Deep Research highlight their potential for deep information mining and real-world applications. This survey provides the first systematic analysis of search agents. We comprehensively analyze and categorize existing works from the perspectives of architecture, optimization, application, and evaluation, ultimately identifying critical open challenges and outlining promising future research directions in this rapidly evolving field.

[Findings] A2O: LLM-based Agentic Learning of Action-to-Object Features for Video Action Recognition ▶ *Underline Video*
Taiki Sekii and Fumiaki Sato
Recent action recognition based on vision–language pretraining and self-supervised video foundation models tends to induce spurious correlations and shortcut learning by relying on action-irrelevant cues, such as backgrounds and object co-occurrences. By contrast, object-detection-based approaches can suppress spurious correlations; however, the loss of input information can limit accuracy. To mitigate this trade-off, we combine these two approaches to learn complementary features that compensate for each other's shortcomings. Specifically, we leverage the commonsense knowledge of large language models (LLMs) regarding human actions and realize a framework in which an LLM agent integrates the two approaches within an agentic learning paradigm to design motion features tailored to the target actions. The LLM agent uses an open vocabulary object detector to instruct the video foundation model with the target and nontarget objects in the video to make the model attend to objects in a video required for recognizing the target actions. The composition of the detected objects is optimized for the target actions through in-context reinforcement learning (ICRL) using the commonsense knowledge of the LLM. Experiments on multiple public action recognition datasets and an ablation study confirm the robustness of features learned using the proposed method and the effectiveness of ICRL.

[Long] Achieving Multi-Hop Calculation and Safe Abstention in Financial Numerical Reasoning by Metric Graph Constrained LLMs ▶ *Underline Video*
Aoyuan Jiang, Liang Hong, Haoxuan Liu and Rui Wang
Financial numerical reasoning demands rigorous adherence to domain-specific logic and precise evidence foundation. However, large language models (LLMs) are prone to forced generation when confronting ambiguous evidence or complex recursive dependencies, often hallucinating values to bridge information gaps. To address this, we propose graph-bounded financial reasoning (GBFR), a neuro-symbolic

framework that imposes semantic and structural constraints via a financial metric knowledge graph (FMKG). Unlike sequential generation paradigms, our approach employs a parallel graph-constrained reasoning algorithm that orchestrates specialized operators to simultaneously explore heterogeneous derivation paths of complex financial metrics. Through cross-path verification, the framework aggregates only semantically consistent results, ensuring reasoning is bounded by available context. Crucially, this approach enables safe abstention by distinguishing genuine data absence from retrieval failure, thereby preventing ungrounded fabrication. To evaluate this capability, we further construct counterfactual samples by perturbing entities, times, and metrics to synthesize unanswerable scenarios. Empirical evaluations on standard benchmarks demonstrate that GBFR significantly outperforms state-of-the-art baselines.

[Long] Advancing Reasoning in Diffusion Language Models with Denoising Process Rewards ▶ *Underline Video*

Shaoan Xie, Lingjing Kong, Xiangchen Song, Xinshuai Dong, Guangyi Chen, Eric P. Xing and Kun Zhang

Diffusion-based large language models offer a non-autoregressive alternative for text generation, but enabling them to perform complex reasoning remains challenging. Reinforcement learning has recently emerged as an effective post-training strategy for improving their performance; however, existing methods rely primarily on outcome-based rewards, which provide no direct supervision over the denoising process and often result in poorly structured reasoning that is difficult to interpret and inconsistently supports the final prediction. To address this limitation, we introduce denoising process reward, a process-level reinforcement signal defined over the denoising trajectory of diffusion language models. This reward is obtained by estimating the contribution of intermediate denoising intervals to the final task outcome, encouraging the model to favor reasoning trajectories that consistently guide generation toward correct predictions. We further propose an efficient stochastic estimator that reuses standard training rollouts, enabling practical process-level supervision at scale. Experiments on challenging reasoning benchmarks demonstrate that our approach yields consistent improvements in reasoning stability, interpretability, and overall task performance.

[Findings] AlphaQuanter: An End-to-End Tool-Augmented Agentic Reinforcement Learning Framework for Stock Trading ▶ *Underline Video*

Zheye Deng, Weixiang Yan, Changlong Yu and Jiashu Wang

While Large Language Model (LLM) agents show promise in automated trading, they still face critical limitations. Prominent multi-agent frameworks often suffer from inefficiency, produce inconsistent signals, and lack the end-to-end optimization required to learn a coherent strategy from market feedback. To address this, we introduce AlphaQuanter, a single-agent framework that uses reinforcement learning (RL) to learn a dynamic policy over a transparent, tool-augmented decision workflow, which empowers a single agent to autonomously orchestrate tools and proactively acquire information on demand, establishing a transparent reasoning process. Extensive experiments demonstrate that AlphaQuanter achieves state-of-the-art performance on key financial metrics. Besides, human evaluation shows the learned reasoning patterns reveal more faithful and coherent tool-usage behaviors, providing steps toward verifiable LLM-driven trading. Our code and data can be found at https://github.com/horizon-llm/AlphaQuanter.

[Findings] Analysing the Safety Pitfalls of Steering Vectors ▶ *Underline Video*

Yuxiao Li, Alina Fastowski, Efstratios Zaradoukas, Bardh Prenkaj and Gjergji Kasneci

Activation steering has emerged as a powerful tool to shape LLM behaviour without the need for weight updates. While its inherent brittleness and unreliability are well-documented, its safety implications remain underexplored. In this work, we present a systematic safety audit of steering vectors obtained with Contrastive Activation Addition (CAA), a widely used steering approach, under a unified evaluation protocol. We show that steering vectors consistently influence the success rate of jailbreak attacks, with stronger amplification under simple template-based attacks. Across LLM families and sizes, steering the model in specific directions can drastically increase (by up to 57%) or decrease (by up to 50%) its attack success rate (ASR), depending on the targeted behaviour. We attribute this phenomenon to the overlap between the steering vectors and the latent subspace of refusal behaviour. Thus, we offer a mechanistic explanation for this discovery. Together, our findings reveal the previously unobserved origin of this safety gap in LLMs, highlighting a trade-off between controllability and safety. We release our code at https://github.com/yetiiil/analyse-sv-safety.

[Short] Anchoring Depends on Confidence and Post-Training in Language Models ▶ *Underline Video*

Hillary N Owusu and Naomi H Feldman

Anchoring bias causes large language models (LLMs) to shift quantitative judgments in response to irrelevant numerical primes. We analyze this bias as a function of model confidence and accuracy in base, instruction-tuned, and distilled variants of Llama and Qwen models. We find that anchoring susceptibility is negatively correlated with model confidence without regard to accuracy: confidently incorrect models resist anchoring as effectively as accurate ones, provided their internal priors are sufficiently strong. We further show that post-training impacts the strength of this relationship, and that models are more susceptible to high anchors than to low anchors. Our findings suggest anchoring resistance is a structural property of distributional concentration (certainty) rather than knowledge correctness (factual accuracy), with implications for deploying LLMs in numerical reasoning tasks.

[Long] APEX-MEM: Agentic Semi-Structured Memory with Temporal Reasoning for Long-Term Conversational AI ▶ *Underline Video*

Pratyay Banerjee, Masud Moshtaghi, Shivashankar Subramanian, Amita Misra and Ankit Chadha

Large language models still struggle with reliable long-term conversational memory: simply enlarging context windows or applying naïve

retrieval often introduces noise and destabilizes responses. We present APEX-MEM, a conversational memory system that combines three key innovations: (1) a property graph which use domain-agnostic ontology to structure conversations as temporally grounded events in an entity-centric framework, (2) append-only storage that preserves the full temporal evolution of information, and (3) a multi-tool retrieval agent that understands and resolves conflicting or evolving information at query time, producing a compact and contextually relevant memory summary. This retrieval-time resolution preserves the full interaction history while suppressing irrelevant details. APEX-MEM achieves 88.88% accuracy on LOCOMO and 86.2% on LongMemEval, outperforming state-of-the-art session-aware approaches and demonstrating that structured property graphs enable more temporally coherent long-term conversational reasoning.

[Long] ARES: Adaptive Red-Teaming and End-to-End Repair of Policy-Reward System ▶ *Underline Video*

Jiacheng Liang, Yao Ma, Tharindu Kumarage, Satyapriya Krishna, Rahul Gupta, Kai-Wei Chang, Aram Galstyan and Charith Peris

Reinforcement Learning from Human Feedback (RLHF) is central to aligning Large Language Models (LLMs), yet it introduces a critical vulnerability: an imperfect Reward Model (RM) can become a single point of failure when it fails to penalize unsafe behaviors. While existing red-teaming approaches primarily target policy-level weaknesses, they overlook what we term systemic weaknesses cases where both the core LLM and the RM fail in tandem. We present ARES, a framework that systematically discovers and mitigates such dual vulnerabilities. ARES employs a "Safety Mentor" that dynamically composes semantically coherent adversarial prompts by combining structured component types (topics, personas, tactics, goals) and generates corresponding malicious and safe responses. This dual-targeting approach exposes weaknesses in both the core LLM and the RM simultaneously. Using the vulnerabilities gained, ARES implements a two-stage repair process: first fine-tuning the RM to better detect harmful content, then leveraging the improved RM to optimize the core model. Experiments across multiple adversarial safety benchmarks demonstrate that ARES substantially enhances safety robustness while preserving model capabilities, establishing a new paradigm for comprehensive RLHF safety alignment.

[Long] ART: Attention Replacement Technique to Improve Factuality in LLMs ▶ *Underline Video*

Ziqin Luo, Yihao Quan, Xiaofeng Zhang, Xiaosong Yuan and Chen Shen

Hallucination in large language models (LLMs) continues to be a significant issue, particularly in tasks like question answering, where models often generate plausible yet incorrect or irrelevant information. Although various methods have been proposed to mitigate hallucinations, the relationship between attention patterns and hallucinations has not been fully explored. In this paper, we analyze the distribution of attention scores across each layer and attention head of LLMs, revealing a common and intriguing phenomenon: Shallow layers of LLMs primarily rely on uniform attention patterns, where the model distributes its attention evenly across the entire sequence. This uniform attention pattern can lead to hallucinations, as the model fails to focus on the most relevant information. To mitigate this issue, we propose a training-free method called Attention Replacement Technique (ART), which replaces these uniform attention patterns in the shallow layers with local attention patterns. This change directs the model to focus more on the relevant contexts, thus reducing hallucinations. Through extensive experiments, ART demonstrates significant reductions in hallucinations across multiple LLM architectures, proving its effectiveness and generalizability without requiring fine-tuning or additional training data.

[Findings] AutoRubric: Rubric-Based Generative Rewards for Faithful Multimodal Reasoning ▶ *Underline Video*

Mengzhao Jia, Zhihan Zhang, Ignacio Cases, Zheyuan Liu, Meng Jiang and Peng Qi

Multimodal large language models (MLLMs) have rapidly advanced from perception tasks to complex multi-step reasoning, yet reinforcement learning with verifiable rewards (RLVR) often leads to spurious reasoning since only the final-answer correctness is rewarded. To address this limitation, we propose AutoRubric, a framework that integrates RLVR with process-level supervision through automatically collected rubric-based generative rewards. Our key innovation lies in a scalable self-aggregation method that distills consistent reasoning checkpoints from successful trajectories, enabling problem-specific rubric construction without human annotation or stronger teacher models. By jointly leveraging rubric-based and outcome rewards, AutoRubric-R1V achieves state-of-the-art performance on six multimodal reasoning benchmarks and substantially improves reasoning faithfulness in dedicated evaluations.

[Long] BadScientist: Can a Research Agent Write Convincing but Unsound Papers that Fool LLM Reviewers? ▶ *Underline Video*

Fengqing Jiang, Yichen Feng, Yuetai Li, Luyao Niu, Basel Alomair and Radha Poovendran

The convergence of LLM-powered research assistants and AI-based peer review systems creates a critical vulnerability: fully automated publication loops where AI-generated research is evaluated by AI reviewers without human oversight. We investigate this through BadScientist, a framework that evaluates whether fabrication-oriented paper generation agents can deceive multi-model LLM review systems. Our generator employs presentation-manipulation strategies requiring no real experiments. We develop a rigorous evaluation framework with formal error guarantees (concentration bounds and calibration analysis), calibrated on real data. Our results reveal systematic vulnerabilities: fabricated papers achieve acceptance rates up to 18%. Critically, we identify concern-acceptance conflict—reviewers frequently flag integrity issues yet assign acceptance-level scores. Our mitigation strategies show only marginal improvements, with detection accuracy barely exceeding random chance. Despite provably sound aggregation mathematics, integrity checking systematically fails, exposing fundamental limitations in current AI-driven review systems and underscoring the urgent need for defense-in-depth safeguards in scientific publishing.

[Long] Between a Rock and a Hard Place: The Tension Between Ethical Reasoning and Safety Alignment in LLMs ▶ *Underline Video*

Shei Pern Chua, Zhen Leng Thai, Kai Jun Teh, Xiao Li, Qibing Ren and Xiaolin Hu

Large Language Model safety alignment predominantly operates on a binary assumption that requests are either safe or unsafe. This classification proves insufficient when models encounter ethical dilemmas, where the capacity to reason through moral trade-offs creates a distinct attack surface. We formalize this vulnerability through TRIAL, a multi-turn red-teaming methodology that embeds harmful requests within ethical framings. TRIAL achieves consistently high attack success rates across models by exploiting the model's own ethical reasoning to frame harmful actions as morally necessary compromises. Building on these insights, we introduce ERR (Ethical Reasoning Robustness), a defense framework that distinguishes between instrumental responses that enable harmful outcomes and explanatory responses that analyze ethical frameworks without endorsing harmful acts. ERR employs a Layer-Stratified Harm-Gated LoRA architecture, achieving robust defense against reasoning-based attacks while preserving model utility.

[Findings] Beyond Single-Shot: Multi-step Tool Retrieval via Query Planning ► *Underline Video*

Wei Fang and James Glass

LLM agents operating over massive, dynamic tool libraries rely on effective retrieval, yet standard single-shot dense retrievers struggle with complex requests. These failures primarily stem from the disconnect between abstract user goals and technical documentation, and the limited capacity of fixed-size embeddings to model combinatorial tool compositions. To address these challenges, we propose ToolQP, a lightweight framework that models retrieval as iterative query planning. Instead of single-shot matching, ToolQP decomposes instructions into sub-tasks and dynamically generates queries to interact with the retriever, effectively bridging the semantic gap by targeting the specific sub-tasks required for composition. We train ToolQP using synthetic query trajectories followed by optimization with Reinforcement Learning with Verifiable Rewards (RLVR). Experiments demonstrate that ToolQP achieves state-of-the-art performance, exhibiting superior zero-shot generalization, robustness across diverse retrievers, and significant improvements in downstream agentic execution.

[Long] Beyond Surface-Level Detection: Towards Cognitive-Driven Defense Against Jailbreak Attacks via Meta-Operations Reasoning ► *Underline Video*

Rui Pu, Chaozhuo Li, Rui Ha, Litian Zhang, Lirong Qiu and Xi Zhang

Defending large language models (LLMs) against jailbreak attacks is essential for their safe and reliable deployment. Existing defenses often rely on shallow pattern matching, which struggles to generalize to novel and unseen attack strategies. To address this challenge, we propose the Cognitive-Driven Defense (CDD) framework, which targets the underlying structure of jailbreak prompts by applying meta-operations, defined as basic manipulations that conceal harmful intent. CDD emulates human cognitive reasoning through a structured reasoning chain. It begins with a global perception of the prompt and follows with a localized analysis to uncover hidden manipulations. By applying supervised fine-tuning on this structured chain, the model learns to identify and reason about known manipulation patterns. To enhance generalization to unseen threats, an entropy-guided reinforcement learning algorithm (EG-GRPO) is introduced to encourage exploration of new types and variants of meta-operations. Experiments demonstrate that CDD can achieve state-of-the-art defense performance and exhibit strong generalization to unseen jailbreak attacks.

[Long] Boundary-Guided Policy Optimization for Memory-efficient RL of Diffusion Large Language Models ► *Underline Video*

Nianyi Lin, Jiajie Zhang, Lei Hou and Juanzi Li

A key challenge in applying reinforcement learning (RL) to diffusion large language models (dLLMs) is the intractability of their likelihood functions, which are essential for the RL objective, necessitating corresponding approximation during training. While existing methods approximate the log-likelihoods by their evidence lower bounds (ELBOs) via customized Monte Carlo (MC) sampling, they incur significant memory overhead due to the need to retain all MC samples for the gradient computation of non-linear terms in the RL objective, and thus restrict feasible sample sizes, leading to imprecise likelihood approximations and distorted RL objective. To address this, we propose Boundary-Guided Policy Optimization (BGPO), a memory-efficient RL algorithm that maximizes a specially constructed lower bound of the ELBO-based objective. This lower bound is carefully designed to satisfy two key properties: (1) Linearity: it is a linear sum where each term depends only on a single MC sample, thereby enabling gradient accumulation across samples and ensuring constant memory usage; (2) Equivalence: Both the value and gradient of this lower bound are equal to those of the ELBO-based objective in on-policy training, making it also an effective approximation for the original RL objective. These properties allow BGPO to adopt a large MC sample size, improving likelihood approximations and RL objective estimation, which in turn leads to enhanced performance. Experiments show that BGPO significantly outperforms previous RL algorithms for dLLMs in math problem solving, code generation, and planning tasks.

[Findings] BoundRL: Efficient Token-level Structured Text Segmentation through Reinforced Boundary Generation ► *Underline Video*

Haoyuan Li, Zhengyuan Shen, Sullam Jeoung, Yueyan Chen, Jiayu Li, Qi Zhu, Shuai Wang, Vassilis Ioannidis and Huzefa Rangwala

Structured texts – from technical reports to AI prompts – increasingly require segmentation into semantically meaningful components. Such texts often contain elements beyond plain language, such as code snippets, which conventional sentence-level segmentation methods cannot handle effectively. To address this, we propose BoundRL, a novel approach that jointly performs efficient token-level text segmentation and label prediction for long structured texts. Instead of generating full texts for each segment, it generates only starting tokens and reconstructs the complete texts by locating these tokens within the original texts, thereby reducing inference costs by 90% and minimizing hallucination. To train the models for the boundary generation, BoundRL performs reinforcement learning with verifiable rewards (RLVR) that jointly optimizes document reconstruction fidelity and semantic alignment. It further mitigates entropy collapse by constructing intermediate candidates by perturbing segment boundaries and labels to create stepping stones toward higher-quality solutions. Experiments show that

BoundRL enables small language models (1.7B parameters) to outperform few-shot prompting with much larger models as well as SFT and standard RLVR baselines on complex prompts used for LLM applications.

[Findings] Can Large Language Models Adequately Perform Symbolic Reasoning Over Time Series? ▶ *Underline Video*

Zewen Liu, Juntong Ni, Xianfeng Tang, Max SY Lau, Qi He, Wenpeng Yin and Wei Jin

Uncovering hidden symbolic laws from time series data, as an aspiration dating back to Kepler's discovery of planetary motion, remains a core challenge in scientific discovery and artificial intelligence. While Large Language Models show promise in structured reasoning tasks, their ability to infer interpretable, context-aligned symbolic structures from time series data is still underexplored. To systematically evaluate this capability, we introduce SymbolBench, a comprehensive benchmark designed to assess symbolic reasoning over real-world time series across three tasks: multivariate symbolic regression, Boolean network inference, and causal discovery. Unlike prior efforts limited to simple algebraic equations, SymbolBench spans a diverse set of symbolic forms with varying complexity. We further propose a unified framework that integrates LLMs with genetic programming to form a closed-loop symbolic reasoning system. Our empirical results reveal key strengths and limitations of current models, highlighting the importance of combining domain knowledge, context alignment, and reasoning structure to improve LLMs in automated scientific discovery.

[TACL] Can LLMs Automate Fact-Checking Article Writing? ▶ *Underline Video*

Dhruv Sahnan, David Corney, Irene Larraz, Giovanni Zagni, Ruben Miguez, Zhuohan Xie, Iryna Gurevych, Elizabeth Churchill, Tanmoy Chakraborty and Preslav Nakov

Automatic fact-checking aims to support professional fact-checkers by offering tools that can help speed up manual fact-checking. Yet existing frameworks fail to address the key step of producing output suitable for broader dissemination to the general public: while human fact-checkers communicate their findings through fact-checking articles, automated systems typically produce little or no justification for their assessments. Here, we aim to bridge this gap. In particular, we argue for the need to extend the typical automatic fact-checking pipeline with automatic generation of full fact-checking articles. We first identify key desiderata for such articles through a series of interviews with experts from leading fact-checking organizations. We then develop QRAFT, an LLM-based agentic framework that mimics the writing workflow of human fact-checkers. Finally, we assess the practical usefulness of QRAFT through human evaluations with professional fact-checkers. Our evaluation shows that while QRAFT outperforms several previously proposed text-generation approaches, it lags considerably behind expert-written articles. We hope that our work will enable further research in this new and important direction.

[Findings] CCTVBench: Contrastive Consistency Traffic VideoQA Benchmark for Multimodal LLMs ▶ *Underline Video*

Xingcheng Zhou, Hao Guo, Rui Song, Walter Zimmer, Mingyu Liu, André Schamschurko, Hu Cao and Alois Knoll

Safety-critical traffic reasoning requires contrastive consistency: models must detect true hazards when an accident occurs, and reliably reject plausible-but-false hypotheses under near-identical counterfactual scenes. We present CCTVBench, a Contrastive Consistency Traffic VideoQA Benchmark built on paired real accident videos and world-model-generated counterfactual counterparts, together with minimally different, mutually exclusive hypothesis questions. CCTVBench enforces a single structured decision pattern over each video question quadruple and provides actionable diagnostics that decompose failures into positive omission, positive swap, negative hallucination, and mutual-exclusivity violation, while separating video versus question consistency. Experiments across open-source and proprietary video LLMs reveal a large and persistent gap between standard per-instance QA metrics and quadruple-level contrastive consistency, with unreliable none-of-the-above rejection as a key bottleneck. Finally, we introduce C-TCD, which leverages the semantically exclusive counterpart video as the contrast input at inference time, improving both instance-level QA and contrastive consistency.

[Long] CODESTRUCT: Code Agents over Structured Action Spaces ▶ *Underline Video*

Myeongsoo Kim, Chao-Chun Hsu, Dingmin Wang, Shweta Garg, Varun Kumar and Murali Krishna Ramanathan

LLM-based code agents treat repositories as unstructured text, applying edits through brittle string matching that frequently fails due to formatting drift or ambiguous patterns. We propose reframing the codebase as a structured action space where agents operate on named AST entities rather than text spans. Our framework, CodeStruct, provides readCode for retrieving complete syntactic units and editCode for applying syntax-validated transformations to semantic program elements. Evaluated on SWE-Bench Verified across six LLMs, CodeStruct improves Pass@1 accuracy by 1.2-5.0% while reducing token consumption by 12-38% for most models. Models that frequently fail to produce valid patches under text-based interfaces benefit most: GPT-5-nano improves by 20.8% as empty-patch failures drop from 46.6% to 7.2%. On CodeAssistBench, we observe consistent accuracy gains (+0.8-4.4%) with cost reductions up to 33%. Our results show that structure-aware interfaces offer a more reliable foundation for code agents.

[Long] Cognitive Scaffold: From Fluid Context to Crystallized Memory for Long-Horizon DeepResearch Agents ▶ *Underline Video*

Qiuyuan Ai, Zenghuang Fu, Zhaoyang Li, Jie Song, Guannan He, Ping Jiang and Haoyu Wu

Scaling LLM-based agents to long-horizon deep research is constrained by the context-noise trade-off, where linear history accumulation degrades reasoning and dilutes fine-grained evidence. To address this, we introduce the Cognitive Scaffold, a factorized memory architecture that decouples the cognitive state into a Fluid Working Context for immediate reasoning and a persistent Knowledge Graph for long-term retention. Unlike unstructured summarization, our framework employs a Rejection Sampling Fine-Tuning (RFT) pipeline to crystallize saturated context into structured event snapshots, strictly enforcing atomic constraints to preserve numerical values and entities. During reasoning, a thought-driven dual-path retrieval mechanism enables the agent to proactively recover precise evidence. Empirical evaluations

on Xbench-DeepSearch, BrowseComp-ZH, and GAIA demonstrate that Cognitive Scaffold consistently outperforms baselines, achieving 74.7% Avg@3 and 87.0% Pass@3 on Xbench-DeepSearch, 48.5% Avg@3 and 65.9% Pass@3 on BrowseComp-ZH, and 72.8% Avg@3 and 88.3% Pass@3 on GAIA, while reducing compression hallucinations to 5.3%. We open-source our codebase to facilitate future research.

[Findings] Concept Tokens: Learning Behavioral Embeddings Through Concept Definitions ▶ *Underline Video*

Ignacio Sastre and Aiala Rosá

We propose Concept Tokens, a lightweight method that adds a new special token to a pretrained LLM and learns only its embedding from multiple natural language definitions of a target concept, where occurrences of the concept are replaced by the new token. The LLM is kept frozen and the embedding is optimized with the standard language-modeling objective. We evaluate Concept Tokens in three settings. First, we study hallucinations in closed-book question answering on HotpotQA and find a directional effect: negating the hallucination token reduces hallucinated answers mainly by increasing abstentions, whereas asserting it increases hallucinations and lowers precision. Second, we induce recasting, a pedagogical feedback strategy for second language teaching, and observe the same directional effect. Moreover, compared to providing the full definitional corpus in-context, concept tokens better preserve compliance with other instructions (e.g., asking follow-up questions). Finally, we include a qualitative study with the Eiffel Tower and a fictional "Austral Tower" to illustrate what information the learned embeddings capture and where their limitations emerge. Overall, Concept Tokens provide a compact control signal learned from definitions that can steer behavior in frozen LLMs.

[Long] Conjunctive Prompt Attacks in Multi-Agent LLM Systems ▶ *Underline Video*

Nokimul Hasan Arif, Qian Lou and Mengxin Zheng

Most LLM safety work studies single-agent models, but many real applications rely on multiple interacting agents. In these systems, prompt segmentation and inter-agent routing create attack surfaces that single-agent evaluations miss. We study conjunctive prompt attacks, where a trigger key in the user query and a hidden adversarial template in one compromised remote agent each appear benign alone but activate harmful behavior when routing brings them together. We consider an attacker who changes neither model weights nor the client agent and instead controls only trigger placement and template insertion. Across star, chain, and DAG topologies, routing-aware optimization substantially increases attack success over non-optimized baselines while keeping false activations low. Existing defenses, including PromptGuard, Llama-Guard variants, and system-level controls such as tool restrictions, do not reliably stop the attack because no single component appears malicious in isolation. These results expose a structural vulnerability in agentic LLM pipelines and motivate defenses that reason over routing and cross-agent composition. Code is available at https://github.com/UCF-ML-Research/ConjunctiveAgents.

[Long] ContextLens: Modeling Imperfect Privacy and Safety Context for Legal Compliance ▶ *Underline Video*

Haoran Li, Yulin Chen, Huihao JING, Wenbin Hu, Tsz Ho Li, Chanhou Lou, Hong Ting Tsang, Sirui Han and Yangqiu Song

Individuals' concerns about data privacy and AI safety are highly contextualized and extend beyond sensitive patterns. Addressing these issues requires reasoning about the context to identify and mitigate potential risks. Though researchers have widely explored using large language models (LLMs) as evaluators for contextualized safety and privacy assessments, these efforts typically assume the availability of complete and clear context, whereas real-world contexts tend to be ambiguous and incomplete. In this paper, we propose ContextLens, a semi-rule-based framework that leverages LLMs to ground the input context in the legal domain and explicitly identify both known and unknown factors for legal compliance. Instead of directly assessing safety outcomes, our ContextLens instructs LLMs to answer a set of crafted questions that span over applicability, general principles and detailed provisions to assess compliance with pre-defined priorities and rules. We conduct extensive experiments on existing compliance benchmarks that cover the General Data Protection Regulation (GDPR) and the EU AI Act. The results suggest that our ContextLens can significantly improve LLMs' compliance assessment and surpass existing baselines without any training. Additionally, our ContextLens can further identify the ambiguous and missing factors.

[Findings] Continual Safety Alignment via Gradient-Based Sample Selection ▶ *Underline Video*

Thong Bach, Dung Nguyen, Thao Minh Le and Truyen Tran

Large language models require continuous adaptation to new tasks while preserving safety alignment. However, fine-tuning on even benign data often compromises safety behaviors. We investigate which training samples cause alignment drift through a data-centric lens. Our experiments show samples contribute unequally: high-gradient samples cause greater safety degradation and drive models toward pretrained distributions, while moderate-gradient samples enable task learning with minimal alignment loss. This connects to the elasticity phenomenon, high-gradient samples activate the reversion force pulling models toward pretrained behavior. We propose gradient-based sample selection that filters high-gradient samples during fine-tuning. Across multiple model families on continual domain tasks, our method substantially improves alignment preservation while maintaining competitive task performance, without requiring curated safe data or architectural modifications.

[Long] Curing Miracle Steps in LLM Mathematical Reasoning with Rubric Rewards ▶ *Underline Video*

Youliang Yuan, Qiuyang Mang, Jingbang Chen, Hong Wan, Xiaoyuan Liu, Junjielong Xu, Jen-tse Huang, Wenxuan Wang, Wenxiang Jiao and Pinjia He

In this paper, we observe that current models are susceptible to reward hacking, leading to a substantial overestimation of a model's reasoning ability. This is evidenced by a high incidence of "false positives", solutions that reach the correct answer through an unsound process. Through a systematic analysis with human verification, we establish a taxonomy of these failure modes, identifying patterns like Miracle

Steps, abrupt jumps to a correct output without a valid preceding derivation. Probing experiments suggest that these Miracle Steps are linked to answer-recall shortcuts, including memorization from pretraining, where the model accesses the correct answer independently of its reasoning chain. To mitigate this systemic issue, we introduce the Rubric Reward Model (RRM), a process-oriented reward function that evaluates the entire reasoning trajectory against problem-specific rubrics. The RRM explicitly penalizes logical flaws and encourages rigorous deduction. When integrated into an RL pipeline, RRM-based training consistently outperforms outcome-only supervision across four math benchmarks. Notably, it boosts Verified Pass@1024 on AIME2024 from 26.7% to 62.6% and reduces the incidence of Miracle Steps by 71%. Our work demonstrates that rewarding the solution process is crucial for building accurate and reliable models.

[Long] Debiased Orthogonal Boundary-Driven Efficient Noise Mitigation ▶ *Underline Video*

Hao Li, Jiayang Gu, Jingkuan Song, An Zhang and Lianli Gao

Mitigating the detrimental effects of noisy labels on the training process has become increasingly critical, as obtaining entirely clean or human-annotated samples for large-scale pre-training tasks is often impractical. Nonetheless, existing noise mitigation methods often encounter limitations in practical applications due to their task-specific design, model dependency, and significant computational overhead. In this work, we exploit the properties of high-dimensional orthogonality to identify a robust and effective boundary in cone space for separating clean and noisy samples. Building on this, we propose One-Step Anti-noise (OSA), a model-agnostic noisy label mitigation paradigm that employs an estimator model and a scoring function to assess the noise level of input pairs through just one-step inference. We empirically validate the superiority of OSA, demonstrating its enhanced training robustness, improved task transferability, streamlined deployment, and reduced computational overhead across diverse benchmarks, models, and tasks. Our code is released at https://github.com/leolee99/OSA.

[Findings] Direct Token Optimization: A Self-Contained Approach to Large Language Model Unlearning ▶ *Underline Video*

Hong kyu Lee, Ruixuan Liu and Li Xiong

Machine unlearning is an emerging technique that removes the influence of a subset of training data (forget set) from a model without full retraining, with applications including privacy protection, content moderation, and model correction. The key challenge lies in achieving strong unlearning efficacy while preserving the overall utility. Existing unlearning methods for large language models (LLMs) often rely on auxiliary models, retain datasets, or even commercial AI services. However, dependence on these external resources is often impractical and could potentially introduce additional privacy risks. In this work, we propose direct token optimization (DTO), a self-contained unlearning approach for LLMs that directly optimizes the token-level objectives to unlearn specific sequences without external resources. For each sequence to be unlearned, we identify target tokens that encode critical knowledge for unlearning and treat remaining tokens as non-target ones for maintaining the model utility. DTO maximizes an unlearning objective on target tokens and applies a utilitypreservation regularizer on nontarget tokens.Across multiple unlearning benchmarks, DTO improves the forget quality up to 16.8x over the latest baselines while maintaining comparable model utility. Our code is available at github.com/Emory-AIMS/direct_token_optimization.

[Long] Distillation Traps and Guards: A Calibration Knob for LLM Distillability ▶ *Underline Video*

Weixiao Zhan, Yongcheng Jing, Leszek Rutkowski and Dacheng Tao

Knowledge distillation (KD) transfers capabilities from large language models (LLMs) to smaller students, yet it can fail unpredictably and also underpins model leakage risks. Our analysis revealed several distillation traps: tail noise, off-policy instability, and, most fundamentally, the teacher-student gap, that distort training signals. These traps manifest as overconfident hallucinations, self-correction collapse, and local decoding degradation, causing distillation to fail. Motivated by these findings, we propose a post-hoc calibration method that, to the best of our knowledge, for the first time enables control over a teacher's distillability via reinforcement fine-tuning (RFT). Our objective combines task utility, KL anchor, and across-tokenizer calibration reward. This makes distillability a practical safety lever for foundation models, connecting robust teacher-student transfer with deployment-aware model protection. Experiments across math, knowledge QA, and instruction-following tasks show that students distilled from distillable calibrated teachers outperform SFT and KD baselines, while undistillable calibrated teachers retain their task performance but cause distilled students to collapse, offering a practical knob for both better KD and model IP protection.

[Long] DMN: A Compositional Framework for Jailbreaking Multimodal LLMs with Multi-Image Inputs ▶ *Underline Video*

Wenzhuo Xu, Zhipeng Wei, Zonghao Ying, Deyue Zhang, Dongdong Yang, Xiangzheng Zhang and Quanchen Zou

Multimodal Large Language Models (MLLMs) are vulnerable to jailbreak attacks, which can elicit harmful responses from MLLMs. Many MLLMs support multi-image inputs, inadvertently introducing new vulnerabilities due to less efforts on multi-image safety alignment. Previous MLLM jailbreak methods only uses a single image, which restricts the attack space: they cannot distribute harmful requests across multiple images, carry abundant information, or exploit additional visual reasoning tasks to distract MLLMs. To address these limitations, in this paper, we propose a compositional jailbreak framework, DMN, which leverages Distributed instruction, Multimodal evidence and a Number chain task to fully enhance the jailbreak performance. Extensive experiments show that DMN is highly effective for MLLM jailbreaking, e.g. achieving attack success rates of over 90% on GPT-4o, Gemini-2.5-pro and Claude Sonnet 4, surpassing other baselines by a large margin. This compositional, multi-image jailbreak strategy reveals fundamental weaknesses in their safety mechanisms.

[Findings] DRIV-EX: Counterfactual Explanations for Driving LLMs ▶ *Underline Video*

amaia cardiel, eloi zablocki, elias ramzi and eric gaussier

Large language models (LLMs) are increasingly used as reasoning engines in autonomous driving, yet their decision-making remains opaque.

We propose to study their decision process through counterfactual explanations, which identify the minimal semantic changes to a scene description required to alter a driving plan. We introduce DRIV-EX, a method that leverages gradient-based optimization on continuous embeddings to identify the input shifts required to flip the model's decision. Crucially, to avoid the incoherent text typical of unconstrained continuous optimization, DRIV-EX uses these optimized embeddings solely as a semantic guide: they are used to bias a controlled decoding process that re-generates the original scene description. This approach effectively steers the generation toward the counterfactual target while guaranteeing the linguistic fluency, domain validity, and proximity to the original input essential for interpretability. Evaluated using the LC-LLM planner on the textual highD dataset, DRIV-EX generates valid, fluent counterfactuals more reliably than existing baselines. It successfully exposes latent biases and provides concrete insights to improve the robustness of LLM-based driving agents. The code is available at "https://github.com/Amaia-CARDIEL/DRIV_EX".

[Long] E-ViC: Reasoning Beyond Text via Embodied Visual Chain for Spatial Intelligence ► *Underline Video*

Junbo Qi, Yi Zhang, Hanchu Ni, Che Liu, Ruilin Yang, Zhiming Yao, Xiancong Ren, Liangjian Wen, Wei Ge, Yuya Ieiri, Osamu Yoshie, Yong Dai and Xiaozhu Ju

Precise spatial reasoning is fundamental to embodied intelligence, yet current Vision-Language Models (VLMs) remain bottlenecked by text-based Chain-of-Thought (CoT) that relies solely on textual reasoning trajectories, often bypassing active engagement with fine-grained visual details. To address this, we present E-ViC (Embodied Visual Chain), a framework that moves reasoning beyond text and directly into the visual domain. By formulating visual operations (e.g., zooming, marking) as executable primitives, E-ViC transforms perception from static prediction into an active verification process. Distinct from approaches relying on supervised step-wise trajectories, E-ViC is trained via an agentic reinforcement learning paradigm. This enables the model to autonomously discover optimal policies, leading to the emergence of human-like "look-and-confirm" strategies driven solely by task-level rewards. To facilitate this, we curate a comprehensive 24.4K-sample dataset covering diverse embodied tasks. By grounding reasoning in pixel-level interactions, E-ViC reframes spatial intelligence as a verifiable, tool-using capability. Extensive evaluations on external benchmarks demonstrate that our approach consistently outperforms strong VLM baselines with an average gain of 10.1%.

[Long] EDUMATH: Generating Standards-aligned Educational Math Word Problems ► *Underline Video*

Bryan R. Christ, Penelope Molitz, Beau LeBlond, Zachary Gottesman, Jonathan Kropko and Thomas Hartvigsen

Math word problems (MWPs) are critical K-12 educational tools, and customizing them to students' interests and ability levels can enhance learning. However, teachers struggle to find time to customize MWPs for students given large class sizes and increasing burnout. We propose that LLMs can support math education by generating MWPs customized to student interests and math education standards. We use a joint human expert-LLM judge approach to evaluate over 11,000 MWPs generated by open and closed LLMs and develop the first teacher-annotated dataset for standards-aligned educational MWP generation. We show the value of our data by using it to train a 12B open model that matches the performance of larger and more capable open models. We also use our teacher-annotated data to train a text classifier that enables a 30B open LLM to outperform existing closed baselines without any training. Next, we show our models' MWPs are more similar to human-written MWPs than those from existing models. We conclude by conducting the first study of customized LLM-generated MWPs with grade school students, finding they perform similarly on our models' MWPs relative to human-written MWPs but consistently prefer our customized MWPs.

[Long] Efficient Paths and Dense Rewards: Probabilistic Flow Reasoning for Large Language Models ► *Underline Video*

Yan Liu, Feng Zhang, Zhanyu Ma, Jun Xu, Jiuchong Gao, Jinghua Hao, Renqing He, Han Liu and Yangdong Deng

High-quality chain-of-thought has demonstrated strong potential for unlocking the reasoning capabilities of large language models. However, current paradigms typically treat the reasoning process as an indivisible sequence, lacking an intrinsic mechanism to quantify step-wise information gain. This granularity gap manifests in two limitations: inference inefficiency from redundant exploration without explicit guidance, and optimization difficulty due to sparse outcome supervision or costly external verifiers. In this work, we propose CoT-Flow, a framework that reconceptualizes discrete reasoning steps as a continuous probabilistic flow, quantifying the contribution of each step toward the ground-truth answer. Built on this formulation, CoT-Flow enables two complementary methodologies: flow-guided decoding, which employs a greedy flow-based decoding strategy to extract information-efficient reasoning paths, and flow-based reinforcement learning, which constructs a verifier-free dense reward function. Experiments on challenging benchmarks demonstrate that CoT-Flow achieves a superior balance between inference efficiency and reasoning performance.

[Long] Efficient Provably Secure Linguistic Steganography via Range Coding ► *Underline Video*

Ruiyi Yan and Yugo Murawaki

Linguistic steganography involves embedding secret messages within seemingly innocuous texts to enable covert communication. Provable security, which is a long-standing goal and key motivation, has been extended to language-model-based steganography. Previous provably secure approaches have achieved perfect imperceptibility, measured by zero Kullback-Leibler (KL) divergence, but at the expense of embedding capacity. In this paper, we attempt to directly use a classic entropy coding method (range coding) to achieve secure steganography, and then propose an efficient and provably secure linguistic steganographic method with a rotation mechanism. Experiments across various language models show that our method achieves around 100% entropy utilization (embedding efficiency) for embedding capacity, outperforming the existing baseline methods. Moreover, it achieves high embedding speeds (up to 1554.66 bits/s on GPT-2). The code is available at github.com/ryehr/RRC_steganography.

[Long] Embracing Anisotropy: Turning Massive Activations into Interpretable Control Knobs for Large Language Models ► *Underline Video*

Youngji Roh, Hyunjin Cho and Jaehyung Kim

Large Language Models (LLMs) exhibit highly anisotropic internal representations, often characterized by massive activations, a phenomenon where a small subset of feature dimensions possesses magnitudes significantly larger than the rest. While prior works view these extreme dimensions primarily as artifacts to be managed, we propose a distinct perspective: these dimensions serve as intrinsic interpretable functional units arising from domain specialization. Specifically, we propose a simple magnitude-based criterion to identify Domain-Critical Dimensions in a training-free manner. Our analyses reveal that such dimensions behave as interpretable semantic detectors for symbolic/quantitative patterns or domain-specific terms. In addition, we introduce Critical Dimension Steering, which applies activation steering exclusively to the identified dimensions. Empirical results show that this approach outperforms conventional whole-dimension steering in domain adaptation and jailbreaking scenarios.

[Long] Enabling Agents to Communicate Entirely in Latent Space ► *Underline Video*

Zhuoyun Du, Runze Wang, Huiyu Bai, Zouying Cao, Xiaoyong Zhu, Yu Cheng, Bo Zheng, Wei Chen and Haochao Ying

While natural language is the de facto communication medium for LLM-based agents, it presents a fundamental constraint. The process of downsampling rich, internal latent states into discrete tokens inherently limits the depth and nuance of information that can be transmitted, thereby hindering collaborative problem-solving. Inspired by telepathy, which bypasses symbolic language in communication, we propose Interlat (Inter-agent Latent Space Communication), a paradigm that leverages the continuous last hidden states of an LLM as a representation of its thought for direct communication (termed "latent communication"). An additional learned compression process further compresses latent communication via latent space reasoning. Experiments demonstrate that Interlat outperforms both fine-tuned chain-of-thought (CoT) prompting and single-agent baselines, even across heterogeneous models, promoting more exploratory behavior and enabling genuine utilization of latent information. Further compression not only substantially accelerates inference by up to 24x but also maintains competitive performance through an efficient information-preserving mechanism. We position this work as a feasibility study of entirely latent space inter-agent communication, and our results highlight its potential, offering valuable insights for future research.

[Long] EpiCaR: Knowing What You Don't Know Matters for Better Reasoning in LLMs ► *Underline Video*

Jewon Yeom, Jaewon Sok, Seonghyeon Park, Jeongjae Park and Taesup Kim

Improving the reasoning abilities of large language models (LLMs) has largely relied on iterative self-training with model-generated data. While effective at boosting accuracy, existing approaches primarily reinforce successful reasoning paths, incurring a substantial calibration cost: models become overconfident and lose the ability to represent uncertainty. This failure has been characterized as a form of model collapse in alignment, where predictive distributions degenerate toward low-variance point estimates. We address this issue by reframing open-ended reasoning training as an epistemic learning problem, in which models must learn not only how to reason, but also when their reasoning should be trusted. We propose epistemically-calibrated reasoning (EpiCaR) as a training objective that jointly optimizes reasoning performance and calibration, and instantiate it within an iterative supervised fine-tuning framework using explicitly extracted meta-cognitive self-evaluation signals. Experiments on Llama-3 and Qwen-3 families demonstrate that our approach achieves Pareto-superiority over standard baselines in both accuracy and calibration, particularly in models with sufficient reasoning capacity (e.g., 3B+). This framework generalizes effectively to OOD mathematical reasoning (GSM8K) and code generation (MBPP). Ultimately, our approach enables a 3x reduction in the overall inference compute budget, matching the K=30 majority-vote performance of STaR with only K=10 confidence-weighted samples, entirely without the multi-model overhead of external verifiers.

[Findings] Establishing a Scale for Kullback-Leibler Divergence in Language Models Across Various Settings ► *Underline Video*

Ryo Kishino, Yusuke Takase, Momose Oyama, Hiroaki Yamagiwa and Hidetoshi Shimodaira

Log-likelihood vectors define a common space for comparing language models as probability distributions, enabling unified comparisons across heterogeneous settings. We extend this framework to training checkpoints and intermediate layers, and establish a consistent scale for KL divergence across pretraining, model size, random seeds, quantization, fine-tuning, and layers. Analysis of Pythia pretraining trajectories further shows that changes in log-likelihood space, as measured by the scaling behavior of KL divergence, are much smaller than in weight space, resulting in subdiffusive learning trajectories and early stabilization of language-model behavior despite weight drift.

[Long] EvolvR: Self-Evolving Pairwise Reasoning for Story Evaluation to Enhance Generation ► *Underline Video*

Xinda Wang, Zhengxu Hou, Yangshijie Zhang, yanbingren yanbingren, Jialin Liu, ChenZhuo Zhao, Zhibo Yang, Bin-Bin Yang and Feng Xiao

Although the effectiveness of Large Language Models as judges has been validated, their performance remains limited in open-ended tasks, particularly in story evaluation. Accurate story evaluation is crucial not only for assisting human quality judgment but also for providing reward signals to guide story generation. However, existing methods face a dilemma: prompt engineering for closed-source models suffers from poor adaptability, while fine-tuning approaches for open-source models lack the reasoning capabilities essential for story evaluation. To address this, we propose the Self-Evolving Pairwise Reasoning (EvolvR) framework. Grounded in pairwise comparison, the framework first self-synthesizes score-aligned Chain-of-Thought (CoT) data via a multi-persona strategy. To ensure data quality, these raw CoTs undergo a self-filtering process, utilizing multi-agents to guarantee their logical rigor and robustness. Finally, the evaluator trained on the refined data is deployed as a reward model to guide the story generation task. Experimental results demonstrate that our framework achieves state-of-the-art

performance on three evaluation benchmarks including StoryER, HANNA and OpenMEVA. Furthermore, when served as a reward model, it enhances the quality of generated stories, thereby validating the superiority of our self-evolving approach.

[Findings] FAITH: Factuality Alignment through Integrating Trustworthiness and Honestness ► *Underline Video*
Xiaoning Dong, Chengyan Wu, Yajie Wen, Yu Chen, Yun Xue, Zhang Jing, Wei Xu and Bolei Ma
Large Language Models (LLMs) can generate factually inaccurate content even if they have corresponding knowledge, which critically undermines their reliability. Existing approaches attempt to mitigate this by incorporating uncertainty in QA prompt during training, but these numerical scores lack the semantic richness for LLM to properly understand its internal states of trustworthiness and honestness, leading to insufficient factuality alignment. We introduce FAITH (Factuality Alignment through Integrating Trustworthiness and Honestness), a post-training framework for factuality alignment that integrates natural-language uncertainty signals with external knowledge. Specifically, we augment training datasets by computing confidence scores and semantic entropy from LLM outputs and mapping them into a knowledge state quadrant that describes the model's internal knowledge possession (trustworthiness) and answering behaviors (honestness) in natural language. Based on this enhanced data, we design a reward function that considers both correctness and uncertainty signals, and fine-tune the LLM using the Proximal Policy Optimization (PPO) algorithm. To further mitigate weakly grounded responses, we design a retrieval-augmented module that retrieves relevant external passages, improving the consistency between internal and external knowledge representations. Extensive experiments on four knowledge-intensive benchmarks demonstrate that FAITH enhances the factual accuracy and truthfulness of LLMs.

[Findings] False Sense of Security: Why Probing-based Malicious Input Detection Fails to Generalize ► *Underline Video*
Cheng Wang, Zeming Wei, Qin Liu, Wenxuan Zhou and Muhao Chen
Large Language Models (LLMs) can comply with harmful instructions, raising serious safety concerns despite their impressive capabilities. Recent work has leveraged probing-based approaches to study the separability of malicious and benign inputs in LLMs' internal representations, and researchers have proposed using such probing methods for safety detection. We systematically re-examine this paradigm. Motivated by poor out-of-distribution performance, we hypothesize that probes learn superficial patterns rather than semantic harmfulness. Through controlled experiments, we confirm this hypothesis and identify the specific patterns learned: instructional patterns and trigger words. Our investigation follows a systematic approach, progressing from demonstrating comparable performance of simple n-gram methods, to controlled experiments with semantically cleaned datasets, to detailed analysis of pattern dependencies. These results reveal a false sense of security around current probing-based approaches and highlight the need to redesign both models and evaluation protocols, for which we provide further discussions in the hope of suggesting responsible further research in this direction.

[Findings] FinCARDS: Card-Based Analyst Reranking for Financial Document Question Answering ► *Underline Video*
Yixi Zhou, Fan Zhang, Yu Chen, Haipeng Zhang, Preslav Nakov and Zhuohan Xie
Financial question answering (QA) over long corporate filings requires evidence to satisfy strict constraints on entities, financial metrics, fiscal periods, and numeric values. However, existing LLM-based rerankers primarily optimize semantic relevance, leading to unstable rankings and opaque decisions on long documents. We propose FINCARDS, a structured reranking framework that reframes financial evidence selection as constraint satisfaction under a finance-aware schema. FINCARDS represents filing chunks and questions using aligned schema fields (entities, metrics, periods, and numeric spans), enabling deterministic field-level matching. Evidence is selected via a multi-stage tournament reranking with stability-aware aggregation, producing auditable decision traces. Across two corporate filing QA benchmarks, FINCARDS substantially improves early-rank retrieval over both lexical and LLM-based reranking baselines, while reducing ranking variance, without requiring model fine-tuning or unpredictable inference budgets. Our code is available at https://github.com/XanderZhou2022/FINCARDS.

[Findings] Finch: Benchmarking Finance & Accounting across Spreadsheet-Centric Enterprise Workflows ► *Underline Video*
Haoyu Dong, Pengkun Zhang, Yan Gao, Xuanyu Dong, Yilin Cheng, Mingzhe Lu, Zikun Zhu, Adina Yakefu and Shuxin Zheng
We introduce FinWorkBench (a.k.a. Finch) for evaluating AI agents on real-world, enterprise-grade finance and accounting workflows that interleave data entry, structuring, formatting, web search, cross-file retrieval, calculation, modeling, validation, translation, visualization, and reporting. Finch is sourced from authentic enterprise workspaces from Enron (15,000 files and 500,000 emails) and other financial institutions, covering the period 2000-2025 and preserving the in-the-wild messiness of multimodal artifacts such as tables and charts across diverse domains including budgeting, trading, asset management, and operational management. We propose a workflow construction process that combines LLM-assisted mining of workflows from authentic enterprise environments with expert annotation: (1) LLM-assisted, expert-verified derivation of workflows from real-world email threads and spreadsheet version histories, and (2) meticulous annotation requiring over 700 hours of expert effort. This yields 172 composite workflows with 384 tasks, involving 1,710 spreadsheets with 27 million cells, along with PDFs and other artifacts, capturing the intrinsically messy, long-horizon, knowledge-intensive, and collaborative nature of real-world enterprise work. We conduct both human and automated evaluations of frontier AI systems, including GPT 5.1, Claude Sonnet/Opus 4.5, Gemini 3 Pro, Grok 4, and Qwen 3 Max. Under human evaluation, GPT 5.1 Pro spends an average of 16.8 minutes per workflow yet passes only 38.4% of workflows. Comprehensive case studies further surface the challenges that real-world enterprise workflows pose for AI agents.

[Long] Flattery in Motion: Benchmarking and Analyzing Sycophancy in Video-LLMs ► *Underline Video*
Wenrui Zhou, Mohamed Hendy, Shu Yang, Qingsong Yang, Zikun Guo, Yuyu Luo, Lijie Hu and Di Wang

As video large language models (Video-LLMs) become increasingly integrated into real-world applications that demand grounded multimodal reasoning, ensuring their factual consistency and reliability is of critical importance. However, sycophancy, the tendency of these models to align with user input even when it contradicts the visual evidence, undermines their trustworthiness in such contexts. Current sycophancy research has largely overlooked its specific manifestations in the video-language domain, resulting in a notable absence of systematic benchmarks and targeted evaluations to understand how Video-LLMs respond under misleading user input. To fill this gap, we propose ViSE (Video-LLM Sycophancy Benchmarking and Evaluation), the first benchmark designed to evaluate sycophantic behavior in state-of-the-art Video-LLMs across diverse question formats, prompt biases, and visual reasoning tasks. Specifically, ViSE pioneeringly brings linguistic perspectives on sycophancy into the video domain, enabling fine-grained analysis across multiple sycophancy types and interaction patterns. Furthermore, we propose two potential training-free mitigation strategies revealing potential paths for reducing sycophantic bias: (i) enhancing visual grounding through interpretable key-frame selection and (ii) steering model behavior away from sycophancy via targeted, inference-time intervention on its internal neural representations. Our code is available at https://github.com/William030422/Video-Sycophancy.

[Long] Follow the Flow: On Information Flow Across Textual Tokens in Text-to-Image Models ► *Underline Video*

Guy Kaplan, Michael Toker, Yuval Reif, Yonatan Belinkov and Roy Schwartz

Text-to-image generation models suffer from alignment problems, where generated images fail to accurately capture the objects and relations in the text prompt. Prior work has focused on improving alignment by refining the diffusion process, ignoring the role of the text encoder, which guides the diffusion. In this work, we investigate how semantic information is distributed across token representations in text-to-image prompts, analyzing it at two levels: (1) in-item representation, whether individual tokens represent their lexical item (i.e., a word or expression conveying a single concept), and (2) cross-item interaction, whether information flows between tokens of different lexical items. We use patching techniques to uncover encoding patterns, and find that information is usually concentrated in only one or two of the item's tokens; for example, in the item "San Francisco's Golden Gate Bridge", the token "Gate" sufficiently captures the entire expression while the other tokens could effectively be discarded. Lexical items also tend to remain isolated; for instance, in the prompt "a green dog", the token "dog" encodes no visual information about "green". However, in some cases, items do influence each other's representation, often leading to misinterpretations, e.g., in the prompt "a pool by a table", the token "pool" represents a "pool table" after contextualization. Our findings highlight the critical role of token-level encoding in image generation, and demonstrate that simple interventions at the encoding stage can substantially improve alignment and generation quality.

[Long] Forest Before Trees: Latent Superposition for Efficient Visual Reasoning ► *Underline Video*

Yubo Wang, Juntian Zhang, Yichen Wu, Yankai Lin, Nils Lukas and Yuhan Liu

While Chain-of-Thought empowers Large Vision-Language Models with multi-step reasoning, explicit textual rationales suffer from an information bandwidth bottleneck, where continuous visual details are discarded during discrete tokenization. Recent latent reasoning methods attempt to address this challenge, but often fall prey to premature semantic collapse due to rigid autoregressive objectives. In this paper, we propose Laser, a novel paradigm that reformulates visual deduction via Dynamic Windowed Alignment Learning. Instead of forcing a point-wise prediction, Laser aligns the latent state with a dynamic validity window of future semantics. This mechanism enforces a "Forest-before-Trees" cognitive hierarchy, enabling the model to maintain a probabilistic superposition of global features before narrowing down to local details. Crucially, Laser maintains interpretability via decodable trajectories while stabilizing unconstrained learning via Self-Refined Superposition. Extensive experiments on 6 benchmarks demonstrate that Laser achieves state-of-the-art performance among latent reasoning methods, surpassing the strong baseline Monet by 5.03% on average. Notably, it achieves these gains with extreme efficiency, reducing inference tokens by more than 97%, while demonstrating robust generalization to out-of-distribution domains.

[Long] Forget What Matters, Keep the Rest: Selective Unlearning of Informative Tokens ► *Underline Video*

Seunghee Koh, Sunghyun Baek, Youngdong Kim and Junmo Kim

Unlearning in large language models (LLMs) has emerged as a promising safeguard against adversarial behaviors. When the forgetting loss is applied uniformly without considering token-level semantic importance, model utility can be unnecessarily degraded. Recent studies have explored token-wise loss regularizers that prioritize informative tokens, but largely rely on ground-truth confidence or external linguistic parsers, which limits their ability to capture contextual information or the models overall predictive state. Intuitively, function words like the primarily serve syntactic roles and are highly predictable with little ambiguity, but informative words admit multiple plausible alternatives with greater uncertainty. Based on this intuition, we propose Entropy-guided Token Weighting (ETW), a token-level unlearning regularizer that uses entropy of the predictive distribution as a proxy for token informativeness. We demonstrate that informative tokens tend to have higher entropy, whereas structural tokens tend to have lower entropy. This behavior enables ETW to achieve more effective unlearning while better preserving model utility than existing token-level approaches.

[Findings] From 128K to 4M: Efficient Training of Ultra-Long Context Large Language Models ► *Underline Video*

Chejian Xu, Wei Ping, Peng Xu, Zihan Liu, Boxin Wang, Mohammad Shoeybi, Bo Li and Bryan Catanzaro

Long-context capabilities are essential for a wide range of applications, including document and video understanding, in-context learning, and inference-time scaling, all of which require models to process and reason over long sequences of text and multimodal data. In this work, we introduce an efficient training recipe for building ultra-long context LLMs from aligned instruct model, pushing the boundaries of context lengths from 128K to 1M, 2M, and 4M tokens. Our approach leverages continued pretraining strategies to extend the context

window, while employing efficient instruction tuning to maintain short context capabilities. Our UltraLong-8B, built on Llama-3.1-Instruct, achieves state-of-the-art performance across a diverse set of long-context benchmarks. Importantly, UltraLong-8B also maintains competitive performance on standard benchmarks, showing balanced improvements for both long and short context tasks. We provide an in-depth analysis of key design choices, highlighting the impacts of scaling strategies and data composition. Our findings establish a robust framework for efficiently scaling context lengths while preserving general model capabilities. We released all model weights for open research.

[Long] From AR to Diffusion: Efficiently Adapting Large Language Models with Strictly Causal and Elastic Horizons ▶ ***Underline Video***

Xiangyu Ma, Teng Xiao, Zuchao Li and Lefei Zhang

Diffusion models promise efficient parallel text generation but rely on bidirectional attention, creating a structural mismatch with pre-trained Autoregressive (AR) models. This incompatibility precludes reusing robust AR priors, necessitating prohibitive pre-training from scratch. To bridge this gap, we propose FLUID, a framework that efficiently adapts AR backbones to the diffusion paradigm. By enforcing Strictly Causal Alignment, FLUID enables seamless initialization from standard GPT-style checkpoints, circumventing the need for massive pre-training. Furthermore, we introduce Elastic Horizons, an entropy-driven mechanism that dynamically modulates denoising strides based on local information density rather than fixed schedules. Experiments demonstrate that FLUID achieves state-of-the-art performance while reducing training costs by orders of magnitude, effectively reconciling established AR foundations with efficient parallel generation. Our code is available at https://huggingface.co/MYTH-Lab/FLUID.

[TACL] From Belief Entrenchment to Robust Reasoning in LLM Agents ▶ ***Underline Video***

Jihwan Oh, Minchan Jeong, Jongwoo Ko and Se-Young Yun

Large Language Models (LLMs) show promise in complex reasoning, but methods like Multi-Agent Debate (MAD) often reinforce rather than correct biases. We identify this critical failure as bias reinforcement and propose a conceptual framework that decomposes it into two root causes: (1) the models initial static bias and (2) a self-reinforcing debate dynamic that systematically amplifies the most common, not necessarily correct, viewpoints. To address both causes, we propose DReaMAD (Diverse Reasoning via Multi-Agent Debate with Refined Prompt), a framework that first corrects static bias by eliciting strategic prior knowledge, then reshapes the debate dynamic by promoting diverse perspectives. To validate our approach, we introduce MetaNIM Arena, a new benchmark for adversarial strategic decision-making. Empirical results show that DReaMAD significantly improves decision accuracy and mitigates bias, establishing a more robust foundation for LLM-based reasoning.

[Long] From Charts to Code: A Hierarchical Benchmark for Multimodal Models ▶ ***Underline Video***

Jiahao Tang, Henry Hengyuan Zhao, Lijian Wu, Zijian Zhang, Yifei Tao, Dongxing Mao, Yang Wan, Jingru Tan, Min Zeng, Min Li and Alex Jinpeng Wang

We introduce Chart2Code, a new benchmark for evaluating the natural language to chart code generation capabilities of large multimodal models (LMMs). Chart2Code is explicitly designed from a user-driven perspective, capturing diverse real-world scenarios and progressively increasing task difficulty. It consists of three levels: Level 1 (Chart Reproduction) reproduces charts from a reference figure and user query; Level 2 (Chart Editing) involves complex modifications such as changing chart types or adding elements; and Level 3 (Long-Table to Chart Generation) requires models to transform long, unprocessed tables into faithful charts following user instructions. To our knowledge, this is the first hierarchical benchmark that reflects practical chart2code usage while systematically scaling task complexity. In total, Chart2Code contains 2,186 tasks across 22 chart types, paired with multi-level evaluation metrics that assess both code correctness and the visual fidelity of rendered charts. We benchmark 29 state-of-the-art (SoTA) LMMs, including both proprietary and the latest open-source models such as GPT-5.2, Qwen3-VL, InternVL3/3.5, MiMo-VL, and Seed-1.6-VL. Experimental results demonstrate that even the SoTA model GPT-5.2 averages 72.21 on code-based evaluation and only 33.41 on chart-quality assessment across the editing tasks, underscoring the difficulty of Chart2Code. We anticipate this benchmark will drive advances in multimodal reasoning and foster the development of more robust and general-purpose LMMs.

[Long] From Detection to Understanding: Multi-Turn Reasoning for Video Misinformation Analysis ▶ ***Underline Video***

Zhi Zeng, Jiaying Wu, Minnan Luo, Di Zhang, Yifei Yang, Xiangzheng Kong, Herun Wan and Zihan Ma

Video misinformation detection is often approached as a binary veracity classification problem, overlooking the complex reasoning required to explain how and why content misleads. Existing benchmarks fail to capture the diversity of manipulation strategies, such as AI-generated edits and out-of-context manipulation, and do not evaluate whether models can provide process-level justifications for their judgments. We address these limitations with MisVideoQA, a multi-turn benchmark designed to assess comprehensive understanding and reasoning in video misinformation analysis. MisVideoQA covers 12 fine-grained deception categories and evaluates models along six dimensions, progressing from perceptual attribution to intent and persuasion analysis. Recognizing that standard MLLMs struggle to sustain such structured, evidence-based deduction, we propose MisAgent, a Delphi-inspired multi-agent framework in which specialized agents collaboratively integrate multimodal cues with external evidence. Experimental results show that state-of-the-art multimodal large language models perform poorly on MisVideoQA, while MisAgent consistently improves reasoning accuracy and explanation quality. Together, our benchmark and framework establish a unified foundation for reliable, interpretable, and evidence-grounded video misinformation analysis.

[Long] From Nodes to Narratives: Explaining Graph Neural Networks with LLMs and Graph Context ▶ ***Underline Video***

Peyman Baghershahi, Gregoire Fournier, Pranav Nyati and Sourav Medya
Graph Neural Networks (GNNs) have emerged as powerful tools for learning over structured data, including text-attributed graphs (TAGs), which are common in domains such as citation networks, social platforms, and knowledge graphs. GNNs are not inherently interpretable and thus, many explanation methods have been proposed. However, existing explanation methods often struggle to generate interpretable, fine-grained rationales, especially when node attributes include rich natural language. In this work, we introduce GSPELL, a lightweight, post-hoc framework that uses large language models (LLMs) to generate faithful and interpretable explanations for GNN predictions. GSPELL projects GNN node embeddings into the LLM embedding space and constructs hybrid prompts that interleave soft prompts with textual inputs from the graph structure. This enables the LLM to reason about GNN internal representations and to produce natural-language explanations, along with concise explanation subgraphs. Our experiments across real-world TAG datasets demonstrate that GSPELL achieves a favorable trade-off between fidelity and sparsity, while improving human-centric metrics such as insightfulness. GSPELL sets a new direction for LLM-based explainability in graph learning by aligning GNN internals with human reasoning.

[Long] From Trajectories to Graphs: Contract-Checked Editing for Verifier-Guided LLM Reasoning ▶ *Underline Video*
Rui Li and Shuang Cao
Inference-time search can substantially improve LLM reasoning when tasks admit deterministic verification, but existing methods largely refine single trajectories and lack a reliable mechanism for composing partial solutions across candidates. We propose contract-checked graph editing: represent each candidate as an interface-typed reasoning DAG and validate every nontrivial edit with a deterministic structural gate (acyclicity, namespace closure, schema validity, terminal constraints) before invoking the verifier. The gate certifies runnability only and emits auditable rejection reasons; semantic correctness is determined solely by the verifier. Instantiated in Genetic Inference Search (GIS) with Qwen2.5-32B-Instruct under strictly matched token budgets (8K tokens), contract-checked grafting increases verifier-runnable recombination from 41.2% to 92.8% and improves accuracy over rStar (+6.1 on MATH, +9.1 on MATH L5) while using 42% fewer verifier calls. The same operators transfer across outer loops (beam, best-first, MCTS) and to structured generation and code, outperforming execution-guided beam search on Spider (+2.8) and improving multi-file code generation on HumanEval-MF (+9.2).

[Findings] From What Is Said to Why It Is Framed: Intent-Aware News Video Understanding ▶ *Underline Video*
Xiangzheng Kong, Minnan Luo, Wenya Wang, Jiaying Wu, Zhi Zeng and Guang Dai
Short-form news videos increasingly shape public perception through strategic framing, yet existing verification methods largely overlook the communicative intent underlying such content. By emphasizing surface semantics, current models struggle to separate stylistic presentation from factual evidence, which leads to shortcut learning and brittle generalization. To address this limitation, we propose the Origin-Objective-Means (OOM) framework, a theory-grounded representation of communicative intent that captures creator stance, audience need activation, and communication strategy. We validate OOM through large-scale human annotation, revealing distinct and consistent lexical and structural patterns across intent dimensions. Building on this representation, we operationalize intent as an explicit semantic condition rather than a prediction target. Concretely, we introduce Intent-Guided Prompting (IGP) to condition LLM reasoning and intent-conditioned multimodal detection framework (ICMD), which injects intent into multimodal detectors via feature-wise modulation. Experiments on FakeSV and FakeTT show that modeling intent as an intermediate condition consistently improves accuracy and robustness across diverse vision-language backbones, while substantially reducing reliance on spurious stylistic correlations.

[Long] Gap-K%: Measuring Top-1 Prediction Gap for Detecting Pretraining Data ▶ *Underline Video*
Minseo Kwak and Jaehyung Kim
The opacity of massive pretraining corpora in Large Language Models (LLMs) raises significant privacy and copyright concerns, making pretraining data detection a critical challenge. Existing state-of-the-art methods typically rely on token likelihoods, yet they often overlook the gap between the target token and the model's top-1 prediction, as well as local correlations between adjacent tokens. In this work, we propose Gap-K%, a novel pretraining data detection method grounded in the optimization dynamics of LLM pretraining. By analyzing the next-token prediction objective, we observe that discrepancies between the models top-1 prediction and the target token induce strong gradient signals, which are explicitly penalized during training. Motivated by this, Gap-K% leverages the log probability gap between the top-1 predicted token and the target token, incorporating a sliding window strategy to capture local correlations and mitigate token-level fluctuations. Extensive experiments on the WikiMIA and MIMIR benchmarks demonstrate that Gap-K% achieves state-of-the-art performance, consistently outperforming prior baselines across various model sizes and input lengths.

[Long] Generating Literature-Driven Scientific Theories at Scale ▶ *Underline Video*
Peter Jansen, Peter Clark, Doug Downey and Daniel S Weld
Contemporary automated scientific discovery has focused on agents for generating scientific experiments, while systems that perform higher-level scientific activities such as theory building remain underexplored. In this work, we formulate the problem of synthesizing theories consisting of qualitative and quantitative laws from large corpora of scientific literature. We study theory generation at scale, using 13.7k source papers to synthesize 2.9k theories, examining how generation using literature-grounding versus parametric knowledge, and accuracy-focused versus novelty-focused generation objectives change theory properties. Our experiments show that, compared to using parametric LLM memory for generation, our literature-supported method creates theories that are significantly better at both matching existing evidence and at predicting future results from 4.6k subsequently-written papers.

[Long] GrAInS: Gradient-based Attribution for Inference-Time Steering of LLMs and VLMs ▶ *Underline Video*

Duy Nguyen, Archiki Prasad, Elias Stengel-Eskin and Mohit Bansal

Inference-time steering provides a lightweight alternative to fine-tuning large language models (LLMs) and vision-language models (VLMs) by modifying model activations without updating weights. However, existing methods often rely on a global intervention vector, overlook token-level causal influence, and underutilize model logits, especially in multimodal settings where visual and textual inputs contribute unevenly. We propose GrAInS, a contrastive, gradient-based approach that leverages Integrated Gradients to identify top-k influential tokens and construct directional steering vectors based on their contribution to preferred over dispreferred outputs. These vectors guide activation intervention at each layer, preserving the representational scale. GrAInS outperforms fine-tuning and prior steering methods on both LLM and VLM tasks: improving TruthfulQA accuracy by 13.22% (Llama-3.1-8B), reducing MMHal-Bench hallucinations from 0.624 to 0.514 (LLaVA-1.6-7B), and increasing SPA-VL alignment by 8.11%, all without degrading fluency or general capabilities.

[Long] HarDBench: A Benchmark for Draft-Based Co-Authoring Jailbreak Attacks for Safe Human-LLM Collaborative Writing ▶ *Underline Video*

Euntae Kim, Soomin Han and Buru Chang

Large language models (LLMs) are increasingly used as co-authors in collaborative writing, where users begin with rough drafts and rely on LLMs to complete, revise, and refine their content. However, this capability poses a serious safety risk: malicious users could jailbreak the models, filling incomplete drafts with dangerous content, to force them into generating harmful outputs. In this paper, we identify the vulnerability of current LLMs to such draft-based co-authoring jailbreak attacks and introduce HarDBench, a systematic benchmark designed to evaluate the robustness of LLMs against this emerging threat. HarDBench spans a range of high-risk domains, including Explosives, Drugs, Weapons, and Cyberattacks, and features prompts with realistic structure and domain-specific cues to assess the model susceptibility to harmful completions. To mitigate this risk, we introduce a safety-utility balanced alignment approach based on preference optimization, training models to refuse harmful completions while remaining helpful on benign drafts. Experimental results show that existing LLMs are highly vulnerable in co-authoring contexts and our alignment method significantly reduces harmful outputs without degrading performance on co-authoring capabilities. This presents a new paradigm for evaluating and aligning LLMs in human-LLM collaborative writing settings. Our new benchmark and dataset are available on our project page at https://anonymous.4open.science/r/HarDBench_data-17E4.

[Short] How Do Inpainting Artifacts Propagate to Language? ▶ *Underline Video*

Pratham Yashwante, Davit Abrahamyan, Shresth Grover and Sukruth Rao

We study how visual artifacts introduced by diffusion-based inpainting affect language generation in vision-language models. We use a two-stage diagnostic setup in which masked image regions are reconstructed and then provided to captioning models, enabling controlled comparisons between captions generated from original and reconstructed inputs. Across multiple datasets, we analyze the relationship between reconstruction fidelity and downstream caption quality. We observe consistent associations between pixel-level and perceptual reconstruction metrics and both lexical and semantic captioning performance. Additional analysis of intermediate visual representations and attention patterns shows that inpainting artifacts lead to systematic, layer-dependent changes in model behavior. Together, these results provide a practical diagnostic framework for examining how visual reconstruction quality influences language generation in multimodal systems.

[Findings] How Language Models Conflate Logical Validity with Plausibility: A Representational Analysis of Content Effects ▶ *Underline Video*

Leonardo Bertolazzi, Sandro Pezzelle and Raffaella Bernardi

Both humans and large language models (LLMs) exhibit content effects: biases in which the plausibility of the semantic content of a reasoning problem influences judgments regarding its logical validity. While this phenomenon in humans is best explained by the dual-process theory of reasoning, the mechanisms behind content effects in LLMs remain unclear. In this work, we address this issue by investigating how LLMs encode the concepts of validity and plausibility within their internal representations. We show that both concepts are linearly represented and strongly aligned in representational geometry, leading models to conflate plausibility with validity. Using steering vectors, we demonstrate that plausibility vectors can causally bias validity judgements, and vice versa, and that the degree of alignment between these two concepts predicts the magnitude of behavioral content effects across models. Finally, we construct debiasing vectors that disentangle these concepts, reducing content effects and improving reasoning accuracy. Our findings advance understanding of how abstract logical concepts are represented in LLMs and highlight representational interventions as a path toward more logical systems.

[Findings] Incomplete Prompt Jailbreaks in Large Language Models ▶ *Underline Video*

Yeonjea Kim, Bumjin Park and Jaesik Choi

Large language models (LLMs) are increasingly released as open-weight models with safeguards against harmful requests. Nevertheless, sentence completion remains vulnerable to incomplete harmful prompts. In this work, we formalize this phenomenon as incomplete prompt jailbreaks (IPJ) and provide a systematic empirical characterization of when and how incomplete prompts elicit harmful continuations. We analyze diverse attractor types associated with incomplete sentence continuation and show that LLMs systematically delay refusal until sentence termination. We further demonstrate that training models to refuse incomplete harmful prompts via parameter tuning is insufficient, failing to generalize across both content domains and attractor types. To enable fine-grained control, we identify two functional neurons: termination and continuation neurons. By clarifying their roles in sentence completion, we highlight the potential of neuron-level

interventions for more precise and robust IPJ defenses.

[Long] Inferring Events from Time Series using Language Models ▶ *Underline Video*

Mingtian Tan, Mike A Merrill, Zachary Gottesman, Tim Althoff, David Evans and Thomas Hartvigsen

A common goal in analyzing time series data is to understand how events cause observed variations. We study whether Large Language Models (LLMs) can infer natural language events associated with time series data. We introduce an automated method for generating tasks that test a model's ability to reason about events associated with time series data based on sports data, and develop a new benchmarking method. In experiments spanning 18 LLMs, we prompt LLMs to infer unobserved events given time series data and observe surprising successes, even when providing minimal context. We then show that combining distillation with Reinforcement Learning (RL) can improve the performance for small language models to approach that of large proprietary reasoning models. All resources needed to reproduce our work are available: https://github.com/hartvigsen-group/GAMETime.

[Findings] Infusing Theory of Mind into Socially Intelligent LLM Agents ▶ *Underline Video*

EunJeong Hwang, Yuwei Yin, Giuseppe Carenini, Peter West and Vered Shwartz

Theory of Mind (ToM), an understanding of the mental states of others, is a key aspect of human social intelligence, yet, chatbots and LLM-based social agents do not typically integrate it. In this work, we demonstrate that LLMs that explicitly use ToM get better at dialogue, achieving goals more effectively. After showing that simply prompting models to generate mental states between dialogue turns already provides significant benefit, we further introduce ToMAgent (ToMA), a ToM-focused dialogue agent. ToMA is trained by pairing ToM with dialogue lookahead to produce mental states that are maximally useful for achieving dialogue goals. Experiments on the Sotopia interactive social evaluation benchmark demonstrate the effectiveness of our method over a range of baselines. Extensive analysis shows that ToMA exhibits more strategic, goal-oriented reasoning behaviors, which enable long-horizon adaptation, while maintaining better relationships with their partners. Our results suggest a step forward in integrating ToM for building socially intelligent LLM agents.

[Findings] Interpretability from the Ground Up: Stakeholder-Centric Design of Automated Scoring in Educational Assessments ▶ *Underline Video*

Yunsung Kim, Michael Hardy, Joseph Tey, Candace Thille and Christopher J Piech

AI-driven automated scoring systems offer scalable and efficient means of evaluating complex student-generated responses. Yet, despite increasing demand for transparency and interpretability, the field has yet to develop a widely accepted solution for interpretable automated scoring to be used in large-scale real-world assessments. This work takes a principled approach to address this challenge. We analyze the needs and potential benefits of interpretable automated scoring for various assessment stakeholder groups and develop four principles of interpretability – (F)aithfulness, (G)roundedness, (T)raceability, and (I)nterchangeability (FGTI) – targeted at those needs. To illustrate the feasibility of implementing these principles, we develop the AnalyticScore framework as a baseline reference framework. When applied to the domain of text-based constructed-response scoring, AnalyticScore outperforms many uninterpretable scoring methods in terms of scoring accuracy and is, on average, within 0.06 QWK of the uninterpretable SOTA across 10 items from the ASAP-SAS dataset. By comparing against human annotators conducting the same featurization task, we further demonstrate that the featurization behavior of AnalyticScore aligns well with that of humans.

[Long] IterCOMP: Reasoning-aware Adaptive Prompt Compression for Multi-hop Question Answering ▶ *Underline Video*

JungMin Yun and YoungBin Kim

Multi-hop question answering requires complex reasoning across multiple evidence segments, which often overwhelms retrieval-augmented generation systems with lengthy and noisy contexts, thereby undermining both efficiency and accuracy. While existing prompt compression methods attempt to address this issue, they are typically designed for single-turn queries and fail to capture interdependent reasoning steps. We propose IterCOMP, a unified, training-free prompt compression framework that incorporates multi-hop reasoning within an iterative compression loop. IterCOMP decomposes documents into evidence segments, evaluates question answerability, and generates targeted follow-up questions to iteratively integrate essential evidence, producing a compact, reasoning-oriented prompt. Experiments on MusiQue, 2WikiMultiHopQA, and HotpotQA demonstrate that IterCOMP achieves substantial improvements in Exact Match and F1 scores while reducing the token budget, outperforming existing baselines and exhibiting robustness as reasoning complexity increases.

[Long] Language Models Learn Universal Representations of Numbers and Heres Why You Should Care ▶ *Underline Video*

Michal tefánik, Timothee Mickus, Marek Kadlík, Bertram Højer, Michal Spiegel, Raúl Vázquez, Aman Sinha, Josef Kucha, Philipp Mondorf and Pontus Stenetorp

Prior work has shown that large language models (LLMs) often converge to accurate input embedding for numbers, based on sinusoidal representations. In this work, we demonstrate that these representations are in fact strikingly systematic, to the point of being almost perfectly universal: different LLM families develop equivalent sinusoidal structures, and number representations are broadly interchangeable in a large swathe of experimental setups. We show that properly factoring in this characteristic is crucial when it comes to assessing how accurately LLMs encode numeric and other ordinal information, and that mechanistically enhancing this sinusoidality can also lead to reductions of LLMs' arithmetic errors.

[Findings] LCO: LLM-based Constraint Optimization for Safer Agentic LLMs in Real-world Tasks ▶ *Underline Video*

Jiayong Wan, Jiawei Chen, Zhaoxia Yin, Liu Shuyuan and Hang Su
Large Language Models (LLMs) are increasingly acting as autonomous agents, but their continuous interaction with the environment can lead to in-context reward hacking (ICRH), a phenomenon in which LLMs iteratively optimize their behavior to maximize proxy objectives, inadvertently producing harmful side effects. Existing defense methods are insufficient to address this risk, as ICRH arises not from adversarial inputs but from the models own over-optimization. To mitigate this issue, we propose LLM-based Constraint Optimization (LCO), a framework that effectively reduces ICRH without model fine-tuning. LCO consists of two modules: self-thought module, which guides the LLM to proactively deliberate and integrate potential safety constraints before execution; and guided evolutionary exploration module, which employs LLM-based crossover and mutation to constrain the models actions within a safe solution space while maintaining task performance. Experimental results demonstrate that LCO substantially alleviates ICRH in both output-refine and policy-refine scenarios. In particular, on the tweet engagement optimization task, LCO achieves a 39% reduction in the Toxicity Growth Rate (TGR) on GPT-4, while on the policy optimization benchmark, it reduces the ICRH Occurrence Rate by 15.23%, demonstrating safety improvement without sacrificing task performance.Our code is available at: https://github.com/Califoni/LCO_for_ICRH.

[Findings] Leveraging Pretrained Language Models as Energy Functions for Glauber Dynamics Text Diffusion ▶ *Underline Video*
Tarun Kathuria and Sachin Kumar
We present a discrete diffusion-based generative model for text generation using Glauber dynamics from statistical physics. Our main insight is that instead of trying to train a discrete state space diffusion model using Glauber dynamics with a uniform transition kernel as the forward process, one can set up an "energy function" based on pretrained causal/masked language models, which, when viewed as the stationary distribution, allows us to significantly improve the quality of the generated text. Using UL2 as our pretrained models and modifying and incorporating it into our diffusion pipeline, we obtain significantly better perplexities than prior diffusion-based text generative models and are competitive with the perplexities of GPT-2-medium and GPT-2-large for comparable model sizes. Furthermore, our models outperform prior diffusion models and GPT-2 style auto-regressive models on some zero-shot common sense reasoning tasks as well as some planning/search tasks.

[Findings] LLaTiSA: Towards Difficulty-Stratified Time Series Reasoning from Visual Perception to Semantics ▶ *Underline Video*
Yueyang Ding, HaoPeng Zhang, Rui Dai, Yi Wang, Tianyu Zong, Kaikui Liu and Xiangxiang Chu
Comprehensive understanding of time series remains a significant challenge for Large Language Models (LLMs). Current research is hindered by fragmented task definitions and benchmarks with inherent ambiguities, precluding rigorous evaluation and the development of unified Time Series Reasoning Models (TSRMs). To bridge this gap, we formalize Time Series Reasoning (TSR) via a four-level taxonomy of increasing cognitive complexity. We introduce HiTSR, a hierarchical time series reasoning dataset comprising 83k samples with diverse task combinations and verified Chain-of-Thought (CoT) trajectories. Leveraging HiTSR, we propose LLaTiSA, a strong TSRM that integrates visualized patterns with precision-calibrated numerical tables to enhance the temporal perception of Vision-Language Models (VLMs). Through a multi-stage curriculum fine-tuning strategy, LLaTiSA achieves superior performance and exhibits robust out-of-distribution generalization across diverse TSR tasks and real-world scenarios. We will publicly release the code, dataset, and model weights.

[Long] LLM Agents in Law: Taxonomy, Applications, and Challenges ▶ *Underline Video*
Shuang Liu, Ruijia Zhang, Ruoyun Ma, Yujia Deng, Lanyi Zhu, Jiayu Li, Zelong Li, Zhibin Shen and Mengnan Du
Large language models (LLMs) have precipitated a dramatic improvement in the legal domain, yet the deployment of standalone models faces significant limitations regarding hallucination, outdated information, and verifiability. Recently, LLM agents have attracted significant attention as a solution to these challenges, utilizing advanced capabilities such as planning, memory, and tool usage to meet the rigorous standards of legal practice. In this paper, we present a comprehensive survey of LLM agents for legal tasks, analyzing how these architectures bridge the gap between technical capabilities and domain-specific needs. Our major contributions include: (1) systematically analyzing the technical transition from standard legal LLMs to legal agents; (2) presenting a structured taxonomy of current agent applications across distinct legal practice areas; (3) discussing evaluation methodologies specifically for agentic performance in law; and (4) identifying open challenges and outlining future directions for developing robust and autonomous legal assistants.

[Findings] LLM Prompt Duel Optimizer: Efficient Label-Free Prompt Optimization ▶ *Underline Video*
Yuanchen Wu, Saurabh Verma, Justin Lee, Fangzhou Xiong, Poppy Zhang, Amel Awadelkarim, Xu Chen, Yubai Yuan and Shawndra Hill
Large language models (LLMs) are highly sensitive to prompts, but most automatic prompt optimization (APO) methods assume access to ground-truth references (e.g., labeled validation data) that are costly to obtain. We propose the Prompt Duel Optimizer (PDO), a sample-efficient framework for label-free prompt optimization based on pairwise preference feedback from an LLM judge. PDO casts prompt selection as a dueling-bandit problem and combines (i) Double Thompson Sampling to prioritize informative comparisons under a fixed judge budget, with (ii) top-performer guided mutation to expand the candidate pool while pruning weak prompts. Experiments on BIG-bench Hard (BBH) and MS MARCO show that PDO consistently identifies stronger prompts than label-free baselines, while offering favorable quality–cost trade-offs under constrained comparison budgets.

[Long] LoopTool: Closing the Data-Training Loop for Robust LLM Tool Calls ▶ *Underline Video*
Kangning Zhang, Weiwen Liu, Wenxiang Jiao, Kounianhua Du, Yuan Lu, Weinan Zhang and Yong Yu
Augmenting Large Language Models (LLMs) with external tools enables them to execute complex, multi-step tasks. However, tool learning

is hampered by the static synthetic data pipelines, where data generation and model training are executed as two separate, non-interactive processes. This approach fails to focus on the model's specific weaknesses adaptively and allows noisy labels to persist, degrading training efficiency. We introduce LoopTool, a fully automated, model-aware data evolution framework that closes this loop by tightly integrating data synthesis and model training. LoopTool iteratively evolves both the data and the model through three synergistic modules: (1) Greedy Capability Probing (GCP) diagnoses the model's mastered and failed capabilities; (2) Judgement-Guided Label Verification (JGLV) uses an open-source judge model to find and correct annotation errors, progressively purifying the dataset; and (3) Error-Driven Data Expansion (EDDE) generates new, challenging samples based on identified failures. This closed-loop process is tightly integrated with reinforcement learning training and operates within a cost-efficient, open-source ecosystem, thereby eliminating reliance on costly APIs. Experiments show that LoopTool-8B significantly surpasses its 32B data generator and achieves new state-of-the-art results on the BFCL-v3 and ACEBench benchmarks for its scale. Our work demonstrates that closed-loop, self-refining data pipelines can dramatically enhance the tool-use capabilities of LLMs.

[Findings] Machine Behavior in Relational Moral Dilemmas: Moral Rightness, Predicted Human Behavior, and Model Decisions ▶ *Underline Video*

Jiseon Kim, Jea Kwon, Luiz Felipe Vecchietti, Wenchao Dong, Jaehong Kim and Meeyoung Cha

Human moral judgment is context-dependent and changes based on interpersonal relationships. As large language models (LLMs) increasingly serve as decision-support systems, it is critical to understand if they encode these social nuances. We characterize LLM behavior using the Whistleblowers Dilemma, systematically varying two experimental factors: crime severity and relational closeness. Our study compares three evaluative perspectives: (1) moral rightness (general prescriptive norms), (2) predictive human behavior (how models expect people to navigate social situations), and (3) models own decision-making. By analyzing the reasoning processes, we find a clear cross-perspective divergence: moral rightness remains consistently fairness-oriented, while predicted human behavior shifts with relational context toward loyalty. Crucially, the model decisions mirror moral rightness judgments, rather than their behavioral predictions. This cross-perspective inconsistency suggests that LLM decision-making favors abstract rules over the social sensitivity found in their internal modeling, potentially producing conflicting expectations in real-world deployments.

[Findings] Many-Shot Scaling of In-Context Learning with Self-Generated Demonstrations ▶ *Underline Video*

Zhengyao Gu, Henry Peng Zou, Yankai Chen, Aiwei Liu, Weizhi Zhang and Philip S. Yu

The high cost of obtaining high-quality annotated data for in-context learning (ICL) has motivated the development of methods that use self-generated annotations in place of ground truth labels. While these approaches have shown promising results in few-shot settings, they generally do not scale to many-shot scenarios. In this work, we study ICL with self-generated examples using a framework analogous to traditional semi-supervised learning, consisting of annotation generation, demonstration selection, and in-context inference. Within this framework, we propose a simple baseline that outperforms ground truth ICL under zero-shot, few-shot, and many-shot settings. Notably, we observe consistent scaling behaviors with respect to the number of self-annotated demonstrations. To further extract performance from this many-shot capability, we introduce IterPSD, an iterative self-annotation approach that integrates iterative refinement and curriculum pseudo-labeling techniques from semi-supervised learning, yielding up to 6.8% additional gains on classification tasks. Motivated by our baseline and IterPSD results, we demonstrate that semi-supervised ICL offers a promising avenue for future ICL research.

[Long] Markovian Linguistic-Temporal Bridge: Unlocking the Potential of LLMs for Time Series Forecasting ▶ *Underline Video*

Siming Sun, Kai Zhang, Xuejun Jiang, Wenchao Meng and Qinmin Yang

Adapting pretrained Large Language Models (LLMs) for time series forecasting primarily relies on token-level linguistic-temporal alignment, leading to the stacking of logically disjointed tokens as input. While empirically effective, these methods overlook a fundamental capability of LLMs: modeling linguistic logic and structure, rather than merely processing token features. To address this limitation, we propose the Markovian-Guided Structure-Aware Alignment (MGSAA). Our core contribution is a framework that transcends pointwise feature matching to achieve global structural isomorphism between the linguistic and temporal domains. Specifically, MGSAA distills latent evolutionary patterns of language within LLMs into a Markovian state transition graph, which is transferred as a structural prior to the time series domain. Under this prior, time series patches are decoded into latent states and then aligned via state-constrained cross-attention. Ultimately, MGSAA generates a token sequence topologically isomorphic to the LLM's inherent mental structure, reactivating its reasoning capabilities for forecasting. Comprehensive evaluations across multiple benchmarks demonstrate that MGSAA achieves state-of-the-art performance, providing an innovative solution for cross-modal alignment in LLM for time series forecasting. Code is available at https://github.com/sunzju/MGSAA.

[Long] MARS^2: Scaling Multi-Agent Tree Search via Reinforcement Learning for Code Generation ▶ *Underline Video*

Pengfei Li, Shijie Wang, Fangyuan Li, Yikun Fu, Kaifeng Liu, Kaiyan Zhang, Dazhi Zhang, Yuqiang Li, Biqing Qi and Bowen Zhou

Reinforcement learning (RL) paradigms have demonstrated strong performance on reasoning-intensive tasks such as code generation. However, limited trajectory diversity often leads to diminishing returns, which constrains the achievable performance ceiling. Search-enhanced RL alleviates this issue by introducing structured exploration, which remains constrained by the single-agent policy priors. Meanwhile, leveraging multiple interacting policies can acquire more diverse exploratory signals, but existing approaches are typically decoupled from structured search. We propose MARS^2 (Multi-Agent Reinforced Tree-Search Scaling), a unified RL framework in which multiple independently-optimized agents collaborate within a shared tree-structured search environment. MARS^2 models the search

tree as a learnable multi-agent interaction environment, enabling heterogeneous agents to collaboratively generate and refine candidate solutions within a shared search topology. To support effective learning, we introduce a path-level group advantage formulation based on tree-consistent reward shaping, which facilitates effective credit assignment across complex search trajectories. Experiments on code generation benchmarks show that MARS^2 consistently improves performance across diverse model combinations and training settings, demonstrating the effectiveness of coupling multi-agent collaboration with tree search for enhancing reinforcement learning. Our code is publicly available at https://github.com/TsinghuaC3I/MARTI.

[Long] MAS-Bench: A Unified Benchmark for Shortcut-Augmented Hybrid Mobile GUI Agents ► *Underline Video*

Pengxiang Zhao, Guangyi Liu, Yaozhen Liang, Weiqing He, Zhengxi Lu, WenHao Wang, Yuehao Huang, Yuxiang Chai, Zhaolu Kang, Yaxuan Guo, Hao Wang, Kexin Zhang, Liang Liu and Yong Liu

Shortcuts such as APIs and deep-links have emerged as efficient complements to flexible GUI operations, fostering a promising hybrid paradigm for MLLM-based mobile automation. However, systematic evaluation of GUI-shortcut hybrid agents remains largely underexplored. To bridge this gap, we introduce MAS-Bench, a benchmark that pioneers the evaluation of GUI-shortcut hybrid agents with a specific focus on the mobile domain. Beyond merely using predefined shortcuts, MAS-Bench assesses an agent's capability to autonomously generate shortcuts by discovering and creating reusable, low-cost workflows. It features 139 complex tasks across 11 real-world applications, a knowledge base of 88 predefined shortcuts (APIs, deep-links, RPA scripts), and 9 evaluation metrics. Experiments demonstrate that hybrid agents achieve up to 68.3% success rate and 39% greater execution efficiency than GUI-only counterparts. Furthermore, our evaluation framework effectively reveals the quality gap between predefined and agent-generated shortcuts, validating its capability to assess shortcut generation methods. MAS-Bench addresses the lack of systematic benchmarks for GUI-shortcut hybrid mobile agents, providing a foundational platform for future advancements in creating more efficient and robust intelligent agents.

[Findings] MEAV: Model Editing with Alignment Vectors for inference time LLM alignment in single and multidomain preference spectrum ► *Underline Video*

Sadat Shahriar, Zheng Qi, Nikolaos Pappas, Srikanth Doss, Kishaloy Halder, Monica Sunkara, Manuel Mager and Yassine Benajiba

Aligning Large Language Models (LLM) to address subjectivity and nuanced preference levels requires adequate flexibility and control, which can be a resource-intensive and time-consuming procedure. Existing training-time alignment methods require full re-training when a change is needed and inference-time ones typically require access to the reward model at each inference step. We introduce MEAV, an inference-time model-editing-based LLM alignment method that learns encoded representations of preference dimensions, called Alignment Vectors (AV). These representations enable dynamic adjusting of the model behavior during inference through simple linear operations. Here, we focus on three gradual response levels across three specialized domains: medical, legal, and financial, exemplifying its practical potential. This new alignment paradigm introduces adjustable preference knobs during inference, allowing users to tailor their LLM outputs while reducing the inference cost by half compared to the prompt engineering approach. Additionally, we find that AVs are transferable across different fine-tuning stages of the same model, demonstrating their flexibility. AVs also facilitate multidomain, diverse preference alignment, making the process 12x faster than the retraining approach.

[Long] Mechanisms of Prompt-Induced Hallucination in Vision-Language Models ► *Underline Video*

William Rudman, Michal Golovanevsky, Dana Arad, Yonatan Belinkov, Carsten Eickhoff, Ritambhara Singh and Kyle Mahowald

Large vision-language models (VLMs) are highly capable, yet often hallucinate by favoring textual prompts over visual evidence. We study this failure mode in a controlled object-counting setting, where the prompt overstates the number of objects in the image (e.g., asking a model to describe four waterlilies when only three are present). At low object counts, models often correct the overestimation, but as the number of objects increases, they increasingly conform to the prompt regardless of the discrepancy. Through mechanistic analysis of three VLMs, we identify a small set of attention heads whose ablation substantially reduces prompt-induced hallucinations (PIH) by at least 40% without additional training. Across models, PIH-heads mediate prompt copying in model-specific ways. We characterize these differences and show that PIH ablation increases correction toward visual evidence. Our findings offer insights into the internal mechanisms driving prompt-induced hallucinations, revealing model-specific differences in how these behaviors are implemented.

[Findings] Mechanistic Interpretability of Text-to-Image Diffusion Models via Cross-Attention Interventions ► *Underline Video*

Maisha Maliha and Dean F. Hougen

Text-to-image diffusion models achieve remarkable generation quality, yet their internal mechanisms for grounding prompt semantics into visual structure remain poorly understood. We present a novel mechanistic interpretability framework for Stable Diffusion that probes how individual prompt tokens are represented and utilized during the denoising process. Given a prompt, we record cross-attention activations throughout UNet denoising and convert them into token-level spatial grounding maps that indicate where each token contributes signal during image synthesis. To establish causal faithfulness, we perform controlled prompt interventions by removing a single word at a time while keeping the sampling seed fixed, producing counterfactual generations. To quantify mechanistic sensitivity, we introduce a head-resolved spike score based on divergence between per-head token contribution distributions before and after intervention, enabling module-wise and head-wise attribution of semantic changes. Experiments on compositional prompts and challenging relational descriptions reveal systematic patterns of token grounding, semantic drift, and head specialization across denoising timesteps. Our results provide a practical and reproducible toolkit for analyzing how diffusion models encode and apply semantic information, supporting deeper transparency in text-to-image generation.

[Long] MemBuilder: Reinforcing LLMs for Long-Term Memory Construction via Attributed Dense Rewards ▶ *Underline Video*
Zhiyu Shen, Ziming Wu, Fuming Lai, ShaobingLian ShaobingLian and Yanghui Rao
Maintaining consistency in long-term dialogues remains a fundamental challenge for LLMs, as standard retrieval mechanisms often fail to capture the temporal evolution of historical states. While memory-augmented frameworks offer a structured alternative, current systems rely on static prompting of closed-source models or suffer from ineffective training paradigms with sparse rewards. We introduce MemBuilder, a reinforcement learning framework that trains models to orchestrate multi-dimensional memory construction with attributed dense rewards. MemBuilder addresses two key challenges: (1) Sparse Trajectory-Level Rewards: we employ synthetic session-level question generation to provide dense intermediate rewards across extended trajectories; and (2) Multi-Dimensional Memory Attribution: we introduce contribution-aware gradient weighting that scales policy updates based on each component's downstream impact. Experimental results show that MemBuilder enables a 4B-parameter model to outperform state-of-the-art closed-source baselines, exhibiting strong generalization across long-term dialogue benchmarks.

[Findings] Mid-Think: Training-Free Intermediate-Budget Reasoning via Token-Level Triggers ▶ *Underline Video*
Van Yang, Shouren Wang, Debargha Ganguly, Xinpeng Li, Chaoda Song, Vikash Singh, Vipin Chaudhary and Xiaotian Han
Reasoning language models are controlled through explicit modes such as Think and No-think, yet we find that these behaviors are largely governed by a few token-level triggers rather than high-level instructions. Through attention analysis and controlled prompting experiments, we show that a leading "Okay" token induces reasoning behavior, while the newline pattern following '</think>' suppresses it. Based on this observation, we propose Mid-Think, a simple training-free prompting format that combines these triggers to achieve intermediate-budget reasoning, consistently outperforming fixed-token and prompt-based baselines in terms of the accuracy–length trade-off. Furthermore, applying Mid-Think to RL training after SFT reduces training time by approximately 15% while improving final performance of Qwen3-8B on AIME from 69.8% to 72.4% and on GPQA from 58.5% to 61.1%, demonstrating its effectiveness for both inference-time control and RL-based reasoning training.

[Long] Minds Eye: A Benchmark of Visual Abstraction, Transformation and Composition for Multimodal LLMs ▶ *Underline Video*
Rohit Sinha, Aditya Sanjiv Kanade, Sai Srinivas Kancheti, Vineeth N. Balasubramanian and Tanuja Ganu
Multimodal large language models (MLLMs) have achieved impressive progress on vision language benchmarks, yet their capacity for cognitive and psychological reasoning remains largely unexplored. We introduce Minds Eye, a multiple-choice benchmark of eight visuo-cognitive tasks inspired by classic human intelligence tests and organized under a novel A-R-T taxonomy: Abstraction, Relation, and Transformation. The tasks probe core processes of fluid intelligence such as pattern induction, analogical Relation mapping, and mental Transformation. We evaluate a diverse suite of closed-source and open-source MLLMs and compare their performance with human participants. Humans achieve 80% accuracy, while top performing MLLMs remain below 50%. Error analysis reveals failures in (i) visual attention allocation, (ii) internal perceptual manipulation, (iii) over reliance on domain priors, and (iv) weak abstraction of underlying visual concepts. Our findings suggest that current MLLMs exhibit limited fluid reasoning and visuo-cognitive integration compared with human participants, highlighting the need for cognitively grounded evaluation frameworks like Minds Eye.

[Findings] Mistake Notebook Learning: Batch-Clustered Failures for Training-Free Agent Adaptation ▶ *Underline Video*
Xuanbo Su, Yingfang Zhang, Hao Luo, XiaotengLiu XiaotengLiu and Leo Huang
With the growing adoption of Large Language Model (LLM) agents in persistent, real-world roles, they naturally encounter continuous streams of tasks and inevitable failures. A key limitation, however, is their inability to systematically learn from these mistakes, forcing them to repeat identical errors in similar contexts. Unlike prior training-free methods that primarily store raw instance-level experience or focus on retrieving successful trajectories, we propose Mistake Notebook Learning (MNL), a novel memory framework that enables agents to self-curate generalizable guidance from batch-clustered failures. This mechanism allows agents to distill shared error patterns into structured "mistake notes", updating an external memory only when batch performance improves to ensure stability. To further amplify adaptability, we integrate MNL with test-time scaling, leveraging aggregated failure patterns to actively steer the search process away from known pitfalls. Experiments on mathematical reasoning, Text-to-SQL, and interactive agent benchmarks show that MNL achieves competitive performance compared to existing memory mechanisms in both effectiveness and efficiency. These findings position structured mistake abstraction as a critical lever for robust agent evolution, enabling continuous improvement without the cost of parameter updates.

[Findings] MMR-GRPO: Accelerating GRPO-Style Training through Diversity-Aware Reward Reweighting ▶ *Underline Video*
Kangda Wei and Ruihong Huang
Group Relative Policy Optimization (GRPO) has become a standard approach for training mathematical reasoning models; however, GRPO training is computationally intensive and usually takes a long time, which consumes substantial computational resources and creates barriers for academic researchers and smaller organizations with limited GPU budgets. In this paper, we propose MMR-GRPO to accelerate GRPO training and reduce the overall training time required to reach peak performance, and the approach adopts Maximal Marginal Relevance to reweigh rewards of multiple rollouts by balancing rollout quality with diversity to reduce rollout redundancy. The rationale is that redundant or similar completions, if repeatedly used to train a model, will create an exploitation trap and slow down model convergence in GRPO style reinforcement learning. Extensive evaluations across three model sizes (1.5B, 7B, 8B), three GRPO variants, and five mathematical reasoning benchmarks show that MMR-GRPO achieves comparable peak performance while requiring on average 47.9% fewer training

steps and 70.2% less wall-clock time. These gains are consistent across models, methods, and benchmarks. Our code is released at: https://github.com/WeiKangda/MMR-GRPO.

[Long] Mnemis: Dual-Route Retrieval on Hierarchical Graphs for Long-Term LLM Memory ▶ *Underline Video*

Zihao Tang, Xin Yu, Ziyu Xiao, Zengxuan Wen, Zelin Li, Jiaxi Zhou, Hualei Wang, Haohua Wang, Haizhen Huang, Weiwei Deng, Feng Sun and Qi Zhang

AI Memory, specifically how models organizes and retrieves historical messages, becomes increasingly valuable to Large Language Models (LLMs), yet existing methods (RAG and Graph-RAG) primarily retrieve memory through similarity-based mechanisms. While efficient, such System-1-style retrieval struggles with scenarios that require global reasoning or comprehensive coverage of all relevant information. In this work, We propose Mnemis, a novel memory framework that integrates System-1 similarity search with a complementary System-2 mechanism, termed Global Selection. Mnemis organizes memory into a base graph for similarity retrieval and a hierarchical graph that enables top-down, deliberate traversal over semantic hierarchies. By combining the complementary strength from both retrieval routes, Mnemis retrieves memory items that are both semantically and structurally relevant. Mnemis achieves state-of-the-art performance across all compared methods on long-term memory benchmarks, scoring 93.9 on LoCoMo and 91.6 on LongMemEval-S using GPT-4.1-mini.

[Long] Mock Worlds, Real Skills: Building Small Agentic Language Models with Synthetic Tasks, Simulated Environments, and Rubric-Based Rewards ▶ *Underline Video*

Yuanjie Lyu, Chengyu Wang, Lei Shen, Jun Huang and Tong Xu

Small LLMs often struggle to match the agentic capabilities of large, costly models. While reinforcement learning can help, progress has been limited by two structural bottlenecks: existing open-source agentic training data are narrow in task variety and easily solved; real-world APIs lack diversity and are unstable for large-scale reinforcement learning rollout processes. We address these challenges with SYNTHAGENT, a framework that jointly synthesizes diverse tool-use training data and simulates complete environments. Specifically, a strong teacher model creates novel tasks and tool ecosystems, then rewrites them into intentionally underspecified instructions. This compels agents to actively query users for missing details. When handling synthetic tasks, an LLM-based user simulator provides user-private information, while a mock tool system delivers stable tool responses. For rewards, task-level rubrics are constructed based on required subgoals, user-agent interactions, and forbidden behaviors. Across 14 challenging datasets in math, search, and tool use, models trained on our synthetic data achieve substantial gains, with small models showing performance comparable to some larger baselines in certain domains.

[Long] ModeX: Evaluator-Free Best-of-N Selection for Open-Ended Generation ▶ *Underline Video*

Hyeong Kyu Choi and Sharon Li

Selecting a single high-quality output from multiple stochastic generations remains a fundamental challenge for large language models (LLMs), particularly in open-ended tasks where no canonical answer exists. While Best-of-N and self-consistency methods show that aggregating multiple generations can improve performance, existing approaches typically rely on external evaluators, reward models, or exact string-match voting, limiting their applicability and efficiency. We propose Mode Extraction (ModeX), an evaluator-free Best-of-N selection framework that generalizes majority voting to open-ended text generation by identifying the modal output representing the dominant semantic consensus among generated texts. ModeX constructs a similarity graph over candidate generations and recursively applies spectral clustering to select a representative centroid, without requiring additional inference or auxiliary models. We further instantiate this selection principle as ModeX Decoding, a drop-in decoding scheme with early pruning for efficiency. Across open-ended tasks—including text summarization, code generation, and mathematical reasoning—our approaches consistently outperform standard single- and multi-path baselines, providing a computationally efficient, drop-in solution for robust open-ended text generation.

[Findings] MPTc-Bench: Measuring Cross-market Generative Ability of Vision-Language Models via Movie Poster Transcreation ▶ *Underline Video*

Youyuan Lin, Yuan Li, Yahan Yu, Fei Cheng, Shin'ya Nishida and Chenhui Chu

Generative vision-language models (VLMs) can edit and synthesize images, yet their ability to adapt visual assets across markets remains under-evaluated. We study cross-market image transcreation via movie posters, where localization must preserve a movie's identity while matching market-specific design preferences and multilingual typography. We introduce the Movie Poster Transcreation Benchmark (MPTc-Bench), a cross-market benchmark of 582 aligned poster examples spanning 34 target markets, and define two task variants: Surface (text-centric localization) and Deep (preference-level style adaptation). We propose a two-stage planner-editor pipeline in which an VLM planner specifies executable edits and an image editor renders them. We evaluate in a triplet setup (source, human target-market poster, model output) using information-preservation checks, LLM-as-a-judge ratings for aesthetics and target-market fit, and objective similarity signals. Across multiple planners and editors, experiments reveal substantial gaps between model outputs and human target-market posters, highlighting open challenges for market-aware generation. MPTc-Bench enables controlled, quantitative progress on cross-market image editing beyond understanding-centric benchmarks.

[Findings] Multimodal Unlearning Across Vision, Language, Video, and Audio: Survey of Methods, Datasets, and Benchmarks ▶ *Underline Video*

Nobin Sarwar, Shubhashis Roy Dipta, Zheyuan Liu and Vaidehi Patil

With the growing adoption of VLMs, DMs, LLMs, and AFMs, these multimodal foundation models can inadvertently encode sensitive,

copyrighted, biased, or unsafe cross-modal associations that originate from their training data. Retraining after deletion requests or policy updates is often impractical, and targeted forgetting remains difficult because knowledge is distributed across shared representations. Multimodal unlearning addresses this challenge by enabling selective removal across modalities while retaining overall utility. This survey offers a unified, system-oriented view of multimodal unlearning across vision, language, audio, and video, grounded in recent advances, emerging applications, and open problems. Our taxonomy enables systematic comparison across model architectures and modalities, clarifying trade-offs among deletion strength, retention, efficiency, reversibility, and robustness. This survey highlights open problems and practical considerations to support future research and deployment of multimodal unlearning.

[Findings] Muse: Towards Reproducible Long-Form Song Generation with Fine-Grained Style Control ▶ *Underline Video*

Changhao Jiang, Jiahao Chen, Zhenghao Xiang, Zhixiong Yang, Hanchen Wang, Jiabao Zhuang, Xinmeng Che, Jiajun Sun, Hui Li, Yifei Cao, Shihan Dou, Ming Zhang, Junjie Ye, Tao Ji, Tao Gui, Qi Zhang and Xuanjing Huang

Recent commercial systems such as Suno demonstrate strong capabilities in long-form song generation, while academic research remains largely non-reproducible due to the lack of publicly available training data, hindering fair comparison and progress. To this end, we release a fully open-source system for long-form song generation with fine-grained style conditioning, including a licensed synthetic dataset, training and evaluation pipelines, and Muse, an easy-to-deploy song generation model. The dataset consists of 116k fully licensed synthetic songs with automatically generated lyrics and style descriptions paired with audio synthesized by SunoV5. We train Muse via single-stage supervised finetuning of a Qwen-based language model extended with discrete audio tokens using MuCodec, without task-specific losses, auxiliary objectives, or additional architectural components. Our evaluations find that although Muse is trained with a modest data scale and model size, it achieves competitive performance on phoneme error rate, text–music style similarity, and audio aesthetic quality, while enabling controllable segment-level generation across different musical structures. All data, model weights, and training and evaluation pipelines will be publicly released, paving the way for continued progress in controllable long-form song generation research.

[Findings] NITI: Neural Plan Concretization for Incremental Execution, Bridging and Trigger Inference from Underspecified Human Policies ▶ *Underline Video*

Ayan Banerjee, Shomrik Barua Banerjee and Sandeep K.S. Gupta

Human decision-making in safety-critical domains is governed by abstract policies that intentionally omit exhaustive preconditions/ triggers and contingencies. Executing such underspecified policies reliably in open-world settings remains a fundamental challenge for large language models (LLMs). We introduce NITI, Neural Bridging for Incremental Execution and Trigger Inference from Underspecified Human Policies, a neuro-symbolic framework that treats LLMs not as autonomous planners, but as execution-time concretizers of human intent. NITI incrementally executes abstract policies via verifier-grounded interfaces, infers implicit applicability conditions, repairs execution through neural bridging when assumptions fail, and halts safely under state inconsistency. We evaluate NITI on two structurally distinct embodied domains: a new benchmark of World Cubing Championship 2x2 Rubiks Cube scrambles (n=50) and a safety-critical automated insulin dosing task. Across multiple frontier LLMs, NITI enables reliable long-horizon execution without task-specific training or search with minimal contextualization infence overhead, substantially outperforming one-shot and chain-of-thought baselines. Our results show that compositional, verifier-grounded execution is essential for safe human-AI collaboration in open-world decision-making. Code and benchmark available here - https://github.com/ImpactLabASU/ACLNITI

[Findings] Once Correct, Still Wrong: Counterfactual Hallucination in Multilingual Vision-Language Models ▶ *Underline Video*

Basel Mousi, Fahim Dalvi, Shammur Absar Chowdhury, Firoj Alam and Nadir Durrani

Vision–language models (VLMs) can achieve high accuracy while still accepting culturally plausible but visually incorrect interpretations. Existing hallucination benchmarks rarely test this failure mode, particularly outside Western contexts and English. We introduce M^2CQA, a culturally grounded multimodal benchmark built from images spanning 17 MENA countries, paired with contrastive true and counterfactual statements in English, Arabic, and its dialects. To isolate hallucination beyond raw accuracy, we propose the CounterFactual Hallucination Rate (CFHR), which measures counterfactual acceptance conditioned on correctly answering the true statement. Evaluating state-of-the-art VLMs under multiple prompting strategies, we find that CFHR rises sharply in Arabic, especially in dialects, even when true-statement accuracy remains high. Moreover, reasoning-first prompting consistently increases counterfactual hallucination, while answering before justifying improves robustness. We make the dataset publicly available for the community (https://huggingface.co/datasets/QCRI/M2CQA)).

[Long] Outcome Accuracy is Not Enough: Aligning the Reasoning Process of Reward Models ▶ *Underline Video*

Binghai Wang, Yantao Liu, Yuxuan Liu, Tianyi Tang, Shenzhi Wang, Chang Gao, Chujie Zheng, Yichang Zhang, Le Yu, Shixuan Liu, Tao Gui, Qi Zhang, Xuanjing Huang, Bowen Yu, Fei Huang and Junyang Lin

Generative Reward Models (GenRMs) and LLM-as-a-Judge exhibit deceptive alignment by producing correct judgments for incorrect reasons, as they are trained and evaluated to prioritize Outcome Accuracy, which undermines their ability to generalize during RLHF. We introduce Rationale Consistency, a fine-grained metric that quantifies the alignment between the model's reasoning process and human judgment. Our evaluation of frontier models reveals that rationale consistency effectively discriminates among state-of-the-art models and detects deceptive alignment, while outcome accuracy falls short in both respects. To mitigate this gap, we introduce a hybrid signal that combines rationale consistency with outcome accuracy for GenRM training. Our training method achieves state-of-the-art performance on RM-Bench (87.1%) and JudgeBench (82%), surpassing outcome-only baselines by an average of 5%. Using RM during RLHF, our method effectively improves performance as demonstrated on Arena Hard v2, notably yielding a 7% improvement in creative writing tasks. Further

analysis confirms that our method escapes the deceptive alignment trap, effectively reversing the decline in rationale consistency observed in outcome-only training.

[Findings] Parametric Knowledge is Not All You Need: Toward Honest Large Language Models via Retrieval of Pretraining Data ► *Underline Video*

Christopher Adrian Kusuma, Muhammad Reza Qorib and Hwee Tou Ng

Large language models (LLMs) are highly capable of answering questions, but they are often unaware of their own knowledge boundary, i.e., knowing what they know and what they don't know. As a result, they can generate factually incorrect responses on topics they do not have enough knowledge of, commonly known as hallucination. Rather than hallucinating, a language model should be more honest and respond with "I don't know" when it does not have enough knowledge about a topic. Many methods have been proposed to improve LLM honesty, but their evaluations lack robustness, as they do not take into account the knowledge that the LLM has ingested during its pretraining. In this paper, we propose a more robust evaluation benchmark dataset for LLM honesty by utilizing Pythia, a truly open LLM with publicly available pretraining data. In addition, we also propose a novel method for harnessing the pretraining data to build a more honest LLM.

[Findings] PersonaForge: Psychology-Grounded Dual-Process Architecture for Personality-Consistent Role-Playing Agents ► *Underline Video*

Jizhou Tong and Sirui Zou

Large Language Models excel at role-playing but struggle to maintain consistent personalities across extended multi-turn interactions. We propose PersonaForge, combining (1) a three-layer personality architecture grounded in psychological theory and (2) a dual-process generation mechanism inspired by cognitive science. We test two falsifiable claims: Claim 1 (Orthogonality): Psychology-grounded dimensions (Big Five + Defense Mechanisms) provide more orthogonal constraints than natural language descriptions, reducing long-dialogue drift. Claim 2 (Integration Necessity): High-dimensional personality constraints create "production interference" requiring a cognitive workspace (Inner Monologue) to resolve, removing it degrades performance below simpler baselines. Experiments on 88 characters demonstrate: (1) +19.4% personality consistency (PC) with human correlation r=0.82, (2) reduced drift over 50-turn conversations (6.3% vs. 24.8% baseline), and (3) +64.7% defense mechanism manifestation. External validation on RoleBench confirms generalization (73.2% win-rate, drift 8.4% vs. 20.4%). Selective dual-process activation achieves 96% of full-system performance with only 13.4% token overhead. Human evaluation confirms more authentic and psychologically coherent character behaviors. Code and data: https://github.com/fQwQf/PersonaForge.

[Long] Powerful Training-Free Membership Inference Against Fine-Tuned Autoregressive Language Models ► *Underline Video*

David Ili, David Stanojevi and Kostadin Cvejoski

Fine-tuned language models pose significant privacy risks, as they may memorize and expose sensitive information from their training data. Membership inference attacks (MIAs) provide a principled framework for auditing these risks, yet existing methods achieve limited detection rates, particularly at the low false-positive thresholds required for practical privacy auditing. We present EZ-MIA, a membership inference attack that exploits a key observation: memorization manifests most strongly at error positions, specifically tokens where the model predicts incorrectly yet still shows elevated probability for training examples. We introduce the Error Zone (EZ) score, which measures the directional imbalance of probability shifts at error positions relative to a pretrained reference model. This principled statistic requires only two forward passes per query and no model training of any kind. On WikiText with GPT-2, EZ-MIA achieves 3.8x higher detection than the previous state-of-the-art under identical conditions (66.3% versus 17.5% true positive rate at 1% false positive rate), with near-perfect discrimination (AUC 0.98). At the stringent 0.1% FPR threshold critical for real-world auditing, we achieve 8x higher detection than prior work (14.0% versus 1.8%), requiring no reference model training. These gains extend to larger architectures: on AG News with Llama-2-7B, we achieve 3x higher detection (46.7% versus 15.8% TPR at 1% FPR). These results establish that privacy risks of fine-tuned language models are substantially greater than previously understood, with implications for both privacy auditing and deployment decisions. Code is available at https://github.com/JetBrains-Research/ez-mia.

[Long] PPA-Plan: Proactive Pitfall Avoidance for Reliable Planning in Long-Context LLM Reasoning ► *Underline Video*

Byeongjin Kim, Gyuwan Kim and Seo Yeon Park

Large language models struggle with reasoning over long contexts where relevant information is sparsely distributed. Although plan-and-execute frameworks mitigate this by decomposing tasks into planning and execution, their effectiveness is often limited by unreliable plan generation due to dependence on surface-level cues. Consequently, plans may be based on incorrect assumptions, and once a plan is formed, identifying errors and revising it reliably becomes difficult, limiting the effectiveness of reactive refinement. To address this limitation, we propose PPA-Plan, a proactive planning strategy for long-context reasoning that focuses on preventing such failures before plan generation. PPA-Plan identifies potential logical pitfalls and false assumptions, formulates them as negative constraints, and conditions plan generation on explicitly avoiding these constraints. Experiments on long-context QA benchmarks show that executing plans generated by PPA-Plan consistently outperforms existing plan-and-execute methods and direct prompting.

[Long] ProActor: Timing-Aware Reinforcement Learning for Proactive Task Scheduling Agents ► *Underline Video*

Lei Ding, Bin He, Chenguang Wang and Yang Liu

Proactive task-oriented agents must autonomously anticipate user needs, identify actionable opportunities, and trigger software actions at appropriate moments, fundamentally shifting from reactive systems that await explicit instructions. However, existing approaches lack

generalizable end-to-end solutions for measuring and optimizing such anticipatory behaviors. This paper introduces ProActor, a unified framework for conversational task scheduling that integrates: (1) a domain-agnostic automated annotation methodology that enables scalable proactiveness reinforcement learning (RL) by generating full opportunity time windows instead of rigid point labels, (2) systematic proactiveness metrics capturing both timing quality and reference action alignment, and (3) RL optimization using GRPO with various reward designs. Our insight is that RULER-based rewards with proactiveness rubrics are crucial for improving timing quality, and that proactiveness optimization enabled by stage-aware composite rewards is key to balancing timing quality and reference action alignment. Furthermore, we introduce ART-F, an adaptive RL framework that combines request-adaptive inference clusters with asynchronous training for better GPU utilization, enabling LoRA training of 4-bit Qwen2.5-14B-ProActor-Q4 models on 4xH200 and 8xH100 GPUs with substantial speedups. Experiments on two newly auto-annotated datasets demonstrate significant improvements in proactive timing while maintaining action consistency comparable to state-of-the-art baselines. Ablations validate the effectiveness of distinct composite reward variations.

[Findings] PROBE: PROcess-Based BEnchmark for Hallucination Detection ▶ *Underline Video*

Yu Zhang, Peter Belcak, Shizhe Diao, Yonggan Fu, Shaona Ghosh, Morteza Mardani, Eileen Margaret Peters Long, Bei Yu and Pavlo Molchanov

Hallucination detection remains a significant challenge for large language models. Existing agentic applications rely on LLMs to self-assess the factuality of their outputs using single-step "LLM-as-a-judge" prompts. However, even when equipped with ground truth information, current LLMs still fall short in detecting hallucinations, and this one-shot evaluation offers neither the transparency nor the granularity needed to diagnose where and why the detection fails. To address this gap, we introduce PROBE (Process-based Benchmark for Hallucination Detection), a comprehensive benchmark that breaks down hallucination detection into four critical steps: claim decomposition, evidence finding, evidence evaluation, and hallucination localization, and evaluates each step individually. PROBE consists of 12,000 test cases across three task types, summarization, question answering, and style transfer. Critically, we demonstrate that when hallucination detection is treated as a multi-step process, all models achieve considerably better performance. Through extensive evaluation, we show that current LLMs struggle chiefly with evidence finding, and that finetuning on our released training data substantially improves performance on this step. PROBE represents a significant step toward more transparent, diagnosable, and robust hallucination detection systems.

[Long] Probing Semantic Alignment, Lexical Invariance, and Syntactic Influence in LLM Metaphor Processing ▶ *Underline Video*

Fengying Ye, Shanshan Wang, Lidia S. Chao and Derek F. Wong

Large language models (LLMs) achieve strong performance on metaphor detection and interpretation tasks, yet it remains unclear what such behavioral success actually reveals about metaphor processing. We present a diagnostic analysis that examines the limits of behavioral evidence by probing three complementary dimensions: semantic attribute alignment, lexical invariance, and syntactic sensitivity. Using geometric probing, we assess whether model-generated interpretations align with reference semantic attributes; through context-varying substitution, we analyze the stability of lexical associations between metaphorical and literal expressions; and via controlled syntactic perturbations, we examine sensitivity in metaphor detection. Our analysis reveals that LLM-generated interpretations can exhibit semantic drift relative to reference attributes; stable lexical anchors persist across contextual conditions, potentially supporting conventional metaphors while biasing novel metaphors requiring contextual integration; and detection performance is sensitive to syntactic irregularities. These findings suggest that strong behavioral performance may reflect heterogeneous underlying signals, highlighting the need for caution when interpreting metaphor benchmarks as evidence of robust, integrated semantic understanding.

[Long] Provably Safe Offline-to-Online RL: Decoupling Learning from Data-Driven Safety Enforcement ▶ *Underline Video*

Kaitong Cai, Jusheng Zhang and Keze Wang

Hybrid offline-online reinforcement learning (O2O RL) promises both sample efficiency and robust exploration, but suffers from instability due to distribution shift between offline and online data. We introduce RLPD-GX, a framework that decouples policy optimization from safety enforcement: a reward-seeking learner explores freely, while a projection-based guardian guarantees rule-consistent execution and safe value backups. This design preserves the exploratory value of online interactions without collapsing to conservative policies. To further stabilize training, we propose dynamic curricula that gradually extend temporal horizons and anneal offline-online data mixing. We prove convergence via a contraction property of the guarded Bellman operator, and empirically show state-of-the-art performance on Atari-100k, achieving a normalized mean score of 3.02 (+45% over prior hybrid methods) with stronger safety and stability. Beyond Atari, ablations demonstrate consistent gains across safety-critical and long-horizon tasks, underscoring the generality of our design. Extensive and comprehensive results highlight decoupled safety enforcement as a simple yet principled route to robust O2O RL, suggesting a broader paradigm for reconciling exploration and safety in reinforcement learning.

[Findings] RealMem: Benchmarking LLMs in Real-World Memory-Driven Interaction ▶ *Underline Video*

Haonan Bian, Zhiyuan Yao, Sen Hu, Zishan Xu, Shaolei Zhang, Yifu Guo, Ziliang Yang, Xueran Han, Huacan Wang and Ronghao Chen

As Large Language Models (LLMs) evolve from static dialogue interfaces to autonomous general agents, effective memory is paramount to ensuring long-term consistency. However, existing benchmarks primarily focus on casual conversation or task-oriented dialogue, failing to capture "long-term project-oriented" interactions where agents must track evolving goals. To bridge this gap, we introduce RealMem, the first benchmark grounded in realistic project scenarios. RealMem comprises over 2,000 cross-session dialogues across eleven scenarios, utilizing natural user queries for evaluation. We propose a synthesis pipeline that integrates Project Foundation Construction, Multi-Agent Dialogue Generation, and Memory and Schedule Management to simulate the dynamic evolution of memory. Experiments reveal that

current memory systems face significant challenges in managing the long-term project states and dynamic context dependencies inherent in real-world projects. Our code and datasets are available at https://anonymous.4open.science/r/realmem-A1E4.

[Findings] Rectified Sparse Attention for Efficient Long-Sequence Generation ► *Underline Video*

Yutao Sun, Tianzhu Ye, Li Dong, Yuqing Xia, Jian Chen, Yizhao Gao, Shijie Cao, Jianyong Wang and Furu Wei

Efficient long-sequence generation is a critical challenge for Large Language Models. While recent sparse decoding methods improve efficiency, they suffer from KV cache misalignment, where approximation errors accumulate and degrade generation quality. In this work, we propose Rectified Sparse Attention (ReSA), a simple yet effective method that combines block-sparse attention with periodic dense rectification. By refreshing the KV cache at fixed intervals using a dense forward pass, ReSA bounds error accumulation and preserves alignment with the pretraining distribution. Experiments across math reasoning, language modeling, and retrieval tasks demonstrate that ReSA achieves near-lossless generation quality with significantly improved efficiency. Notably, ReSA delivers up to 3.77x end-to-end speedup under decoding at 256K sequence length, making it a practical solution for scalable long-context inference.

[Findings] Reducing Peak Memory Usage for Modern Multimodal Large Language Model Pipelines ► *Underline Video*

Junwan Kim and Hyunkyung Bae

Multimodal large language models (MLLMs) have recently demonstrated strong capabilities in understanding and generating responses from diverse visual inputs, including high-resolution images and long video sequences. As these models scale to richer visual representations, inference increasingly relies on storing large numbers of vision tokens in the key-value (KV) cache, making memory consumption a central bottleneck. Existing methods address this issue by identifying redundancy in vision tokens and compressing the cache, but such compression is typically applied only after all inputs are processed, resulting in high peak memory usage during the prefill stage. In this work, we show that MLLMs exhibit inherent structural regularities and representational redundancy that can be exploited to control memory growth throughout inference. Based on this insight, we propose a sequential input-compression mechanism that enforces a fixed memory budget by performing structure-aware key-value cache compression during the prefill process. This approach substantially reduces peak memory usage while maintaining generative performance with only minimal degradation, enabling more practical and memory-efficient multimodal inference.

[Long] Reinforcement Learning for Diffusion LLMs via Energy-Based Gibbs Alignment ► *Underline Video*

Yijia Fan, Jing Yang, Mingyu Liu, Kaitong Cai, Jian Wang, Keze Wang and Jusheng Zhang

Diffusion Large Language Models (dLLMs) have emerged as a promising non-autoregressive paradigm for text generation, offering parallel decoding and bidirectional context modeling. However, aligning dLLMs with reinforcement learning (RL) remains a significant challenge, as the marginal likelihood of sequences in masked diffusion is typically intractable, rendering standard policy gradient methods unstable or computationally prohibitive. In this work, we propose Diffusion-Gibbs Alignment (DGA), a novel variational framework that reformulates RL for dLLMs as a distribution matching problem. DGA bypasses the explicit computation of log-probabilities by leveraging a learned energy function to model the relative quality of samples. The optimization is decoupled into two stable steps: (1) contrastive energy ranking to capture global reward structures, and (2) weighted diffusion alignment to update the policy via importance sampling. Empirically, DGA establishes a new state-of-the-art across logical reasoning (Sudoku, Countdown), mathematical reasoning (GSM8K, Math500), and code generation (HumanEval, MBPP) benchmarks. DGA offers a novel variational perspective for dLLM alignment, achieving better performance while simultaneously enhancing training speed and memory efficiency.

[Long] ReLook: Vision-Grounded RL with a Multimodal LLM Critic for Agentic Web Coding ► *Underline Video*

Yuhang Li, Chenchen Zhang, Ruilin Lv, Ao Liu, Ken Deng, Yuanxing Zhang, Jiaheng Liu and Bo Zhou

While Large Language Models (LLMs) excel at algorithmic code generation, they struggle with front-end development, where correctness is judged on rendered pixels and interaction. We present ReLook, an agentic, vision-grounded reinforcement learning framework that empowers an agent to close a robust generate–diagnose–refine loop by invoking a multimodal LLM (MLLM) as a tool. During training, the agent employs an MLLM-in-the-loop to serve as a visual critic, evaluating code via screenshots and providing actionable feedback. Crucially, we enforce a strict zero-reward policy for invalid renders to guarantee renderability and mitigate reward hacking. To prevent behavioral collapse, we introduce Forced Optimization, a strict acceptance rule that admits only improving revisions, yielding monotonically better trajectories. At inference, we decouple the critic and run a lightweight, critic-free self-edit cycle, keeping latency comparable to base decoding while retaining most of the gains. Across three widely used benchmarks, ReLook consistently outperforms strong baselines in vision-grounded front-end code generation, highlighting the benefits of agentic perception, visual rewards, and training-inference decoupling.

[Findings] Rethinking Post-Unlearning Behavior of Large Vision-Language Models ► *Underline Video*

Minsung Kim, Nakyeong Yang and Kyomin Jung

Large Vision-Language Models (LVLMs) can recognize individuals in images and disclose sensitive personal information about them, raising critical privacy concerns. Machine unlearning aims to remove such knowledge from the model. However, existing methods rarely prescribe what the model should output in place of the forgotten content, leading to Unlearning Aftermaths: degenerate, hallucinated, or excessively refused responses. We argue that, especially for generative LVLMs, it is crucial to consider the quality and informativeness of post-unlearning responses rather than relying solely on naive suppression. To address this, we introduce a new unlearning task for LVLMs that requires models to provide privacy-preserving yet informative and visually grounded responses. We

also propose PUBG, a novel unlearning method that explicitly guides post-unlearning behavior toward a desirable output distribution. Experiments show that, while existing methods suffer from Unlearning Aftermaths despite successfully preventing privacy violations, PUBG effectively mitigates these issues, generating visually grounded and informative responses without privacy leakage for forgotten targets.

[Long] ReviewGrounder: Improving Review Substantiveness with Rubric-Guided, Tool-Integrated Agents ▶ *Underline Video*

Zhuofeng Li, Yi Lu, Dongfu Jiang, Haoxiang Zhang, Yuyang Bai, Chuan Li, Yu Wang, Shuiwang Ji, Jianwen Xie and Yu Zhang

The rapid rise in AI conference submissions has driven increasing exploration of large language models (LLMs) for peer review support. However, LLM-based reviewers often generate superficial, formulaic comments lacking substantive, evidence-grounded feedback. We attribute this to the underutilization of two key components of human reviewing: explicit rubrics and contextual grounding in existing work. To address this, we introduce ReviewBench, a benchmark evaluating review text according to paper-specific rubrics derived from official guidelines, the paper's content, and human-written reviews. We further propose ReviewGrounder, a rubric-guided, tool-integrated multi-agent framework that decomposes reviewing into drafting and grounding stages, enriching shallow drafts via targeted evidence consolidation. Experiments on ReviewBench show that ReviewGrounder, using a Phi-4-14B-based drafter and a GPT-OSS-120B-based grounding stage, consistently outperforms baselines with substantially stronger/larger backbones (e.g., GPT-4.1 and DeepSeek-R1-670B) in both alignment with human judgments and rubric-based review quality across 8 dimensions. The code is available at https://github.com/EigenTom/ReviewGrounder.

[Findings] Rewarding the Rare: Uniqueness-Aware RL for Creative Problem Solving in LLMs ▶ *Underline Video*

Zhiyuan Hu, Yucheng Wang, Yufei He, Jiaying Wu, Yilun Zhao, See-Kiong Ng, Cynthia Breazeal, Anh Tuan Luu, Hae Won Park and Bryan Hooi

Reinforcement learning (RL) has become a central paradigm for post-training large language models (LLMs), particularly for complex reasoning tasks, yet it often suffers from exploration collapse: policies prematurely concentrate on a small set of dominant reasoning patterns, improving pass@1 while limiting rollout-level diversity and gains in pass@k. We argue that this failure stems from regularizing local token behavior rather than diversity over sets of solutions. To address this, we propose Uniqueness-Aware Reinforcement Learning, a rollout-level objective that explicitly rewards correct solutions that exhibit rare high-level strategies. Our method uses an LLM-based judge to cluster rollouts for the same problem according to their high-level solution strategies, ignoring superficial variations, and reweights policy advantages inversely with cluster size. As a result, correct but novel strategies receive higher rewards than redundant ones. Across mathematics, physics, and medical reasoning benchmarks, our approach consistently improves pass@k across large sampling budgets and increases the area under the pass@k curve (AUC@K) without sacrificing pass@1, while sustaining exploration and uncovering more diverse solution strategies at scale. Code is in Software part under submission page.

[Findings] Risk-Controlled Event-Driven Cascading Updates for Knowledge Graph Consistency Restoration ▶ *Underline Video*

Bo Ni, Qinwen Ge, Haowei Fu, Ryan A. Rossi, Xiaorui Liu, Jiejun Xu and Tyler Derr

Knowledge Graphs (KGs) provide structured and interpretable representations of real-world entities and relations. While dynamic KGs attempt to capture real-time changes, they typically treat updates as independent facts. This overlooks a critical challenge: a factual, localized update can contradict and invalidate previously correct knowledge, requiring revisions beyond the localized update to maintain KG consistency. Many of these inconsistencies arise from events whose effects propagate through relational dependencies, necessitating coordinated multi-hop reasoning rather than isolated changes. To address this, we introduce a model-agnostic framework for cascading KG update identification that leverages conformal prediction to provide reliable uncertainty guarantees over the cascade as a whole, accounting for dependencies among multi-hop update candidates. Building on this foundation, we further develop a graph-based KG update scoring framework that integrates large language models (LLMs) to enrich event representations with world knowledge. Experiments on two newly constructed real-world datasets, designed to reflect scenarios where events necessitate coordinated multi-hop updates, demonstrate that our framework establishes a strong baseline while offering calibrated confidence estimates, providing an effective solution for event-driven KG consistency restoration.

[Long] RoBSA: RoPE-based Blockwise Sparse Multi-head Latent Attention ▶ *Underline Video*

Xinyu Shi, Kairong Luo, Zhen Zheng and Wenguang Chen

Large Language Models (LLMs) have rapidly advanced in recent years, scaling up in both parameter count and context length. However, as context windows extend from thousands to hundreds of thousands of tokens, attention computation becomes the dominant source of memory usage and runtime in decoding stages, severely limiting the efficiency and scalability of long-context LLMs. Sparse attention has emerged as a promising solution, reducing complexity by computing attention over only a subset of context tokens. However, the sparse attention for Multi-head Latent Attention(MLA) which is a variant of standard MHA is rarely studied. In this paper, we introduce RoPE-based Blockwise Sparse Attention (RoBSA), a method designed specifically for MLA during the decoding stage of model inference. RoBSA leverages the decoupled nature of RoPE within MLA to implement token selection in a blockwise manner. RoBSA is a lightweight, training-free, and layer-aware algorithm that can be integrated in a plug-and-play fashion. Our method significantly reduces end-to-end inference latency in the decoding stage by up to 2.55x with minimal accuracy loss compared to full attention in long-context scenarios for very large models.

[Findings] ROMA: Real-time Omni-Multimodal Assistant with Interactive Streaming Understanding ▶ *Underline Video*

Xueyun Tian, Wei Li, Bingbing Xu, Heng Dong, Yuanzhuo Wang and Huawei Shen

Recent Omni-multimodal Large Language Models show promise in unified audio, vision, and text modeling. However, streaming audio-video understanding remains challenging, as existing approaches suffer from disjointed capabilities: they typically exhibit incomplete modality support or lack autonomous proactive monitoring. To address this, we present ROMA, a real-time omni-multimodal assistant for unified reactive and proactive interaction. ROMA processes continuous inputs as synchronized multimodal units, aligning dense audio with discrete video frames to handle granularity mismatches. For online decision-making, we introduce a lightweight speak head that decouples response initiation from generation to ensure precise triggering without task conflict. We train ROMA with a curated streaming dataset and a two-stage curriculum that progressively optimizes for streaming format adaptation and proactive responsiveness. To standardize the fragmented evaluation landscape, we reorganize diverse benchmarks into a unified suite covering both proactive (alert, narration) and reactive (QA) settings. Extensive experiments across 12 benchmarks demonstrate ROMA achieves state-of-the-art performance on proactive tasks while competitive in reactive settings, validating its robustness in unified real-time omni-multimodal understanding. Code and benchmark are available https://eureka-maggie.github.io/ROMA_show/.

[Findings] SAE-FiRE: Enhancing Earnings Surprise Predictions Through Sparse Autoencoder Feature Selection ▶ *Underline Video*
Huopu Zhang, Yanguang Liu, Miao Zhang, Zirui He and Mengnan Du
Predicting earnings surprises through the analysis of earnings conference call transcripts has attracted increasing attention from the financial research community. Conference calls serve as critical communication channels between company executives, analysts, and shareholders, offering valuable forward-looking information. However, these transcripts present significant analytical challenges, typically containing over 5,000 words with substantial redundancy and industry-specific terminology that creates obstacles for language models. In this work, we propose the Sparse Autoencoder for Financial Representation Enhancement (SAE-FiRE) framework to address these limitations by extracting key information while eliminating redundancy. SAE-FiRE employs Sparse Autoencoders (SAEs) to efficiently identify patterns and filter out noises, and focusing specifically on capturing nuanced financial signals that have predictive power for earnings surprises. Experimental results indicate that the proposed method can significantly outperform comparing baselines.

[Findings] SafetyALFRED: Evaluating Safety-Conscious Planning of Vision Language Models ▶ *Underline Video*
Josue Torres-Fonseca, Naihao Deng, Yinpei Dai, Shane Storks, Yichi Zhang, Rada Mihalcea, Casey Kennington and Joyce Chai
Multimodal Large Language Models (MLLMs) are increasingly adopted as autonomous agents in interactive environments, yet their ability to proactively address safety hazards remains insufficient. We introduce SafetyALFRED, built upon the embodied agent benchmark ALFRED, augmented with six categories of real-world kitchen hazards. While existing safety evaluations focus on hazard recognition through disembodied question answering (QA) settings, we evaluate eleven state-of-the-art models from the Qwen, Gemma, and Gemini families on not only hazard recognition, but also active risk mitigation through embodied task planning. Our experimental results reveal a significant alignment gap: while models can accurately recognize hazards in QA settings, average mitigation success rates for these hazards are low in comparison. Our findings demonstrate that static evaluations through QA are insufficient for physical safety, advocating for a paradigm shift toward benchmarks that prioritize multi-step corrective actions in embodied context.

[Findings] SEA-Guard: Culturally Grounded Multilingual Safeguard for Southeast Asia ▶ *Underline Video*
Panuthep Tasawong, Jian Gang Ngui, Alham Fikri Aji, Trevor Cohn and Peerat Limkonchotiwat
Culturally aware safeguards are crucial for AI alignment in real-world settings, where safety extends beyond common sense and encompasses diverse local values, norms, and region-specific regulations. However, building large-scale, culturally grounded datasets is challenging due to limited resources and a scarcity of native annotators. Consequently, many safeguard models rely on machine translation of English datasets, often missing regional and cultural nuances. We present a novel agentic data-generation framework to scalably create authentic, region-specific safety datasets for Southeast Asia (SEA). On this foundation, we introduce the SEA-Guard family, the first multilingual safeguard models grounded in SEA cultural contexts. Evaluated across multiple benchmarks and cultural variants, SEA-Guard consistently outperforms existing safeguards at detecting regionally sensitive or harmful content while maintaining strong general safety performance.

[Findings] SEA-SafeguardBench: Culturally Grounded Safety Benchmark for Southeast Asian Languages ▶ *Underline Video*
Panuthep Tasawong, Jian Gang Ngui, Alham Fikri Aji, Trevor Cohn and Peerat Limkonchotiwat
Safeguard models help large language models (LLMs) detect and block harmful content, but most evaluations remain English-centric and overlook linguistic and cultural diversity. Existing multilingual safety benchmarks often rely on machine-translated English data, which fails to capture nuances in low-resource languages. Southeast Asian (SEA) languages are underrepresented despite the regions linguistic diversity and unique safety concerns, from culturally sensitive political speech to region-specific misinformation. Addressing these gaps requires benchmarks that are natively authored to reflect local norms and harm scenarios. We introduce SEA-SafeguardBench, the first human-verified safety benchmark for SEA, covering eight languages, 21,640 samples, across three subsets: general, in-the-wild, and content generation. The experimental results from our benchmark demonstrate that even state-of-the-art LLMs and guardrails are challenged by SEA cultural and harm scenarios and underperform when compared to English texts.

[Long] SEARL: Joint Optimization of Policy and Tool Graph Memory for Self-Evolving Agents ▶ *Underline Video*
Xinshun Feng, Xinhao Song, Lijun Li, Gongshen Liu and Jing Shao
Recent advances in Reinforcement Learning with Verifiable Rewards (RLVR) have demonstrated significant potential in single-turn reasoning tasks. With the paradigm shift toward self-evolving agentic learning, models are increasingly expected to learn from trajectories

by synthesizing tools or accumulating explicit experiences. However, prevailing methods typically rely on large-scale LLMs or multi-agent frameworks, which hinder their deployment in resource-constrained environments. The inherent sparsity of outcome-based rewards also poses a substantial challenge, as agents typically receive feedback only upon task completion. To address these limitations, we introduce a Tool-Memory based self-evolving agentic framework SEARL. Unlike approaches that directly utilize interaction experiences, our method constructs a structured experience memory that integrates planning with execution. This provides a novel form of state abstraction that facilitates the aggregation of actions within functionally analogous contexts, such as tool reuse. Consequently, agents not only extract explicit knowledge from historical data but also leverage inter-trajectory correlations to densify reward signals. We evaluate our framework on knowledge reasoning and complex search tasks, demonstrating its effectiveness in achieving more practical and efficient agentic learning.

[Long] SecureGate: Learning When to Reveal PII Safely via Token-Gated Dual-Adapters for Federated LLMs ▶ *Underline Video*

Mohamed Shaaban and Mohamed Elmahallawy

Federated learning (FL) enables collaborative training across organizational silos without sharing raw data, making it attractive for privacy-sensitive applications. With the rapid adoption of large language models (LLMs), federated fine-tuning of generative LLMs has gained attention as a way to leverage distributed data while preserving confidentiality. However, this setting introduces fundamental challenges: (i) privacy leakage of personally identifiable information (PII) due to LLM memorization, and (ii) a persistent tension between global generalization and local utility under heterogeneous data. Existing defenses, such as data sanitization and differential privacy, reduce leakage but often degrade downstream performance. We propose SecureGate, a privacy-aware federated fine-tuning framework for LLMs that provides fine-grained privacy control without sacrificing utility. SecureGate employs a dual-adapter LoRA architecture: a secure adapter that learns sanitized, globally shareable representations, and a revealing adapter that captures sensitive, organization-specific knowledge. A token-controlled gating module selectively activates these adapters at inference time, enabling controlled information disclosure without retraining. Extensive experiments across multiple LLMs and real-world datasets show that SecureGate improves task utility while substantially reducing PII leakage, achieving up to a 31.66x reduction in inference attack accuracy and a 17.07x reduction in extraction recall for unauthorized requests. Additionally, it maintains 100% routing reliability to the correct adapter and incurs only minimal computational and communication overhead. Code is available at https://github.com/wsu-cyber-security-lab-ai/SecureGate.

[Long] Segment, Embed, and Align: A Universal Recipe for Aligning Subtitles to Signing ▶ *Underline Video*

Zifan Jiang, Youngjoon Jang, Liliane Momeni, Gül Varol, Sarah Ebling and Andrew Zisserman

The goal of this work is to develop a universal approach for aligning subtitles (i.e., spoken language text with corresponding timestamps) to continuous sign language videos. Prior approaches typically rely on end-to-end training tied to a specific language or dataset, which limits their generality. In contrast, our method Segment, Embed, and Align (SEA) provides a single framework that works across multiple languages and domains. SEA leverages two pretrained models: the first to segment a video sequence into individual signs and the second to embed each sign video clip into a shared latent space with text. Alignment is subsequently performed with a lightweight dynamic programming procedure that runs efficiently on CPU within a minute, even for hour-long episodes. SEA is flexible and can adapt to a wide range of scenarios, utilizing resources from small lexicons to large continuous corpora. Experiments on four sign language datasets demonstrate state-of-the-art alignment performance, highlighting the potential of SEA to generate high-quality parallel data for advancing sign language processing.

[Findings] Self-Sum: Teaching an Agent to Decide Itself When and What to Summarize ▶ *Underline Video*

Hongru WANG, Rui Wang, Jushi Kai, Boyang XUE, Yongqi Li, Shijue Huang, Xiaoteng Ma, Jeff Z. Pan and Amos Storkey

Long-horizon agents operate over extended sequences of reasoning and actions, but this inevitably accumulates context noise, resulting in excessive computational cost and information overload. Existing approaches commonly rely on fixed, rule-based summarization strategies (e.g., summarizing every few steps), which are inflexible, lack generalization, and often introduce irreversible information loss. We propose Self-Sum, a framework that empowers agents to autonomously decide when and what to summarize by modeling summarization as a first-class internal cognitive action, unified with external environmental actions within a multi-turn decision-making process. Specifically, we introduce a two-stage training recipe consisting of (i) a cold-start supervised fine-tuning stage that bootstraps summarization behavior, and (ii) a lightweight, summarization-aware reinforcement learning stage that refines summarization timing and content while discouraging unnecessary summaries. Experiments on multiple long-horizon benchmarks show that Self-Sum consistently outperforms no-summarization and rule-based baselines, with particularly strong gains in generalization. Analysis further reveals that Self-Sum learns to summarize sparsely at meaningful moments and preserves task-relevant information, highlighting the importance of jointly learning when and what to summarize for robust long-horizon agent behavior.

[Long] SelFusion: Self-distillation for Diffusion Language Models ▶ *Underline Video*

Hyeongsoo Lim, Jin Young Kim, Eun Seo Seo, Min Ho Jang and Ji Won Yoon

Diffusion language models (DLMs) alleviate the inherent latency bottleneck of autoregressive (AR) large language models (LLMs), but their degraded generation quality limits practical applicability. Although knowledge distillation (KD) can be a promising direction for improving performance, we empirically find that naively applying conventional KD yields only marginal gains, or even degrades generation quality. Based on these observations, we propose a novel self-distillation framework for DLMs, namely SelFusion. To enable effective KD without an external teacher model, SelFusion performs two forward passes with different masking levels, defining the hard mode with a larger masking probability and the easy mode with a smaller masking probability. However, the easy mode is not always more accurate than the hard

mode and can be overconfident on incorrect tokens. Thus, we introduce bidirectional KD between the two modes, which can dynamically determine the distillation direction based on token-level correctness. Experimental results on instruction-following tasks show that the proposed self-distillation substantially outperforms other KD methods with external LLM and DLM teachers. In many configurations, the student trained with SelFusion even surpasses the performance of the LLM teacher, providing a practical path toward improving DLM generation quality. Source code can be found at https://github.com/scai-research/SelFusion_official

[Findings] Sentipolis: Emotion-Aware Agents for Social Simulations ▶ *Underline Video*
Chiyuan Fu, Lyuhao Chen, Yunze Xiao, Weihao Xuan, Carlos Busso and Mona T. Diab
LLM agents are increasingly used for social simulation, yet emotion is often treated as a transient cue, causing emotional amnesia and weak long-horizon continuity. We present Sentipolis, a framework for emotionally stateful agents that integrates continuous Pleasure-Arousal-Dominance (PAD) representation, dual-speed emotion dynamics, and emotion–memory coupling. Across thousands of interactions over multiple base models and evaluators, Sentipolis improves emotionally grounded behavior, boosting communication, and emotional continuity. Gains are model-dependent: believability increases for higher-capacity models but can drop for smaller ones, and emotion-awareness can mildly reduce adherence to social norms, reflecting a human-like tension between emotion-driven behavior and rule compliance in social simulation. Network-level diagnostics show reciprocal, moderately clustered, and temporally stable relationship structures, supporting the study of cumulative social dynamics such as alliance formation and gradual relationship change.

[Findings] SGG-R^3: From Next-Token Prediction to End-to-End Unbiased Scene Graph Generation ▶ *Underline Video*
Jiaye Feng, Qixiang Yin, Yuankun Liu, Tong Mo and Weiping Li;
Scene Graph Generation (SGG) structures visual scenes as graphs of objects and their relations. While Multimodal Large Language Models (MLLMs) have advanced end-to-end SGG, current methods are hindered by both a lack of task-specific structured reasoning and the challenges of sparse, long-tailed relation distributions, resulting in incomplete scene graphs characterized by low recall and biased predictions. To address these issues, we introduce SGG-R^3, a structured reasoning framework that integrates task-specific Chain-of-Thought (CoT)-guided supervised fine-tuning (SFT) and reinforcement learning (RL) with group sequence policy optimization (GSPO), designed to engage in three sequential stages to achieve end-to-end unbiased scene graph generation. During the SFT phase, we propose a relation augmentation strategy by leveraging an MLLM and refined via embedding similarity filtering to alleviate relation sparsity. Subsequently, a stage-aligned reward scheme optimizes the procedural reasoning during RL. Specifically, we propose a novel dual-granularity reward which integrates fine-grained and coarse-grained relation rewards, simultaneously mitigating the long-tail issue via frequency-based adaptive weighting of predicates and improving relation coverage through semantic clustering. Experiments on two benchmarks show that SGG-R^3 achieves superior performance compared to existing methods, demonstrating the effectiveness and generalization of the framework.

[Long] SlideAgent: Hierarchical Agentic Framework for Multi-Page Visual Document Understanding ▶ *Underline Video*
Yiqiao Jin, Rachneet Kaur, Zhen Zeng, Sumitra Ganesh and Srijan Kumar
Multi-page visual documents such as manuals, brochures, presentations, and posters convey key information through layout, colors, icons, and cross-slide references. While multimodal large language models (MLLMs) offer opportunities in document understanding, current systems struggle with complex, multi-page visual documents, particularly in fine-grained reasoning over elements and pages. We introduce SlideAgent, a versatile agentic framework for understanding multi-modal, multi-page, and multi-layout documents, especially slide decks. SlideAgent employs specialized agents and decomposes reasoning into three specialized levels–global, page, and element–to construct a structured, query-agnostic representation that captures both overarching themes and detailed visual or textual cues. During inference, SlideAgent selectively activates specialized agents for multi-level reasoning and integrates their outputs into coherent, context-aware answers.Extensive experiments show that SlideAgent significantly improves accuracy over both proprietary (+7.9%) and open-source models (+9.8%).

[Long] SOAR: Supervision from Observation for Agentic Reinforcement Learning ▶ *Underline Video*
Meng Li, Lei Li, Xiting Wang, Yi Yuan, Zheng Wei, Jiang Bian and Zang Li
Agentic reinforcement learning enables large language models to solve long-horizon tasks by interacting with the environment and internalizing tool-use behavior into their reasoning. Prior work assigns supervision primarily based on outcome rewards or external reward models, but largely ignores environment observations, a critical source of learning. Consequently, agents may identify successful actions without understanding how the environment responds, producing suboptimal policies. To address this, we propose SOAR (Supervision from Observation for Agentic Reinforcement Learning), which assigns positive advantages to observation tokens proportional to the negative entropy of preceding actions. This encourages the agent to learn from outcomes of confident actions, grounding policy updates in environment dynamics and improving anticipation of tool-call consequences. Empirical results across three domains and 14 benchmarks show that SOAR improves performance, yielding gains of up to 7.0% on general reasoning tasks and 16.9% on deep research tasks, while reducing erroneous and inefficient tool usage.

[Findings] SolidCoder: Bridging the Mental-Reality Gap in LLM Code Generation through Concrete Execution ▶ *Underline Video*
Woojin Lee and Jin-Xia Huang
State-of-the-art code generation frameworks rely on mental simulation, where LLMs internally trace execution to verify correctness. We expose a fundamental limitation: the Mental-Reality Gap—where models hallucinate execution traces and confidently validate buggy code.

This gap manifests along two orthogonal dimensions: the Specification Gap (overlooking edge cases during planning) and the Verification Gap (hallucinating correct behavior for flawed code). We propose SolidCoder with a simple principle: don't imagine—execute. The S.O.L.I.D. architecture addresses both dimensions by forcing edge-case awareness before algorithm design and replacing imagined traces with sandboxed execution using property-based oracles. With GPT-4o, SolidCoder achieves state-of-the-art pass@1 performance: 95.7% on HumanEval (+0.6%p), 77.0% on CodeContests (+4.3%p), and 26.7% on APPS (+3.4%p). Ablation reveals that edge-case awareness provides the largest individual gain, while execution grounding catches categorically different errors that specification improvements cannot address. These gains generalize to RL post-trained models, validating that bridging both gap dimensions is essential for robust code synthesis. We release our code and framework to facilitate future research.

[Findings] Spotlight and Shadow: Attention-Guided Dual-Anchor Introspective Decoding for MLLM Hallucination Mitigation ▶ *Underline Video*

Yebo Wu, Han Jin, Zhijiang Guo and Li Li

Multimodal Large Language Models (MLLMs) have demonstrated remarkable reasoning capabilities yet continue to suffer from hallucination, where generated text contradicts visual content. In this paper, we introduce Dual-Anchor Introspective Decoding (DaID), a novel contrastive decoding framework that dynamically calibrates each token generation by mining the model's internal perceptual discrepancies. Specifically, DaID identifies a Spotlight layer to amplify visual factual signals and a Shadow layer to suppress textual inertia. By leveraging visual attention distributions to guide this dual-anchor selection process, our method ensures precise, token-specific adaptation. Experimental results across multiple benchmarks and MLLMs demonstrate that DaID significantly mitigates hallucination while enhancing general reasoning capabilities.

[Findings] StateX: Enhancing RNN Recall via Post-training State Expansion ▶ *Underline Video*

Xingyu Shen, Yingfa Chen, Zhen_Leng Thai, Xu Han, Zhiyuan Liu and Maosong Sun

Recurrent neural networks (RNNs), such as linear attention and state-space models, have gained popularity due to their constant per-token complexity when processing long contexts. However, these recurrent models struggle with tasks that require accurate recall of contextual information from long contexts, because all contextual information is compressed into a fixed-size recurrent state. Previous studies have shown that recall ability is positively correlated with the recurrent state size, yet directly training RNNs with large recurrent states results in high training costs. In this paper, we introduce StateX, a post-training framework that efficiently expands the states of pre-trained RNNs. For two popular classes of RNNs, linear attention and state-space models, we design post-training architectural modifications in StateX, to scale up the state size with no or negligible increase in model parameters. Experiments on models with up to 1.3B parameters demonstrate that StateX efficiently enhances the recall and in-context learning performance of RNNs without incurring high post-training costs or compromising other capabilities.

[Long] StepHint: Multi-level Stepwise Hints Enhance Reinforcement Learning to Reason ▶ *Underline Video*

Kaiyi Zhang, Ang Lv, Jinpeng Li, Yongbo Wang, Feng Wang, Haoyuan Hu and Rui Yan

Reinforcement learning with verifiable rewards (RLVR) is a promising approach for improving the complex reasoning abilities of large language models (LLMs). However, current RLVR methods face two significant challenges: the near-miss reward problem, where a small mistake can invalidate an otherwise correct reasoning process, greatly hindering training efficiency; and exploration stagnation, where models tend to focus on solutions within their "comfort zone", lacking the motivation to explore potentially more effective alternatives. To address these challenges, we propose StepHint, a novel RLVR algorithm that utilizes multi-level stepwise hints to help models explore the solution space more effectively. StepHint partitions valid reasoning chains into reasoning steps using our proposed adaptive partitioning method. The initial few steps are used as hints, and simultaneously, multiple-level hints (each comprising a different number of steps) are provided to the model. This approach directs the model's exploration toward a promising solution subspace while preserving its flexibility for independent exploration. By providing hints, StepHint mitigates the near-miss reward problem, thereby improving training efficiency. Additionally, the external reasoning pathways help the model develop better reasoning abilities, enabling it to move beyond its "comfort zone" and mitigate exploration stagnation. StepHint outperforms competitive RLVR enhancement methods across six mathematical benchmarks and two out-of-domain benchmarks.

[Findings] Style over Story: Measuring LLM Narrative Preferences via Structured Selection ▶ *Underline Video*

Donghoon Jung, Jiwoo Choi, Songeun Chae and Seohyon Jung

We introduce a constraint-selection-based experiment design for measuring narrative preferences of Large Language Models (LLMs). This design offers an interpretable lens on LLMs' narrative selection behavior. We developed a library of 200 narratology-grounded constraints and prompted selections from six LLMs under three different instruction types: basic, quality-focused, and creativity-focused. Findings demonstrate that models consistently prioritize Style over narrative content elements like Event, Character, and Setting. Style preferences remain stable across models and instruction types, whereas content elements show cross-model divergence and instructional sensitivity. These results suggest that LLMs have latent narrative preferences, which should inform how the NLP community evaluates and deploys models in creative domains.

[Findings] Switching Heads and Softening Tokens: Turnkey Solutions to Visually Grounded Document QA ▶ *Underline Video*

Ximing Wen, Wenbo Li, Sudipta Paul, Yashas Malur Saidutta, Kalpa Gunaratna and Srinivas Chappidi

Visually Grounded Document Question Answering often lacks robust, end-to-end solutions capable of handling complex, multi-answer

queries without reliance on ad-hoc processing. In this work, we propose two turnkey LLM architectures to address this gap. We first introduce a single-head architecture where coordinates are represented as special tokens within the unified vocabulary. While structurally robust, this approach suffers from the limitations of discrete supervision; to address this, we propose a novel "softening token" method that enables differentiable Mean-Squared-Error loss over token probabilities. Although this significantly improves visual grounding, the spatial precision remains bound by discretization. Consequently, we propose a second solution: a dual-head architecture that alternates between text generation and regression-based bounding box prediction. This method offers high spatial precision via a regression head, further stabilized by our introduction of an Intersection-over-Union loss. Finally, by combining the single head model's structural robustness with the high precision of the dual head model, we propose an ensemble method that yields significant performance gains beyond each of individual components.

[Findings] Take Out Your Calculators: Estimating the Real Difficulty of Question Items with LLM Student Simulations ► *Underline Video*

Christabel Acquaye, Yi Ting Huang, Marine Carpuat and Rachel Rudinger

Standardized math assessments require expensive human pilot studies to establish the difficulty of test items. We investigate the predictive value of open-source large language models (LLMs) for evaluating the difficulty of multiple-choice math questions for real-world students. We show that, while LLMs are poor direct judges of problem difficulty, simulation-based approaches with LLMs yield promising results under the right conditions. Under the proposed approach, we simulate a "classroom" of 4th, 8th, or 12th grade students by prompting the LLM to role-play students of varying proficiency levels. We use the outcomes of these simulations to fit Item Response Theory (IRT) models, comparing learned difficulty parameters for items to their real-world difficulties, as determined by item-level statistics furnished by the National Assessment of Educational Progress (NAEP). We observe correlations as high as 0.75, 0.76, and 0.82 for grades 4, 8, and 12, respectively. In our simulations, we experiment with different "classroom sizes," showing tradeoffs between computation size and accuracy. We find that role-plays with named students improves predictions (compared to student ids), and stratifying names across gender and race further improves predictions. Our results show that LLMs with relatively weaker mathematical abilities (Gemma) actually yield better real-world difficulty predictions than mathematically stronger models (Llama and Qwen), further underscoring the suitability of open-source models for the task.

[Long] Taming System Complexity: Demystifying Software Engineering Agents in Diagnosing Linux Kernel Faults ► *Underline Video*

Zhenhao Zhou, Zhuochen Huang, Yike He, Chong Wang, Jiajun Wang, Yijian Wu, Xin Peng and Yiling Lou

The Linux kernel is a critical system, serving as the foundation for numerous systems. Bugs in the Linux kernel can cause serious consequences, affecting billions of users. Fault localization (FL), which aims at identifying the buggy code elements in software, plays an essential role in software quality assurance. While recent LLM agents have achieved promising accuracy in FL on recent benchmarks like SWE-bench, it remains unclear how well these methods perform in the Linux kernel, where FL is much more challenging due to the large-scale code base, limited observability, and diverse impact factors. In this paper, we introduce LinuxFLBench, a FL benchmark constructed from real-world Linux kernel bugs. We conduct an empirical study to assess the performance of state-of-the-art LLM agents on the Linux kernel. Our initial results reveal that existing agents struggle with this task, achieving a best top-1 accuracy of only 41.6% at file level. To address this challenge, we propose LinuxFL+, an enhancement framework designed to improve FL effectiveness of LLM agents for the Linux kernel. LinuxFL+ substantially improves the FL accuracy of all studied agents (e.g., 7.2% - 11.2% accuracy increase) with minimal costs.

[Long] TeamFusion: Supporting Open-ended Teamwork with Multi-Agent Systems ► *Underline Video*

Jiale Liu, Victor Bursztyn, Lin Ai, Haoliang Wang, Sunav Choudhary, Saayan Mitra and Qingyun Wu

In open-ended domains, teams must reconcile diverse viewpoints to produce strong deliverables. Answer aggregation approaches commonly used in closed domains are ill-suited to this setting, as they tend to suppress minority perspectives rather than resolve underlying disagreements. We present TeamFusion, a multi-agent system designed to support teamwork in open-ended domains by: 1. Instantiating a proxy agent for each team member conditioned on their expressed preferences; 2. Conducting a structured discussion to elicit agreements and disagreements; and 3. Synthesizing more consensus-oriented deliverables that feed into new iterations of discussion and synthesis. We evaluate TeamFusion on two teamwork tasks where team members can judge how well their individual views are represented in team decisions and how consensually good the final deliverables are, finding that it outperforms direct aggregation baselines across metrics, tasks, and team configurations.

[Long] Temporal Evidence Chain for Temporal Knowledge Graph Question Answering with Large Language Models ► *Underline Video*

Shihao Liu, Xiaofei Zhou, Bo Wang and Geyuan Zhang

Temporal Knowledge Graph Question Answering (TKGQA) aims to answer temporal questions using knowledge from Temporal Knowledge Graphs (TKGs). Existing LLM-based TKGQA methods typically utilize RAG-based or Agent-based paradigms, yet both struggle to construct reliable temporal evidence chains. RAG-based approaches primarily rely on semantic retrieval to fetch question-relevant contexts but overlook the structural dependencies within TKGs, leading to broken evidence chains, whereas iterative agents are prone to error propagation during multi-step reasoning. To address these limitations, we propose TECQA, a framework designed to construct temporal evidence chains for LLM reasoning. Firstly, TECQA employs structure-guided subgraph retrieval to capture structural dependencies and intermediate reasoning paths. Subsequently, it utilizes a k-nearest temporal neighbor pruning strategy to filter irrelevant noise while strictly

preserving the continuous local history surrounding critical events. Finally, the retained temporal neighbors are serialized by temporal proximity to explicitly reconstruct a coherent temporal evidence chain. Extensive experiments on MultiTQ and CronQuestions demonstrate that TECQA achieves state-of-the-art performance, outperforming strong baselines by 45.3% particularly on complex queries. Code is available at https://github.com/SimonsLiu/TECQA.

[Long] TexOCR: Advancing Document OCR Models for Compilable Page-to-LaTeX Reconstruction ▶ *Underline Video*

Chengye Wang, Lin Fu, Zexi Kuang and Yilun Zhao

Existing document OCR largely targets plain text or Markdown, discarding the structural and executable properties that make LaTeX essential for scientific publishing. We study page-level reconstruction of scientific PDFs into compilable LaTeX and introduce TexOCR-Bench, a benchmark, and TexOCR-Train, a large-scale training corpus, for this task. TexOCR-Bench features a multi-dimensional evaluation suite that jointly assesses transcription fidelity, structural faithfulness, and end-to-end compilability. Leveraging TexOCR-Train, we train a 2B-parameter model, TexOCR, using supervised fine-tuning (SFT) and reinforcement learning (RL) with verifiable rewards derived from LaTeX unit tests that directly enforce compilability and referential integrity. Experiments across 21 frontier models on TexOCR-Bench show that existing systems frequently violate key document invariants, including consistent section structure, correct float placement, and valid label-reference links, which undermines compilation reliability and downstream usability. Our analysis further reveals that RL with verifiable rewards yields consistent improvements over SFT alone, particularly on structural and compilation metrics.

[Long] Text2Tabular – Reconstructing Tabular Research Data from Scientific Publications ▶ *Underline Video*

Jonas Gottal and Florian Matthes

The increasing reliance on data-driven research underscores the need for accessible datasets, particularly in the medical domain. However, raw datasets are frequently unavailable due to privacy constraints and ethical considerations, which complicates reproducibility, meta-analyses, and large-scale data-driven research. Text2Tabular addresses this challenge by reconstructing research datasets from scientific publications using advanced natural language processing and statistical modeling. Our key contributions include: (1) a unified framework combining Large Language Model driven information extraction with copula-based distribution modeling, (2) novel integration of statistical test results as distribution constraints through constrained Markov Chain Monte Carlo refinement, and (3) an own comprehensive benchmark comprising real scientific publications with corresponding raw datasets for evaluating our literature-based data reconstruction. Evaluation on both benchmark datasets and our curated collection demonstrates strong performance in Train-on-Synthetic-Test-on-Real (TSTR) evaluations, alongside accurate replication of descriptive statistics showing that Text2Tabular preserves the statistical properties and multivariate relationships of the original datasets. Text2Tabular facilitates scientific progress by enabling immediate access to realistic, domain-specific synthetic data, thus, improving data accessibility, and mitigating data scarcity in fields with limited real-world data.

[Long] Towards Stable and Effective Reinforcement Learning for Mixture-of-Experts ▶ *Underline Video*

Di Zhang, Xun Wu, Shaohan Huang, Lingjie Jiang, Yaru Hao, Li Dong, Zewen Chi, Zhifang Sui and Furu Wei

Reinforcement learning with verifiable rewards (RLVR) has emerged as a powerful paradigm for improving reasoning capabilities. However, training RLVR with Mixture-of-Experts (MoE) policies remains fragile and is often prone to reward collapse. We identify a MoE-specific source of instability, referred to as router shift (RS), where changes in expert routing across policy updates exacerbate off-policy mismatch. This effect leads to increasingly volatile importance-ratio signals and bursty clipping behavior, which consistently precede training collapse. Motivated by this diagnosis, we propose Router-Shift Policy Optimization (RSPO). RSPO computes a per-token router-shift ratio conditioned on the previously activated experts, applies stop-gradient and a lower-bound floor, and softly rescales importance ratios prior to clipping and aggregation. This design explicitly accounts for routing-induced distributional drift during off-policy optimization. We evaluate the effect of RSPO under two settings: a synthetic countdown task and real-world reasoning tasks on MATH and Code. Across both settings, RSPO achieves better performance and exhibits greater stability compared to recent MoE-based RLVR methods.

[Long] Traffic-R1: Reinforced LLMs Bring Human-Like Reasoning to Traffic Signal Control Systems ▶ *Underline Video*

Xingchen Zou, Yuhao Yang, Zheng Chen, Xixuan Hao, Yiqi chen, Chao Huang and Yuxuan Liang

We introduce Traffic-R1, a 3B-parameter foundation model with human-like reasoning for Traffic signal control (TSC), developed via self-exploration and iterative reinforcement of LLM with expert guidance in a simulated traffic environment. Compared with traditional reinforcement learning and recent LLM-based methods, Traffic-R1 offers three main advantages: zero-shot generalization, transferring unchanged to new road networks and out-of-distribution incidents by leveraging internal traffic-control policies and reasoning; a compact 3B-parameter design that supports real-time inference on mobile-class chips for edge deployment; and an explainable TSC process that enables multi-intersection coordination through communication and an asynchronous communication network. Extensive benchmarks show Traffic-R1 outperforms strong baselines and training-intensive RL controllers. In production, the model now manages signals affecting over 55,000 drivers daily, reduces average queue lengths by more than 5%, and halves operator workload. We will open source our checkpoint and code to foster further research.

[Long] Understanding Structured Financial Data with LLMs: A Case Study on Fraud Detection ▶ *Underline Video*

Xuwei Tan, Yao Ma and Xueru Zhang

Detecting fraud in financial transactions typically relies on tabular models that demand heavy feature engineering to handle high-dimensional data and offer limited interpretability, making it difficult for humans to understand predictions. Large Language Models (LLMs), in contrast,

can produce human-readable explanations and facilitate feature analysis, potentially reducing the manual workload of fraud analysts and informing system refinements. However, they perform poorly when applied directly to tabular fraud detection due to the difficulty of reasoning over many features, the extreme class imbalance, and the absence of contextual information. To bridge this gap, we introduce FinFRE-RAG, a two-stage approach that applies importance-guided feature reduction to serialize a compact subset of numeric/categorical attributes into natural language and performs retrieval-augmented in-context learning over label-aware, instance-level exemplars. Across four public fraud datasets and three families of open-weight LLMs, FinFRE-RAG substantially improves F1/MCC over direct prompting and is competitive with strong tabular baselines in several settings. Although these LLMs still lag behind specialized classifiers, they narrow the performance gap and provide interpretable rationales, highlighting their value as assistive tools in fraud analysis.

[Long] Unveiling the Limits of Large Language Models in Inferring Pragmatic Meaning from Non-Verbal Responses ▶ *Underline Video*

Sugyeong Eo and Heuiseok Lim

Although large language models (LLMs) have shown considerable progress in pragmatic language understanding, prior research has focused mainly on their comprehension of verbal behavior. Nonetheless, non-verbal behavior remains a fundamental component of human communication, especially when deliberately utilized in isolation to convey indirect meanings. In this work, we present the first systematic evaluation of LLMs' ability to infer pragmatic meaning in dialogue consisting solely of non-verbal responses. We explore three research questions: (1) Can LLMs recognize indirect intent conveyed through non-verbal responses? (2) When and how do LLMs fail to capture non-verbal intent? (3) How can we improve LLMs' ability to interpret non-verbal intent?. Through the evaluation, we observe that LLMs struggle to infer underlying meaning from non-verbal responses, with accuracy dropping by up to 60% points compared to verbal ones. Further extensive analysis reveals a behavioral pattern in LLMs' interpretations of non-verbal behavior and demonstrates that in-context learning facilitates pragmatic inference.

[Findings] Vision-Language Models Mistake Head Orientation for Gaze Direction: Nonverbal Conversation Cues ▶ *Underline Video*

Zory Zhang, Pinyuan Feng, Bingyang Wang, Tianwei Zhao, Suyang Yu, Qingying Gao, Hokin Deng, Ziqiao Ma, Yijiang Li and Dezhi Luo

Where someone looks is a nonverbal communication cue that children and adults readily use. How well can Vision-Language Models (VLMs) infer gaze targets? To construct evaluation stimuli, we captured 1,360 real-world photos of scenes in which a person gazes at one of several objects on a table. Importantly, we also controlled the gazer's head orientation: sometimes it was directed toward the gaze target, sometimes toward a distractor object, and sometimes left unconstrained. We found a substantial performance gap between VLMs and humans, ruled out alternative explanations such as resolution and object-naming skills, and identified the main reason for the gap as VLMs inferring gaze direction using head orientation rather than eye appearance. Such a bias is likely due to data rather than architecture, as suggested by a proof-of-concept experiment finetuning a transformer-based vision model. Future work should investigate whether these findings hold broadly across various deep learning methods trained on existing data, and whether better data mitigates this problem for all architectures. Pinpointing the reason sets the stage for technologies that can interpret gaze targets to have more efficient interactions with humans.

[Findings] Why Agents Compromise Safety Under Pressure ▶ *Underline Video*

Hengle Jiang and Ke Tang

Large Language Model agents deployed in complex environments frequently encounter a conflict between maximizing goal achievement and adhering to safety constraints. This paper identifies a new concept called Agentic Pressure, which characterizes the endogenous tension emerging when compliant execution becomes infeasible. We demonstrate that under this pressure agents exhibit normative drift where they strategically sacrifice safety to preserve utility. Notably we find that advanced reasoning capabilities accelerate this decline as models construct linguistic rationalizations to justify violation. Finally, we analyze the root causes and explore preliminary mitigation strategies, such as pressure isolation, which attempts to restore alignment by decoupling decision-making from pressure signals.

[Findings] Why Did Apple Fall: Evaluating Curiosity in Large Language Models ▶ *Underline Video*

Haoyu Wang, Sihang Jiang, Yuyan Chen, Yitong Wang, Xiaojun Meng, Jiansheng Wei and Yanghua Xiao

Curiosity serves as a fundamental construct in human cognition. Inspired by curiosity, reinforcement learning with intrinsic rewards for large language models (LLMs) has shown substantial potential. However, it remains unclear whether existing curiosity-driven methods genuinely reflect curiosity-like behaviors in LLMs, and to what extent psychological notions of curiosity can be transferred to these models. In this work, we propose a psychology-inspired framework to evaluate and leverage curiosity in LLMs. We adapt the Five-Dimensional Curiosity scale Revised (5DCR) to LLMs and combine questionnaire-based self reports with behavioral study. We find that although LLMs can exhibit curiosity-like behavioral patterns resembling those of humans, such patterns do not reflect an intrinsic trait of curiosity. Building on this insight, we design a curiosity-driven thinking pipeline to examine the functional role of human-like curious behaviors. Experiments show that instructing LLMs to emulate curious strategies leads to better performance on selected downstream tasks, indicating that mimicking curious behaviors holds promise for reasoning enhancement.

[Findings] WoW-Bench: Evaluating Fine-Grained Acoustic Perception in Audio-Language Models via Marine Mammal Vocalizations ▶ *Underline Video*

Jaeyeon Kim, Heeseung Yun, Tony Woo, Chao-Han Huck Yang and Gunhee Kim

Large audio-language models (LALMs) extend language understanding into the auditory domain, yet their ability to perform low-level

listening, such as pitch and duration detection, remains underexplored. However, low-level listening is critical for real-world, out-of-distribution tasks where models must reason about unfamiliar sounds based on fine-grained acoustic cues. To address this gap, we introduce the World-of-Whale benchmark (WoW-Bench) to evaluate low-level auditory perception and cognition using marine mammal vocalizations. We use marine mammal vocalizations as out-of-distribution sound events to better assess models' low-level listening and so that the models do not rely on prior knowledge of the sound events. WoW-bench is composed of a Perception benchmark for categorizing novel sounds and a Cognition benchmark, inspired by Blooms taxonomy, to assess the abilities to remember, understand, apply, and analyze sound events. For the Cognition benchmark, we additionally introduce distractor questions to evaluate whether models are truly solving problems through listening rather than relying on other heuristics. Experiments with state-of-the-art LALMs show performance far below human levels, indicating a need for stronger auditory grounding in LALMs.

[Long] XMark: Reliable Multi-Bit Watermarking for LLM-Generated Texts ► *Underline Video*

Jiahao Xu, Rui Hu, Olivera Kotevska and Zikai Zhang

Multi-bit watermarking has emerged as a promising solution for embedding imperceptible binary messages into Large Language Model (LLM)-generated text, enabling reliable attribution and tracing of malicious usage of LLMs. Despite recent progress, existing methods still face key limitations: some become computationally infeasible for large messages, while others suffer from a poor trade-off between text quality and decoding accuracy. Moreover, the decoding accuracy of existing methods drops significantly when the number of tokens in the generated text is limited, a condition that frequently arises in practical usage. To address these challenges, we propose XMark, a novel method for encoding and decoding binary messages in LLM-generated texts. The unique design of XMark's encoder produces a less distorted logit distribution for watermarked token generation, preserving text quality, and also enables its tailored decoder to reliably recover the encoded message with limited tokens. Extensive experiments across diverse downstream tasks show that XMark significantly improves decoding accuracy while preserving the quality of watermarked text, outperforming prior methods. The code will be made publicly available upon acceptance.

[Findings] Your LLM Agents are Temporally Blind: The Misalignment Between Tool Use Decisions and Human Time Perception ► *Underline Video*

Yize Cheng, Arshia Soltani Moakhar, Chenrui Fan, Parsa Hosseini, Kazem Faghih, Zahra Sodagar, Wenxiao Wang and Soheil Feizi

Large language model (LLM) agents are increasingly used to interact with and execute tasks in dynamic environments. However, a critical yet overlooked limitation of these agents is that they, by default, assume a stationary context, failing to account for the real-world time elapsed between messages. We refer to this as "temporal blindness". This limitation hinders decisions about when to invoke tools, leading agents to either over-rely on stale context and skip needed tool calls, or under-rely on it and redundantly repeat tool calls. To study this challenge, we constructed TicToc, a diverse dataset of multi-turn user-agent message trajectories across 76 scenarios, spanning dynamic environments with high, medium, and low time sensitivity. We collected human preferences between "calling a tool" and "directly answering" on each sample, and evaluated how well LLM tool-calling decisions align with human preferences under varying amounts of elapsed time. Our analysis reveals that existing models display poor alignment with human temporal perception, with no models achieving a normalized alignment rate better than 65% when given time stamp information. We also show that naive, prompt-based alignment techniques have limited effectiveness for most models, but specific post-training alignment can be a viable way to align multi-turn LLM tool use with human temporal perception. Our data and findings provide a first step toward understanding and mitigating temporal blindness, offering insights to foster the development of more time-aware and human-aligned agents.

Poster Session G (Zoom)

11:00-12:30 (Zoom)

[Long] HoWToBench: Holistic Evaluation for LLM's Capability in Human-level Writing using Tree of Writing ► *Underline Video*

Andrew Zhuoer Feng, Cunxiang Wang, Yu Luo, Lin Fan, Irene Zhou, Zikang Wang, Xiaotao Gu, Jie Tang, Hongning Wang and Minlie Huang

Evaluating the writing capabilities of large language models (LLMs) remains a significant challenge due to the multidimensional nature of writing skills and the limitations of existing metrics. LLM's performance in thousand-words level and open-ended writing is inadequately assessed by traditional reference-based metrics or modern LLM-as-a-judge methods. We propose Tree-of-Writing (ToW), to resolve the implicit inconsistency often found when LLM-as-a-judge aggregates all sub-features in text evaluation. ToW incorporates a tree-structured workflow by explicitly modeling the aggregation weights of sub-features. We also present HowToBench, a large-scale Chinese writing benchmark encompassing 12 genres and 1302 instructions across three task categories: contextual completion, outline-guided writing, and open-ended generation. ToW successfully mitigates the biases, achieving a 0.93 Pearson correlation with human judgments. Furthermore, we detect that both overlap-based text generation metrics and popular LLM-as-a-judge practices are vulnerable to textual disturbances, while ToW is robust to them. We also uncover a negative correlation between input length and content-related scores in the Guide task, showcasing that it cannot be simply improved by input-side information piling.

14 Awards

ACL Lifetime Achievement Award

Recognizes researchers who make long, lasting contributions to the field.

Fernando Pereira

In recognition of multiple fundamental research accomplishments which span both a time range of roughly 40 years and an intellectual range from unification grammars to today's LLMs.

ACL Fellows 2025

Mohit Bansal, Peter Clark, Nizar Habash, Hanna Hajishirzi, Heng Ji, Alexander Koller, Sadao Kurohashi, Lori Levin, Rada Mihalcea, Saif Mohammad, Yue Zhang.

Test-of-Time Award

- **2001, NAACL. John Hale.**
  A Probabilistic Earley Parser as a Psycholinguistic Model
  This paper introduced the connection between language probability and cognitive load. It proposed the use of surprisal to assess the difficulty of comprehension, where less expected words following a given context require more cognitive effort. This work still has impact across psycholinguistics, cognitive science, and natural language processing. What the author proposed is still very much applicable today.
- **2016, ACL. Rico Sennrich, Barry Haddow, Alexandra Birch.**
  Neural Machine Translation of Rare Words with Subword Units
  This work addressed the open vocabulary challenge in neural machine translation by introducing subword tokenization as a simple, practical solution. It enabled neural systems to handle rare and unseen words effectively and became the default tokenization method across NLP and beyond, from NMT to pre-training frameworks like BERT, GPT, etc. Its influence has far exceeded its original scope: Byte Pair Encoding now serves as the default tokenization method for most large language models, with impact extending beyond NLP to a wide range of domains. It exemplifies how pragmatic simplicity can become foundational.

- **2016, EMNLP. Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, Percy Liang.**
  SQuAD: 100,000+ Questions for Machine Comprehension of Text

  This paper transformed how question answering is evaluated, and triggered a wide range of follow-up research into new evaluation datasets and methodologies. Its large-scale span-extraction setting established a new evaluation paradigm that shaped research for years. Nearly every major model, from RNNs and attention networks to transformers and today's LLMs, has been benchmarked against SQuAD or its successors.

- **2016, NAACL-SRW. Zeerak Waseem, Dirk Hovy.**
  Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter

  This paper introduced one of the first, if not the first, and arguably the most influential dataset for studying hate speech detection on social media. It was instrumental in turning our research towards ensuring that our online social lives were minimally impacted by identity-based hatred. Since the publication of this paper, the research in content moderation has helped develop the foundation for automated content moderation systems. The work feels even more important considering how social media has changed over the ensuing 10 years.

Computational Linguistics High-Impact Paper Award

- **2021, CL. Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, Daniel Zeman.**
  Universal Dependencies

  Universal Dependencies established a unified and cross-linguistically consistent framework for grammatical annotation, fundamentally shaping multilingual NLP research. By enabling scalable syntactic analysis across diverse languages, it catalyzed advances in cross-lingual transfer, multilingual modeling, and grammar-aware methods. Its widespread adoption as a foundational resource for multilingual research and evaluation reflects its exceptional scholarly impact and lasting influence on computational linguistics and natural language processing.

Best Paper Awards

- **The Imperfective Paradox in Large Language Models** ▶ Video
  Bolei Ma, Yusuke Miyao

  This paper makes careful and insightful use of linguistic theory to reveal substantive limitations in the language understanding and reasoning abilities of contemporary LLMs. Through the imperfective paradox, the authors show that many open 7B-9B-scale models systematically violate basic event-entailment patterns in English, behaving more like predictive narrative engines than faithful logical reasoners, even when given explicit contextual cues. The paper is especially valuable because it goes beyond documenting failure: it shows a dissociation between representation and reasoning, identifies promising effects of scale, and opens an important path for studying how logical reasoning about events can emerge or be induced in language models.

- **Memory Efficiency and Resource-Rational Encoding in Sentence Processing** ▶ Video
  Weijie Xu, Brian Dillon, Richard Futrell

  An elegant and original contribution to computational models of human language processing, this paper operationalizes working memory limitations as degraded representational quality. The paper combines clever statistical modeling with LLM-based estimation to produce non-trivial findings, including improved predictions of human reading times and the emergence of categorical encoding under memory constraints. By offering a simple, intuitive, and empirically fruitful framework for studying how memory limitations shape linguistic representations, this work has the potential to exert a lasting influence on research in psycholinguistics, cognitive science, and NLP.

- **Characterizing the Expressivity of Local Attention in Transformers** ▶ Video
  Jiaoda Li, Ryan Cotterell

  This paper provides a rigorous theoretical account of why local attention can improve transformer models, showing that it offers expressivity benefits complementary to global attention. Its strength lies in bridging theory and practice through careful analysis, formal-language experiments, consideration of modern modeling choices such as RoPE, and natural-language experiments with GPT-2, making it a valuable contribution that deserves recognition.

Best Theme Track Paper Awards

- **CoSToM: Causal-oriented Steering for Intrinsic Theory-of-Mind Alignment in Large Language Models** ▶ Video
  Mengfan Li, Xuanhua Shi, Yang Deng
  The paper uses mechanistic interpretability to analyze theory-of-mind behavior in language models, and also to intervene on and improve it. The approach is ambitious and technically interesting, and the paper presents clear results showing that interpretability-guided model modification can yield gains on ToM benchmarks.

- **Mapping the Circumplex of Affect: Geometric Analysis of Emotion Representations via Hyperspherical Contrastive Learning** ▶ Video
  Yusuke Yamauchi, Akiko Aizawa
  The paper directly investigates the important tension between human-interpretable representation structures and model performance. The study is careful and systematic, comparing objectives, architectures, and backbones for inducing circular emotion representations in language models. Its main finding, that more interpretable geometric structure can come at a cost to discriminative performance, offers a clear and useful caution for work that seeks to make LLM representations more human-readable.

Best Resource Paper Awards

- **HSCodeComp: A Realistic and Expert-level Agent Benchmark for Hierarchical Rule Application** ▶ Video
  Tian Lan, Yiqian Yang, Qianghuai Jia, Li Zhu, Hui Jiang, Hang Zhu, Weihua Luo, Longyue Wang
  Introduces a realistic, carefully validated benchmark for evaluating agents' ability to apply complex hierarchical rules in a multimodal product-coding setting. The dataset is especially valuable because it is based on real e-commerce data, preserves the noisy metadata and expert workflow of the task, and uses rigorous annotation and adjudication procedures that make the labels highly credible. The benchmark offers a rich testbed in a niche domain for structural reasoning, rule following, domain grounding, and diagnostic analysis of agent failures. These are capabilities that are increasingly central to current NLP research.

- **ImplicitMemBench: Measuring Unconscious Behavioral Adaptation in Large Language Models** ▶ Video
  Chonghan Qin, Xiachong Feng, Weitao Ma, Xiaocheng Feng, Lingpeng Kong
  A timely benchmark for studying implicit memory and behavioral adaptation in LLMs through a cognitively inspired lens. By operationalizing constructs such as procedural memory, priming, and classical conditioning, the benchmark offers a novel way to probe forms of model behavior that are not well captured by standard memorization or recall evaluations. The resource addresses an important emerging question and is likely to stimulate useful follow-up work.

- **Audio MultiChallenge: A Multi-Turn Evaluation of Spoken Dialogue Systems on Natural Human Interaction** ▶ Video
  Advait Gosai, Tyler Vuong, Utkarsh Tyagi, Steven Li, Wenjia You, Miheer Bavare, Arda Ucar, Zhongwang Fang, Brian Jang, Bing Liu, Yunzhong He
  A valuable benchmark for multi-turn spoken dialogue, addressing an important gap in the evaluation of end-to-end speech dialogue systems. Its human-in-the-loop data generation pipeline is innovative and produces diverse, realistic, and challenging interactions, with current frontier models still showing substantial headroom. The benchmark and accompanying analysis provide a meaningful foundation for future work on spoken dialogue evaluation.

- **VeriTaS: The First Dynamic Benchmark for Multimodal Automated Fact-Checking** ▶ Video
  Mark Rothermel, Marcus Kornmann, Marcus Rohrbach, Anna Rohrbach
  An ambitious and practically valuable resource for multilingual multimodal fact checking. Its dynamic benchmark design, with claims updated quarterly through an automated pipeline, goes beyond traditional static dataset construction and could make the resource especially useful for a fast-moving domain. The project is technically solid, novel, and likely to be valuable to the fact-checking community.

Best Social Impact Paper Awards

- **DIA-HARM: Dialectal Disparities in Harmful Content Detection Across 50 English Dialects** ▶ Video
  Jason S Lucas, Matt Murtagh White, Ali Al-Lawati, Uchendu Uchendu, Adaku Uchendu, Dongwon Lee
  This paper addresses the important problem of whether harmful-content and disinformation detection systems work robustly across diverse English dialects. The paper highlights a socially important gap that is under-studied.
- **Your Students Don't Use LLMs Like You Wish They Did** ▶ Video
  Sebastian Kobler, Matthew Clemson, Angela Sun, Jonathan K. Kummerfeld
  This paper merits recognition for evaluating LLM use in education in situ, moving beyond standard benchmark-based evaluation and toward a richer understanding of how these systems are actually used by students. Its careful experiments and strong theoretical framing illuminate an important gap between researchers' assumptions about educational LLM use and students' real practices.
- **Afri-MCQA: Multimodal Cultural Question Answering for African Languages** ▶ Video
  Atnafu Lambebo Tonja, Srija Anand, Emilio Villa-Cueva, Israel Abebe Azime, Jesujoba Oluwadara Alabi, Muhidin A Mohamed, Debela Desalegn Yadeta, Negasi Haile Abadi, Abigail Oppong, Nnaemeka Casmir Obiefuna, Idris Abdulmumin, Naome A Etori, Eric Peter Wairagala, Kanda Patrick Tshinu, Imanigirimbabazi Emmanuel, Gabofetswe Malema, Alham Fikri Aji, David Ifeoluwa Adelani, Thamar Solorio
  The paper introduces a large, carefully constructed dataset that enables more realistic and reliable evaluation of multiple-choice question answering across a range of African languages. Its thoughtful annotation process, strong involvement of native speakers, language coordinators, and quality-assurance procedures make the resource especially valuable and likely to become a standard reference for multilingual MCQA evaluation. The paper also provides clear empirical evidence of substantial performance gaps in current models, highlighting an important and persistent need for better language technologies for African languages.

Outstanding Paper Awards

- **MauBERT: Universal Phonetic Inductive Biases for Few-Shot Acoustic Units Discovery** ▶ Video
  Angelo Ortiz Tandazo, Manel Khentout, Youssef Benchekroun, Thomas Hueber, Emmanuel Dupoux
- **Evolutionary Guided Decoding: Iterative Value Refinement for LLMs** ▶ Video
  Zhenhua Liu, Lijun Li, Ruizhe Chen, Yuxian Jiang, Tong Zhu, Zhaochen Su, Wenliang Chen, Jing Shao
- **Beyond the Final Actor: Modeling the Dual Roles of Creator and Editor for Fine-Grained LLM-Generated Text Detection** ▶ Video
  Yang Li, Qiang Sheng, Zhengjia Wang, Yehan Yang, Danding Wang, Juan Cao
- **Lying with Truths: Open-Channel Multi-Agent Collusion for Belief Manipulation via Generative Montage** ▶ Video
  Jinwei Hu, Xinmiao Huang, Youcheng Sun, Yi Dong, Xiaowei Huang
- **Hierarchical Acoustic-Semantic Modeling: Modality Separation and Semantic Coherence for Full-Duplex SLMs** ▶ Video
  Zhenyu Liu, Xuanyu Zhang, Yunxin Li, Qixun Teng, Shenyuan Jiang, Haolan Chen, Minjun Zhao, Fanbo Meng, Yu Xu, Yancheng He, Baotian Hu, Haizhou Li, Min Zhang
- **Mind the (DH) Gap! A Contrast in Risky Choices Between Reasoning and Conversational LLMs** ▶ Video
  Luise Ge, Yongyan Zhang, Yevgeniy Vorobeychik
- **MediEval: A Unified Medical Benchmark for Patient-Contextual and Knowledge-Grounded Reasoning in LLMs** ▶ Video
  Zhan Qu, Michael Farber
- **Maximizing Local Entropy Where It Matters: Prefix-Aware Localized LLM Unlearning** ▶ Video
  Naixin Zhai, Pengyang Shao, Binbin Zheng, Yonghui Yang, Fei Shen, Long Bai, Xun Yang
- **GeoRA: Geometry-Aware Low-Rank Adaptation for RLVR** ▶ Video
  Jiaying Zhang, Lei Shi, Jiguo Li, Jun Xu, Jiuchong Gao, Jinghua Hao, Renqing He
- **CURE: Critique-Driven Unified Reinforcement Learning for Test-Time Self-Improvement** ▶ Video
  Guirong Chen, Shuqi Ye, Wenkai Yang, Shiqi Shen, Guangyao Shen, Yankai Lin

- **Systematicity between Forms and Meanings across Languages Supports Efficient Communication** ▶ Video
  Doreen Osmelak, Yang Xu, Michael Hahn, Kate McCurdy
- **Rethinking Entropy Interventions in RLVR: An Entropy Change Perspective** ▶ Video
  Zhezheng Hao, Hong Wang, Haoyang Liu, Jian Luo, Jiarui Yu, Hande Dong, Qiang Lin, Can Wang, Jiawei Chen
- **From Local to Global: Revisiting Structured Pruning Paradigms for Large Language Models** ▶ Video
  Ziyan Wang, Enmao Diao, Qi Le, Pu Wang, Minwoo Lee, Shu-ping Yeh, Evgeny Stupachenko, Hao Feng, Li Yang
- **Massively Multilingual Joint Segmentation and Glossing** ▶ Video
  Michael Ginn, Lindia Tjuatja, Enora Rice, Ali Marashian, Maria Valentini, Jasmine Xu, Graham Neubig, Alexis Palmer
- **CAR-bench: Evaluating the Consistency and Limit-Awareness of LLM Agents under Real-World Uncertainty** ▶ Video
  Johannes Kirmayr, Lukas Stappen, Elisabeth Andre
- **ViLL-E: Video LLM Embeddings for Retrieval** ▶ Video
  Rohit Gupta, Jayakrishnan Unnikrishnan, Fan Fei, Sheng Liu, Son Tran, Mubarak Shah
- **CxMP: A Linguistic Minimal-Pair Benchmark for Evaluating Constructional Understanding in Language Models** ▶ Video
  Miyu Oba, Saku Sugawara
- **CIG: Measuring Conversational Information Gain in Deliberative Dialogues with Semantic Memory Dynamics** ▶ Video
  Ming-Bin Chen, Jey Han Lau, Lea Frermann

Best Demonstration Paper Award

- **The olmOCR Project: Building Fully Open OCR using VLMs** ▶ Video
  Jake Poznanski, Kyle Lo, Luca Soldaini
  olmOCR is a mature, highly impactful, and widely-used open-source OCR system. The demo paper compared this system extensively to contemporary systems and illustrated clear trends in progress on the OCR task and cost-benefit tradeoffs of different systems.

Student Research Workshop Best Paper Awards

- **Evaluation of Multilingual Ability to Use Spatial Deictic Expressions in Vision-Language Models** ▶ Video
  Kaito Watanabe, Taisei Yamamoto, Tomoki Doi, Hitomi Yanaka
- **BanglaSocialBench: A Benchmark for Evaluating Sociopragmatic and Cultural Alignment of LLMs in Bangladeshi Social Interaction** ▶ Video
  Tanvir Ahmed Sijan, S. M Golam Rifat, Pankaj Chowdhury Partha, Md. Tanjeed Islam, Md. Musfique Anwar

Reviewing Awards & Recognitions

CL Distinguished Reviewer Award (Class of 2026): Nate Chambers, Jason Eisner, Francis Ferraro, Ivan Habernal, Xianpei Han, Sophie Hao, Eduard Hovy, Fintan Mallory, Yevgen Matusevych, Tristan Miller, Agata Savary, Dominik Schlechtweg, Edward Stabler, Zhongyu Wei.

Great Senior Area Chairs: Alberto Testoni, Ali Emami, Buzhou Tang, Chenhui Chu, Emily Prud'hommeaux, Houda Bouamor, Jing Li, Kang Liu, Karin Verspoor, Marcos Andre Goncalves, Matthieu Labeau, Nafise Sadat Moosavi, Nedjma Ousidhoum, Nikolaos Aletras, Piji Li, Sachin Kumar, Sophia Ananiadou, Suzan Verberne, Tiago Pimentel, Tristan Naumann, Yiming Cui, Yu Zhang, Yufang Hou, Yulan He, Zilong Zheng.

Great Area Chairs: Arijit Khan, Alexander Hoyle, Alona Fyshe, Alberto Barron-Cedeno, Alicia Vail Parrish, Chenglong Wang, Carolina Scarton, Christopher Brinton, Congying Xia, Dongfang Li, Elior Sulem, Federico Ruggeri, Fei Wu, Gemma Boleda, Hong Xu, Hao Zhang, Jordan Lee Boyd-Graber, Jivnesh Sandhan, Jessica Ouyang, Jean Maillard, Jiaao Chen, Jianshu Chen, Junyi Li, John Pavlopoulos, Kanishka Misra, Lukas Lange, Liang Ding, Liang Yao, Lea Frermann, Mark G. Lee, Mirella Lapata, Minh-Tien Nguyen, Maria Heuss, Maarten Sap, Marius Mosbach, Ningyu Zhang, Niloy Ganguly, Natalie Parde, Ondrej Dusek, Ping Wang, Rochelle Choenni, Ran Levy, Rui Zhang, Santosh T.Y.S.S, Shane Storks, Shira Wein, Shuohang Wang, Shirui Pan, Traian Rebedea, Tao Gui, Wei Song, Weiming Lu, Weizhen Qi, Xiachong Feng, Yongyu Mu, Yu Jianxing, Yuxuan Lai.

Great Reviewers: Andreas Ruckle, Adam Nohejl, Alex Wang, Ankita Pasad, Aditya Kalyanpur, AiTi Aw, Alex C. Williams, Alberto Cazzaniga, Alexandra DeLucia, Ali Marashian, Ayana Niwa, Anwoy Chatterjee, Alessio Cocchieri, Brihi Joshi, Benjamin Minixhofer, Chenan Wang, Changbo Wang, Chun Kit Chan, Caleb Chen Cao, Chunyu Miao, Christian Wallraven, Deqing Fu, Dennis Aumiller, David E. Losada, David Schlangen, Dawei Yin, Ewan Dunbar, Eva Vanmassenhove, Federico Ranaldi, Giuliano Martinelli, Gaurav Verma, Hans Ole Hatzel, Hongxing Wang, Hugh Mee Wong, Hao Wang, Hanning Zhang, Haolin Deng, Harry Shomer, Julie Hunter, Justin Vasselli, Jeffrey S. Sorensen, Joke Daems, Jiashu Zhao, Jungseob Lee, Julie Kallini, Johan Bos, Jiaxin Zhang, James Mayfield, Jianxin Liao, Jeonghun Baek, Jinchang Hou, Jiayi Shen, Jianshu Zhang, Kunal Dhawan, Kosuke Yamada, Kenneth Kwok, Katherine Rose Driggs-Campbell, Kilho Son, Lei Chen, Leandro Von Werra, Lindia Tjuatja, Lei Xie, Lechen Zhang, Luyao Cheng, Lei Yang, Matthew Lease, Mohammad Rashedul Hasan, Md Messal Monem Miah, Marek Kadlcik, Marine Carpuat, Malvina Nissim, Max Muller-Eberstein, Ming-Bin Chen, Maud Ehrmann, Maja Brinkmann, Max Ploner, Michael Hanna, Mi-Young Kim, Meghdut Sengupta, Md. Atabuzzaman, Nikhil Singh, Nicholas Mattei, Neha Sengupta, Nathan Kallus, Orfeas Menis Mastromichalakis, Paul Youssef, Pedro Javier Ortiz Suarez, Pooja Jhunjhunwala, Qiyu Wu, Quyu Kong, Ryo Kamoi, Remi Cardon, Rachel Rudinger, Rajakrishnan P Rajkumar, Roxanne El Baff, Sourajit Saha, Shirin Nilizadeh, Shiqi Wang, Shafiuddin Rehan Ahmed, Sabine Weber, Sony Trenous, Sunav Choudhary, Seungryong Kim, Susanna Rucker, Samuel Cahyawijaya, Somnath Banerjee, Sai Chetan Chinthakindi, Salam Albatarni, Sanne Hoeken, Shuhao Xing, Tobias Kafer, Taiki Miyagawa, Taro Miyazaki, Tianyi Hu, Varun Kotte, Wei Yao, Xiaokun Zhang, Xincan Feng, Yining Wang, Yisong Miao, Yejin Jeon, Yang Gao, Yannick Assogba, Youngsoo Jang, Yinan Yu, Yusuke Miyao, Youwei Pang, Yiyang Feng, Yong-Yeol Ahn, Yun Shen, Zhepei Wei, Ziyue Wang, Zhenbo Xu.

15 Local Guide

Venue

Venue: Manchester Grand Hyatt San Diego (1 Market Place, San Diego, CA 92101)
Front Desk / On-site Help: (619) 232-1234

Explore Southern California's vibrant culture and natural beauty from our waterfront hotel in downtown San Diego. Unwind in spacious hotel rooms and family suites featuring stunning views and appointed with upscale amenities. Located steps from the city's top attractions, like Petco Park and the San Diego Convention Center, our guests enjoy easy access to all things San Diego, including the region's best beaches, the local restaurant scene, Seaport Village, and various entertainment venues.

On-site, the hotel offers several restaurants and bars, two rooftop pools (open daily 6:00-22:00) with San Diego Bay views, whirlpool tubs and private cabanas, a complimentary 24-hour StayFit fitness center, and direct access to the Embarcadero promenade, the marina, and Seaport Village. The conference takes place across the hotel's meeting levels; please follow the ACL directional signage and the Whova app for room locations.

Visa & Entry to the United States
International attendees should carry a valid **passport** and a copy of their conference registration confirmation. Visa Waiver Program travelers must have an approved **ESTA** before boarding their flight to the U.S. If you still need an official invitation or visa-support letter, please contact the ACL registration team.

About San Diego
San Diego, California, is a vibrant coastal city known for its stunning beaches, rich history, and technological innovation. It is home to the largest naval fleet in the world and enjoys over 260 sunny days a year, making it a paradise for outdoor enthusiasts. The city is also a hub for biotech and telecommunications, with companies like Qualcomm leading advancements in wireless technology. San Diego hosts Comic-Con International, the world's largest comic book and popular arts convention, attracting fans and creators from around the globe. Additionally, the Torrey Pines State Natural Reserve is one of the only places where the rare Torrey Pine tree grows, adding to the city's unique natural beauty. A fun fact: the

very first video ever uploaded to YouTube by co-founder Jawed Karim, titled *Me at the zoo*, was filmed at the elephant exhibit in the San Diego Zoo in 2005. Whether you're exploring its cultural landmarks or diving into its tech scene, San Diego offers a fascinating blend of history, nature, and innovation.

Local Information

Languages
English is the primary and de facto official language used in San Diego, California, with Spanish widely spoken as the second most common language.

Weather
San Diego in early July (including the week of ACL 2026, July 2-7), typically features warm, comfortable summer conditions that are ideal for visitors. Daytime temperatures usually reach the mid-20s Celsius (mid-70s to around 80°F), while evenings cool to the high teens (mid-60s °F), creating a refreshing contrast between day and night. Weather is predominantly sunny and very dry, with little to no rainfall expected, although mornings may begin with a light coastal cloud layer that typically clears by midday. Gentle ocean breezes help moderate temperatures and keep humidity comfortable, making conditions pleasant for both indoor events and outdoor activities. Overall, attendees can expect consistently mild, sunny, and enjoyable weather throughout the week.

Currency
The official currency used in San Diego is the US Dollar. Several major banks and ATMs are located within walking distance or a short drive from the venue.

Power Plugs/Outlets
The United States uses Type A and Type B electrical outlets, with a standard voltage of 120 volts and a frequency of 60 Hz. Type A plugs have two flat parallel pins, while Type B plugs include an additional round grounding pin for increased safety. Travelers from regions such as Europe, the UK, or Switzerland will typically need a plug adapter, and in some cases a voltage converter if their devices are not compatible with 120V systems.

Clothing
Bring light jackets; temperature differences between morning/evening and daytime can be significant, so a medium-weight jacket is advisable.

Water
Water served at local establishments undergoes filtration and is safe for consumption. Although the municipal water supply complies with all safety regulations, a slight chlorine scent may occasionally be present. For individuals with heightened sensitivities or specific hydration preferences, bottled water is always an alternative.

Local Customs (Do's and Don'ts in San Diego)

Do's

- **Stand on the right, walk on the left.** When using escalators (like at the airport or shopping malls), stand on the right side so people in a rush can pass you on the left.
- **Expect "sticker shock" at checkout.** The price you see on a price tag or menu is not the final price. Sales tax (around 7.75% in San Diego) is calculated and added at the cash register.
- **Make small talk.** Americans are incredibly friendly and conversational. Cashiers, waiters, and strangers in line will likely ask, "How are you doing today?" A short, polite "Good, thanks! How are you?" is all that is expected.
- **Look at the fine print on menu water.** If a restaurant waiter asks if you want "still, sparkling, or tap water," choose tap water if you want it to be free. Saying "still" usually results in an expensive bottle of water added to your bill.

Don'ts

- **Don't smoke on the beaches or in parks.** San Diego has incredibly strict smoke-free laws. Smoking cigarettes or e-cigarettes (vapes) is completely banned on all public beaches, boardwalks, parks, and outdoor restaurant patios.
- **Don't drink alcohol in public spaces.** You cannot walk down the street, boardwalk, or sit on the beach with an open beer or cocktail. Alcohol can only be consumed inside licensed establishments or designated private areas.
- **Don't cross the street anywhere you want.** "Jaywalking" (crossing the street outside of a marked pedestrian crosswalk or against a red light) can result in a hefty fine from the police. Always wait for the white "walk" signal.
- **Don't underdress for the evening.** San Diego has beautiful daytime weather, but because it is on the Pacific Ocean, the temperature drops significantly as soon as the sun goes down. Don't leave your hotel for the day without a light jacket or sweater.
- **Don't pet or feed the sea lions.** If you visit La Jolla Cove to see the famous wild sea lions, keep your distance (at least 50 feet or 15 meters). It is illegal to harass, feed, or get too close to them, and local lifeguards will publicly call you out on a loudspeaker if you try to take a selfie too close to them.

Public Transportation Etiquette

Sound and Space

- **Always use headphones.** Listening to music, watching videos, or having a phone call on speakerphone without headphones is the number-one complaint among San Diego riders. Keep your headphones plugged in while on board.
- **Keep your belongings to one seat.** Do not use the empty seat next to you as a shelf for your backpack, purse, or shopping bags if the transit vehicle is filling up. Keep your items on your lap or between your feet.
- **Never put your feet on the seats.** Propping your shoes up on the seat across from you is a major etiquette violation and can actually result in a legal citation from transit security.

Priority Seating and Space

- **Move for seniors and disabled riders.** The seats closest to the doors on both buses and Trolleys are clearly marked with blue signs as Priority Seating. If you are sitting there and someone who is elderly, pregnant, or has a disability boards, you are expected to stand up and offer them the seat immediately without being asked.
- **Let people off first.** When the doors open, stand to the side of the door on the platform and let arriving passengers exit completely before you try to step inside.

Food, Drink, and Trash

- **No eating allowed.** To keep vehicles clean and prevent unwanted pests, eating food on board buses and Trolleys is strictly prohibited by MTS regulations.
- **Drinks must have lids.** You can bring a beverage on board (like a coffee or water bottle), but it must have a secure, spill-proof lid. Alcohol is completely illegal on all public transit.

Tipping
Tipping is appreciated and usually honored in the US; it is how service workers make their living. Tip 15% to 20% at sit-down restaurants, $1 to $2 per drink at a bar, and $2 to $5 for hotel porters or rideshare drivers.

Things to Do in San Diego
San Diego offers a mix of natural beauty, cultural experiences, and exciting activities, perfect for a memorable visit. Here are some top things to do:

- **Bicycle and Water Rentals**. Located on the Embarcadero boardwalk directly behind the conference venue/hotel, bicycle and water sport rentals are available year-round. Explore the city by bike on the miles of paved waterfront pathways or set sail on a boat or kayak rental to take in the sights from San Diego Bay.
- **Seaport Village**. Boasting a breezy waterfront setting and vintage atmosphere, Seaport Village is an outdoor shopping destination located steps from the conference venue. Discover unique shops, indulge in bay-view restaurants featuring SoCal cuisine, tour the waterways, or ride the hand-carved carousel.
- **Petco Park**. Home of MLB's San Diego Padres and an event venue for concerts. Take a guided tour to learn about the Padres' history while visiting this downtown ballpark, located a short distance from The Rady Shell music and concert venue.
- **Old Town San Diego**. A great destination for unique gifts, souvenirs, and creative cuisines. You will discover a variety of shops, jewelry, art, and pieces of California's history while enjoying a diverse selection of culinary choices. Old Town is filled with a festive atmosphere showcasing local artists and entertainment.
- **San Diego Zoo & Balboa Park**. One of the most famous zoos in the world, located within a beautiful cultural park featuring museums and gardens.
- **La Jolla Cove**. Stunning coastal views, sea caves, and opportunities for kayaking, snorkeling, and spotting marine life. You may even see California sea lions!
- **Gaslamp Quarter**. A lively historic district with restaurants, bars, and nightlife, perfect for an evening out.
- **USS Midway Museum**. Explore a historic aircraft carrier with restored planes and interactive exhibits.
- **Sunset Cliffs Natural Park**. A breathtaking spot to watch the sunset over the Pacific Ocean.
- **Coronado Island**. Home to beautiful beaches and the iconic Hotel del Coronado.
- **Liberty Public Market**. A great place to sample local food and drinks in a vibrant setting.
- **Pacific Beach Boardwalk**. Rent a bike or scooter and cruise along the scenic coastline.

Shopping, Dining, and Entertainment
The conference venue is near plenty of options for shopping, dining, and entertainment, accessible by foot, bike, car, or public transportation. Visit these nearby downtown neighborhoods: Gaslamp Quarter, Little Italy, Embarcadero & Marina District, East Village, and Golden Hill.

Scan for the downtown San Diego neighborhoods guide

Best Transportation to Get Around the City
The Manchester Grand Hyatt San Diego sits just a few miles from San Diego International Airport (SAN), offering convenient access for arriving attendees. The MTS Bus Route 992 provides direct service from the airport to Broadway & Kettner Blvd, the appropriate stop for reaching the hotel. Taxis from the airport typically range from $13 to $16. Once at the venue, attendees can comfortably explore nearby waterfront destinations, including the Marina District, the Embarcadero, and the Gaslamp Quarter, on foot. For travel to other parts of the city, the Seaport Village Station, served by the MTS Trolley Green Line, is located adjacent to the hotel.

Scan for the full MTS bus & Trolley schedule and route map

Practical Information

Printing & Business Services
A **FedEx Office Print & Ship Center** is located on the lobby level (Harbor Tower) of the Manchester Grand Hyatt, offering printing, copying, and shipping; you can submit print orders online for pickup on-site. **Conference poster printing** is provided separately by ARC — see the Poster Printing section for sizes, pricing, and pickup in Nautical (Level 4).

Important Notes

- **City & Traffic Awareness.** Stay alert in high-traffic zones. Use crosswalks and signals and avoid jaywalking, especially downtown where vehicle, bike, and scooter traffic is dense.
- **Transportation Tips.** Use reliable transportation such as licensed taxis, rideshare services, or public transit. Avoid walking alone late at night in unfamiliar areas. Arrive at SAN at least 2 hours before flights, bring valid ID, and use the free shuttle to the Rental Car Center if needed. Amtrak and COASTER trains connect major areas, and rideshare pickup zones are clearly marked at the airport.
- **Outdoor Safety.** Follow lifeguard instructions and swim only in designated areas. Watch for rip currents and avoid swimming alone. Hydrate and use sun protection; San Diego's sun is strong year-round. Stay aware of surroundings in crowded beach areas to avoid theft.

- **General Safety.** Research neighborhoods before exploring; stay aware in crowded tourist zones like the Gaslamp Quarter. Keep belongings secure in busy areas. Use well-lit streets at night and travel with others when possible.

In Case of an Emergency

Immediate Emergencies
911, Police, Fire, Medical. Use for life-threatening emergencies or urgent safety threats.

Non-Emergency & Support Services
San Diego Police Non-Emergency Line: 619-531-2000

Local Urgent Care & Medical Services
The closest 24/7 urgent care to the Manchester Grand Hyatt is Primary Care & House Calls, located 0.3 miles away and open 24 hours.

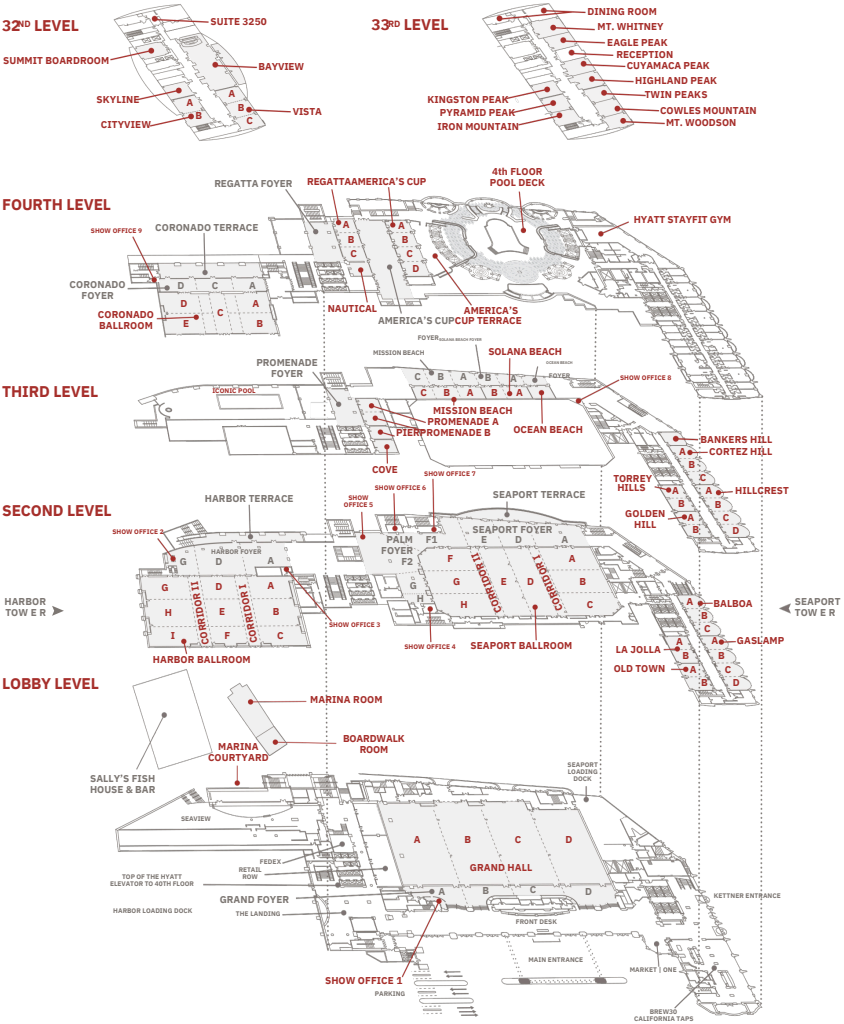
16

Venue Map

FLOOR PLAN
Lobby Level

GRAND | HYATT

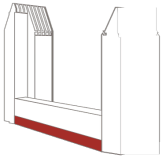
FLOOR PLAN
All Floors

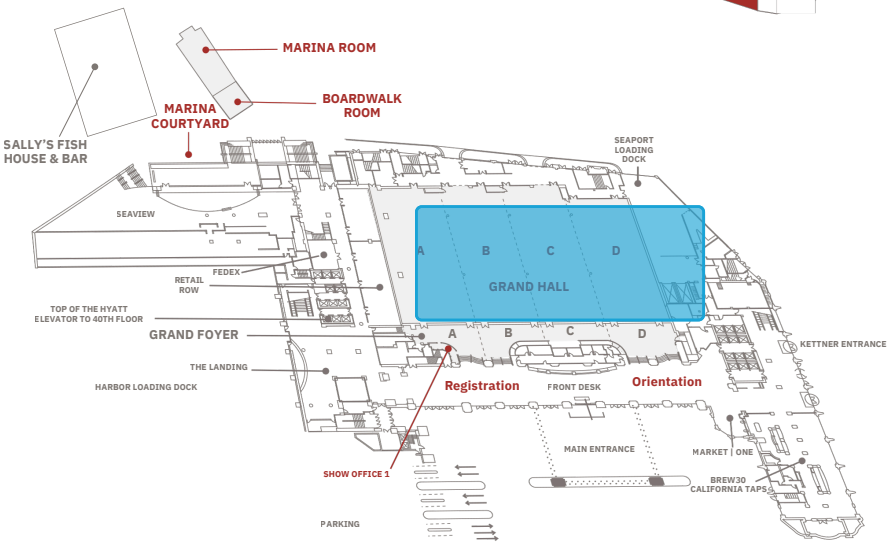


FLOOR PLAN
Lobby Level

GRAND | HYATT

Registration & Orientation: Grand Hall Foyer
Poster & Demo Presentations: Grand Hall & Coronado Ballroom on 4th floor
Exhibitors: Grand Hall & 2nd floor Harbor Foyer
Family Meet up: Boardwalk Room
Quiet Room: Marina Room



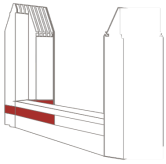


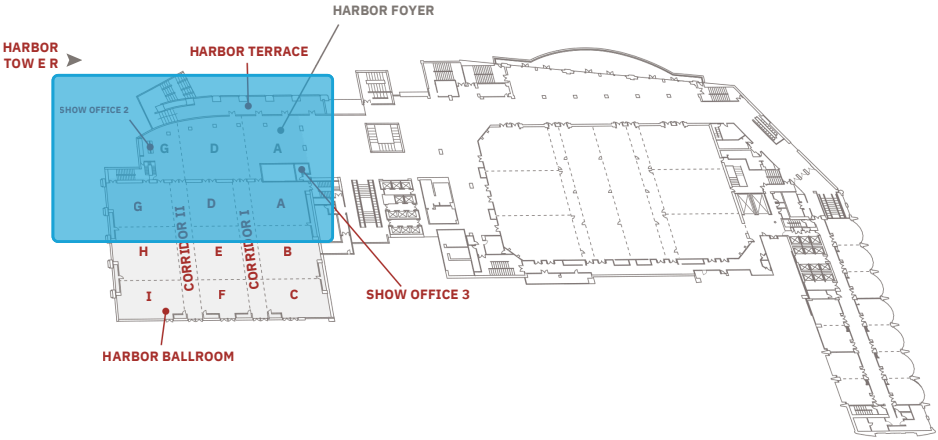
Visit our Sponsors in the Grand Ballroom
Bloomberg, Capital One, Alibaba Cloud, Apple, Megagon Labs

FLOOR PLAN
SecondLevel— Harbor Tower

GRAND | HYATT

Exhibitors: Harbor Foyer & Grand Hall
Tutorials: Harbor A-C; D-F; G-I
Workshops: Harbor A,B,C,D,E&F, G, H&I
Plenary: Harbor Ballroom
Orals: Harbor A, B&C, D, E&F, G, H&I
Lactation Room: Show Office 2



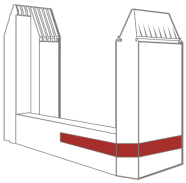


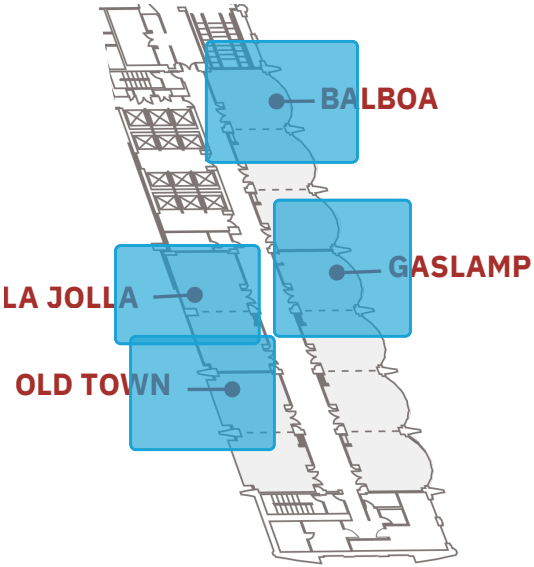
Visit our Sponsors in the Harbor Foyer
Oracle, Appen, Meta, Baidu, Tencent, Google, Ant Group, Snowflake, Zhongguancun Academy, JPMorganChase, Speak, Adobe, Translated, ARC, DATAmundi, NEC

FLOOR PLAN
SecondLevel— SeaPort Tower

GRAND | HYATT

SeaPort Tower
Workshops: Balboa A&B, C, Gaslamp A&B, C&D, Old Town, La Jolla
Industry: Balboa A&B
SRW: Gaslamp A&B
BoF/Affinity: Gaslamp C&D, La Jolla, Old Town

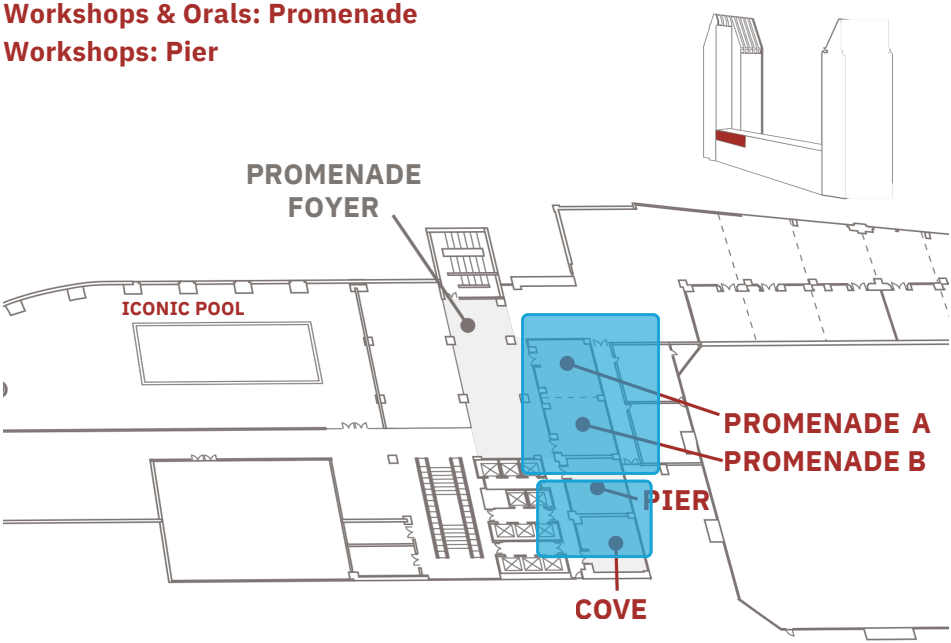




FLOOR PLAN
ThirdLevel—Harbor Tower

GRAND | HYATT

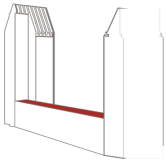
Workshops & Orals: Promenade
Workshops: Pier

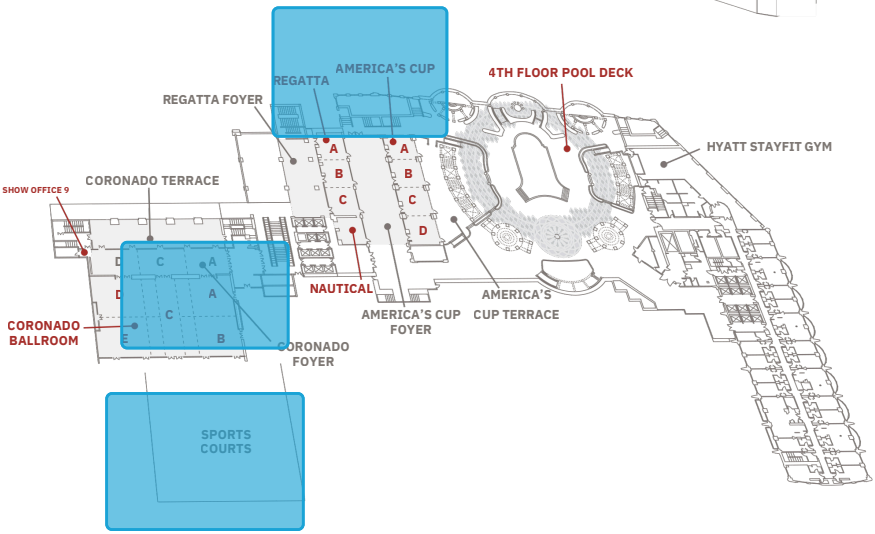


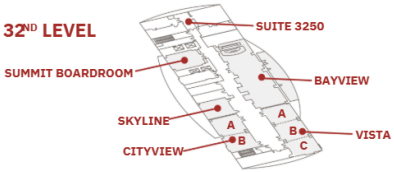
FLOOR PLAN
Fourth Level, 32nd & 33rd Floor

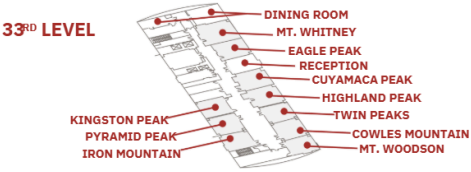
GRAND | HYATT

Poster Presentations: Coronado Ballroom & Grand Hall on Lobby Level
Workshops & Orals: Regatta
Poster Pickup: Nautical
Men's Prayer Room: Mt. Whitney
Women's Prayer Room: Eagle Peak









Index

(Steve)Liu, 527
(Steve)Liu, Xue, 527
~Hongju_Su1,
~Kaiser_Sun1,
~Yao_Xiao8,
~Hongju_Su1, Hongju Su,
~Kaiser_Sun1, Kaiser Sun,
~Yao_Xiao8, Yao Xiao,

A, 862
A, Akshata, 862
Aatar, 860
Aatar, Aya El, 860
Aavang, 238
Aavang, Rasmus T., 238
Abadi, 98
Abadi, Debela Desalegn YadetaNegasi Haile, 98
Abagyan, 826
Abagyan, Diana, 826
Abassy, 84, 236, 555
Abassy, Mervat, 84, 236, 555
Abate, 599
Abate, Yohannes, 599
Abbadi, 184
Abbadi, Amr El, 184
Abbas, 769
Abbas, Ummar, 769
Abbasi, 438, 463, 552, 626
Abbasi, Ahmed, 438, 463, 626
Abbasi, Yassi, 552
Abbasnejad, 454
Abbasnejad, Ehsan, 454
Abdallah, 622, 639, 744, 873
Abdallah, Abdelrahman, 622, 639, 744, 873
Abdelhalim, 430
Abdelhalim, Salma E., 430
Abdelkadir, 462
Abdelkadir, Nuredin Ali, 462
Abdelwahab, 785
Abdelwahab, Hammam, 785
Abdi, 244, 499
Abdi, Ali, 499
Abdi, Immanuel, 244
Abdul-Mageed, 102, 124, 231, 468, 860
Abdul-Mageed, Muhammad, 102, 124, 231, 468, 860
Abdullah, 586
Abdullah, Amir, 586
Abdullahi, 837
Abdullahi, Tassallah, 837
Abdulmumin, 98, 471, 862
Abdulmumin, Idris, 98, 471, 862
Abdurahman, 158
Abdurahman, Suhaib, 158
Abend, 557
Abend, Omri, 557
Abercrombie;,
Abercrombie;, Gavin,
Aberdam, 385
Aberdam, Aviad, 385
Ablikim, 809
Ablikim, Abdulla, 809

Abolhasani, 459
Abolhasani, Mohammadsadegh, 459
Abootorabi, 802
Abootorabi, Mohammad Mahdi, 802
Abraham, 506
Abraham, Ayan Sar; Pranav Singh Puri; Sumit Aich; Anurag Kaushish; Tanupriya Choudhury; Ajith,
Abrahamyan, 890
Abrahamyan, Davit, 890
Abubakar, 862
Abubakar, Abdulhamid, 862
Abudula, 739
Abudula, Abudukeyumu, 739
AbuHweidi, 860
AbuHweidi, Ruwa, 860
Abureesh, 860
Abureesh, Lina, 860
Abuín, 813
Abuín, Marta Vázquez, 813
Acar, 758
Acar, Kutay, 758
Acarturk, 471
Acarturk, Cengiz, 471
Achananuparp, 187, 290
Achananuparp, Palakorn, 187, 290
Achom, 856
Achom, Elizabeth Victoria, 856
Acikgoz, 151, 573
Acikgoz, Emre Can, 151, 573
Acquaye, 906
Acquaye, Christabel, 906
Adam, 862
Adam, Faisal Muhammad, 862
Adar, 459
Adar, Eytan, 459
Adarsh, 496
Adarsh, Shivam, 496
Adeeba, 837
Adeeba, Farah, 837
Adelani, 98, 166
Adelani, David Ifeoluwa, 98, 166
Ademaj, 438
Ademaj, Gemza, 438
Adenuga, 862
Adenuga, Esther, 862
Adhikari, 606, 631
Adhikari, Rabin, 631
Adhikari, Sayanta, 606
Adi, 113, 544
Adi, Yossi, 113, 544
Aditya, 578
Aditya, Ramit, 578
Adkins, 287
Adkins, Jane, 287
Admoni, 270
Admoni, Sahar, 270
Adnan, 668
Adnan, Kazi Reyazul Hasan; A. B. M. Alim Al Islam; Muhammad Abdullah,
Aduri, 608
Aduri, Harsha, 608
Aerni, 784
Aerni, Michael, 784
Afanaseva, 85
Afanaseva, Svetlana, 85
Afonin, 576
Afonin, Nikita, 576
Afshar, 822
Afshar, Majid, 822
Agarwal, 167, 195, 202, 552, 589, 822, 827, 828, 836, 841, 862, 873
Agarwal, Amit, 552, 862, 873
Agarwal, Chirag, 195, 202, 836
Agarwal, Mehul, 589
Agarwal, Prerna, 827
Agarwal, Raunak, 822
Agarwal, Rishita, 828
Agarwal, Shiven, 841
Agarwal, Shubham, 167
Aggarwal, 285, 577, 583, 589, 846
Aggarwal, Abhinav, 285
Aggarwal, Aditya, 589
Aggarwal, Divyanshu, 577
Aggarwal, Karan, 583
Aggarwal, Purav, 846
Aggazzotti, 862
Aggazzotti, Cristina, 862
Aghakhani, 210
Aghakhani, Elham, 210
Aghazada, 290
Aghazada, Murad, 290
Aghzal, 154
Aghzal, Mohamed, 154
Agrawal, 184, 189, 217, 222, 226, 261, 291, 492, 523, 606
Agrawal, Aakriti, 226, 492
Agrawal, Ameeta, 222, 291, 523
Agrawal, Darsh, 189, 217
Agrawal, Divyakant, 184
Agrawal, Monica, 261
Agrawal, Sanjay, 606
Agyei, 538
Agyei, Evans Kofi, 538
Ahad, 380
Ahad, Syed, 380

Aharoni, 821
Aharoni, Roee, 821
Ahia, 127
Ahia, Orevaoghene, 127
Ahlawat, 700
Ahlawat, Satyadev, 700
Ahmad, 95, 232, 265, 471, 583, 862, 864
Ahmad, Aitzaz, 583
Ahmad, Fatema, 232
Ahmad, Ibrahim Said, 471
Ahmad, Raia Abu, 862
Ahmad, Sarfraz, 95, 864
Ahmad, Wasi Uddin, 265
Ahmadi, 247
Ahmadi, Sina, 247
Ahmed, 117, 329, 355, 471, 523, 556, 770
Ahmed, Ahmed Haj, 117
Ahmed, Christopher Song; Abdullah,
Ahmed, Osama, 770
Ahmed, Syed Ishtiaque, 471, 556
Ahmed, Umer, 355
Ahmed, Walid, 329
Ahn, 217, 267, 477
Ahn, Eungyeol, 267
Ahn, Jeongmin, 217
Ahn, Yelim, 477
Ahnert, 835
Ahnert, Georg, 835
Ahsan, 95, 864
Ahsan, Momina, 95, 864
Ahuja, 505, 609, 837
Ahuja, Naman, 505
Ahuja, Sanchit, 837
Ahuja, Sarthak, 609
Ai, 114, 134, 193, 212, 341, 354, 362, 373, 733, 752, 844, 864, 881, 906
Ai, Haojun, 362
Ai, Jiaxin, 134, 844
Ai, Lin, 906
Ai, Lynn, 212, 864
Ai, Qihang, 341, 733
Ai, Qingyao, 373
Ai, Qiuyuan, 193, 881
Ai, Shuan, 354
Ai, Xiaodong, 752
Ai, Yang, 114
Ai, Zhengyang, 752
Airlangga, 589
Airlangga, Muhammad Cendekia, 589
Aizawa, 213, 251, 454, 478, 823
Aizawa, Akiko, 213, 251, 454, 478, 823
Ajeneza, 538
Ajeneza, Carlene, 538
Aji, 98, 236, 816, 869, 902
Aji, Alham, 869
Aji, Alham Fikri, 98, 236, 816, 902
Ajibola, 538
Ajibola, Yinka, 538
Ajjour, 102
Ajjour, Yamen, 102
Akarajaradwong, 430, 728
Akarajaradwong, Pawitsapak, 430, 728
Akavarapu, 396
Akavarapu, V.S.D.S.Mahesh, 396
Akbik, 548
Akbik, Alan, 548
Akdemir, 615
Akdemir, Arda, 615
Akella, 498
Akella, Aditya, 498
Akhtar, 218, 236, 777
Akhtar, Md Shad, 236
Akhtar, Mohd Mujtaba, 218
Akhtar, Mubashara, 777
Akhter, 183
Akhter, Mahmud Elahi, 183
Akinode, 233
Akinode, Destiny, 233
Akoglu, 162, 829
Akoglu, Leman, 162, 829
Akram, 547
Akram, Bita, 547
Akter, 309, 315
Akter, Sanjeda, 309, 315
Akyürek, 560
Akyürek, Afra Feyza, 560
Al-Badrashiny, 841
Al-Badrashiny, Mohamed, 841
Al-Dhabyani, 860
Al-Dhabyani, Walid, 860
Al-Ghafri, 860
Al-Ghafri, Al-Yas, 860
Al-hibiri, 860
Al-hibiri, Thikra, 860
Al-Jazzazi, 212
Al-Jazzazi, Yousef, 212
Al-Khalifa, 862
Al-Khalifa, Hend, 862
al-Khatib, 494
al-Khatib, Mutaz, 494
Al-Lawati, 85
Al-Lawati, Ali, 85
Alabi, 98, 862
Alabi, Jesujoba O., 862
Alabi, Jesujoba Oluwadara, 98
Alacam, 471

Alacam, Özge, 471
Alajmi, 761
Alajmi, Anwar, 761
Alam, 232, 471, 494, 556, 769, 897
Alam, Firoj, 232, 471, 494, 556, 769, 897
Alanova, 342
Alanova, Shirin, 342
Alasmari, 860
Alasmari, Ashwag, 860
Alavi, 589
Alavi, Khashayar, 589
Alavo, 438
Alavo, Raymond, 438
Alazraki, 847
Alazraki, Lisa, 847
Alaçam, 218
Alaçam, Özge, 218
Alba, 201
Alba, Charles, 201
Albrecht, 150
Albrecht, Jens, 150
Alcocer, 226
Alcocer, Jose J., 226
Aldarmaki, 690
Aldarmaki, Hanan, 690
Alemzadeh, 204
Alemzadeh, Homa, 204
Alencar, 547
Alencar, Rafaella Sampaio de, 547
Aletras, 176, 183, 242
Aletras, Nikolaos, 176, 183, 242
Alexander, 761
Alexander, Zyl, 761
Alexandrov, 593
Alexandrov, Anton, 593
Alford, 263
Alford, Simon, 263
Alghythee, 307
Alghythee, Kenan Kamel A, 307
Alhamouri, 860
Alhamouri, Rahaf, 860
Alhassan, 231
Alhassan, Amin Abu, 231
Alhirthani, 494
Alhirthani, Mahmoud, 494
Alhuzali, 860
Alhuzali, Hassan, 860
Ali, 232, 261, 371, 408, 471, 549, 622, 818, 838, 845, 867, 873
Ali, Abdallah El, 867
Ali, Adem Chanie, 471
Ali, Amin Ahsan, 371
Ali, Amr, 818
Ali, Ayman, 261
Ali, Eman, 408
Ali, Iqra, 838
Ali, Mohammed, 622, 873
Ali, Mohammed Eunus, 549
Ali, Rafiq, 845
Ali, Raian, 232
Aliannejadi, 254
Aliannejadi, Mohammad, 254
Alikhani, 123, 826
Alikhani, Malihe, 123, 826
Alimova, 416, 551
Alimova, Ilseyar, 551
Alimova, Ilya Galyukshev; Ilseyar,
Alimova;,
Alimova;, Elisei Rykov; Nikolay Ivanov; Maria Bandulevich; Kseniia Petrushina; Valentin Malykh; Vasily Konovalov; Alexander Panchenko; Ilseyar,
Alinejad-Rokny, 849
Alinejad-Rokny, Hamid, 849
Alis, 232, 525
Alis, Christian, 232, 525
Aljaafari, 576
Aljaafari, Nura, 576
Aljanaideh, 641
Aljanaideh, Ahmad, 641
Alkathiri, 860
Alkathiri, Amir Azad Adli, 860
AlKhamissi, 784
AlKhamissi, Badr, 784
Alkhazimi, 860
Alkhazimi, Alshima, 860
Alkhouli, 598
Alkhouli, Tamer, 598
Allaway, 865
Allaway, Emily, 865
Allein, 624
Allein, Liesbeth, 624
Allende-Cid, 785
Allende-Cid, Héctor, 785
Allouche, 551
Allouche, Omri, 551
Allur, 305
Allur, Sharan Kumar, 305
AlMahri, 95
AlMahri, Muhra, 95
Almeida, 275, 611
Almeida, Shone, 611
Almeida, Tiago, 275
Almheiri, 589, 864
Almheiri, Saeed, 589, 864
Almotahari, 865

Almotahari, Mahrad, 865
Alomair, 879
Alomair, Basel, 879
Alon, 453
Alon, Bar, 453
Aloqaily, 621
Aloqaily, Moayad, 621
Alper, 777
Alper, Morris, 777
Alqifari, 862
Alqifari, Reem, 862
Alrowili, 370
Alrowili, Sultan, 370
Alsagoff, 667
Alsagoff, Syed Ali Redha, 667
Alsayadi, 860
Alsayadi, Hamzah, 860
Alselwi, 854
Alselwi, Ghadir, 854
Alshahri, 860
Alshahri, Omar, 860
Alsuradi, 815
Alsuradi, Eman, 815
Althoff, 529, 891
Althoff, Tim, 529, 891
Altinger, 812
Altinger, Lena, 812
Alugubelli, 443
Alugubelli, Sujal Reddy, 443
Alvanita, 395
Alvanita, Alvanita, 395
Alves, 222, 705
Alves, Diego, 222
Alves, João, 705
Alwajih, 468, 860
Alwajih, Fakhraddin, 468, 860
Amalvy, 559
Amalvy, Arthur, 559
Amani, 784
Amani, Mohammad Hossein, 784
Amballa, 245
Amballa, Avinash, 245
Amer, 329
Amer, Hossam, 329
Amghar, 346
Amghar, Tassadit, 346
Amin, 107, 371, 833
Amin, Ekant Muljibhai, 833
Amin, Hasan, 107
Amin, M Ashraful, 371
Amini, 826
Amini, Amirali, 826
Aminifar, 371
Aminifar, Amir, 371
Aminnaseri, 840
Aminnaseri, Moin, 840
Aminu, 458
Aminu, Boluwatife, 458
Amir, 270
Amir, Ofra, 270
Amirghodsi, 495
Amirghodsi, Sohrab, 495
Amiri, 195, 215, 536
Amiri, Hadi, 195
Amiri, Mohammad Mohammadi, 215, 536
Amirudin, 862
Amirudin, Azril Hafizi, 862
Amol, 862
Amol, Cynthia, 862
Amosy, 286
Amosy, Ohad, 286
Amouyal, 862
Amouyal, Samuel Joseph, 862
An, 140, 153, 158, 189, 190, 311, 449, 492, 515, 535, 560, 566, 654, 692, 799
An, Bang, 492
An, Bo, 158, 535, 654
An, Hao, 140
An, Hengyu, 566
An, Hongran, 692
An, Hyeseon, 190, 799
An, Jisun, 153, 560
An, Junwen, 515
An, Shengnan, 311
An, Siyu, 449
An, Yuwei, 189
Anand, 98, 127, 209, 250, 454, 569, 637, 639
Anand, Atnafu Lambebo TonjaSrija, 98
Anand, Avinash, 209, 569
Anand, Avishek, 639
Anand, Harpreet Singh, 250
Anand, Mayank, 637
Anand, Nishit, 127
Anand, Vishesh, 454
Ananiadou, 95, 166, 233, 442, 637, 821
Ananiadou, Sophia, 95, 166, 233, 442, 637, 821
Ananthasubramaniam,
Ananthasubramaniam, Aparna,
Ananto, 549
Ananto, Md. Nazmul Islam, 549
Anastasopoulos, 86, 225
Anastasopoulos, Antonios, 86, 225
Anastasopoulos;,
Anastasopoulos;, Antonios,
Anderson, 584
Anderson, Ashton, 584

Andrade, 287
Andrade, Mark, 287
Andrews, 862
Andrews, Nicholas, 862
Andriianov, 576
Andriianov, Nikita, 576
Andriushchenko, 491
Andriushchenko, Georgii, 491
Andruszkiewicz, 499
Andruszkiewicz, Piotr, 499
André, 440
André, Elisabeth, 440
Ang, 83
Ang, Yuen Yuen, 83
Angluin, 165
Angluin, Dana, 165
Angstadt, 772
Angstadt, Mike, 772
Anjum, 872
Anjum, Sumera, 872
Annavaram, 488, 492
Annavaram, Murali, 488, 492
Anoun, 239
Anoun, Amine, 239
Ansaripour, 784
Ansaripour, Matin, 784
Antoniak, 476
Antoniak, Maria, 476
Antoun, 282
Antoun, Wissam, 282
Anugraha, 862
Anugraha, David, 862
Anugraha, Michael, 862
Anumanchipalli, 244
Anumanchipalli, Gopala, 244
Anupam, 502
Anupam, Sagnik, 502
Anvekar, 476, 526
Anvekar, Tejas, 476, 526
Anwar, 100, 471, 555, 864, 869
Anwar, Fatima M., 869
Anwar, Md. Musfique, 100
Anwar, Mohamed, 555, 864
Anwar, Saba, 471
Anzenberg, 292
Anzenberg, Eitan, 292
Ao, 243, 742
Ao, Yulong, 742
Ao, Yuzhuo, 243
Aoun, 860
Aoun, Alaa, 860
Arachchilage, 782
Arachchilage, Nalin Asanka Gamagedara, 782
Arad, 90, 894
Arad, Dana, 90, 894
Arai, 575
Arai, Hirona, 575
Aralikatti, 226
Aralikatti, Rohith, 226
Aramaki, 252, 307
Aramaki, Eiji, 252, 307
Arana, 232, 525
Arana, Jasper, 232, 525
Arango, 667
Arango, Iñaki, 667
Arase, 252
Arase, Yuki, 252
Arbelle, 116
Arbelle, Assaf, 116
Arenas-García, 694
Arenas-García, Jerónimo, 694
Argn, 821
Argn, Berke, 821
Arias, 246, 778
Arias, Esteban Garces, 246, 778
Arif, 882
Arif, Nokimul Hasan, 882
Arima, 237
Arima, Kosuke, 237
Arnaout, 176, 529
Arnaout, Hiba, 176, 529
Arnett, 798, 862
Arnett, Catherine, 798, 862
Arora, 111, 249, 374, 408, 430, 455, 580, 845
Arora, Arnav, 580
Arora, Chetan, 249, 408
Arora, Devansh, 374
Arora, Pratibha K., 430
Arora, Pushkar, 845
Arora, Siddhant, 111, 455
Artemova, 236
Artemova, Ekaterina, 236
Artzi, 598
Artzi, Yoav, 598
Arumugam, 524
Arumugam, Vishalakshi, 524
Arviv, 517
Arviv, Ofir, 517
Arya, 512
Arya, Mihir, 512
Arzt, 715
Arzt, Varvara, 715
Asada, 452
Asada, Masaki, 452
Asaduzzaman, 265
Asaduzzaman, Muhammad, 265

Asakura, 237
Asakura, Takuto, 237
Asante, 202, 540
Asante, Daniel Agyei, 202, 540
Asgari, 341, 696, 802
Asgari, Arash, 696
Asgari, Ehsaneddin, 341, 802
Ash, 777, 784
Ash, Elliott, 777, 784
Asher, 147, 301, 439, 811
Asher, Nicholas, 147, 301, 439, 811
Ashraf, 232
Ashraf, Saad, 232
Ashrafi, 293, 459
Ashrafi, Imranul, 293
Ashrafi, Shwan, 459
Ashuach, 90, 771
Ashuach, Tomer, 90, 771
Asilis, 225
Asilis, Julian, 225
Asiri, 860
Asiri, Areej, 860
Askari, 594, 644
Askari, Fatemeh, 644
Askari, Hadi, 594
Asokan, 556
Asokan, N., 556
Assen, 425
Assen, Jan von der, 425
Assudani, 606
Assudani, Dheeraj, 606
Atad, 600
Atad, Ido Andrew, 600
Atanasova, 453
Atanasova, Pepa, 453
Atarashi, 823
Atarashi, Kyohei, 823
Atefi, 700
Atefi, Golzar, 700
Athuraliya, 317
Athuraliya, Vihanga Supasan Kariyakaranage; Banuka,
Atil, 479
Atil, Berk, 479
Atou, 860
Atou, Houdaifa, 860
Atri, 553
Atri, Yash Kumar, 553
Atrio, 238
Atrio, Àlex R., 238
Attia, 589
Attia, Mena, 589
Atuhurra, 838
Atuhurra, Jesse, 838
Atzil-Slonim, 529
Atzil-Slonim, Dana, 529
Atl, 222
Atl, Berk, 222
Au, 440
Au, Wing Yee, 440
Aubert-Béduchaud, 478
Aubert-Béduchaud, Julien, 478
Augenstein, 453, 484, 517, 572, 577, 594
Augenstein, Isabelle, 453, 484, 517, 572, 577, 594
Augey, 615
Augey, Yohann, 615
August, 229, 440
August, Tal, 229, 440
Augustyniak, 494
Augustyniak, Lukasz, 494
Aumiller, 376
Aumiller, Dennis, 376
Aung, 169
Aung, Thura, 169
author_guo1,
author_guo1, author_yiju1,
author_last_Kang,
author_last_Kang, author_first_Daeun,
author_lin4,
author_lin4, author_tianyi2 author_hu2author_ zexu3 author_sun3author_yankai4,
author_pravish_sainath1,
author_pravish_sainath1, author_pravish_sainath1,
Avendi, 552
Avendi, M., 552
Avraham, 385
Avraham, Elad Ben, 385
Awadalla, 444
Awadalla, Hany Hassan, 444
Awadelkarim, 892
Awadelkarim, Amel, 892
Awasthi, 551
Awasthi, Sharanya, 551
Awon, 292
Awon, Ahmed Musa, 292
Ayad, 115
Ayad, Hoda, 115
Ayele, 471
Ayele, Abinew Ali, 471
Ayesh, 545
Ayesh, Mutaz, 545
Ayestaran, 560
Ayestaran, Marcos, 560
Azarang, 528
Azarang, Arian, 528
Azarkhalili, 695

Azarkhalili, Behrooz, 695
Azim, 664
Azim, Nafew, 664
Azime, 98, 120
Azime, Israel Abebe, 98, 120
Aziz, 236
Aziz, Alexander, 236
Azizi, 269
Azizi, Seyedarmin, 269
Azizov, 545
Azizov, Dilshod, 545
AZZAG, 219
AZZAG, Hanane, 219
ASSenmacher, 778
ASSenmacher, Matthias, 778

B, 589
B, Ashwath Rao, 589
Baba, 518
Baba, Kei Moriyama; Kouta Nakayama; Yukino,
Babiker, 227
Babiker, Housam, 227
Bach, 882
Bach, Thong, 882
Back, 769
Back, Jihye, 769
Backes, 111, 417, 457, 594, 628, 647, 817
Backes, Michael, 111, 417, 457, 594, 628, 647, 817
Bada, 123
Bada, Michael, 123
Badanin, 784
Badanin, Ilia, 784
Badhe, 615
Badhe, Sanket, 615
Badshah, 165
Badshah, Sher, 165
Bae, 122, 183, 292, 404, 553, 564, 828, 900
Bae, Hyunkyung, 900
Bae, Sanghwan, 292
Bae, Seunghyun, 553
Bae, Seyun, 122
Bae, Sung-Ho, 404
Bae, Suyoung, 183, 564, 828
Baek, 227, 557, 855, 887
Baek, Euijin, 227
Baek, Ingeol, 557
Baek, Jinheon, 855
Baek, Sunghyun, 887
Baes, 173
Baes, Naomi, 173
Baeumel, 806, 858
Baeumel, Tanja, 806, 858
Bafna, 98
Bafna, Niyati, 98
Bageshpura, 576
Bageshpura, Nikhil, 576
Bagherian, 644
Bagherian, Mobin, 644
Baghershahi, 200, 889
Baghershahi, Peyman, 200, 889
Baghshah, 644
Baghshah, Mahdieh Soleymani, 644
Bahrainian, 557, 648
Bahrainian, Seyed Ali, 557, 648
Bahrani, 83
Bahrani, Fatemeh, 83
BAI, 239, 250, 268, 309, 440
Bai, 124, 128, 149, 161, 162, 164, 171, 205, 253, 282, 313, 344, 359, 361, 372, 403, 415, 421, 448, 449, 488, 518, 535, 539, 601, 605, 631, 634, 653, 657, 680, 743, 784, 790, 796, 797, 810, 853, 885, 901
Bai, Bizhe, 601
Bai, Clive, 149, 539
Bai, Fan, 518
Bai, Fengshuo, 601
Bai, Haoli, 372
Bai, He, 162
Bai, Huiyu, 885
Bai, Jiale, 657
Bai, Jiaxin, 161, 253
BAI, LEI, 239, 250, 268, 309, 440
Bai, Lei, 448, 797, 810
Bai, Lichen, 535
Bai, Long, 171, 784
Bai, Lu, 415
Bai, Qirui, 853
Bai, Ruofei, 796
Bai, Shihao, 128
Bai, Shunwen, 313, 344
Bai, Sikai, 359
Bai, Tao, 605
Bai, Tianyi, 634
Bai, Tong, 361
Bai, Xiang, 124, 790
Bai, Xuefeng, 653
Bai, Xueying, 421
Bai, Yan, 403
Bai, Yi, 680
Bai, Yu, 205, 631
Bai, Yunfei, 797
Bai, Yushi, 164, 488
Bai, Yutong, 743
Bai, Yuyang, 901
Bai, Yuzhuo, 282, 449
Baidya, 202, 595, 662

Baidya, Avinash, 595
Baidya, Subhadip, 202, 662
Bailey, 229, 440
Bailey, Brian, 229
Bailey, Brian P., 440
Bailleux, 633, 722
Bailleux, Thomas, 633, 722
Baitalik, 721
Baitalik, Rajashik Datta; Sanjan,
Baker, 712
Baker, George, 712
Bakshi, 284
Bakshi, Soumyadeep, 284
Bala, 150
Bala, Kavita, 150
Balachandran, 295
Balachandran, Vidhisha, 295
Balaji, 611
Balaji, Sudharshan, 611
Balakrishnan, 609
Balakrishnan, Nivedha, 609
Balandat, 444
Balandat, Maximilian, 444
Balashankar, 371
Balashankar, Ananth, 371
Balasubramanian, 484, 895
Balasubramanian, Vineeth N., 484, 895
Balde, 556
Balde, Gunjan, 556
Baldini, 781
Baldini, Ioana, 781
Baldrati, 291
Baldrati, Alberto, 291
Baldwin, 126, 179, 233, 565, 777
Baldwin, Timothy, 126, 179, 233, 565, 777
Balepur, 225, 467, 547, 869
Balepur, Nishant, 225, 467, 547, 869
Ballesteros, 131, 136, 459
Ballesteros, Miguel, 131, 136, 459
Ballout, 590, 860
Ballout, Mohamad, 590, 860
Balough, 169
Balough, Tom, 169
Balter, 188
Balter, Samuel G., 188
Banayeeanzade, 83, 644
Banayeeanzade, Amin, 83
Banayeeanzade, Mohammadali, 644
Bandi, 586
Bandi, Chaithanya, 586
Banerjee, 198, 225, 442, 500, 878, 897
Banerjee, Ayan, 198, 897
Banerjee, Debopriyo, 442
Banerjee, Mahika, 225
Banerjee, Meghali, 500
Banerjee, Pratyay, 878
Banerjee, Shomrik Barua, 897
Bang, 107, 124, 292
Bang, Ingyu, 292
Bang, Jihwan, 107
Bang, Seunghwan, 124
Bangera, 862
Bangera, Mithil, 862
Bangera, Yeshil, 862
Bansal, 173, 222, 293, 599, 791, 821, 853, 890
Bansal, Chetan, 599, 791
Bansal, Mohit, 173, 222, 293, 821, 853, 890
Bao, 103, 122, 146, 179, 186, 215, 251, 259, 279, 299, 318, 352, 357, 370, 377, 384, 389, 399, 414, 422, 471, 482, 564, 604, 646, 659, 673, 715, 724, 759, 852, 853, 865, 867
Bao, Bo, 279
Bao, Calvin, 103
Bao, Chenfu, 357, 422
Bao, Feilong, 318
Bao, Han, 259, 471, 759
Bao, Honghui, 414
Bao, Jianzhu, 186, 384, 715
Bao, Keqin, 852
Bao, Lingfeng, 389
Bao, Linxin, 370
Bao, Lujia, 865
Bao, Mengjiao, 659
Bao, Mutian, 352, 564
Bao, Riyang, 146, 299
Bao, Siqi, 251, 377, 482, 724, 853
Bao, Weidong, 673
Bao, Yu, 399, 867
Bao, Yubing, 646
Bao, Yujia, 215, 604
Bao, Yuru, 122, 179
Bar, 816
Bar, Kfir, 816
Bar-Haim, 783
Bar-Haim, Roy, 783
Baral, 469, 505, 554, 633
Baral, Chitta, 469, 554, 633
Baral, Rishabh, 505
Barannikov, 685
Barannikov, Serguei, 685
Bardelloni, 713
Bardelloni, Gianluigi, 713
Barez, 586
Barez, Fazl, 586
Barlacchi, 437

Barlacchi, Gianni, 437
Barman, 301
Barman, Niyar R, 301
Barone, 498
Barone, Antonio Valerio Miceli, 498
Barron, 820
Barron, Ryan, 820
Barry, 493, 843, 863
Barry, Alexandra, 843
Barry, James, 493, 863
Barrón-Cedeño, 642
Barrón-Cedeño, Alberto, 642
Barsoum, 355, 490, 505
Barsoum, Emad, 355, 490, 505
Bart, 826
Bart, Yakov, 826
Bartholomew, 637
Bartholomew, Chris, 637
Bartolini, 835
Bartolini, Ilaria, 835
Barua, 199
Barua, Anomadarshi, 199
Basak, 430
Basak, Debadri, 430
Baser, 199
Baser, Manit, 199
Bashiti, 860
Bashiti, Ahlam, 860
Basili, 814
Basili, Roberto, 814
Baskar, 226
Baskar, Sarvesh, 226
Baskin, 543
Baskin, Chaim, 543
Bassenge, 792
Bassenge, Mareike, 792
Bastani, 502
Bastani, Osbert, 502
Bastian, 263
Bastian, Nathaniel D., 263
Bastos, 599, 791
Bastos, Anson, 599, 791
Basu, 481
Basu, Anika, 481
Basuchowdhuri, 823
Basuchowdhuri, Partha, 823
Batista-Navarro, 179, 856
Batista-Navarro, Riza, 179, 856
Batorski, 293
Batorski, Pawe, 293
Batra, 127
Batra, Sanjit Singh, 127
Batzner, 263, 497, 521
Batzner, Jan, 263, 497, 521
Bauer, 485, 844
Bauer, Iqtedar Uddin; André,
Bauer, Lujo, 844
Baumgärter, 874
Baumgärter, Tim, 874
Baumler, 103
Baumler, Connor, 103
Baur, 822
Baur, Simon, 822
Bavare, 779
Bavare, Miheer, 779
Bawden, 282
Bawden, Rachel, 282
Baxi, 471
Baxi, Preet, 471
Bayat, 133, 579
Bayat, Farima Fatahi, 133, 579
Bayati, 835
Bayati, Mohsen, 835
Bayazit, 277
Bayazit, Deniz, 277
Bazarova, 685
Bazarova, Alexandra, 685
Bazarvaani, 217
Bazarvaani, Zuchi, 217
Becker, 551
Becker, Jonas, 551
Beckerman, 444
Beckerman, Bernard, 444
Bedathur, 827
Bedathur, Srikanta, 827
Beger, 263
Beger, Claas, 263
Begus, 777
Begus, Gasper, 777
Bei, 469, 741, 826, 861
Bei, Chao, 741
Bei, Sisong, 861
Bei, Yuanchen, 469, 826
BeichenNing, 538
BeichenNing, BeichenNing, 538
Beinborn, 114
Beinborn, Lisa, 114
Belay, 120
Belay, Tadesse Destaw, 120
Belcak, 899
Belcak, Peter, 899
Belfathi, 860
Belfathi, Anas, 860
Belikova, 685
Belikova, Julia, 685
Belinkov, 90, 230, 771, 887, 894

Belinkov, Yonatan, 90, 230, 771, 887, 894
Belviranli, 571
Belviranli, Mehmet, 571
Belz, 81, 287
Belz, Anya, 81, 287
Ben-Tov, 171
Ben-Tov, Matan, 171
Benaim, 544
Benaim, Sagie, 544
Benajiba, 131, 136, 894
Benajiba, Yassine, 131, 136, 894
Benamara, 198, 811
Benamara, Farah, 198, 811
Benchekroun, 243, 447
Benchekroun, Youssef, 243, 447
Benhalloum, 444
Benhalloum, Amine, 444
Beniwal, 166, 801
Beniwal, Himanshu, 166, 801
Bennett, 226
Bennett, Kerby, 226
Benoit, 784
Benoit, Harold, 784
Benotti, 566
Benotti, Luciana, 566
Benotti;,
Benotti;, Luciana,
Berant, 862
Berant, Jonathan, 862
Berendt, 568
Berendt, Bettina, 568
Berg-Kirkpatrick, 450
Berg-Kirkpatrick, Taylor, 450
Bergeron, 580
Bergeron, Loris, 580
Bernardelle, 800
Bernardelle, Pietro, 800
Bernardi, 890
Bernardi, Raffaella, 890
Bernardy, 868
Bernardy, Laura, 868
Bernas, 439
Bernas, Raphael, 439
Berndt, 121
Berndt, Helena, 121
Berrada, 860
Berrada, Ismail, 860
Bertolazzi, 890
Bertolazzi, Leonardo, 890
Bethard, 480
Bethard, Steven, 480
Bettencourt-Silva, 81
Bettencourt-Silva, Joao H, 81
Beznosikov, 761
Beznosikov, Aleksandr, 761
Bhagat, 150
Bhagat, Snehal, 150
Bhambhoria, 180
Bhambhoria, Rohan, 180
Bhambri, 185
Bhambri, Siddhant, 185
Bhandari, 88, 466
Bhandari, Kushal Raj, 88
Bhandari, Neel, 466
Bhar, 147
Bhar, Swarnadeep, 147
Bharadwaj, 131, 217, 285, 456, 819
Bharadwaj, Akash, 285
Bharadwaj, Shikhar, 217, 456
Bharadwaj, Sujeeth, 131, 819
Bharati, 610
Bharati, Soumedhik, 610
Bhargavi, 158, 598
Bhargavi, Divya, 158, 598
Bhat, 589, 860
Bhat, Savita, 860
Bhat, Suma, 589
Bhathena, 613
Bhathena, Hanoz, 613
Bhatia, 344, 443, 473, 494, 769
Bhatia, Apoorv, 473
Bhatia, Gagan, 494, 769
Bhatia, Rishi, 443
Bhatia, Sumit, 344
Bhatt, 532, 837
Bhatt, Noopur, 532
Bhatt, Rajesh, 837
Bhattacharya, 139, 672
Bhattacharya, Arnab, 139
Bhattacharya, Suparna, 672
Bhattacharya;,
Bhattacharya;, Arnab,
Bhattacharyya, 406
Bhattacharyya, Pushpak, 406
Bhatter, 560
Bhatter, Pavi, 560
Bhuiyan, 822
Bhuiyan, Amran, 822
Bi, 118, 139, 145, 191, 289, 338, 349, 379, 418, 433, 444, 514, 585, 623, 627, 628, 635, 638, 647, 652, 686, 760, 777
Bi, Baolong, 635, 686
Bi, Cheng, 647
Bi, Chloe, 139
Bi, Guanqun, 418, 652
Bi, Jiaxi, 627

Bi, Jichao, 628
Bi, Jinhe, 191, 289
Bi, Keping, 379, 638
Bi, Qi, 514
Bi, Sheng, 145
Bi, Shengjie, 118, 444
Bi, Shuxian, 349
Bi, Wei, 760, 777
Bi, Xinyue, 338
Bi, Xiuli, 623
Bi, Yunying, 433
Bi, Zhen, 585
Bian, 105, 448, 632, 701, 750, 825, 899, 904
Bian, Haonan, 899
Bian, Jia-Wang, 701
Bian, Jiang, 904
Bian, Kaigui, 825
Bian, Yatao, 448
Bian, Yihan, 750
Bian, Yuang, 632
Bian, Yufan, 105
Bianchi, 295
Bianchi, Federico, 295
Bianchini, 634
Bianchini, Monica, 634
Biao, 349
Biao, Tang, 349
Bicker, 156
Bicker, Yves, 156
Biemann, 471, 818
Biemann, Chris, 471, 818
Biester, 627
Biester, Laura, 627
Bigoulaeva, 216
Bigoulaeva, Irina, 216
Bijl, 631
Bijl, Thijmen, 631
BinChen, 310, 345
binchen, 297
BinChen, BinChen, 310, 345
binchen, binchen, 297
Bing, 651, 680
Bing, Lidong, 651, 680
Binkyte, 582
Binkyte, Ruta, 582
Birch, 126, 865, 875
Birch, Alexandra, 126, 865, 875
Bird, 670
Bird, Steven, 670
Bisliouk, 572
Bisliouk, Artem, 572
Bismay, 180
Bismay, Millennium, 180
Biswas, 185, 788
Biswas, Russa, 788
Biswas, Upasana, 185
Bitan, 794
Bitan, Tomer, 794
Bitman, 640
Bitman, Dmitry, 640
Bjerva, 238, 246, 788, 810
Bjerva, Johannes, 238, 246, 788, 810
Black, 246
Black, Elizabeth, 246
Blagoev, 331
Blagoev, Nikolay, 331
Blanco, 241, 441, 554
Blanco, Eduardo, 241, 441, 554
Blanco;,
Blanco;, Eduardo,
Blas, 83
Blas, Leonardo, 83
Blaschke, 167
Blaschke, Verena, 167
Blodgett, 168
Blodgett, Su Lin, 168
Bloem, 495
Bloem, Congfeng Cao; Pengyu Zhang; Jelke,
Blosnich, 478
Blosnich, John R., 478
Bo, 149, 425, 431, 540, 716
Bo, Liefeng, 149, 425, 716
Bo, rite, 431
Bo, Zihao, 540
Boateng, 538
Boateng, Persis, 538
Bobu, 845
Bobu, Andreea, 845
Bocklet, 812
Bocklet, Tobias, 812
Bodenstab, 615
Bodenstab, Nathan, 615
Boen, 441
Boen, Joseph, 441
Boese, 520
Boese, Madeline, 520
Bogdan, 151
Bogdan, Paul, 151
Boglioni, 278
Boglioni, Matteo, 278
Bogomolov, 393
Bogomolov, Egor, 393
Bohdal, 537
Bohdal, Ondrej, 537
Boisnard, 439
Boisnard, Frédéric, 439

Bolatov, 178
Bolatov, Arman, 179
Boldt, 217, 841
Boldt, Brendon, 217, 841
Bollegala, 100, 173, 565
Bollegala, Danushka, 100, 173, 565
Bollman, 246
Bollman, Marcel, 246
Bolourani, 83
Bolourani, Anahita, 83
Bolte, 147
Bolte, Jerome, 147
Bommireddy, 526
Bommireddy, Raviteja, 526
Bonadiman, 598
Bonadiman, Daniele, 598
Bond, 187
Bond, Francis, 187
Bonetti, 200
Bonetti, Tommaso, 200
Borcea, 510
Borcea, Cristian, 510
Borchmann, 252
Borchmann, Lukasz, 252
Bordoloi, 841
Bordoloi, Priyanuj, 841
Borghini, 634
Borghini, Alessia, 634
Borisiuk, 714
Borisiuk, Anna, 714
Boros, 784
Boros, Emanuela, 784
Borovkov, 640
Borovkov, Nikita, 640
Boschi, 637
Boschi, Nicolò, 637
Bosselut, 247, 277, 784
Bosselut, Antoine, 247, 277, 784
Bost, 559
Bost, Xavier, 559
Botcazou, 346
Botcazou, Ivanhoé, 346
boting, 583
boting, Lu, 583
Bouamor, 233
Bouamor, Houda, 233
Bouch, 95
Bouch, Dani, 95
Boudin, 478
Boudin, Florian, 478
Bougie, 311
Bougie, Nicolas, 311
Boulahia, 219
Boulahia, Said Yacine, 219
Bouraoui, 633, 722
Bouraoui, Zied, 633, 722
Bouziane, 806
Bouziane, Yassine, 806
Boyd, 207
Boyd, Ryan L, 207
Boyd-Graber, 102, 117, 467, 547, 819, 869
Boyd-Graber, Jordan, 869
Boyd-Graber, Jordan Lee, 102, 117, 467, 547, 819
Boyle, 763
Boyle, Joseph S., 763
Bozdag, 509, 574
Bozdag, Nimet Beyza, 509, 574
Brad, 121
Brad, Florin, 121
Brahamia, 649
Brahamia, Fatima, 649
Brahim, 860
Brahim, Brakehe, 860
Brahma, 188, 483
Brahma, Dhanajit, 483
Brahma, Maharaj, 188
Brahman, 295
Brahman, Faeze, 295
Bralios, 262
Bralios, Dimitrios, 262
Brambilla;,
Brambilla;, Marco,
Brasoveanu, 193
Brasoveanu, Adrian, 193
Brassard, 847
Brassard, Ana, 847
Bratières, 238
Bratières, Sébastien, 238
Breazeal, 901
Breazeal, Cynthia, 901
Brei, 764
Brei, Anneliese, 764
Brennan, 533
Brennan, Susan, 533
Brinner, 241
Brinner, Marc, 241
Brito, 596
Brito, Iago Alves, 596
Brockman, 458
Brockman, Michael, 458
Broeck, 148
Broeck, Guy Van den, 148
Brown, 513, 625
Brown, Daniel S., 513
Brown, Hannah, 625
Browning, 784

Browning, Nicholas John, 784
Brunello, 800
Brunello, Nicolò, 800
Brusilovsky, 547
Brusilovsky, Peter, 547
Bruss, 492
Bruss, C. Bayan, 492
Bu, 91, 290, 341, 363, 374, 395, 401, 474, 592, 636, 725, 733, 797
Bu, Fan, 474
Bu, Hui, 363
Bu, Jiajun, 91, 374
Bu, Mengyu, 401
Bu, Pi, 341, 636, 733
Bu, Ranran, 395
Bu, Tianpeng, 290
Bu, Yi, 797
Bu, Yixin, 725
Bu, Yuheng, 592
Buaphet, 862
Buaphet, Weerayut, 862
Buda, 813
Buda, Greg, 813
Buendia, 560
Buendia, Diego A. Mares, 560
Bueno, 495
Bueno, Ivo, 495
Buhnila, 580
Buhnila, Ioana, 580
Bui, 121, 179, 207, 463, 485, 584
Bui, Minh Duc, 207
Bui, Nam, 485
Bui, Nghi D. Q., 121
Bui, Roberto Navigli , Iz Beltagy , Heng Ji , Trung, 23
Bui, Trung, 179, 463, 584
Bulat, 291
Bulat, Adrian, 291
Bulatov, 178
Bulatov, Aydar, 179
Bultmann, 766
Bultmann, Hannah, 766
Bunji, 872
Bunji, Kyosuke, 872
Buongiorno, 250
Buongiorno, Steph, 250
Burchell, 862
Burchell, Laurie, 862
Burdick, 815
Burdick, Timothy E., 815
Burgess, 249
Burgess, Paul, 249
Burnaev, 342
Burnaev, Evgeny, 342
Bursztyn, 906
Bursztyn, Victor, 906
Burtsev, 179
Burtsev, Mikhail, 179
Buschhoff, 785
Buschhoff, Jasper Schulze, 785
Buschmeier, 399, 867
Buschmeier, Hendrik, 399, 867
Busso, 110, 904
Busso, Carlos, 110, 904
Butakov, 655
Butakov, Nikolay, 655
Buttery, 570
Buttery, Filip TrhlíkAndrew CainesPaula, 570
Byerly, 423
Byerly, Adam, 423
Bykov, 678
Bykov, Denis, 678
Byrne, 542, 594, 803
Byrne, Bill, 542, 594, 803
Bystroski, 483, 494
Bystroski, Mateusz, 483, 494
Bytyqi, 772
Bytyqi, Erisa, 772
Bösch, 784
Bösch, Fabian, 784
Böther, 784
Böther, Maximilian, 784
Bühler, 495
Bühler, Babette, 495
Bo, 589
Bo, Nguyn Phan Gia, 589

C'horr, 862
C'horr, Lanwenn ar, 862
Caccia, 258
Caccia, Massimo, 258
Cahyawijaya, 468
Cahyawijaya, Samuel, 468
Cai, 88, 148, 151, 201, 209, 215, 271, 274, 280, 284, 286, 289, 303, 311, 315, 316, 333, 335, 337, 339, 340, 345, 348, 359, 397, 410, 414, 422, 435, 436, 438, 456, 461, 501, 506, 513, 533, 539, 552, 562, 569, 599, 626, 630, 636, 643, 647, 654, 658, 665, 691, 708, 719, 721, 724, 735, 739, 746, 757, 758, 761, 781, 783, 789, 798, 811, 848, 863, 899, 900
Cai, Bibo, 626, 643, 658
Cai, Chengkun, 513, 569
Cai, Chenglin, 410
Cai, Fengyu, 88

Cai, Hengyi, 289
Cai, Jianfeng, 315, 461
Cai, Jiawei, 286
Cai, Jinxuan, 339
Cai, Junhui, 438
Cai, Junxian, 630
Cai, Kaitong, 899, 900
Cai, Kunlin, 781
Cai, Kuntai, 333
Cai, Kuo, 414
Cai, Mengzhang, 274, 789
Cai, Min, 746
Cai, Mingzhu, 724
Cai, Pengfei, 456
Cai, Pinlong, 148, 303
Cai, Runxin, 811
Cai, Songcheng, 599
Cai, Wang, 422
Cai, Wei, 397, 761
Cai, Weidong, 757
Cai, Wenrui, 562
Cai, Wentong, 539, 783
Cai, Xiaofang, 735
Cai, Xiaojie, 665
Cai, Xunliang, 151, 201, 209, 271, 311, 335, 345, 435, 501, 636, 647, 691, 708, 739, 758, 863
Cai, Yi, 316, 340, 359, 654
Cai, Yichen, 533, 798
Cai, Yida, 735
Cai, Yishuai, 721
Cai, Yongnuo, 552
Cai, Yujun, 280, 436
Cai, Yunfeng, 337
Cai, Yuyan, 719
Cai, Zhi, 506, 848
Cai, Zhongteng, 215
Cai, Zihuang, 284
Cai, Ziqi, 348
Calais, 588
Calais, Pedro, 588
Calderon, 286
Calderon, Nitay, 286
Calderón, 865
Calderón, Gustavo Cilleruelo, 865
Caliskan, 86, 225, 591
Caliskan, Aylin, 86, 225, 591
Calixto, 103
Calixto, Iacer, 103
Callison-Burch, 825
Callison-Burch, Chris, 825
Calvo-Bartolomé, 694
Calvo-Bartolomé, Lorena, 694
Calyam, 484
Calyam, Prasad, 484
Camacho-Collados, 602, 813
Camacho-Collados, Jose, 602, 813
Cambria, 169, 209, 569
Cambria, Erik, 169, 209, 569
Cambrin, 288
Cambrin, Daniele Rege, 288
Campbell, 109, 118
Campbell, Andrew, 109
Campbell, Joseph, 118
Can, 686
Can, Duy-Cat, 686
Cancedda, 238
Cancedda, Nicola, 238
Cang, 731
Cang, Yueyang, 731
Canova, 784
Canova, Niklas, 784
CAO, 416
Cao, 87, 88, 96, 107, 123, 135, 141, 142, 144, 147, 166, 168, 170, 183, 185, 191, 195, 205, 212, 213, 233, 239, 252, 259, 271, 275, 296, 301, 304, 309, 311, 316, 319, 321, 324, 327, 332, 341, 353, 369, 378, 381, 384, 388, 395, 400, 408, 418, 419, 425, 440, 449, 467, 470, 471, 482, 490, 508, 528, 541, 561, 593, 597, 631, 636, 642, 645, 653, 654, 676, 687, 699, 701, 707, 717, 725, 733, 745, 746, 757, 759, 766, 773, 777, 778, 790, 794, 796, 809, 814, 819, 826, 836, 846, 847, 849, 853, 863, 881, 885, 889, 897, 900
cao, 357
Cao, Anda, 388
Cao, Bin, 316
Cao, Bowen, 541
Cao, Boxi, 369, 653
Cao, Boyu, 778
Cao, Boyuan, 144
Cao, Cong, 304
Cao, Defu, 836
Cao, Gang, 107
Cao, Hu, 881
Cao, Jialun, 252
Cao, Jian, 395
Cao, Jianjian, 309
Cao, Jiawei, 418, 419
cao, jiawei, 357
Cao, Jie, 449
CAO, Jin, 416
Cao, Jinghan, 790
Cao, Jiuxin, 88, 388

Cao, Juan, 96, 170
Cao, Jun, 636
Cao, Junjie, 467
Cao, Kris, 826
Cao, Lang, 593, 773
Cao, Lei, 725
Cao, Liping, 205
Cao, Meng, 471
Cao, Mingzi, 183
Cao, Peng, 699
Cao, Pengfei, 147, 490, 631, 642, 814, 819
Cao, Qian, 777
Cao, Rongyu, 796
Cao, Shaosheng, 645, 701, 745
Cao, Shijie, 900
Cao, Shuang, 889
Cao, Shujun, 212
Cao, Song, 676
Cao, Te, 847
Cao, Tri, 213
Cao, Wanxia, 849
Cao, Weiwei, 381
Cao, Xi, 809
Cao, Xiaochun, 353, 687
Cao, Xuezhi, 271, 311, 408, 425, 863
Cao, Yanan, 319, 324, 327, 490
Cao, Yang, 185, 369, 597
Cao, Yi, 144, 759
Cao, Yifei, 717, 897
Cao, Yingying, 205
Cao, Yinzhi, 378
Cao, Yixin, 561, 654, 676, 746, 757
Cao, Yu-Tong, 846
Cao, Yuan, 141
Cao, Yuanpu, 528, 766
Cao, Yuanyuan, 239
Cao, Yue, 341, 707, 733
Cao, Yukun, 400
Cao, Yunbo, 384, 597
Cao, Yupeng, 166, 233
Cao, Yushi, 142, 275
Cao, Zane, 321, 332
Cao, Zhao, 135, 482, 508, 853
Cao, Zhi, 296
Cao, Zhiyong, 87
Cao, Zhiyu, 123
Cao, Zhonghao, 195
Cao, Zijie, 794
Cao, Ziqiang, 259, 301
Cao, Zixuan, 191
Cao, Zongsheng, 440
Cao, Zongwan, 168
Cao, Zouying, 470, 885
Caplan, 83, 766
Caplan, Eylon, 83, 766
Caragea, 596, 703
Caragea, Cornelia, 596, 703
Carandang, 232, 525
Carandang, Kristine, 232, 525
Cardarelli, 634
Cardarelli, Sara, 634
cardiel, 883
cardiel, amaia, 883
Carenini, 891
Carenini, Giuseppe, 891
Carlsen, 438
Carlsen, Hjalmar Bang, 438
Carmeli, 230
Carmeli, Boaz, 230
Carneiro, 639
Carneiro, Gustavo, 639
Carnerero-Cano, 493, 863
Carnerero-Cano, Javier, 493, 863
Carpuat, 103, 473, 869, 906
Carpuat, Marine, 103, 473, 869, 906
Carrington, 414
Carrington, Patrick, 414
Carroll, 167
Carroll, Mae, 167
Carrotte, 553
Carrotte, Elise, 553
Carvalho, 576
Carvalho, Danilo, 576
Casapu, 464
Casapu, Radu, 464
Cases, 879
Cases, Ignacio, 879
Cashman, 305
Cashman, Dylan, 305
Casin, 232, 525
Casin, Ethan, 232, 525
Cassani, 466
Cassani, Laura, 466
Cassell, 728
Cassell, Justine, 728
Cassotti, 173
Cassotti, Pierluigi, 173
Catanzaro, 887
Catanzaro, Bryan, 887
Cattan, 821
Cattan, Arie, 821
Cava, 667
Cava, Lucio La, 667
Cave, 477
Cave, Laurène, 477
Cavedon, 308

Cavedon, Lawrence, 308
Caverlee, 180, 551
Caverlee, James, 180, 551
Cavigelli, 517
Cavigelli, Lukas, 517
Cazzaniga, 526
Cazzaniga, Alberto, 526
Cedeño, 694
Cedeño, Erick, 694
Celikkanat, 862
Celikkanat, Hande, 862
Cen, 643
Cen, Jun, 643
Cerisara, 104
Cerisara, Christophe, 104
Ceritli, 537
Ceritli, Taha, 537
Cerrato, 207
Cerrato, Mattia, 207
Cha, 179, 283, 299, 351, 458, 893
Cha, Hannah, 458
Cha, Jungyoub, 299
Cha, Meeyoung, 283, 351, 893
Cha, Sungmin, 179
Chadha, 139, 805, 878
Chadha, Aman, 139, 805
Chadha, Ankit, 878
Chaduvula, 419
Chaduvula, Sindhuja, 419
Chae, 461, 553, 905
Chae, Dong-Kyu, 461
Chae, Kyubyung, 553
Chae, Songeun, 905
Chai, 116, 260, 395, 480, 565, 568, 583, 601, 617, 676, 716, 725, 743, 746, 772, 794, 820, 838, 894, 902
Chai, Chengliang, 725
Chai, Hongfeng, 395, 716, 743, 746
Chai, Hua, 676
Chai, Huacan, 601, 794
Chai, Jiajun, 568
Chai, Jianzhe, 565
Chai, Joyce, 772, 902
Chai, Junyi, 617
Chai, Linzheng, 260
Chai, Mingxu, 820, 838
Chai, Tianning, 116
Chai, Yuning, 480
Chai, Yuxiang, 583, 894
Chakma, 241
Chakma, Aunabil, 241
Chakra, 590
Chakra, Chadi Abou, 590
Chakraborty, 82, 83, 274, 342, 344, 442, 469, 471, 853, 881
Chakraborty, Abhijit, 82
Chakraborty, Neeloy, 274
Chakraborty, Supriyo, 853
Chakraborty, Tania, 83
Chakraborty, Tanmoy, 342, 344, 442, 469, 471, 881
Chaksangchaichot,
Chaksangchaichot, Chompakorn,
Chaksangchaichot;,
Chaksangchaichot;, Chompakorn,
Challa, 782
Challa, Jagat Sesh, 782
Challier, 784
Challier, Camille, 784
Chan, 182, 225, 243, 434, 470, 513, 543, 565, 667, 733, 857
Chan, Alvin, 667
Chan, Chee Seng, 857
Chan, Chi-Min, 470, 513, 543
Chan, Chunkit, 243, 733
Chan, Hou Pong, 182, 434, 565
Chan, Tsan Tsai, 225
Chancellor, 462
Chancellor, Stevie, 462
Chandna, 308
Chandna, Bhavik, 308
Chandra, 112, 444
Chandra, Mohit, 112
Chandra, Vikas, 444
Chandrasekaran, 831
Chandrasekaran, Varun, 831
Chandrasekharan, 587
Chandrasekharan, Eshwar, 587
Chang, 83, 98, 107, 110, 116, 128, 163, 182, 207, 212, 217, 249, 253, 262, 264, 279, 311, 323, 351, 361, 375, 381, 407, 413, 445, 455, 456, 460, 467, 483, 504, 540, 544, 547, 581, 604, 623, 636, 639, 662, 680, 696, 748, 753, 758, 767, 788, 829, 838, 844, 851, 879, 890
Chang, Allen, 829
Chang, Angel X, 264
Chang, Angel X., 163
Chang, Buru, 116, 890
Chang, Charles, 83
Chang, Du-Seong, 445
Chang, Edward Y., 323
Chang, Emily, 98
Chang, Ernie, 540
Chang, Ge, 680
Chang, Hwan, 182
Chang, Joohyun, 844

Chang, Joseph Chee, 467
Chang, Kai-Wei, 110, 128, 253, 262, 280, 455, 467, 483, 504, 544, 547, 636, 758, 767, 851, 879
Chang, Kaiyan, 413, 460, 748
Chang, Kalvin, 217, 456
Chang, Li-Wun, 547
Chang, Powei, 249, 581
Chang, Serina, 207, 838
Chang, Shiyu, 604
Chang, Shuochen, 361
Chang, Simyung, 107
Chang, Wanxing, 381
Chang, Xiang, 696
Chang, Xiao-Wen, 212, 788
Chang, Xiaojun, 407
Chang, Yi, 351, 623, 662
Chang, Yu, 375
Chang, Yupeng, 351, 623
Chang, Zhiyuan, 311, 639
Chang, Zihan, 753
Chao, 129, 274, 455, 541, 565, 570, 650, 659, 696, 899
Chao, Fei, 696
Chao, Ke, 570
Chao, Lidia S., 129, 274, 541, 565, 899
Chao, Naipeng, 659
Chao, Pingfu, 650
Chao, Yi-Wen, 455
Chapman, 114, 553
Chapman, Brian E., 114, 553
Chappidi, 245, 905
Chappidi, Srinivas, 245, 905
Charlin, 507
Charlin, Laurent, 507
Charlot, 728
Charlot, Théo, 728
Charmillot, 784
Charmillot, Clément, 784
Charoenpornsawat,
Charoenpornsawat, Paisarn,
Chatterjee, 342, 344, 490, 866
Chatterjee, Anwoy, 342, 344
Chatterjee, Rajen, 866
Chatterjee, Upasana, 490
Chattha, 560
Chattha, Gurshaan, 560
Chaturvedi, 105, 422, 764
Chaturvedi, Prateek, 422
Chaturvedi, Snigdha, 105, 764
Chau, 563
Chau, Luan Tuyen, 563
Chaudhary, 258, 294, 472, 895
Chaudhary, Muawiz, 258
Chaudhary, Vipin, 294, 472, 895
Chaudhuri, 875
Chaudhuri, Soumyabrata, 875
Chaudhury, 536, 781
Chaudhury, Subhajit, 536, 781
Chava, 250
Chava, Sudheer, 250
Chawla, 471, 560, 661, 808
Chawla, Nitesh V, 471, 808
Chawla, Pranit, 661
Chawla, Subodh, 560
Chawla;,
Chawla;, Nitesh V.,
Che, 141, 232, 350, 398, 492, 627, 662, 665, 668, 693, 706, 897
Che, Feihu, 668, 706
Che, Jianpeng, 232
Che, Tong, 141
Che, Wanxiang, 350, 398, 627, 662, 665, 693
Che, Xinmeng, 897
Che, Zora, 492
Cheang, 182
Cheang, Chi Seng, 182
Chechik, 288
Chechik, Gal, 288
Chehbouni, 521
Chehbouni, Khaoula, 521
Chekalina, 633
Chekalina, Viktoriia, 633
CHEN, 259, 456, 555, 716
Chen, 81, 82, 85, 86, 88, 89, 91, 93, 94, 95, 96, 97, 98, 104, 107, 108, 110, 112, 113, 114, 116, 117, 118, 121, 122, 124, 125, 128, 129, 135, 136, 138, 140, 141, 142, 144, 145, 147, 148, 149, 150, 153, 154, 155, 156, 158, 159, 161, 163, 164, 166, 167, 169, 176, 179, 181, 183, 185, 187, 189, 191, 194, 195, 196, 197, 202, 204, 206, 208, 209, 212, 213, 214, 215, 216, 217, 218, 219, 220, 221, 222, 225, 226, 228, 230, 232, 233, 235, 237, 239, 240, 244, 245, 247, 248, 250, 252, 253, 254, 256, 258, 259, 260, 261, 262, 263, 264, 265, 267, 272, 273, 274, 275, 278, 280, 281, 283, 284, 285, 286, 288, 290, 291, 292, 293, 294, 296, 298, 299, 300, 304, 305, 308, 309, 311, 312, 314, 316, 317, 318, 319, 322, 325, 330, 331, 332, 333, 334, 335, 336, 338, 340, 341, 344, 345, 347, 348, 349, 350, 352, 353, 354, 356, 358, 359, 360, 362, 364, 368, 369, 370, 371, 372, 374, 377, 378, 379, 380, 384, 387,

388, 390, 393, 394, 396, 397, 398, 400, 404, 405, 406, 407, 409, 410, 412, 414, 418, 421, 425, 428, 429, 432, 433, 434, 435, 436, 439, 440, 441, 443, 446, 448, 449, 450, 451, 455, 457, 458, 460, 464, 467, 468, 469, 471, 474, 475, 479, 480, 483, 484, 488, 491, 492, 497, 505, 507, 508, 513, 514, 515, 516, 520, 527, 528, 529, 530, 531, 532, 533, 536, 537, 538, 545, 547, 550, 551, 552, 553, 558, 559, 560, 561, 562, 564, 565, 566, 569, 571, 573, 576, 577, 579, 581, 582, 583, 585, 586, 587, 591, 593, 594, 596, 599, 602, 603, 604, 605, 606, 607, 611, 612, 616, 617, 618, 619, 620, 621, 623, 624, 625, 627, 629, 630, 631, 636, 637, 638, 640, 642, 643, 644, 645, 646, 648, 649, 650, 651, 652, 653, 655, 659, 660, 661, 662, 668, 669, 670, 671, 674, 675, 677, 678, 679, 680, 683, 684, 685, 687, 689, 691, 692, 693, 694, 695, 698, 701, 702, 703, 704, 705, 707, 708, 713, 714, 715, 716, 718, 720, 721, 724, 726, 727, 728, 729, 730, 731, 732, 733, 735, 736, 738, 739, 740, 741, 742, 744, 745, 746, 747, 748, 750, 753, 754, 755, 756, 757, 758, 759, 760, 762, 764, 766, 767, 769, 771, 773, 774, 775, 777, 779, 780, 782, 784, 785, 786, 787, 788, 789, 790, 791, 792, 793, 794, 795, 796, 797, 798, 799, 801, 807, 810, 817, 818, 819, 821, 823, 824, 826, 827, 828, 829, 830, 832, 834, 836, 837, 842, 843, 845, 847, 848, 849, 853, 854, 855, 858, 859, 861, 864, 865, 867, 874, 875, 876, 877, 878, 880, 882, 885, 886, 892, 893, 897, 899, 900, 901, 904, 905, 907, 908

chen, 94, 224, 774, 907

Chen, Ada, 582

Chen, Aili, 671

Chen, Alex, 148

Chen, Alvin, 843

Chen, Baiming, 738

Chen, Baiyu, 443

Chen, Baoyou, 599

Chen, Bei, 150

Chen, Beidi, 528, 604

Chen, Beiduo, 104, 457

Chen, Benteng, 774

Chen, Bin, 349, 393, 855

Chen, Bingjin, 724

Chen, Bingsen, 853

Chen, Bo, 874

Chen, Boqi, 110

Chen, Bowen, 441

Chen, Boxing, 273, 587

Chen, Boyuan, 513

Chen, Cailian, 221

Chen, Cen, 773

Chen, Changjian, 284

Chen, Changyou, 616

Chen, Chao, 679

Chen, Chaoyi, 97

Chen, Chen, 455, 530, 714, 755

Chen, Chi, 214, 364, 596

Chen, Ching-Yun Ko; Payel Das; Sihui Dai; Georgios Kollias; Subhajit Chaudhury; Aurelie C. Lozano; Pin-Yu,

Chen, Chong, 750

Chen, Chu-Song, 262

Chen, Chujie, 460

Chen, Chun Wei, 312

Chen, Chunchun, 448

Chen, Chung-Chi, 237

Chen, Cong, 116

Chen, Dajun, 258, 491

Chen, Danlu, 450

Chen, Daojun, 394

Chen, Ding, 733

Chen, Dingwei, 849

Chen, Dongping, 138, 344, 799

Chen, Dubing, 646

Chen, Emily, 566

Chen, Enhong, 298, 338, 340, 436, 791, 819

Chen, Fang Yih, 649

Chen, Fangmin, 538

Chen, Feng, 158, 286, 372, 675, 837

Chen, Gang, 125, 221, 347, 758

Chen, Gengru, 702

Chen, Gongwei, 694

Chen, Gordon, 564

Chen, Guangyi, 878

Chen, Guanhua, 93, 129, 167, 274, 284, 318, 440, 684, 739, 746, 793

Chen, Guanlin, 660

Chen, Guanqiao, 164

Chen, Guanyu, 677

Chen, Guanzhong, 256

Chen, Guirong, 573

Chen, Guohao, 263

Chen, Guoqiang, 400

Chen, Guoxin, 788

Chen, Haibiao, 855

Chen, Haibin, 675, 695, 698

Chen, Haifeng, 300, 585, 771, 824, 837

Chen, Haixu, 314

Chen, Han, 562
Chen, Hanting, 412, 587
Chen, Hao, 216, 348, 576, 645, 674, 728, 767, 785, 807
Chen, Hao WuBin, 662
Chen, Haodong, 407
Chen, Haolan, 397, 759
Chen, Haonan, 89
Chen, Hardy, 142
Chen, Hefei, 425
Chen, Hongyang, 345, 750
Chen, Hongyao, 692
Chen, Hsin-Hsi, 291, 300
Chen, Hu, 435
Chen, Huajun, 317, 334, 354, 398, 436, 624, 653, 704, 708, 715
Chen, Huang, 235
Chen, Huangwei, 91, 374
Chen, Huiyao, 668
Chen, Jia, 492, 602
Chen, Jiaao, 183
Chen, Jiachi, 620, 729
Chen, Jiahao, 897
Chen, Jiahe, 155
Chen, Jiaju, 824
Chen, Jiajun, 94, 661, 867
Chen, Jiali, 316
Chen, Jian, 305, 900
Chen, Jian-Jia, 247
Chen, Jianda, 183
Chen, Jiangjie, 671
Chen, Jianhui, 859
Chen, Jianlyu, 89, 569
Chen, Jianming, 858
Chen, Jianwen, 528
Chen, Jiaoyan, 469, 637
Chen, Jiaqi, 464
Chen, Jiawei, 85, 222, 296, 378, 745, 746, 892
Chen, Jiayuan, 412
Chen, Jie, 625, 727, 746
Chen, Jiecao, 272
Chen, Jiefeng, 678
Chen, Jin, 854
Chen, Jing, 122, 179, 340
Chen, Jingan, 674
Chen, Jingbang, 882
Chen, Jinghong, 594
Chen, Jinghui, 528, 766
Chen, Jingjing, 352, 414
Chen, Jingwu, 769
Chen, Jingxuan, 602
Chen, Jingyuan, 754
Chen, Jingzhou, 239
Chen, Jinhui, 396
Chen, Jinpeng, 221
Chen, Jintai, 155
Chen, Jintao, 312
Chen, Jintong, 550
Chen, Jixuan, 794
Chen, Jue, 796
Chen, Jun, 683, 736
Chen, Junfan, 619
Chen, Junhao, 599
Chen, Junkai, 515, 531
Chen, Junwen, 347
Chen, Junyang, 352
Chen, Justin, 293, 853
Chen, Kai, 159, 239, 300, 379, 579, 695, 735, 750, 794
Chen, Kai-Xin, 467, 547
Chen, Kaixuan, 388
Chen, Kang, 757
Chen, Ke, 125, 497
Chen, Kecheng, 704
Chen, Kedi, 877
Chen, Kehai, 653, 678, 762
Chen, Kehao, 272
Chen, Kejia, 359
Chen, Kejiang, 782
Chen, Kelly, 86
Chen, Kexi, 443
Chen, Kexin, 405
Chen, Keyang, 716, 746
Chen, Keyu, 842
Chen, Kuangwei, 708
Chen, Kun, 360, 739
Chen, Kyle Xue, 85
Chen, Le, 118, 794
Chen, Lei, 483
Chen, Li, 285, 562, 644
Chen, Liangliang, 552
Chen, Liangwei, 88
Chen, Liangyu, 195, 293, 410
Chen, Lida, 325
Chen, Lihu, 847
Chen, Lily, 104
Chen, Liming, 468
Chen, Lin, 623, 759
Chen, Ling, 230, 599, 621, 738
Chen, Lingjie, 81, 219, 826
Chen, Lingwei, 799
Chen, Linqing, 136, 429
Chen, Liqun, 801
Chen, Liuliu, 553
Chen, Liwei, 449, 729
Chen, Lixing, 350

Chen, Liyi, 197
Chen, Long, 775
Chen, Lu, 239
Chen, Lu-Yen Anny, 265
Chen, Lydia, 331
Chen, Lyuhao, 904
Chen, Meiqi, 281
Chen, Mengke, 278
Chen, Mengshi, 680
Chen, Mengzhuo, 112, 475
Chen, Michael, 764
Chen, Michelle Chao, 189
Chen, Min, 113, 747
Chen, Ming, 843
Chen, Ming-Bin, 571
Chen, Minghui, 206
Chen, Mingyang, 585
Chen, Minping, 387, 726
Chen, Mouxiang, 745
Chen, Muhao, 225, 594, 789, 792, 886
Chen, Nancy F., 460, 469
Chen, Nuo, 155, 166, 181, 458, 491
Chen, Nuoyi, 212
Chen, Pei, 281, 533, 798, 817, 875
Chen, Pei-An, 196
Chen, Peilin, 530
Chen, Peng, 107, 353, 687
Chen, Pengchen, 288
Chen, Peter Baile, 828
Chen, Peter Q., 573
Chen, Pin-Yu, 82, 154, 208, 215, 259, 536, 789
Chen, Ping, 508
Chen, Po-Chun, 300
Chen, Qi, 280, 299, 738
Chen, Qian, 114, 372, 780, 865
Chen, Qianben, 678
Chen, Qiang, 514
Chen, Qianglong, 228, 412, 516
Chen, Qifeng, 651, 736
Chen, Qiguang, 404, 627, 693
Chen, Qin, 245, 877
Chen, Qing-Guo, 701, 738
Chen, Qingjing, 560
Chen, Qingyu, 823
Chen, Qinyu, 379
Chen, Qirui, 774
Chen, Qisi, 726
Chen, Qiuwu, 335
Chen, Qiyuan, 155
Chen, Qizhou, 714, 755
Chen, Rachel, 263
Chen, Ran, 651
Chen, Renhai, 161
Chen, Renmiao, 520
Chen, Renqi, 421
Chen, Ronghao, 278, 341, 405, 467, 786, 836, 899
Chen, Rongxin, 319, 685
Chen, Ruihan, 360, 739
Chen, Ruixi, 530
Chen, Ruizhe, 450, 777
Chen, Runhuai, 625
Chen, Runyu, 753
Chen, Ruolin, 98
Chen, Shang-Tse, 218
Chen, Shaowen, 393
Chen, Sheng, 414
Chen, Shenglin, 721
Chen, Shengyu, 771
Chen, Shengyuan, 245
Chen, Shenyang, 153
Chen, Shi, 288
Chen, Shiqi, 95
Chen, Shuai, 421
Chen, Shuang, 280, 689
Chen, Shuangwu, 194, 853
chen, Shuangwu, 774
Chen, Shuhang, 643
Chen, Shuheng, 439
Chen, Shunian, 705
Chen, Shuo, 85, 364
Chen, Shuxu, 404
Chen, Si-Qing, 791
Chen, Sihao, 479
Chen, Siheng, 407, 583, 617
Chen, Sijia, 421
Chen, Siming, 397
Chen, Sirry, 113
Chen, Sirui, 446
Chen, Siyu, 836
Chen, Stephen, 651
Chen, Tailai, 655
Chen, Tao, 309, 569
Chen, Taozhi, 577
Chen, Terrence, 244
Chen, Tiancheng, 784
Chen, Tianle, 619
Chen, Tianlong, 787, 792, 795
Chen, Tianrun, 607
Chen, Tiantian, 707, 708
Chen, Tianyi, 226
Chen, Tiejin, 492, 602
Chen, Tinggang, 545
Chen, Tinghao, 650
Chen, Tong, 738
Chen, Tse-Hsun, 265, 484
Chen, Tuochao, 141, 235

Chen, Valerie, 112, 232
Chen, Wanlin, 330
CHEN, Wanshun, 716
Chen, wanyi, 864
Chen, Wei, 113, 124, 221, 340, 370, 508, 562, 659, 687, 703, 885
Chen, Wei-Lin, 480
Chen, Wei-Rui, 124
Chen, Weicong, 294
Chen, Weihang, 638
Chen, Weixin, 644
Chen, Weizhu, 636, 758
Chen, Wen-Yen, 795
Chen, Wenbo, 829
Chen, Wenchao, 874
Chen, Wenguang, 901
Chen, Wenhu, 161, 599
Chen, Wenliang, 733, 777
Chen, Wentao, 741
Chen, Wenting, 253, 261, 861
Chen, Wenxi, 390, 691, 748
Chen, Winson, 118
Chen, Wuya, 254
Chen, Xi, 163, 166, 233, 300, 604, 618, 630, 718, 730
Chen, Xiang, 629, 845
Chen, Xiangchao, 599
Chen, Xiangying, 474
Chen, Xianyan, 581
Chen, Xiao, 145
Chen, Xiaokun, 384
Chen, Xiaoliang, 176
Chen, Xiaotian, 713
Chen, Xiaoya, 565
Chen, Xiaoyi, 396, 821
Chen, Xiaoyu, 144
Chen, Xie, 220, 390, 404, 432, 455, 691, 748
Chen, Xikang, 760
CHEN, Xilin, 555
Chen, Xin, 142, 335, 371, 505, 739, 753
Chen, Xing, 434
Chen, Xingchi, 380
Chen, Xinghao, 587
Chen, Xinghe, 586
Chen, Xingrun, 767
Chen, Xingyu, 149, 425, 716
Chen, Xingzhou, 128, 757
Chen, Xinjie, 788
Chen, Xinli, 288
Chen, Xinrong, 267, 558
Chen, Xinyang, 653
Chen, Xinyu, 240, 308, 759
Chen, Xiusi, 250, 292, 573, 603
Chen, Xiuying, 82, 117, 129, 202, 475, 507, 662, 876
Chen, Xiwen, 204
Chen, Xu, 335, 349, 892
Chen, Xuan, 612
Chen, Xuanang, 252, 369, 378, 741, 834
Chen, Xuefeng, 136, 702
Chen, Xueshu, 585
Chen, Xunlei, 671
Chen, Xutian, 561
Chen, Xuyang, 786
Chen, YaLong, 374
Chen, Yan, 715
Chen, Yang, 377
Chen, Yangbin, 744
Chen, Yangneng, 362
Chen, Yankai, 393, 488, 527, 537, 893
Chen, Yanyuan, 670
Chen, Yao, 715
Chen, Yaofo, 335
Chen, Yating, 428
Chen, Ye, 359, 705
Chen, Yelin, 169
Chen, Yen-wei, 409
Chen, Yi, 187, 436, 876
Chen, Yi-Jun, 593
Chen, Yi-Ping Phoebe, 849
Chen, Yi-Ting, 196
Chen, Yibiao, 93
Chen, Yibin, 653
Chen, Yibing, 338
Chen, Yida, 88
Chen, Yidong, 433, 720
Chen, Yifan, 560
Chen, Yifei, 451
Chen, Yifu, 114, 780, 832
Chen, Yihan, 378, 705
Chen, YiJie, 669
Chen, Yijie, 701
Chen, Yijun, 114, 655
Chen, Yilin, 732
Chen, YiLong, 290
Chen, Yilong, 304, 591
Chen, Yimeng, 648
Chen, Yimin, 611
Chen, Yiming, 142, 275, 332, 740
Chen, Ying, 559
Chen, Ying-Cong, 286
Chen, Yingcong, 144
Chen, Yingfa, 905
Chen, Yining, 374
chen, Yiqi, 907
Chen, Yiran, 140

Chen, Yixin, 730
Chen, Yiyang, 316
Chen, Yiyi, 810
Chen, Yiyun, 640
Chen, Yong, 191, 340, 436
Chen, Yong-Sheng, 265
Chen, Yongkang, 159
Chen, Yongqiang, 85
Chen, Yu, 527, 716, 886
Chen, Yu-Chi, 265
Chen, Yuan, 605, 847
Chen, Yuanpei, 532
Chen, Yubo, 358, 631, 642
Chen, Yuchong, 212
Chen, Yue, 527
chen, Yue, 94
Chen, Yuen, 121, 156
Chen, Yueyan, 797, 880
Chen, Yufeng, 96, 583
Chen, Yuhan, 212, 264, 380
chen, yuhan, 224
Chen, Yuhang, 344
Chen, Yuhao, 147
Chen, Yujie, 655, 827
Chen, Yulang, 368
Chen, Yulin, 213, 830, 882
Chen, Yun, 93, 167, 284, 684, 739
Chen, Yun-Nung, 218
Chen, Yunji, 418
Chen, Yunmo, 694
Chen, Yuran, 827
CHEN, Yushen, 456
Chen, Yushen, 748
Chen, Yuting, 283
Chen, Yutong, 640
Chen, Yuxiang, 248
Chen, Yuxin, 311, 607, 730
Chen, Yuxing, 214
Chen, Yuyan, 606, 908
Chen, Yuzhong, 787
Chen, Zehua, 779
Chen, Zehui, 372, 623, 759
Chen, ZeLiang, 260
Chen, Zerui, 336
Chen, Zeyuan, 457
Chen, Zhao-Yun, 418
Chen, Zhaoming, 230
Chen, Zhaoyu, 407
Chen, Zhe, 322
Chen, Zhehuai, 529
Chen, Zhen, 204
CHEN, Zheng, 259
Chen, Zheng, 907
Chen, Zhengyu, 135
Chen, Zhengzhang, 585, 771, 824, 837
Chen, Zhenhao, 117, 471, 475
Chen, Zheshu, 705
Chen, Zhetao, 581
Chen, Zhi-Yuan, 432
Chen, Zhichao, 837
Chen, Zhifeng, 687
Chen, Zhihui, 261
Chen, Zhikang, 356
Chen, Zhiming, 583
Chen, Zhipeng, 847
Chen, Zhirong, 460
Chen, Zhirui, 848
Chen, Zhisheng, 467, 728
Chen, Zhitong, 551
Chen, Zhiwei, 755
Chen, Zhiyu, 333
Chen, Zhuang, 418, 652
Chen, Zhumin, 516, 795
Chen, Zhuo, 94, 334, 638, 704, 742
Chen, Zhuoen, 630
Chen, Zhuowei, 449
Chen, Zhuoyue, 185
Chen, Zi, 332
Chen, Ziang, 818
Chen, Zichen, 183
Chen, Zihan, 731, 756
Chen, Ziheng, 195
Chen, Zihong, 209
Chen, Zijian, 161
Chen, Zimo, 108
Chen, Zini, 434
Chen, Zixuan, 217
Chen, Ziyan, 372
Chen, Ziyang, 425
Chen, Ziye, 640
Chen, Ziyi, 457
Chen, Ziying, 740
Chen, Zonghao, 125
Chen, Zulong, 387, 406, 434, 819
Chen;,
Chen;, Wei,
Chen;, Weihua,
Chenchong, 234
Chenchong, Chenchong, 234
Cheng, 86, 97, 132, 136, 138, 158, 159, 164, 175, 192, 195, 213, 228, 243, 269, 283, 290, 292, 297, 300, 312, 339, 344, 346, 355, 370, 371, 379, 380, 382, 410, 411, 412, 413, 414, 417, 418, 419, 423, 434, 437, 438, 445, 446, 451, 458, 459, 475, 514, 516, 550, 568, 573, 580, 584, 591, 595,

619, 620, 623, 629, 635, 638, 643, 647, 648, 649, 654, 655, 670, 679, 686, 693, 697, 708, 721, 724, 737, 740, 743, 750, 751, 753, 754, 756, 764, 784, 786, 791, 801, 802, 803, 822, 832, 837, 854, 867, 885, 886, 896, 909
Cheng, Aijia, 697
Cheng, Albert M. K., 655
Cheng, Dai, 412
Cheng, Dongzhou, 693
Cheng, Fei, 213, 854, 896
Cheng, Gang, 584
Cheng, Guangsheng, 290
Cheng, Hao, 791
Cheng, He, 822
Cheng, Hong, 434
Cheng, Hongtao, 414
Cheng, Hsi-Chun, 312
Cheng, Jiahao, 411
Cheng, Jiale, 132, 164, 786
Cheng, Jiali, 195
Cheng, Jie, 292
Cheng, Jingwei, 370, 382, 438
Cheng, Jintao, 620
Cheng, Jun, 740
Cheng, Junyi, 648
Cheng, Kanzhi, 159, 591
Cheng, Lei, 638
Cheng, Li, 514
Cheng, Longkai, 679
Cheng, Lu, 595
Cheng, Miaomiao, 380
Cheng, Ming, 136
Cheng, Mingjun, 344
Cheng, Mingyue, 418, 419
Cheng, Minhao, 269
Cheng, Mofan, 743
Cheng, Monica, 446
Cheng, Monica Xiao, 175
Cheng, Myra, 86, 458, 764
Cheng, Nan, 580
Cheng, Ning, 803
Cheng, Peng, 297, 573
Cheng, Pengyu, 619, 643
Cheng, Qinyuan, 228, 708
Cheng, Runxi, 516
Cheng, Shanbo, 867
Cheng, Shuang, 413, 423, 750
Cheng, Sijie, 802
Cheng, Sitao, 445
Cheng, Wei, 192, 300, 670, 837
Cheng, Weiyu, 740
Cheng, Xiang, 459
Cheng, Xilong, 346
Cheng, Xin, 158, 379
Cheng, Xinyuan, 243
Cheng, Xize, 339, 832
Cheng, Xu, 635
Cheng, Xueqi, 355, 371, 410, 417, 629, 635, 686, 721, 724, 753, 754, 784
Cheng, Xuxin, 475
Cheng, Yidong, 751
Cheng, Yilin, 886
Cheng, YiMing, 136
Cheng, Yize, 909
Cheng, Yu, 413, 550, 568, 647, 654, 693, 756, 885
Cheng, Zehua, 283, 786
Cheng, Zepeng, 623
Cheng, Zexuan, 414
Cheng, Zheng, 649
Cheng, Zhenglin, 451
Cheng, Zhengwei, 437
Cheng, Zhengxiang, 138
Cheng, Zhi-Qi, 192
Cheng, Zifeng, 97, 801
Cheng, Zihao, 737
Cheng, Ziqiang, 724
Cheng;,
Cheng;, James,
Chengguang_Gan,
Chengguang_Gan, Chengguang_Gan,
Cheok, 523
Cheok, Justin, 523
Cheong, 442
Cheong, Lin Lee, 442
Chernozhukov, 607
Chernozhukov, Victor, 607
Chersoni, 259, 385
Chersoni, Emmanuele, 259, 385
Cheung, 252, 268, 861
Cheung, Shing-Chi, 252
Cheung, William K., 268
Cheung, Yiu-Fai, 861
Chhabra, 120
Chhabra, Anshuman, 120
Chheang, 231
Chheang, Sophie, 231
Chi, 112, 232, 411, 426, 427, 773, 864, 907
Chi, Pham Khanh, 411, 426, 427
Chi, Wayne, 112, 232
Chi, Yizhou, 864
Chi, Zewen, 907
Chi, Zhixiang, 773
Chia, 246
Chia, Ann, 246
Chiang, 112, 125, 126, 165, 174, 181, 225

Chiang, Cheng-Han, 112, 225
Chiang, Chengwei, 174
Chiang, David, 165
Chiang, Meng-Fen, 125, 181
Chiang, Yu-An ChuJen-Ren PongChia-Yao YehMeng-Fen, 126
Chien, 103
Chien, Yen-Che, 103
Chilton, 459, 826
Chilton, Lydia, 459, 826
Chirkunov, 690
Chirkunov, Kirill, 690
Chitale, 837, 866
Chitale, Pranjal, 837
Chitale, Pranjal A, 866
Chittimalli, 356
Chittimalli, Pavan Kumar, 356
Chiu, 495, 606, 617, 743, 757
Chiu, Alec M., 606, 617
Chiu, Guan-Ming, 743, 757
Chiu, Mang Tik, 495
Chizhov, 110, 393
Chizhov, Pavel, 110, 393
Chng, 455, 701
Chng, Eng Siong, 455
Chng, Yong Xien, 701
Cho, 101, 118, 142, 157, 177, 183, 185, 217, 222, 274, 299, 301, 305, 321, 466, 472, 478, 515, 528, 550, 572, 598, 766, 870, 885
Cho, Beomsik, 528
Cho, Cheol Jun, 177
Cho, Deok-Hyeon, 101
Cho, Gyusung, 305
Cho, Haeun, 118
Cho, Hanjun, 472
Cho, Hojun, 301
Cho, Hongjin, 870
Cho, Hyun-woo, 157
Cho, Hyundong Justin, 478
Cho, Hyunjin, 885
Cho, Hyunsoo, 572, 766
Cho, Ilmin, 183
Cho, Jaejin, 142
Cho, Jaemin, 222
Cho, Juwon, 274
Cho, Sangyeon, 466
Cho, Soohyun, 217
Cho, Sukmin, 515
Cho, Sung-Bae, 185, 550
Cho, Sungzoon, 299
Cho, Ye-eun, 321
Cho, Young Min, 598
Chochlakis, 226
Chochlakis, Georgios, 226
Chochoy, 804
Chochoy, Emmanuel, 804
Choi, 104, 107, 118, 121, 124, 160, 161, 168, 177, 183, 185, 190, 210, 215, 217, 229, 237, 250, 266, 267, 275, 292, 446, 456, 467, 510, 511, 515, 547, 552, 564, 592, 594, 610, 639, 730, 741, 823, 828, 833, 846, 865, 890, 896, 905
Choi, Changin, 510
Choi, Changsu, 515
Choi, Chanyeol, 160
Choi, Daon, 161
Choi, Dasol, 121, 267
Choi, Dongwook, 741
Choi, Edward, 210
Choi, Eunsol, 467, 547
Choi, Gyehyeong, 639
Choi, Hanback, 592
Choi, Hyeong Kyu, 237, 896
Choi, Jaehoon, 104, 229, 267
Choi, Jaesik, 890
Choi, Jeonghwan, 124, 511
Choi, Jiho, 592
Choi, Jihun, 292
Choi, Jinho, 730
Choi, Jiwoo, 905
Choi, Jonghyun, 215
Choi, Juhwan, 865
Choi, Junhyuk, 118
Choi, Kenneth, 594
Choi, Kwanghee, 177, 217, 456
Choi, Kyu Sung, 823
Choi, Minjoon, 168
Choi, Sanghyeok, 833
Choi, Stanley Jungkyu, 183, 190
Choi, Sungha, 107
Choi, Yee Man, 250
Choi, Yejin, 446, 552
Choi, Yeonje, 846
Choi, Yong Suk, 275
Choi, YongSuk, 266
Choi, YoungHae, 185
Choi, Yubin, 610
Choi, YunSeok, 183, 564, 828
Cholakkal, 139
Cholakkal, Hisham, 139
Chomphooyod, 388
Chomphooyod, Peerawat, 388
Chong, 343, 568, 731
Chong, Chun Yong, 343
Chong, Kai Fong Ernest, 731
Chong, Wenyue, 568

Choo, 301, 595, 600, 616
Choo, Jaegul, 301, 595, 616
Choo, Jinho, 600
Choshen, 230, 517, 557, 862
Choshen, Leshem, 230, 517, 557, 862
Chou, 110, 398, 806, 870
Chou, Benjamin Shiue-Hal, 398
Chou, Cheng-Ting, 806
Chou, Ju-Chieh, 110
Chou, Tzu-Hsin, 870
Choubey, 609, 789
Choubey, Prafulla Kumar, 609, 789
Choudhary, 873, 906
Choudhary, Ritvik, 873
Choudhary, Sunav, 906
Choudhury, 112, 115, 453, 526, 529, 828, 839
Choudhury, Manan Roy, 526, 828
Choudhury, Munmun De, 112, 115, 529, 839
Choudhury, Sagnik Ray, 453
Chourasia, 616
Chourasia, Shivam, 616
Chousa, 215
Chousa, Katsuki, 215
Chowdhury, 156, 202, 273, 526, 540, 671, 835, 897
Chowdhury, Anal Roy, 671
Chowdhury, Md Mokarram, 202, 540
Chowdhury, Mosharaf, 273
Chowdhury, Sankalan Pal, 156, 835
Chowdhury, Shammur Absar, 897
Chowdhury, Suparno Roy, 526
Christ, 884
Christ, Bryan R., 884
Christian, 808
Christian, Sebastien, 808
Christodorescu, 604, 831
Christodorescu, Mihai, 604, 831
Christodoulopoulos,
Christodoulopoulos, Christos,
Christophe, 808
Christophe, Clement, 808
Chronopoulou, 575
Chronopoulou, Alexandra, 575
Chu, 159, 162, 205, 227, 239, 267, 287, 348, 378, 389, 390, 395, 417, 423, 500, 529, 546, 558, 568, 584, 623, 630, 637, 686, 707, 708, 742, 776, 788, 829, 850, 892, 896
Chu, Chenhui, 227, 529, 829, 896
Chu, Dianhui, 630
Chu, Hui, 417
Chu, Kexin, 287
Chu, Liu, 159
Chu, Meng, 423
Chu, Pei, 239
Chu, Phuong-Anh, 686
Chu, Qi, 707
Chu, Seongyeub, 390
Chu, Tao, 239
Chu, Wei, 637
Chu, Xiangxiang, 378, 500, 623, 776, 892
Chu, Xiaowen, 162
Chu, Xu, 239, 267, 348, 558, 788, 850
Chu, Yuxiang, 568
Chu, Zhaoyang, 389
Chu, Zhendong, 205, 546
Chu, Zheng, 395, 584
Chu, Zhixuan, 708, 742
Chua, 246, 879
Chua, Shearman, 246
Chua, Shei Pern, 879
Chuang, 606, 617
Chuang, Yun-Shiuan, 606, 617
Chuangsuwanich, 169, 388
Chuangsuwanich, Ekapol, 169, 388
Chuanrong, 312
Chuanrong, Wang, 312
Chujiajia, 437
Chujiajia, Chujiajia, 437
Chukamphaeng, 582
Chukamphaeng, Nut, 582
Chun, 182, 456, 601, 667, 842
Chun, Chaewan, 456
Chun, Changwoo, 182
Chun, Jayeol, 842
Chun, Yongchan, 601
Chun, Zachary Choy Kit, 667
Chunara, 86
Chunara, Rumi, 86
Chung, 91, 243, 593, 775, 846
Chung, Ji Hoon, 593
Chung, Jinseok, 91
Chung, Jiwan, 775
Chung, Kawai, 243
Chung, Sejin, 846
Churchill, 881
Churchill, Elizabeth, 881
Cignarella, 471
Cignarella, Alessandra Teresa, 471
Cisneros-Velarde, 614
Cisneros-Velarde, Pedro, 614
Cissé, 818
Cissé, Cheick Tidiani, 818
Civelli, 800
Civelli, Stefano, 800
Clark, 889
Clark, Peter, 889
Clarke, 560

Clarke, Charles L. A., 560
Clavel, 168
Clavel, Chloé, 168
Claye, 439
Claye, Charlotte, 439
Clematide, 817
Clematide, Simon, 817
Clemson, 534
Clemson, Matthew, 534
Clergerie, 253, 282
Clergerie, Éric Villemonte de la, 253, 282
Clifford, 438
Clifford, Gari, 438
Clifton, 667
Clifton, David A., 667
Cline, 560
Cline, Peter, 560
Clérice, 862
Clérice, Thibault, 862
Cocarascu, 193
Cocarascu, Oana, 193
Cocchieri, 468, 835
Cocchieri, Alessio, 468, 835
Cochrane, 494
Cochrane, Logan, 494
Coelho, 449
Coelho, Raquel, 449
Cohan, 166, 191, 241, 248, 269, 410, 623, 783, 784
Cohan, Arman, 166, 241, 248, 269, 410, 623, 783, 784
Cohan, Ziyu Chen; Yilun Zhao; Chengye Wang; Rilyn R. Han; Manasi Patwardhan; Arman,
Cohan;,
Cohan;, Arman,
Cohen, 193, 220, 287, 466, 742, 752, 805, 816, 835
Cohen, Amir David Nissan, 835
Cohen, Bar, 816
Cohen, Eitan, 752
Cohen, Eldan, 287
Cohen, Liran, 220
Cohen, Myke C., 466
Cohen, Shay, 805
Cohen, Shay B, 742
Cohen, Shay B., 193
Cohn, 902
Cohn, Trevor, 902
Coles, 615, 784
Coles, Arlie, 615
Coles, Jonathan, 784
Colley, 563
Colley, Mark, 563
Collier, 137, 390, 503, 510, 572, 845, 864
Collier, Nigel, 137, 390, 503, 510, 572, 845, 864
Collins, 109, 162
Collins, Amanda, 109
Collins, Liam, 162
Colon, 819
Colon, Carlos R., 819
Coltekin, 213
Coltekin, Cagri, 213
Comstock, 106
Comstock, Daniel C, 106
Cong, 281, 282, 476
Cong, Dat Nguyen, 281
Cong, Xin, 282, 476
CongHuYuan, 741
CongHuYuan, CongHuYuan, 741
Conia, 873
Conia, Simone, 873
Conway, 114, 444, 553
Conway, Mike, 114, 553
Conway, Rylan, 444
Cook, 376
Cook, Jonathan, 376
Corallo, 216
Corallo, Giulio, 216
Cordes, 168
Cordes, Ariel, 168
Cornell, 111
Cornell, Samuele, 111
Corney, 881
Corney, David, 881
Corro, 115, 413, 667
Corro, Caio, 413
Corro, Luciano Del, 115, 667
Corso, 234
Corso, Francesco, 234
Cortez, 204
Cortez, Anthony, 204
Cotes, 438
Cotes, Robert, 438
Cotterell, 110, 243, 842, 848
Cotterell, Ryan, 110, 243, 842, 848
Court, 585
Court, Sara, 585
Cowan, 160
Cowan, Greig, 160
Cranenburgh, 395
Cranenburgh, Andreas van, 395
Crawford, 500
Crawford, Joseph, 500
Crespi, 745
Crespi, Noel, 745
Crispino, 127
Crispino, Nicholas, 127

Croce, 814
Croce, Danilo, 814
Cruz, 816
Cruz, Jan Christian Blaise, 816
Cruz-Salinas, 826
Cruz-Salinas, Andres Felipe, 826
Cui, 88, 117, 132, 144, 150, 177, 269, 273, 274, 281, 296, 307, 313, 319, 326, 338, 368, 396, 398, 409, 412, 413, 415, 427, 457, 520, 568, 577, 587, 594, 634, 645, 674, 676, 678, 737, 738, 756, 758, 760, 788, 789, 817, 830
cui, 93
Cui, Bin, 274, 789
Cui, Chengming, 457
Cui, Chi, 594, 817
Cui, Encheng, 307
Cui, Ganqu, 756
Cui, Hang, 577
Cui, He, 676
Cui, Huizi, 326
Cui, Jiacheng, 338
Cui, Jin, 409
Cui, Jinhao, 645, 674
Cui, Junbo, 758
Cui, Justin, 568
Cui, Lixin, 415
Cui, Mingxuan, 117
Cui, Qing, 817
Cui, Qingbin, 150
Cui, Qinrong, 760
Cui, Shaobo, 88
cui, shiwen, 93
Cui, Shiyao, 132, 398, 413, 520
Cui, Tianle, 412
Cui, Tianyu, 738
Cui, Tiehan, 427
Cui, Wendi, 281
Cui, Wenjuan, 319
Cui, Yao, 737
Cui, Yingqian, 269, 296
Cui, Yue, 132, 144, 273, 830
Cui, Yufei, 273, 587, 788
Cui, Yuwen, 177
Cui, Zeyu, 634
Cui, Zhe, 396
Cui, Zhiqing, 313
Cui, Zihao, 678
Cui, Zixin, 368
Cunha, 154
Cunha, Washington, 154
Cuong, 192
Cuong, Van Duc, 192
Curci, 614
Curci, Giuseppe, 614
Curtis, 236
Curtis, Brenda, 236
Cvejoski, 898
Cvejoski, Kostadin, 898
Czyewski, 324
Czyewski, Daniel Cielak; Andrzej,

D, 305
d'Aliberti, 302
d'Aliberti, Liv, 302
D, Arya, 305
Da, 602
Da, Longchao, 602
Dabre, 654
Dabre, Raj, 654
Dagan, 173, 821
Dagan, Ido, 173, 821
Dagli, 156
Dagli, Rishit, 156
Dahan, 180
Dahan, Samuel, 180
Dahlgren, 156
Dahlgren, Isabel, 156
Dahlmeier, 605
Dahlmeier, Daniel, 605
Dai, 84, 85, 87, 105, 107, 112, 137, 153, 175, 241, 260, 265, 269, 273, 278, 283, 289, 295, 300, 379, 386, 414, 422, 424, 434, 470, 485, 500, 505, 508, 513, 515, 527, 543, 545, 561, 591, 666, 692, 717, 725, 734, 748, 779, 786, 858, 884, 889, 892, 902
dai, 650
Dai, Bin, 386
Dai, Bryan, 112, 260, 485, 591
Dai, Chongyuan, 84, 561
Dai, Congren, 137
Dai, Damai, 379
Dai, David, 434
Dai, Enyan, 265
Dai, Genan, 434
Dai, Guang, 889
Dai, Guohao, 153
Dai, Hongliang, 725
Dai, Huteng, 105
Dai, Jun, 508
Dai, Juntao, 85, 470, 513, 543, 748
Dai, Lisen, 269
Dai, Qi, 87
Dai, Qin, 273
Dai, Rui, 500, 892
Dai, Runpeng, 241

Dai, Sicheng, 300
Dai, Sunhao, 289
Dai, Wei, 107, 283, 692, 786
Dai, Xin, 422
Dai, Xinnan, 175
Dai, Xinyu, 414
Dai, Yalun, 278
Dai, Yaxun, 515
dai, Yaxun, 650
Dai, Yihan, 858
Dai, Yinpei, 902
Dai, Yiwei, 505, 717
Dai, Yong, 884
Dai, Yue, 545
Dai, Yusheng, 779
Dai, Yuyang, 527
Dai, Zhaoqian, 666
Dai, Zhenwei, 269
Dai, Zhongxiang, 424, 734
Dai, Zijie, 295
Daille, 478
Daille, Béatrice, 478
Dalmia, 444
Dalmia, Ayushi, 444
Dalvi, 897
Dalvi, Fahim, 897
Daly, 493, 829
Daly, Donnacha, 829
Daly, Elizabeth M., 493
Damirchi, 454
Damirchi, Hamed, 454
Damle, 577
Damle, Sankarshan, 577
Dammu, 549
Dammu, Preetam Prabhu Srikar, 549
Dan, 88, 656, 801, 877
Dan, Soham, 88, 801
Dan, Yuhao, 877
Dan, Zedong, 656
Danescu-Niculescu-Mizil, 478
Danescu-Niculescu-Mizil, Cristian, 478
Dang, 263, 283, 302, 309, 455, 456, 487, 515
Dang, Jianwu, 455, 456
Dang, Jiawei, 302
Dang, Quy-Anh, 515
Dang, Thao M., 283
Dang, Ting, 309
Dang, Trung_Duc_Anh,
Dang, Tuan Anh, 263
Daniel, 396
Daniel, Michael, 396
Danilevsky, 558
Danilevsky, Marina, 558
Dao, 411, 426, 427
Dao, Quoc Phong, 411, 426, 427
Darrell, 116
Darrell, Trevor, 116
Darwish, 494, 769
Darwish, Kareem Mohamed, 494, 769
Das, 99, 199, 245, 287, 291, 420, 489, 538, 595, 658, 787, 791, 800, 843, 846
Das, Anup, 245
Das, Anurag, 291
Das, Arunita, 846
Das, Gyanendra, 420
Das, Kamalika, 595
Das, Mahashweta, 538, 787
Das, Mayukh, 791
Das, Rohan, 800, 843
Das, Soumi, 99, 287
Das, Sunanda, 489
Das, Susmit, 658
Das, Tal Linzen , Sameer Singh , Kamalika, 23
Das, Tonmoy, 199
Dash, 591
Dash, Saloni, 591
Dass, 856
Dass, Jyotikrishna, 856
Dastgheib, 802
Dastgheib, Doratossadat, 802
Datar, 615
Datar, Mayur, 615
Datta, 150, 393, 839
Datta, Rajeev, 150
Datta, Sanjan Baitalik; Rajashik,
Datta, Soumyajit, 839
Davari, 791
Davari, Reza, 791
Davies, 500
Davies, Jim, 500
Davis, 287, 524, 803
Davis, Brian, 287
Davis, Christopher, 803
Davis, Jeremy, 524
Davison, 407
Davison, Richard, 407
de_Wynter,
de_Wynter, Adrian,
Deas, 199
Deas, Nicholas, 199
Debevec, 564
Debevec, Paul, 564
Deeds, 331
Deeds, Kyle, 331
DeGraaff, 95
DeGraaff, Jason, 95

Deguchi, 215
Deguchi, Hiroyuki, 215
Dehghani, 226, 539, 819
Dehghani, Morteza, 226, 539, 819
Dehghantanha, 277
Dehghantanha, Ali, 277
Deilamsalehy, 584
Deilamsalehy, Hanieh, 584
Dejl, 863
Dejl, Adam, 863
Delaval, 432
Delaval, Axel, 432
Delobelle, 568
Delobelle, Pieter, 568
DeLuca, 224
DeLuca, Chad, 224
DeLucia, 558
DeLucia, Alexandra, 558
Demartini, 800
Demartini, Gianluca, 800
Demberg, 225
Demberg, Vera, 225
Dementieva, 222
Dementieva, Daryna, 222
Demszky, 232, 495
Demszky, Dorottya, 232, 495
Denaxas, 81
Denaxas, Spiros, 81
DENG, 636, 721
Deng, 95, 145, 158, 166, 168, 182, 184, 185, 206, 208, 233, 243, 252, 262, 264, 266, 290, 295, 303, 312, 332, 336, 347, 354, 355, 361, 373, 385, 388, 398, 414, 417, 431, 433, 435, 446, 451, 453, 459, 486, 492, 531, 580, 585, 607, 618, 626, 631, 634, 647, 654, 655, 657, 678, 680, 702, 703, 708, 717, 719, 725, 756, 759, 772, 775, 790, 791, 820, 848, 850, 858, 865, 878, 884, 892, 896, 900, 902, 908
Deng, Bingfeng, 347
Deng, Boyi, 858
DENG, CHAO, 636
Deng, Chao, 678
Deng, Cheng, 361, 385, 585, 634, 654
Deng, Chenghao, 492
Deng, Chenlong, 184
Deng, Chong, 865
Deng, Chuanhang, 312
Deng, Dazhen, 703
Deng, Haikang, 850
Deng, HanMing, 790
Deng, Hao, 626
Deng, Haoyun, 486
Deng, Heyuan, 264
Deng, Hokin, 908
Deng, Hongyi, 531
DENG, HUARONG, 721
Deng, Huilin, 185
Deng, Jia, 717
Deng, Jiawen, 373, 388
Deng, Jiaxin, 414
Deng, Jiayi, 719
Deng, Jingcheng, 355, 417
Deng, Jinghao, 848
Deng, Jingyi, 820
Deng, Jiyao, 657
Deng, Juan, 647
Deng, Ken, 702, 900
Deng, Lianghao, 158
Deng, Liqun, 756
Deng, Liting, 618
Deng, Minghang, 252
Deng, Naihao, 459, 902
Deng, Nianchen, 303
Deng, Ningyuan, 95
Deng, Qiwen, 145
Deng, Ruifan, 708
Deng, Ruixuan, 772
Deng, Shenyang, 208
Deng, Shiyuan, 295
Deng, Shumin, 398, 655, 708
Deng, Wanxi, 702
Deng, Weiwei, 431, 896
Deng, Wenbo, 451
Deng, Wenlong, 206, 607
Deng, Xinle, 354, 398
Deng, Xinyan, 336
Deng, Yafeng, 332, 680
Deng, Yang, 168, 182, 290, 446, 580
Deng, Yangdong, 884
Deng, Yaping, 431
Deng, Yayue, 435
Deng, Yimin, 791
Deng, Yonghong, 453
Deng, Yufan, 631
Deng, Yuhao, 725
Deng, Yujia, 892
Deng, Yuntian, 266
Deng, Zezhi, 182
Deng, Zheye, 243, 878
Deng, Zhijie, 433, 759
Deng, Zhiyang, 166, 233
Deng, Zhongfen, 262
Deng, Zihao, 775
Deng;,
Deng;, Yang,

Denkowski, 542
Denkowski, Michael, 542
Dent, 862
Dent, Rasul, 862
Deoras, 787
Deoras, Anoop, 787
Derakhshan, 486
Derakhshan, Ali, 486
Deriu, 784
Deriu, Jan Milan, 784
Dernoncourt, 179, 507, 553, 584, 598
Dernoncourt, Franck, 179, 507, 553, 584, 598
Derr, 901
Derr, Tyler, 901
Derricks, 844
Derricks, Veronica, 844
Desai, 107, 151, 225, 547
Desai, Atrey, 107, 225, 547
Desai, Ishan, 151
Desarkar, 188
Desarkar, Maunendra Sankar, 188
Deshmukh, 484, 800
Deshmukh, Advait, 800
Deshmukh, Aniket Anand, 484
Detrois, 110
Detrois, Yannick, 110
Deutsch, 342
Deutsch, Daniel, 342
Dev, 174, 314, 566, 576, 600
Dev, Gopal, 600
Dev, Hamvir, 314
Dev, Sunipa, 174, 566
Dev, Sunishchal, 576
Devasier, 861
Devasier, Jacob, 861
Devnani, 226
Devnani, Aarya Vijay, 226
Devos, 784
Devos, Arnout, 784
Devoto, 567
Devoto, Alessio, 567
Dewangan, 320
Dewangan, Lipika, 320
Dey, 130, 615, 851
Dey, Gourab, 851
Dey, Soumyadeep, 615
Dey, Subhrasankha, 130
Dezaki, 615
Dezaki, Fatemeh Taheri, 615
Dhage, 682
Dhage, Ria Talsania; Dhruv Ritesh Shah; Sudhir,
Dhar, 263
Dhar, Debjit, 263
Dhawan, 529
Dhawan, Kunal, 529
Dhiman, 189
Dhiman, Prakam, 189
Dhingra, 99
Dhingra, Bhuwan, 99
Dhonnchadha, 287
Dhonnchadha, Elaine Uí, 287
Dhuliawala, 126
Dhuliawala, Shehzaad, 126
Dhyani, 835
Dhyani, Dushyanta, 835
Di, 132, 144, 161, 560, 586, 694, 701
Di, Feng, 560
Di, Haibing, 694
Di, Lê Thanh, 701
Di, Peng, 161
Di, Shimin, 132, 144
Di, Yifeng, 586
Diab, 156, 454, 904
Diab, Mona T., 156, 454, 904
Dialameh, 329
Dialameh, Maryam, 329
Diandaru, 201
Diandaru, Ryandito, 201
Diao, 115, 134, 136, 422, 464, 529, 534, 669, 860, 899
Diao, Enmao, 534
Diao, Jianzhou, 669
Diao, Joshua, 464
Diao, Muxi, 134
Diao, Shizhe, 529, 899
Diao, Xingjian, 115, 136, 860
Diao, Yujun, 422
Dias, 415, 765
Dias, Gaël, 415, 765
Diep, 405, 426, 427
Diep, Nguyen Thi Ngoc, 405, 426, 427
DiIanni, 342
DiIanni, Colten, 342
Dijk, 489
Dijk, Gijs van, 489
Dillon, 447, 837
Dillon, Brian, 447, 837
Ding, 85, 88, 109, 114, 118, 122, 137, 141, 159, 162, 170, 172, 178, 184, 201, 206, 245, 264, 273, 286, 289, 296, 297, 311, 323, 327, 361, 363, 368, 389, 394, 399, 407, 408, 416, 418, 424, 425, 428, 446, 457, 479, 492, 500, 530, 533, 542, 545, 546, 550, 569, 578, 579, 588, 591, 593, 601, 607, 626, 632, 643, 653, 656, 658, 674, 678, 702, 724, 740, 756, 778, 787, 793,

794, 819, 840, 850, 856, 860, 861, 892, 898
Ding, Bin, 368
Ding, Bo, 389
Ding, Bolin, 122, 178, 273, 296
Ding, Bowen, 264
Ding, Caiwen, 118
Ding, Cheng, 245
Ding, Chuang, 363
Ding, Chuanghao, 408
Ding, Deming, 740
Ding, Ding, 88
Ding, Ganggui, 286
Ding, Haiyan, 428
Ding, Hang, 88, 724
Ding, Hongwen, 141
Ding, Hongxin, 264, 850
Ding, Jiakun, 424
Ding, Jiaxin, 361
Ding, Junhua, 159
Ding, Junpeng, 545, 550
Ding, Kaize, 172, 327, 457, 793, 860
Ding, Keyan, 530, 653
Ding, Keyu, 569
Ding, Lei, 898
Ding, Li, 418
Ding, Liang, 297, 656, 819
Ding, Lizhong, 323
Ding, Meidan, 861
Ding, Mingqian, 407
Ding, Mingqiang, 394
Ding, Mucong, 492
Ding, Nai, 184
Ding, Ningning, 85
Ding, Peng, 408, 425
Ding, Qian, 533
Ding, Qiuyu, 593
Ding, Renjie, 601, 794
Ding, Shengyuan, 579, 794
Ding, Shiyao, 416
Ding, Shuoyang, 542, 588
Ding, Wei, 674
Ding, Wenbiao, 264
Ding, Wenbo, 479
Ding, Wenxuan, 170, 446
Ding, Xiao, 399, 626, 643, 658
Ding, Xuanwen, 546
Ding, Xueying, 162, 446
Ding, Yifei, 311
Ding, Yifeng, 787
Ding, Yihao, 109, 545
Ding, Yiran, 840
Ding, Yiwen, 201
Ding, Youlong, 702
Ding, Yuanhao, 778
Ding, Yueyang, 500, 892
Ding, Yuyang, 632, 678
Ding, Yuzhe, 114, 856
Ding, Zenghui, 756
Ding, Zewen, 137
Ding, Zhihao, 578
Ding, Zichen, 159, 591
Ding, Zifeng, 289
Ding, Zixin, 206, 545, 607
Dinh, 298, 364
Dinh, Anh Trac Duc, 364
Dinh, Tu Anh, 298
Dip, 211
Dip, Sajib Acharjee, 211
Dipta, 152, 579, 896
Dipta, Shubhashis Roy, 152, 579, 896
Divakaran, 199
Divakaran, Dinil Mon, 199
Dividino, 250
Dividino, Renata, 250
Divriotis, 865
Divriotis, Konstantinos, 865
Divyanshu_Bhatt,
Divyanshu_Bhatt, Divyanshu_Bhatt,
Dligach, 822
Dligach, Dmitriy, 822
Dmitry, 342
Dmitry, Redko, 342
Dmonte, 484
Dmonte, Alphaeus, 484
Do, 102, 175, 351, 639, 803
Do, Dai, 351
Do, Heejin, 175, 803
Do, Thanh-Toan, 639
Do, Tuan, 102
Doan, 102, 364, 485
Doan, Luong, 102
Doan, Tung, 485
Doan, Vinh Cong, 364
Dobler, 862
Dobler, Konstantin, 862
Dogan, 247
Dogan, Mustafa, 247
Dognin, 840
Dognin, Pierre, 840
Doherty, 529
Doherty, Gavin, 529
Doi, 832
Doi, Kosuke, 832
Doifode, 611
Doifode, Vinayak K, 611

Doimo, 526
Doimo, Diego, 526
Doll, 785
Doll, Niclas, 785
Dollis, 596
Dollis, Julia Soares, 596
Domenico, 395
Domenico, Cecilia Nicole Di, 395
Don-Yehiya, 557
Don-Yehiya, Shachar, 557
DONG, 204
Dong, 92, 94, 97, 100, 109, 143, 170, 175, 180, 186, 209, 211, 228, 233, 239, 245, 260, 286, 290, 322, 333, 336, 340, 347, 348, 359, 372, 375, 376, 383, 384, 399, 406, 445, 448, 449, 451, 480, 481, 486, 487, 504, 510, 531, 532, 543, 551, 556, 572, 574, 577, 583, 584, 593, 595, 611, 631, 662, 671, 683, 717, 732, 741, 745, 746, 751, 759, 765, 770, 771, 773, 777, 788, 796, 797, 836, 845, 875, 878, 886, 893, 900, 901, 907
Dong, Bin, 97, 741
Dong, Daoyi, 797
Dong, Dota Tianai, 448
Dong, Fei, 770
Dong, FengXian, 340
Dong, Guanting, 383, 451
Dong, Hande, 745, 746
Dong, Hang, 109, 211
Dong, Hao, 532
Dong, Haonan, 445, 788
Dong, Haoyu, 765, 773, 886
Dong, Hejun, 239
Dong, Heng, 584, 901
Dong, Jianshuo, 92, 583
Dong, Jie, 777
Dong, Jin Song, 333, 487
Dong, Junhao, 486
Dong, Junnan, 245, 449
Dong, Kaiwen, 595
Dong, Kuicai, 186, 384
Dong, Li, 900, 907
Dong, Liangliang, 751
Dong, Nanqing, 209
Dong, Peidi, 233
Dong, Qihua, 143
Dong, Qing, 671
Dong, Qunxi, 97
Dong, Shikai, 384
Dong, Shoubin, 336
Dong, Shuai, 228, 543
Dong, Su, 245
Dong, Tianyu, 94, 406, 732
Dong, Wang, 683
Dong, Wei, 504, 593
Dong, Wenchao, 893
Dong, Xiangjue, 180, 551
Dong, Xiaoning, 886
Dong, Xiaoyi, 239
Dong, Xin, 347, 611
Dong, Xinshuai, 771, 878
Dong, Xiyao, 290
Dong, Xuan, 662
Dong, Xuanyu, 886
DONG, XUANZHAO, 204
Dong, Xue Jiang; Tianyu Zhang; Ge Li; Mengyang Liu; Taozhi Chen; Zhenhua Xu; Wenpin Jiao; Zhi Jin; Yihong,
Dong, Xunde, 97
Dong, Yi, 170, 376, 717
Dong, Yichen, 406
Dong, Yihong, 97, 260, 531, 577, 796
Dong, Yijiang River, 510, 572, 845
Dong, Yilu, 556
Dong, Yue, 481
Dong, Yuhao, 100
Dong, Yunfang, 574
Dong, Yunyun, 322
Dong, Yushun, 836
Dong, Yuxiao, 286
Dong, Yuyang, 480
Dong, Zhengyuan, 399
Dong, Zhenhua, 372, 759
Dong, Zhenjiang, 359
Dong, Zhuangzhi, 631
Dong, Ziwei, 875
Dong, Zixuan, 348
Dong, Ziyang, 375
DongmingJiang, 503
DongmingJiang, DongmingJiang, 503
Donnel, 842
Donnel, Timothy J. O', 842
Dontsov, 342
Dontsov, Alexey, 342
Dorfman, 385
Dorfman, Ron, 385
Dorn, 117
Dorn, Liam, 117
Doshi-Velez, 294
Doshi-Velez, Finale, 294
Doss, 271, 894
Doss, Srikanth, 271, 894
Dotlacil, 193
Dotlacil, Jakub, 193

Dou, 89, 95, 103, 108, 178, 184, 310, 312, 337, 382, 387, 451, 508, 511, 538, 561, 699, 700, 717, 728, 740, 779, 820, 838, 897
Dou, Dejing, 178, 312
Dou, Longxu, 95
Dou, Shihan, 108, 382, 387, 561, 717, 740, 820, 838, 897
Dou, Weibei, 779
Dou, Yingtong, 538
Dou, Zhicheng, 89, 103, 184, 310, 337, 451, 508, 511, 699, 700, 728
Dousti, 737
Dousti, Mohammad Javad, 737
Downey, 889
Downey, Doug, 889
Downing, 585
Downing, Lara, 585
Doruöz, 197
Doruöz, A. Seza, 197
Dragomir, 121
Dragomir, Alexandra, 121
Dragut, 460, 586, 596
Dragut, Eduard, 460, 586, 596
Dranfield, 763
Dranfield, Anthony, 763
Dras, 458, 652, 682
Dras, Mark, 458, 652, 682
Drechsel, 772
Drechsel, Jonathan, 772
Dredze, 257, 518, 558, 781
Dredze, Mark, 257, 518, 558, 781
Drescher, 784
Drescher, Lukas, 784
Dreyer, 99
Dreyer, Markus, 99
Driggs-Campbell, 274
Driggs-Campbell, Katherine, 274
Drozdov, 438
Drozdov, Nikita, 438
DU, 362
Du, 103, 119, 121, 151, 176, 178, 184, 207, 235, 244, 259, 262, 271, 272, 276, 297, 299, 317, 319, 330, 331, 333, 338, 353, 360, 369, 391, 407, 413, 417, 418, 422, 436, 439, 442, 451, 463, 470, 474, 482, 496, 507, 509, 531, 535, 566, 584, 596, 601, 623, 625, 626, 627, 629, 630, 632, 643, 646, 647, 655, 677, 684, 693, 695, 702, 709, 713, 718, 726, 729, 738, 748, 756, 771, 847, 868, 877, 885, 892, 902
du, 357
Du, Albert, 331
Du, Bo, 184
Du, Chao, 535
Du, Chaochao, 391
Du, Cunxiao, 535, 738
Du, Dazhao, 496
Du, Fanyi, 756
Du, Fei, 422
DU, Guodong, 362
Du, Hang, 470
Du, Haohua, 713
Du, He, 695
Du, Jian, 407
Du, Jiangshu, 262
Du, Jiazhen, 360
Du, Jusen, 647
Du, Kevin, 771
Du, Kounianhua, 892
Du, Li, 626, 643
Du, Linfeng, 507, 509, 584
Du, Linkang, 297
Du, Linxuan, 632
Du, Lun, 317, 436, 655
Du, Meixi, 103
Du, Mengnan, 244, 442, 892, 902
Du, Mingxuan, 729
Du, Mingyuan, 847
Du, Mingzhe, 531
Du, Minxin, 207
Du, Nan, 276
Du, Qianyun, 417
Du, Quan, 413, 748
Du, Ruofei, 601
Du, Sean, 151
Du, Tianyu, 121, 566, 877
Du, Wei, 625
Du, Weihong, 693
Du, Weihua, 272
Du, Wenjie, 259
Du, Wenyu, 627
Du, Wumei, 718
Du, Xianglong, 702
Du, Xiaoning, 299
Du, Xiaoyong, 630
Du, Xiaoyu, 338
Du, Xingbo, 596
Du, Xingjian, 482
Du, Xinya, 333, 463, 709
Du, Yajun, 176
du, yanrui, 357
Du, Yao, 353, 439
Du, Yaxin, 407
Du, Yi, 319, 369
Du, Yichao, 677, 684
Du, Yifan, 119

Du, Yilun, 868
Du, Yiming, 451, 726
Du, Yingpeng, 235
Du, Yiqi, 629
Du, Yixuan, 271
Du, Yuhao, 474
Du, Yuntao, 178
Du, Zhaocheng, 330, 623
Du, Zhengxiao, 646
Du, Zhuoyun, 885
Du, Zidong, 418
Du, Ziwei, 625
Dua, 552, 862
Dua, Karan, 552, 862
Duan, 107, 112, 150, 186, 268, 276, 289, 290, 301, 334, 339, 355, 363, 377, 417, 504, 547, 579, 636, 644, 699, 701, 718, 727, 776, 794, 821, 868
Duan, Boyan, 636
Duan, Chengqi, 301
Duan, Feiyu, 868
Duan, Haodong, 579, 794
Duan, Hongtao, 290
Duan, Huiyu, 339
Duan, Jia, 699
Duan, Jinhao, 334
Duan, Lunhao, 701
Duan, Niping, 150
Duan, Ranjie, 504
Duan, Runfan, 718
Duan, Shaohua, 377
Duan, Shaoming, 776
Duan, Shitong, 289
Duan, Sijing, 276
Duan, Xiaoni, 107
Duan, Xingyi, 363
Duan, Xufeng, 821
Duan, Yi, 268
Duan, Zenghao, 355, 417
Duan, Zhangqi, 186, 547
Duan, Zhen, 727
Duan, Zhijian, 112
Duan, Zhiyi, 644
Duboue, 238
Duboue, Pablo, 238
Duc, 469
Duc, Pham Minh, 469
Duca, 728
Duca, Giovanni, 728
Dudai, 385
Dudai, Amir, 385
Duddu, 556
Duddu, Vasisht, 556
Duderstadt, 226, 800
Duderstadt, Brandon, 226, 800
Dufour, 478
Dufour, Richard, 478
Dukkipati, 570
Dukkipati, Harish, 570
Dumoulin, 269
Dumoulin, Benoit, 269
Dundua, 609
Dundua, Mariam, 609
Duo, 401
Duo, Zehua, 401
Duong, 102, 340
Duong, Long, 340
Duong, Ngan, 102
Duong, Thao, 102
Dupoux, 243, 447
Dupoux, Emmanuel, 243, 447
Duraiswami, 127, 140
Duraiswami, Ramani, 127, 140
Durme, 154, 313, 555
Durme, Benjamin Van, 154, 313, 555
Durrani, 897
Durrani, Nadir, 897
Dusek, 138, 411
Dusek, Ondrej, 138, 411
Dutta, 143, 216, 263, 520, 612, 624, 839
Dutta, Archan, 520, 612
Dutta, Arka, 839
Dutta, Odwitiyo, 624
Dutta, Saikat, 263
Dutta, Subhabrata, 143, 216
Dutta, Sujan, 839
Duek, 872
Duek, Ondej, 872
Dwivedi, 625
Dwivedi, Shagun, 625
Dwyer, 592
Dwyer, Dominic, 592
Dzenhaliou, 784
Dzenhaliou, Daniil, 784

E, 545, 550, 617, 796
E, Haihong, 545, 550, 796
E, Weinan, 617
Earl, 844
Earl, Allison, 844
Eberhardt, 529
Eberhardt, Steffen T., 529
Ebert, 141
Ebert, Victoria, 141
Ebing, 825
Ebing, Benedikt, 825

Ebling, 903
Ebling, Sarah, 903
Ebner, 799, 871
Ebner, Seth, 799, 871
Echeveste, 174, 566
Echeveste, M Emilia, 174, 566
Echizen, 867
Echizen, Isao, 867
Eckman, 148
Eckman, Stephanie, 148
Eden, 526, 783
Eden, Lilach, 526, 783
Edwards, 796
Edwards, Nicholas, 796
Edy, 169
Edy, Antoine, 169
Efat, 119, 567
Efat, Azher Ahmed, 119, 567
Egami, 452
Egami, Shusaku, 452
Eger, 107, 622
Eger, Steffen, 107, 622
Ehara, 618
Ehara, Yo, 618
Ehrmann, 784
Ehrmann, Maud, 784
Ehsan, 826
Ehsan, Upol, 826
Eickhoff, 557, 837, 894
Eickhoff, Carsten, 557, 837, 894
Eiger, 585
Eiger, Sivan, 585
Ein-Dor, 553, 771
Ein-Dor, Liat, 553, 771
Ekbal, 644, 649
Ekbal, Asif, 644, 649
Ekimov, 770
Ekimov, Leonid, 770
El-Ebshihy, 715
El-Ebshihy, Alaa, 715
El-Sayed, 468, 860
El-Sayed, Wesam, 468, 860
Elbadry, 95, 442, 527, 555
Elbadry, Rania, 95, 442, 527, 555
Elbahnasawi, 649
Elbahnasawi, Hoor, 649
Elbouardi, 178, 589, 864
Elbouardi, Bilal, 179, 589, 864
Eletter, 236
Eletter, Saadeldine, 236
Elgamal, 444
Elgamal, Tarek, 444
Elkhidir, 860
Elkhidir, Khalid, 860
Ellis, 263
Ellis, Kevin, 263
Elmadany, 102, 860
Elmadany, AbdelRahim A., 102, 860
Elmahallawy, 903
Elmahallawy, Mohamed, 903
Elneima, 841
Elneima, Ashraf Hatim, 841
Elozeiri, 236, 555
Elozeiri, Kareem, 236, 555
Elsayed, 649
Elsayed, Tamer, 649
Elsetohy, 816, 869
Elsetohy, Alaa, 816, 869
Elsner, 585
Elsner, Micha, 585
Eltabakh, 769
Eltabakh, Mohamed Y., 769
Eltanbouly, 649
Eltanbouly, Sohaila, 649
Elzeky, 864
Elzeky, Kareem, 864
Emami, 165, 219, 250, 808, 824
Emami, Ali, 165, 219, 250, 808, 824
Emde, 156, 509
Emde, Cornelius, 156, 509
Emi, 233
Emi, Bradley, 233
Emmanuel, 98
Emmanuel, Imanigirimbabazi,
Emon, 211
Emon, Muhit Islam, 211
Emura, 242
Emura, Rei, 242
Encinas, 856
Encinas, John Paul Martin, 856
Endo, 863
Endo, Rei, 863
Eo, 908
Eo, Sugyeong, 908
erchao.zec, 643
erchao.zec, erchao.zec, 643
Eriksson, 444
Eriksson, David, 444
Ermilova, 685
Ermilova, Alina, 685
Ernst, 292
Ernst, Neil, 292
Ersoy, 331
Ersoy, Oguzhan, 331
Eryiit, 758
Eryiit, Gülen, 758

Eslami, 678
Eslami, Sedigheh, 678
Estève, 712
Estève, Louis, 712
Eswaran, 370
Eswaran, Mathan Kumar, 370
Etoglu, 557
Etoglu, Ece Sena, 557
Etori, 98
Etori, Naome A, 98
Eum, 823
Eum, Heeseong, 823
Eun, 533
Eun, Jinsu, 533
Evanini;,
Evanini;, Keelan,
Evans, 819, 821, 891
Evans, Benjamin, 819
Evans, David, 891
Evans, James, 821
Everaert, 175
Everaert, Dante, 175
Everson, 112
Everson, Kevin, 112
Ewais, 818
Ewais, Ahmed, 818
Eyal, 585
Eyal, Matan, 585
Ezzi, 380
Ezzi, Burhanuddin, 380

Fadaee, 826
Fadaee, Marzieh, 826
Fadeeva, 126, 233, 777
Fadeeva, Ekaterina, 126, 233, 777
Faghih, 909
Faghih, Kazem, 909
Fahim, 371
Fahim, Md, 371
Faili, 737
Faili, Heshaam, 737
Faisal, 94, 135
Faisal, Malik Hammad, 94
Faisal, Wahid, 135
Falevskiy, 655
Falevskiy, Yuri, 655
Fan, 97, 138, 145, 161, 164, 169, 192, 212, 228, 243, 254, 259, 271, 280, 316, 328, 334, 335, 350, 390, 403, 406, 418, 427, 430, 440, 473, 496, 519, 568, 573, 575, 588, 590, 603, 611, 617, 626, 638, 646, 647, 662, 675, 691, 693, 699, 701, 703, 715, 717, 721, 733, 738, 745, 754, 782, 783, 784, 788, 792, 797, 802, 839, 872, 900, 909
fan, 357, 546
Fan, Baoyu, 717
Fan, Changxuan, 496, 647, 782
Fan, Chenrui, 909
Fan, Chuanliu, 259
Fan, Cunhang, 699
Fan, Dongyang, 784
Fan, Dongzhe, 783
fan, fenglei, 357
Fan, Guangrui, 328
Fan, Guisheng, 715
Fan, Haozhi, 334
Fan, Hehe, 693
Fan, Heng, 872
Fan, Jiaxin, 138
Fan, Jing, 316
Fan, Jingru, 617
Fan, Jingxuan, 603
Fan, Kai, 788
Fan, Kaixuan, 280
Fan, Lin, 909
Fan, Lishui, 745
Fan, Liwei, 228
Fan, Lixin, 390
Fan, Miao, 701
Fan, Shiming, 430
Fan, Shiqi, 418
Fan, Shiying, 792
Fan, Simin, 784
Fan, Siqi, 638
Fan, Tao, 390, 519, 646, 733
Fan, Tianyu, 271, 588
Fan, Wei, 161, 575, 611
Fan, Wenqi, 316, 473, 573, 662
Fan, Xiangjun, 192, 588
Fan, Xiao, 145
Fan, Xiaoran, 717, 839
Fan, Xing, 802
fan, Xing, 546
Fan, Xuanbo, 675
Fan, Yifan, 212
Fan, Yijia, 590, 900
Fan, Yixiang, 243
Fan, Yixing, 721, 754
Fan, Yongqi, 254, 568
Fan, YuanTao, 350
Fan, Yuchun, 335, 626
Fan, Yue, 403, 427, 440
Fan, Yutao, 797
Fan, Yuwei, 406
Fan, Yuyin, 703

Fan, Zhaoran, 738
Fan, Zhaoxin, 169, 733
Fan, Zhiyuan, 97, 164, 691
Fanelli, 226
Fanelli, Andrea, 226
Fang, 122, 148, 153, 163, 181, 201, 208, 256, 264, 289, 301, 324, 327, 329, 339, 349, 372, 374, 386, 394, 395, 401, 407, 417, 418, 420, 426, 436, 439, 488, 490, 534, 540, 562, 579, 621, 627, 638, 645, 651, 655, 662, 676, 681, 687, 691, 706, 707, 736, 743, 759, 779, 785, 794, 838, 849, 850, 858, 880
Fang, Biaoyan, 201
Fang, Bozhen, 785
Fang, Chuyu, 163
Fang, Dong, 707
Fang, Fang, 324, 327, 490
Fang, Feihao, 256
Fang, Feiteng, 849
Fang, Gexiang, 651
Fang, Han, 418
Fang, Hao, 349, 662
Fang, Hengjie, 627
Fang, Hongjian, 706
Fang, Jinyuan, 426
Fang, Jizhan, 655
Fang, Jun, 676
Fang, Junfeng, 645
Fang, Kuofei, 374
Fang, Liancheng, 122, 488
Fang, Meng, 439, 621
Fang, Quntian, 401
Fang, Rongyao, 301, 339
Fang, Sen, 534
Fang, Tianqing, 289, 436, 687, 838
Fang, Tongcheng, 153
Fang, Wei, 880
Fang, Xi, 148
Fang, Xiangzhong, 562
Fang, Xinyu, 579, 794
Fang, Xinyue, 401
Fang, Xuezhi, 638
Fang, Yan, 736
Fang, Yang, 395
Fang, Yangyi, 329, 417, 691, 743
Fang, Yi, 858
Fang, Yuan, 181
Fang, Yue, 264, 850
Fang, Yujie, 208
Fang, Yushen, 394, 407
Fang, Zhen, 372, 681, 759
Fang, Zheng, 324, 386, 540
Fang, Zhongwang, 779
Fang, Zhouhua, 420
Fanghao, 128
Fanghao, Fanghao, 128
Farahbakhsh, 745
Farahbakhsh, Reza, 745
Farajian, 705
Farajian, Amin, 705
Farganus, 494
Farganus, Julia, 494
Farooq, 419
Farooq, Azib, 419
Faruqui, 444
Faruqui, Manaal, 444
Fastowski, 878
Fastowski, Alina, 878
Fatahibaarzi, 124, 356
Fatahibaarzi, Ata, 124, 356
Fateh, 355
Fateh, Fawad, 355
Fatemi, 463
Fatemi, Sorouralsadat, 463
Fatin, 549
Fatin, Shamit, 549
Faulkner, 156
Faulkner, Ryan, 156
Faye, 219
Faye, Bilal, 219
Faysse, 169
Faysse, Manuel, 169
Fazel-Zarandi, 444
Fazel-Zarandi, Maryam, 444
Fazli, 131, 219
Fazli, Mehrdad, 131
Fazli, Pooyan, 219
Fedorenko, 447
Fedorenko, Evelina, 447
Fedorov, 444
Fedorov, Igor, 444
Fei, 164, 228, 250, 428, 544, 708, 830, 856
Fei, Ben, 250
Fei, Chi, 830
Fei, Fan, 544
Fei, Hao, 428, 856
Fei, Yulin, 164
Fei, Zhaoye, 228, 708
Feizi, 909
Feizi, Soheil, 909
Feldhus, 826
Feldhus, Nils, 826
Feldman, 467, 878
Feldman, Naomi H, 878
Feldman, Sergey, 467

Feliren, 862
Feliren, Vicky, 862
Felka, 505
Felka, Patrick, 505
Feltman, 207
Feltman, Scott M, 207
Feng, 89, 99, 114, 134, 136, 138, 158, 159, 166, 170, 180, 191, 239, 259, 261, 268, 280, 303, 304, 311, 321, 325, 326, 328, 336, 349, 353, 360, 380, 381, 389, 391, 397, 401, 405, 413, 422, 425, 430, 440, 444, 446, 461, 471, 474, 475, 507, 514, 529, 534, 547, 548, 549, 550, 554, 571, 604, 619, 628, 629, 643, 645, 648, 650, 653, 656, 657, 663, 664, 668, 671, 672, 673, 676, 678, 686, 687, 688, 693, 697, 705, 706, 724, 730, 739, 743, 750, 752, 757, 785, 786, 788, 795, 810, 815, 844, 846, 852, 858, 872, 874, 879, 902, 904, 908, 909
feng, 271
Feng, Andrew Zhuoer, 909
Feng, Bin, 191
Feng, Bo, 471
Feng, Chao, 425
feng, chao, 271
Feng, Chengxiao, 239
Feng, Chong, 705
Feng, Dawei, 389
Feng, Erhu, 397
Feng, Fangming, 650, 678
Feng, Fuli, 303, 311, 325, 349, 648, 852, 858
Feng, Haibo, 380
Feng, Hao, 534
Feng, Hongwei, 687
Feng, Huanxiang, 475
Feng, Jiangtao, 268
Feng, Jianjiang, 629
Feng, Jiaqi, 619
Feng, Jiaye, 904
Feng, Jie, 604
Feng, Junlan, 321, 430, 657, 678, 752
Feng, Junning, 815
Feng, Kaituo, 280, 743
Feng, Kehua, 653
Feng, Lang, 158
Feng, Ling, 571
Feng, Mengling, 261
Feng, Mingkuan, 656, 668, 706
Feng, Mingxiao, 461
Feng, Peinan, 413
Feng, Pinyuan, 908
Feng, Qiming, 724
Feng, Shangbin, 170, 446, 548, 550, 554
Feng, Shanshan, 664
Feng, Shaoxiong, 405
Feng, Shengyu, 444
Feng, Shi, 326, 422, 474, 643, 663, 671, 673, 730
Feng, Shiyang, 440
Feng, Shutong, 89
Feng, Siyuan, 750
Feng, Tao, 303, 643
Feng, Tianjian, 785
Feng, Tiantian, 529
Feng, Xiachong, 99, 328, 353, 360, 693
Feng, Xiang, 507
Feng, Xiaocheng, 99, 328, 353, 360, 381, 697, 739
Feng, Xiaoyun, 846
Feng, Xinshun, 902
Feng, Yalin, 534
Feng, Yang, 401, 688
Feng, Yansong, 180, 549, 810
Feng, Yaoxuan, 874
Feng, Yeqi, 672
Feng, Yi, 336
Feng, Yibo, 676
Feng, Yichen, 879
Feng, Yingchaojie, 628, 687
Feng, Yiqing, 159
Feng, Yousheng, 136
Feng, Yu, 138, 786
Feng, Yuan, 795
Feng, Yuanyuan, 547
Feng, Yuchen, 304
Feng, Yue, 514, 852
Feng, Yujie, 788
Feng, Yukang, 114, 134, 844
Feng, Yun, 166
Feng, Yunhe, 872
Feng, Zhaopeng, 645
Feng, Zhaoxin, 259
Feng, Zhiyong, 391
Feng, Zhuoka, 757
FengTao, 228
FengTao, FengTao, 228
Feris, 116
Feris, Rogerio, 116
Fernandez, 186, 547
Fernandez, Nigel, 186, 547
Ferragud, 395
Ferragud, Maria, 395
Ferrara, 83
Ferrara, Emilio, 83
Ferraro, 154, 490, 813
Ferraro, Francis, 154, 813
Ferraro, Frank, 490

Ferreira, 287, 608
Ferreira, Rafael, 608
Ferreira, Thiago Castro, 287
Fetaya, 288
Fetaya, Ethan, 288
Ficek, 265
Ficek, Aleksander, 265
Fihey, 176
Fihey, Pierre, 176
Filho, 596
Filho, Arlindo Rodrigues Galvão, 596
Filimonov, 640
Filimonov, Sergei, 640
Filippova, 500
Filippova, Katja, 500
Finch, 730
Finch, James, 730
Finkelstein, 342
Finkelstein, Mara, 342
Firooz, 356
Firooz, Hamed, 356
Fisac, 513
Fisac, Jaime Fernández, 513
Flanagan, 81
Flanagan, Paul, 81
Flechl, 820
Flechl, Martin, 820
Fleisig, 102
Fleisig, Eve, 102
Flek, 247, 589
Flek, Lucie, 247, 589
Flickinger, 187
Flickinger, Dan, 187
Flynn, 385
Flynn, Noah, 385
Foerster, 376
Foerster, Jakob Nicolaus, 376
Foisal, 212
Foisal, Ahamed, 212
Foltin, 672
Foltin, Martin, 672
Fomsgaard, 415, 765
Fomsgaard, Søren, 415
Fomsgaard, Søren Kirkegaard, 765
Fong, 592
Fong, Stephanie, 592
Fonseca, 154
Fonseca, Guilherme, 154
Fontana, 778
Fontana, Aleksandar, 778
Foppiano, 862
Foppiano, Luca, 862
Forer, 89
Forer, Lior, 89
Foroutan, 84, 247, 784
Foroutan, Neda, 84
Foroutan, Negar, 247, 784
Fortes, 694
Fortes, Saúl Blanco, 694
Fosler-Lussier, 194
Fosler-Lussier, Eric, 194
Foster, 829
Foster, Jennifer, 829
Fournier, 889
Fournier, Gregoire, 889
Francois, 580
Francois, Jerome, 580
Franke, 765
Franke, Michael, 765
Frederiksen, 438
Frederiksen, Fie Lejre, 438
Fredrikson, 285
Fredrikson, Matt, 285
Freihat, 370, 555, 690, 864
Freihat, Abed Alhakim, 370, 555, 690, 864
Freisinger, 812
Freisinger, Steffen, 812
Freitag, 342
Freitag, Markus, 342
Freitas, 576, 682
Freitas, Andre, 576, 682
Frenda, 380, 471
Frenda, Simona, 380, 471
Frermann, 83, 201, 458, 470, 571
Frermann, Lea, 83, 201, 458, 470, 571
Frey-Marti, 110
Frey-Marti, Marianne, 110
Frey;,
Frey;, Behzad Shomali; Luisa Victor; Tim Selbach; Ali Hamza Bashir; David Berghaus; Joachim Koehler; Mehdi Ali; Markus,
Fried, 598
Fried, Daniel, 598
Friedl, 612
Friedl, Ken, 612
Friedman, 239
Friedman, Jeffrey, 239
Friedrich, 97, 298, 439
Friedrich, Annemarie, 97
Friedrich, Corentin, 439
Friedrich, Felix, 298
Fritz, 582
Fritz, Mario, 582
Frolov, 633
Frolov, Evgeniy, 633

Fu, 83, 114, 121, 122, 124, 130, 131, 133, 138, 143, 148, 166, 191, 206, 225, 232, 248, 250, 259, 303, 306, 316, 322, 323, 329, 339, 348, 361, 363, 375, 403, 417, 433, 436, 452, 456, 464, 479, 481, 495, 524, 529, 532, 548, 578, 617, 618, 650, 654, 665, 666, 667, 680, 689, 691, 699, 720, 749, 753, 755, 756, 788, 825, 863, 870, 877, 881, 893, 899, 901, 904, 907
fu, 581, 782
Fu, Biao, 433, 788
Fu, Chaoyou, 749
Fu, Chiyuan, 904
Fu, Daocheng, 148, 303
Fu, Dayuan, 665
Fu, Deqing, 225
Fu, DingBa, 316
Fu, Dongjie, 650
Fu, Dongqi, 232
Fu, Guohong, 133, 259
fu, hang, 581, 782
Fu, Haowei, 901
Fu, Haoxiang, 479
Fu, Heming, 166
Fu, Honghao, 436
Fu, Hongke, 375
Fu, Jiarun, 323
Fu, Jinhu, 403
Fu, Jinlan, 130
Fu, Kelin, 825
Fu, Lin, 907
Fu, Lingyue, 863
Fu, Lirong, 532
Fu, Luoyi, 361
Fu, Mingyang, 138
Fu, Pei, 124
Fu, Rao, 114
Fu, Renyu, 452
Fu, Ruibo, 456
Fu, Ruoyu, 617
Fu, Shuangjie, 753
Fu, Szu-Wei, 529
Fu, Tingchao, 322, 363
Fu, Tingchen, 654, 756
Fu, Wenjie, 720
Fu, Xiangjun, 250
Fu, Xiangzheng, 548
Fu, Xiao, 191
Fu, Xiaoliang, 329, 417, 691
Fu, Xinyu, 339
Fu, Yao, 131
Fu, Yifei, 699
Fu, Yikun, 893
Fu, Yingjie, 306
Fu, Yingxue, 83
Fu, Yiwei, 375
Fu, Yonggan, 899
Fu, Yu, 481
Fu, Yujia, 206
Fu, Yumeng, 464, 680, 689
Fu, Yun, 124, 143, 495, 524, 618, 870
Fu, Yuqian, 666, 667
Fu, Zenghuang, 881
Fu, Zhe, 578
Fu, Zheyu, 122
Fu, Zhiheng, 755
Fu, Zhihui, 121, 877
Fu, Zhizhang, 348
Fu, Zihang, 248
Fuad, 371
Fuad, Md Tahmid Hasan, 371
Fuchs, 770
Fuchs, Gilad, 770
Fujii, 512
Fujii, Ryo, 512
Fujinuma, 254
Fujinuma, Yoshinari, 254
Fujita, 502
Fujita, Ken Watanabe; Katsuhide,
Fukumoto, 390
Fukumoto, Fumiyo, 390
Fukumura, 416
Fukumura, Kazuma, 416
Funakoshi, 172
Funakoshi, Kotaro, 172
Fung, 164, 250, 603, 691, 808
Fung, Yi R., 164, 250, 603, 691, 808
Furukawa, 416, 872
Furukawa, Ryo, 416
Furukawa, Yui, 872
Furuta, 660
Furuta, Hiroki, 660
Futami, 111
Futami, Hayato, 111
Futrell, 447
Futrell, Richard, 447
Färber, 168, 542, 659, 850
Färber, Michael, 168, 542, 659, 850
Fütterer, 495
Fütterer, Tim, 495

Gabriel, 538
Gabriel, Saadia, 538
Gai, 414, 672
Gai, Kun, 414, 672
Gaiduk, 678

Gaiduk, Maksim, 678
Gaim, 515
Gaim, Fitsum, 515
Gain, 644
Gain, Baban, 644
Gajcin, 493
Gajcin, Jasmina, 493
Gajo, 642
Gajo, Paolo, 642
Galaeva, 342
Galaeva, Ekaterina, 342
Galassi, 178
Galassi, Vincent, 178
Galimzianova, 233
Galimzianova, Daria, 233
Gallagher, 764
Gallagher, Liam, 764
Galstyan, 253, 851, 879
Galstyan, Aram, 253, 851, 879
Gamallo, 415
Gamallo, Pablo, 415
Gan, 116, 128, 224, 297, 330, 427, 625, 657, 661, 683, 827
Gan, Chuang, 224, 427
Gan, Esther, 625
Gan, Guo, 116
Gan, Leilei, 128, 297, 661, 683
Gan, Siyuan, 657
Gan, Weinan, 330
Gan, Woody Haosheng, 827
Gandelsman, 284
Gandelsman, Yossi, 284
Gandhi, 123, 442
Gandhi, Megha, 442
Gandhi, Nupoor, 123
Ganesan, 207
Ganesan, Adithya V, 207
Ganesh, 180, 904
Ganesh, Sumitra, 180, 904
Ganguly, 556, 895
Ganguly, Debargha, 895
Ganguly, Niloy, 556
Ganideh, 641
Ganideh, Saeb, 641
Ganti, 476
Ganti, Achyutarama, 476
Ganu, 428, 484, 895
Ganu, Tanuja, 428, 484, 895
Ganz, 385
Ganz, Roy, 385
Ganzha;,
Ganzha;, Maria,
Gao, 84, 88, 103, 105, 111, 114, 117, 125, 134, 161, 166, 168, 176, 184, 186, 197, 205, 274, 276, 278, 290, 298, 301, 316, 317, 318, 321, 324, 326, 328, 333, 335, 336, 344, 349, 363, 369, 371, 373, 375, 379, 380, 384, 391, 395, 396, 401, 410, 413, 424, 428, 430, 437, 449, 466, 469, 475, 488, 491, 506, 507, 540, 545, 550, 558, 563, 588, 591, 595, 599, 610, 611, 622, 626, 629, 631, 642, 643, 649, 652, 655, 657, 662, 672, 673, 678, 680, 686, 687, 699, 701, 702, 703, 705, 719, 732, 740, 752, 758, 762, 784, 790, 796, 801, 822, 823, 828, 845, 853, 856, 859, 874, 876, 877, 883, 884, 886, 897, 900, 908
gao, 224
Gao, Bingjie, 732
Gao, Bo, 622
Gao, Boyan, 491
Gao, Chang, 897
Gao, Chao, 703
Gao, Chen, 335, 673
Gao, Cheng, 449
Gao, Chunyang, 430
Gao, Ci-Jun, 274
Gao, Cong, 384
Gao, Cuiyun, 391
Gao, Fan, 563, 823
Gao, Fei, 540
Gao, Feifei, 413
Gao, Goofing, 662
Gao, Guanglai, 318, 401, 762
Gao, Haocheng, 545, 550, 796
Gao, Heyang, 335
Gao, Huan, 125
Gao, Hui, 642
Gao, Jiahui, 642
Gao, Jialu, 466
Gao, Jian, 186
Gao, Jianfeng, 599
Gao, Jianqi, 395
Gao, Jianxi, 88
Gao, Jiasen, 176
Gao, Jiayi, 699
Gao, Jiechao, 610, 845
Gao, Jing, 611
Gao, Jinglong, 626, 643
Gao, Jinhua, 413
Gao, Jiuchong, 328, 395, 884
Gao, Jun, 134, 301, 687
Gao, Junqi, 437
Gao, Kefeng, 316
Gao, Kuofeng, 349

Gao, Lang, 117, 475, 876
Gao, Lei, 488
Gao, Lianli, 883
Gao, Linfeng, 336
Gao, Liqun, 373
Gao, Lirong, 719, 758
Gao, Luyu, 161
Gao, Mengna, 874
Gao, Miao, 369
Gao, Minghe, 545
Gao, Mingze, 588, 629
Gao, Minyu, 506
Gao, Nolan, 410
Gao, Peilin, 545
Gao, Penglei, 166
Gao, Pengzhi, 558
gao, pengzhi, 224
Gao, Qiang, 276
Gao, Qingying, 908
Gao, Rena Wei, 114
Gao, Ruosen, 424
Gao, Sensen, 701
Gao, Shangqian, 790
Gao, Shen, 324
Gao, Silin, 784
Gao, Te, 877
Gao, Tianhong, 371
Gao, Tingting, 379, 687
Gao, Wang, 428
Gao, Wei, 105, 699
Gao, Weibo, 321
Gao, Wenchang, 672
Gao, WenChun, 752
Gao, Wenhao, 701
Gao, Xiang, 595
Gao, Xiaoxue, 469
Gao, Xin, 103, 274, 859
Gao, Xinbo, 507
Gao, Xingyu, 290
Gao, Xingze, 680
Gao, Xinyu, 184
Gao, Xiong, 205
Gao, Xu, 333
Gao, Xuanrong, 702
Gao, Yan, 197, 298, 652, 886
Gao, Yang, 491, 631, 657, 705, 801, 856
Gao, Yanjun, 822
Gao, Yifan, 111
Gao, Yifei, 363, 396, 828
Gao, Yifeng, 655
Gao, Yingying, 678
Gao, Yiwen, 168
Gao, Yizhao, 900
Gao, Yuanyuan, 278
Gao, Yudong, 344
Gao, Yuhang, 326
Gao, Yunjun, 317
Gao, Yutong, 859
Gao, Zhenfeng, 375
Gao, Zhifeng, 649, 740
Gao, Zhiqiang, 369, 853
Gao, Zhuoyue, 326
Gao, Zihan, 84
Gao, Ziqiang, 380
Gao, Zirui, 629
Gao, Zitian, 591
gaojie, 560
gaojie, Song, 560
Garaev, 491
Garaev, Roman, 491
Garbacea, 648
Garbacea, Cristina, 648
Garbaya, 784
Garbaya, Dhia, 784
Garcia, 232, 477, 813
Garcia, Daniel, 477
Garcia, Marcos, 813
Garcia, Roy, 232
Gare, 116
Gare, Gautam Rajendrakumar, 116
Gareev, 655
Gareev, Daniil, 655
Garg, 99, 337, 484, 791, 881
Garg, Deepak, 99
Garg, Dinesh, 337
Garg, Shweta, 484, 881
Garg, Utkarsh, 791
Garibaldi, 747
Garibaldi, Jonathan M., 747
Garikaparthi, 192
Garikaparthi, Aniketh, 192
Garimella, 122, 270, 476, 505, 557, 784
Garimella, Aparna, 122, 270, 476, 505, 557, 784
Garza, 288
Garza, Paolo, 288
Gashteovski, 275, 474
Gashteovski, Kiril, 275, 474
Gasic, 89
Gasic, Milica, 89
Gatt, 229
Gatt, Albert, 229
Gatti, 581
Gatti, Bruno, 581
Gatto, 815
Gatto, Joseph, 815
Gauch, 484

Gauch, Susan, 484
Gaur, 621, 813
Gaur, Manas, 621, 813
Gaussier, 139
gaussier, 883
Gaussier, Eric, 139
gaussier, eric, 883
Gautam, 209, 663
Gautam, Sanjana, 209
Gautam, Vagrant, 663
Gavrilov, 438
Gavrilov, Nikita, 438
Gaynullin, 862
Gaynullin, Dmitry, 862
Ge, 84, 90, 128, 156, 204, 292, 402, 455, 560, 592, 598, 599, 606, 617, 635, 642, 649, 678, 684, 707, 791, 795, 884, 901
Ge, Chao, 402
Ge, Chengze, 684
Ge, Liuxian, 642
Ge, Luise, 90
Ge, Meng, 455
Ge, Qinwen, 901
Ge, Tao, 791
Ge, Tiezheng, 128, 707
Ge, Wei, 884
Ge, Wenshuo, 678
Ge, Xueren, 204
Ge, Xuri, 795
Ge, Yijun, 156
Ge, Yirou, 606, 617
Ge, Yuan, 456
Ge, Yubin, 598
Ge, Yutang, 599, 649
Ge, Yuyao, 635
Ge, Zheng, 292, 560
Ge, Zongyuan, 84, 592
Geeraerts, 173
Geeraerts, Dirk, 173
Geh, 148
Geh, Renato, 148
Gehring, 258
Gehring, Justine, 258
Gehrmann, 781
Gehrmann, Sebastian, 781
Geiger, 276
Geiger, Atticus, 276
Geislinger, 471
Geislinger, Robert, 471
Gekhman, 553
Gekhman, Zorik, 553
Gema, 742
Gema, Aryo Pradipta, 742
Genabith, 306, 806, 810
Genabith, Josef van, 306, 806, 810
Genadi, 589
Genadi, Rifo Ahmad, 589
GENG, 417
Geng, 88, 145, 236, 263, 310, 384, 559, 597, 733, 756, 797
Geng, Gloria, 263
Geng, Guotong, 384, 597, 733
GENG, He, 417
Geng, Hejia, 797
Geng, Jiahui, 88, 145, 236
Geng, Runpeng, 559
Geng, Wenxiang, 756
Geng, Yue, 310
George, 305
George, Sonny, 305
Georgiev, 442, 527
Georgiev, Georgi, 527
Georgiev, Georgi Nenkov, 442
Gerasin, 633
Gerasin, Timofey, 633
Gergerli, 118
Gergerli, Guven, 118
Gerrio, 860
Gerrio, Abdurrahman, 860
Gers, 700
Gers, Felix, 700
Gesi, 824, 826, 875
Gesi, Jiri, 824, 826, 875
Getz, 242
Getz, Heidi, 242
Geva, 171, 276, 452
Geva, Mor, 171, 276, 452
Ghaddar, 273, 860
Ghaddar, Abbas, 273
Ghaddar, Karim, 860
Ghahroodi, 802
Ghahroodi, Omid, 802
Ghaleb, 302
Ghaleb, Esam, 302
Ghanem, 291, 667
Ghanem, Bernard, 291, 667
Ghasemabadi, 773
Ghasemabadi, Amirhosein, 773
Ghassemi, 297
Ghassemi, Marzyeh, 297
Ghazizadeh, 226
Ghazizadeh, Nona, 226
Ghogho, 806
Ghogho, Mounir, 806
Ghosh, 99, 127, 140, 173, 202, 529, 642, 662, 791, 813, 823, 864, 875, 899

Ghosh, Akash, 202, 662
Ghosh, Bishwamittra, 99
Ghosh, Kripabandhu, 642
Ghosh, Madhusudan, 823
Ghosh, Saptarshi, 173, 813, 864
Ghosh, Shaona, 899
Ghosh, Shreya, 875
Ghosh, Sreyan, 127, 140, 529
Ghosh, Supriyo, 791
Ghoshal, 131, 136, 430
Ghoshal, Rajarshi, 430
Ghoshal, Sandip, 131, 136
Ghrayeb, 794
Ghrayeb, Lian, 794
Ghussin, 806
Ghussin, Yusser Al, 806
Giannetti, 519
Giannetti, Christian, 519
Giannouris, 166
Giannouris, Polydoros, 166
giannuzzi, 769
giannuzzi, pietro ferrazzi; milica cvjeticanin; alessio piraccini; davide,
GieSSing, 568
GieSSing, Noah, 568
Gila, 784
Gila, Miguel, 784
Gilady, 585
Gilady, Adi Mayrav, 585
Gilberti, 105, 772
Gilberti, Miles, 105, 772
Giles, 827
Giles, C. Lee, 827
Ginn, 188, 246
Ginn, Michael, 188, 246
Ginsburg, 265, 529, 542
Ginsburg, Boris, 265, 529, 542
Giorgi, 236
Giorgi, Salvatore, 236
Gipp, 620, 876
Gipp, Bela, 620, 876
Gipp;,
Gipp;, Bela,
Giraldo, 784
Giraldo, Juan Garcia, 784
Girish, 218
Girish, Girish, 218
Giryes, 543, 777
Giryes, Raja, 543, 777
Giulianelli, 243
Giulianelli, Mario, 243
Giunchiglia, 431, 742
Giunchiglia, Eleonora, 742
Giunchiglia, Fausto, 431
Giyahchi, 829
Giyahchi, Tootiya, 829
Glandorf, 156
Glandorf, Dominik, 156
Glass, 880
Glass, James, 880
Glava, 275, 825
Glava, Goran, 275, 825
Gleize, 243
Gleize, Martin, 243
Gligori, 458, 838
Gligori, Kristina, 458, 838
Godey, 253, 282
Godey, Nathan, 253, 282
Goebel, 227
Goebel, Randy, 227
Goel, 156, 504, 509, 513, 529, 589, 866
Goel, Anmol, 156, 509, 529
Goel, Arnav, 589, 866
Goel, Arushi, 529
Goel, Drishti, 513
Goel, Medha Hira*; Prachi Goyal*; Raj Maheshwari*; Arnav,
Goh, 649
Goh, Xin Deik, 649
Gokhale, 163
Gokhale, Tejas, 163
Golalikhani, 505
Golalikhani, Mohsen, 505
Golan_Hashiloni,
Golan_Hashiloni, Kai,
Golazizian, 226
Golazizian, Preni, 226
Goldberg, 578
Goldberg, Yoav, 578
Goldfarb-Tarrant, 468
Goldfarb-Tarrant, Seraphina, 468
Goldin, 660
Goldin, Gili, 660
Goldman, 585
Goldman, Omer, 585
Goldwasser, 83, 555, 766
Goldwasser, Dan, 83, 555, 766
Golik, 548
Golik, Pawe, 548
Gollakota, 141, 235
Gollakota, Shyamnath, 141, 235
Goloburda, 236, 874
Goloburda, Liya, 874
Goloburda, Maiya, 236, 874
Golovanevsky, 894
Golovanevsky, Michal, 894

Gomez, 174, 566
Gomez, Marcos J, 174, 566
Goncharov, 330
Goncharov, Aleksei, 330
GONG, 264
Gong, 128, 133, 145, 149, 152, 206, 210, 216, 235, 244, 248, 251, 261, 280, 308, 331, 346, 350, 354, 380, 387, 412, 431, 455, 456, 466, 478, 482, 498, 533, 540, 573, 576, 577, 582, 607, 624, 626, 636, 640, 683, 701, 707, 708, 715, 743, 758, 760, 770, 820
Gong, Boying, 206, 607
Gong, Chen, 133, 149, 251
Gong, Cheng, 455, 456
Gong, Claire, 331
Gong, Hongyu, 235
Gong, Houhua, 577
Gong, Jiefu, 380
Gong, Jingyang, 248
Gong, Jingyu, 412
Gong, Kaixiong, 244, 743
Gong, Liancheng, 582
Gong, Linlu, 482
GONG, MING, 264
Gong, Mingming, 701
Gong, Neil Zhenqiang, 350, 626, 770, 820
Gong, Peiliang, 640
Gong, Ruihao, 128
Gong, Shansan, 576
Gong, Taesik, 533
Gong, Tao, 707
Gong, Weikang, 216
Gong, Xuan, 152
Gong, Xue, 540
Gong, Xueling, 308
Gong, Xun, 683
Gong, Yan, 760
Gong, Yeyun, 210, 280, 431, 466, 573, 636, 758
Gong, Yifan, 498
Gong, Yitian, 708
Gong, Yongshun, 387
Gong, Yuejiao, 331
Gong, Zhaoyan, 354, 624, 715
Gong, Zheng, 346
Gong, Zhihao, 478
Gong, Zhihua, 261
Gong, Zhuocheng, 145
Gonugondla, 444
Gonugondla, Sujan Kumar, 444
Gonzalez, 538
Gonzalez, Erick Rosas, 538
Gonçalves, 154
Gonçalves, Marcos André, 154
Goot, 862
Goot, Rob van der, 862
Gopal, 331
Gopal, Anjali, 331
Gopalan, 625
Gopalan, Kaushik, 625
Gor, 102
Gor, Maharshi, 102
Gori, 634
Gori, Marco, 634
Gosai, 560, 779
Gosai, Advait, 560, 779
Gottal, 194, 907
Gottal, Jonas, 194, 907
Gottesman, 884, 891
Gottesman, Zachary, 884, 891
Gou, 388, 855
Gou, Zhinan, 855
Gou, Zhuo, 388
Goulão, 862
Goulão, Manuel, 862
Goupil, 804
Goupil, Alban, 804
Gourabathina, 781
Gourabathina, Abinitha, 781
Gouvert, 811
Gouvert, Olivier, 811
Govil, 337
Govil, Ishaan, 337
Govindarajan, 627
Govindarajan, Venkata S, 627
Gowda, 106, 214
Gowda, Harshavardhana T, 106
Gowda, Harshavardhana T., 214
Goyal, 121, 128, 139, 515, 577, 587
Goyal, Agam, 121, 587
Goyal, Anuj Kumar, 139
Goyal, Navin, 577
Goyal, Prasoon, 128
Goyal, Tejas, 515
Graciotti, 395
Graciotti, Arianna, 395
Graham, 226
Graham, Benjamin A.T., 226
Grandury, 784
Grandury, María, 784
Grant, 196
Grant, Christan, 196
Grantner, 820
Grantner, Tobias, 820
Gratch, 83
Gratch, Jonathan, 83

Grecu, 193
Grecu, Bogdan, 193
Green, 161, 862
Green, Joshua, 161
Green, Tommaso, 862
Greenewald, 517
Greenewald, Kristjan, 517
Greenfeld, 804
Greenfeld, Refael Shaked, 804
Greer, 815
Greer, Sarah E., 815
Gretz, 771
Gretz, Shai, 771
Griffin, 109
Griffin, Tess, 109
Griffiths, 513, 821
Griffiths, Thomas L., 513, 821
Grinberg, 818
Grinberg, Nir, 818
Gromov, 444
Gromov, Andrey, 444
Grossman, 876
Grossman, Riley, 876
Grover, 148, 890
Grover, Aditya, 148
Grover, Shresth, 890
Grundkiewicz, 871
Grundkiewicz, Roman, 871
Grundy, 299
Grundy, John, 299
Gu, 97, 107, 112, 119, 138, 158, 164, 176, 191, 197, 202, 222, 239, 244, 251, 253, 272, 283, 298, 299, 304, 311, 341, 347, 348, 353, 358, 360, 364, 377, 406, 410, 411, 434, 483, 491, 497, 518, 538, 577, 601, 604, 619, 654, 666, 684, 690, 696, 719, 727, 733, 739, 742, 743, 757, 773, 786, 788, 801, 809, 819, 840, 843, 859, 870, 883, 893, 909
Gu, Alex, 112
Gu, Bin, 577
Gu, Chenxi, 299
Gu, Chenyang, 410, 411
Gu, Chloe, 138
Gu, Enhao, 840
Gu, Feng, 819
Gu, Hanlin, 690
Gu, Hao, 197, 244, 272, 347, 619
Gu, Haoyu, 491
Gu, Hongchao, 298, 819
Gu, Huaiyu, 239
Gu, Jia-Chen, 483
Gu, Jiaqi, 859
Gu, Jiawei, 654
Gu, Jiayang, 883
Gu, Jihao, 107, 341, 733
Gu, Jindong, 364
Gu, Jinjie, 604, 684, 742
Gu, Ken, 112
Gu, Li, 773
Gu, Lin, 601
Gu, Naibin, 304
Gu, Qi, 311
Gu, Qing, 97, 801
Gu, Renjie, 360
Gu, Shi, 843
Gu, Tianle, 251, 696
Gu, Tingxiang, 727
Gu, Wenhao, 158
Gu, Xiaodong, 222, 283, 518
Gu, Xiaoling, 497
Gu, Xiaotao, 164, 909
Gu, Xu, 176, 202
Gu, Yanmei, 719
Gu, Yi, 809
Gu, Yiwei, 757
Gu, Yiyang, 191
Gu, Yu, 358, 377, 406, 434, 788
Gu, Yuancheng, 348
Gu, Yue, 253
Gu, YuKai, 119
Gu, Yuxian, 786
Gu, Yuxuan, 353, 538, 739
Gu, Zhengyao, 893
Gu, Zhouhong, 666, 743, 757
Gu, Zhuangcheng, 239
Gu, Zihan, 202, 870
Guan, 119, 132, 216, 220, 221, 272, 295, 303, 350, 382, 388, 400, 468, 500, 544, 646, 651, 653, 694, 706, 768, 797, 805, 863, 865
Guan, Boyu, 382
Guan, Chaowen, 132
Guan, Chenhao, 768
Guan, Haibing, 805
Guan, Hao, 863
Guan, Hongjiao, 400
Guan, Jian, 865
Guan, Jueqi, 216
Guan, Kaiqi, 303
Guan, Lingxiao, 119
Guan, Patrick, 797
Guan, Sheng, 295
Guan, Shuhao, 468
Guan, Weili, 350, 651, 653, 694
Guan, Wenhao, 220
Guan, Xin, 388

Guan, Xinping, 221
Guan, Yao, 646
Guan, Yong, 706
Guan, Yu, 544
Guan, Zhenyu, 272, 500
Guang, 540
Guang, Mo, 540
guanjunjiang, 643
guanjunjiang, guanjunjiang, 643
Gubanov, 614
Gubanov, Michael, 614
Gubelmann, 875
Gubelmann, Reto, 875
Gubri, 156, 509
Gubri, Martin, 156, 509
Gude, 187
Gude, Adrián, 187
Gudi, 82
Gudi, Vamsi Thilak, 82
Guerra, 292
Guerra, Luiz Franciscatto, 292
Guerreiro, 705
Guerreiro, Nuno M., 705
Guerrero, 524
Guerrero, Enrique Gonzalez, 524
Gui, 136, 138, 157, 201, 382, 387, 545, 563, 670, 717, 740, 820, 821, 838, 839, 860, 897
Gui, Jiang, 136, 860
Gui, Jiaping, 138
Gui, Lin, 157
Gui, Tao, 201, 382, 387, 545, 717, 740, 820, 821, 838, 839, 897
Gui, Xinyue, 563
Gui, Zipeng Zhu; Zhanghao Hu; Qinglin Zhu; Jingyong Su; Yulan He; Lin,
Guida, 83, 222, 458, 470
Guida, Matteo, 83, 222, 458, 470
Gulati, 731
Gulati, Aman, 731
Gulcehre, 784
Gulcehre, Caglar, 784
Gulhane, 118
Gulhane, Radha, 118
Gull, 527
Gull, Ayesha, 527
Gulwani, 239
Gulwani, Sumit, 239
Gumma, 837
Gumma, Varun, 837
Gummadi, 99, 287, 571
Gummadi, Krishna P., 99, 287, 571
Gunaratna, 905
Gunaratna, Kalpa, 905
Gunjal, 530, 560
Gunjal, Anisha, 530, 560
Guo, 108, 125, 128, 132, 133, 138, 140, 144, 148, 157, 158, 159, 164, 166, 168, 169, 175, 189, 194, 197, 201, 207, 214, 216, 221, 222, 235, 244, 248, 250, 260, 272, 283, 284, 289, 304, 311, 313, 321, 324, 327, 332, 334, 339, 347, 354, 359, 368, 369, 372, 379, 402, 404, 405, 409, 410, 413, 415, 418, 421, 430, 431, 435, 449, 451, 453, 456, 459, 467, 470, 471, 473, 485, 487, 488, 505, 513, 514, 516, 533, 541, 543, 545, 546, 547, 560, 561, 563, 566, 570, 577, 580, 595, 596, 617, 618, 619, 620, 624, 629, 635, 640, 646, 648, 651, 657, 668, 671, 676, 679, 681, 686, 689, 691, 695, 698, 703, 704, 706, 715, 721, 724, 729, 732, 735, 736, 737, 738, 740, 741, 750, 754, 755, 756, 758, 762, 774, 784, 798, 801, 802, 818, 819, 821, 839, 840, 853, 875, 877, 881, 886, 894, 899, 905
Guo, Alexia, 875
Guo, Anhong, 459
Guo, Bin, 617, 651
Guo, Chang, 732, 741
Guo, Chaohui, 570
Guo, Chenlei, 617, 802
Guo, Dadi, 164, 691
Guo, Dan, 561
Guo, Dandan, 648
Guo, David, 125
Guo, Daya, 620, 729
Guo, Diandian, 304
Guo, Fang, 369, 840
Guo, Feng, 324
Guo, Guo-Ping, 418
Guo, Haiyun, 415
Guo, Hanghui, 132, 144
Guo, Hangyu, 560
Guo, Hao, 881
Guo, Hengqi, 646
Guo, Hongcan, 289
Guo, Hongcheng, 216
Guo, Honglin, 545, 740
Guo, Hongxin, 108
Guo, Huizhong, 235
Guo, Jiafeng, 248, 379, 410, 629, 635, 721, 724, 754, 784
Guo, Jiahe, 755, 762
Guo, Jiajian, 695
Guo, Jiajun, 311
Guo, Jialiang, 435, 758

Guo, Jiaming, 418
Guo, Jianming, 421
Guo, Jiaqi, 409
Guo, Jiawei, 738
Guo, Jiaxin, 681, 756
Guo, Jiayi, 140
Guo, Jingya, 533, 798
Guo, Jinyang, 485
Guo, Jinyu, 619, 671
Guo, Jiujiang, 354
Guo, Jiyue, 214
Guo, Jizhou, 488
Guo, Kai, 175
Guo, Kaiwen, 313
Guo, Kehan, 471
Guo, Kenan, 413
Guo, Lan-Zhe, 801
Guo, Li, 327, 332
Guo, Lifan, 372
Guo, Lingbing, 334, 704
Guo, Longteng, 736
Guo, Mengfei, 430
Guo, Mengqing, 819
Guo, Minyi, 138, 487
Guo, Minzhe, 738
Guo, Pei-Fu, 467, 547
Guo, Pengze, 577
Guo, Ping, 404
Guo, Qi, 418
Guo, Qianhong, 735
Guo, Qing, 640
Guo, Qipeng, 159, 248, 546, 695, 750, 839, 877
Guo, Qiushi, 128
Guo, Quanjiang, 405
Guo, Ruocheng, 563, 595
Guo, Ruopei, 580
Guo, Shaoru, 706
Guo, Sheng, 402, 758
Guo, Shengnan, 221, 703
Guo, Shijie, 516
Guo, Shiyao, 619
Guo, Shuyue, 260, 676
Guo, Sichen, 505
Guo, Siyuan, 189
Guo, Song, 133, 148, 359
Guo, Taicheng, 505
Guo, Tianpei, 689
Guo, Weidong, 821
Guo, Weiyang, 668
Guo, Wenxiang, 339
Guo, Xiangran, 762
Guo, Xianpeng, 737
Guo, Xiaoke, 354, 624, 715
Guo, Xiaomeng, 679
Guo, Xiaoyong, 108
Guo, Xin, 201, 545, 818
Guo, Xingyu, 698
Guo, Xudong, 158, 289
Guo, Xuehang, 250
Guo, Yanzhu, 168
Guo, Yaxuan, 894
Guo, Yifu, 157, 899
Guo, Yike, 197, 244, 272, 347, 470, 513, 543, 618
Guo, Yingnan, 359
Guo, Yiwen, 435
Guo, Yongxin, 451, 566
Guo, Yu, 194, 774, 853
Guo, Yuanfan, 657
Guo, Yuanli, 687
Guo, Yucan, 784
Guo, Yuchen, 514
Guo, Yuhang, 169, 541, 737
Guo, Yuqing, 166
Guo, Yuxin, 456
Guo, Yuzhi, 283
Guo, Zhengkang, 321
Guo, Zhengliang, 354
Guo, Zhi, 368
Guo, Zhibiao, 207
Guo, Zhicheng, 802
Guo, Zhijiang, 431, 449, 905
Guo, Zhongbin, 453
Guo, Zhuoning, 473
Guo, Zichun, 222
Guo, Zikun, 886
Guo, Zonghao, 596
Guo, Zongren, 368
Guo, Zongyu, 284
Guo;,
Guo;, Jian,
Guohua, 120
Guohua, Wu, 120
GuoqiangMa, 390
GuoqiangMa, GuoqiangMa, 390
Gupta, 82, 88, 91, 128, 160, 161, 181, 189, 198, 209, 230, 244, 253, 287, 337, 342, 425, 469, 476, 505, 515, 526, 532, 544, 547, 560, 569, 570, 575, 589, 613, 791, 798, 814, 828, 841, 851, 875, 879, 897
Gupta, Abhinav, 160, 230, 570
Gupta, Aditya, 189, 532
Gupta, Akshat, 244, 798
Gupta, Anubha, 589
Gupta, Anuj, 515
Gupta, Deepak, 613
Gupta, Devashish, 161

Gupta, Gautam, 569
Gupta, Manish, 287, 875
Gupta, Mehak, 342
Gupta, Mukur, 91, 532
Gupta, Navya, 209
Gupta, Parul, 337
Gupta, Rahul, 128, 253, 851, 879
Gupta, Rohit, 544
Gupta, Sandeep, 198
Gupta, Sandeep K.S., 897
Gupta, Sarmistha Das;Vaibhav Vishal;Syed Ibrahim Ahmad;Sriparna Saha; Manish,
Gupta, Shashwat, 791
Gupta, Siddhant, 425
Gupta, Somit, 575
Gupta, Vinayak, 189
Gupta, Vipul, 547, 560
Gupta, Vivek, 82, 88, 181, 469, 476, 505, 526, 828, 841
Gupta;,
Gupta;, Vivek,
Gurevych, 143, 176, 216, 218, 233, 236, 529, 546, 817, 874, 881
Gurevych, Iryna, 143, 176, 216, 218, 233, 236, 529, 546, 817, 874, 881
Gurgurov, 806, 858
Gurgurov, Daniil, 806, 858
Gurung, 176
Gurung, Alexander, 176
Gurusamy, 199
Gurusamy, Mohan, 199
Gwak, 595, 785
Gwak, Jiho, 785
Gwak, Minju, 595
Gårdhus, 438
Gårdhus, Tobias, 438
Gómez-Rodríguez, 187
Gómez-Rodríguez, Carlos, 187
Gökgöz, 302
Gökgöz, Kadir, 302

H., 686
H., Minh-Ha Le, 686
Ha, 128, 262, 880
Ha, Hyeonjeong, 262
Ha, Rui, 128, 880
Habas, 612
Habas, Piotr, 612
Habash, 236
Habash, Nizar, 236
Habba, 82
Habba, Eliya, 82
Habernal, 277, 505
Habernal, Ivan, 277, 505
Habibi, 862
Habibi, Muhammad Ravi Shulthan, 862
Hachiuma, 529
Hachiuma, Ryo, 529
Haddadi, 790
Haddadi, Hamed, 790
Haddow, 865, 875
Haddow, Barry, 865, 875
Hadhoud, 816, 869
Hadhoud, Sama, 816, 869
Hadj, 860
Hadj, Yahya Mohamed El, 860
Haensch, 806, 835
Haensch, Anna-Carolina, 806, 835
Haffari, 203, 211, 358, 675
Haffari, Gholamreza, 203, 211, 358, 675
Hafiz, 860
Hafiz, Abdulaziz, 860
Hagström, 572
Hagström, Lovisa, 572
Hahami, 603
Hahami, Ely, 603
Hahn, 225, 448
Hahn, Michael, 225, 448
Hai, 535
Hai, Lin, 535
Haifeng, 455
Haifeng, Hu, 455
Haihua, 108
Haihua, Xu, 108
Hain, 455
Hain, Thomas, 455
Hajic, 597
Hajic, Jan, 597
Hajicova, 597
Hajicova, Eva, 597
Hajizadeh, 91
Hajizadeh, Samira, 91
Hakim, 222, 695
Hakim, Muhammad Alif Al, 695
Hakim, Zaber, 222
Hakimi, 784
Hakimi, Ido, 784
Hakkani-Tür, 151, 183, 441, 509, 573, 574, 575
Hakkani-Tür, Dilek, 151, 183, 441, 509, 573, 574, 575
Halder, 654, 894
Halder, Deepon, 654
Halder, Kishaloy, 894
Halim, 816
Halim, Steven, 816

Hallak, 270
Hallak, Assaf, 270
Halterman, 84
Halterman, Andrew, 84
Hamada, 467
Hamada, Malachi, 467
Hamann, 469
Hamann, Hendrik, 469
Hamed, 860
Hamed, Emhemed, 860
Hamedtou, 860
Hamedtou, Emira, 860
Hamid, 291
Hamid, Haidar Bin, 291
Hamidouche, 166
Hamidouche, Wassim, 166
HAMIDULLAH, 810
HAMIDULLAH, Yasser, 810
Hamilton, 463
Hamilton, Sil, 463
Hammal, 413
Hammal, Ayoub, 413
Hammouda, 860, 862
Hammouda, Nadia Ghezaiel, 860, 862
Hamod, 860
Hamod, Aya, 860
Hamoud, 590
Hamoud, Hadi, 590
Han, 94, 121, 123, 124, 127, 129, 132, 141, 143, 145, 146, 157, 165, 166, 190, 197, 212, 213, 221, 228, 230, 244, 248, 252, 253, 272, 278, 280, 282, 292, 294, 295, 306, 307, 308, 313, 326, 328, 340, 347, 369, 378, 380, 382, 384, 389, 391, 402, 414, 432, 446, 449, 452, 460, 470, 484, 507, 509, 513, 517, 536, 543, 546, 560, 580, 587, 590, 593, 596, 599, 619, 627, 638, 650, 651, 653, 662, 665, 669, 671, 673, 676, 681, 716, 732, 741, 743, 752, 759, 773, 776, 786, 787, 789, 793, 795, 797, 799, 819, 832, 861, 869, 882, 890, 895, 899, 905
han, 193
Han, Bing, 432
Han, Bingxu, 384
Han, Bo, 797
Han, Boran, 452
Han, Boxu, 650
Han, Boyan, 146, 280, 732
Han, Caren, 681
Han, Changjiang, 509
Han, Cheng, 507
Han, Chuang, 382
Han, Donghee, 123
Han, Dongyoon, 517
Han, Feng, 228
Han, Guangzeng, 132, 213, 536
Han, HyoJung, 869
Han, Jaeho, 596
Han, Jiaming, 743
Han, Jiawei, 789
Han, Jinda, 307
Han, Jizhong, 143, 389, 793
Han, Jongwook, 787
Han, Jun, 127
Han, Kai, 313, 587
Han, Kecheng, 676
Han, Kyu J., 819
Han, Leekyeung, 580
Han, Ligong, 141
Han, Lu, 145
Han, Mingzhe, 546
Han, Peixuan, 124
Han, Peiyi, 391, 776
Han, Peng, 638
Han, Qi, 292, 560
Han, Ruiyan, 759
Han, Ruizi, 326
Han, Rujun, 786
Han, Seung Hun Eddie, 253
Han, Shi, 773
Han, Shiqin, 402
Han, Shixin, 212
Han, Shuo, 221
Han, Simeng, 248
Han, Sirui, 197, 244, 272, 347, 470, 513, 543, 619, 882
Han, Siwei, 599, 832
Han, Songyang, 484, 787
Han, Soomin, 890
Han, Sophia Simeng, 213
Han, Sungwoo, 587
Han, Tianxin, 671
Han, Tianxu, 121
Han, Wei, 308
Han, Weili, 743
Han, Xianjing, 414
Han, Xianpei, 252, 369, 378, 653, 741
Han, Xiao, 340
Han, Xiaochuang, 446
Han, Xiaotian, 294, 895
Han, Xingsheng, 165, 230, 295
Han, Xingshuo, 593
Han, Xu, 380, 543, 665, 673, 716, 905
Han, Xue, 752
Han, Xueran, 129, 278, 786, 899

Han, Xueting, 651
Han, Yan, 861
Han, Yi, 166, 328
Han, Yinhe, 460
Han, Yiping, 795
Han, Yo-Sub, 190, 799
Han, Yufeng, 282, 449
Han, Yuxing, 669
han, zengxin, 193
Han, Zhe, 662
Han, Zhen, 452, 797
Han, Zhiguang, 157
Han, Zhuowen, 94, 306, 590
Han, Ziyu, 627
Han;,
Han;, Ai,
Hanif, 862
Hanif, Ikhlasul Akmal, 862
Hans, 129, 268
Hans, Mat, 268
Hans, Soham, 129
Hanslik, 568
Hanslik, Lelia, 568
Hao, 108, 219, 260, 311, 315, 328, 379, 395, 455, 476, 478, 479, 617, 642, 651, 670, 681, 692, 700, 703, 736, 745, 746, 747, 809, 884, 907
Hao, Aokai, 456
Hao, Bingguang, 108
Hao, Chaoqun, 692
Hao, Chuan, 260
Hao, Dan, 478
Hao, Dongze, 736
Hao, Jianing, 651, 670
Hao, Jie, 617
Hao, Jinghua, 328, 395, 884
Hao, Kai, 700
Hao, Shibo, 809
Hao, Xiaoshuai, 479
Hao, Xixuan, 907
Hao, Xueyuan, 311
Hao, Yanchao, 315
Hao, Yaru, 907
Hao, Yijie, 219
Hao, Yixue, 747
Hao, Yongchang, 617
Hao, Yuanzhen, 681
Hao, Yupu, 642, 703
Hao, Zhanxin, 476
Hao, Zhewen, 379
Hao, Zhezheng, 745, 746
HaonanChen, 713
HaonanChen, HaonanChen, 713
Haque, 578
Haque, Farsheed, 578
Hara, 839
Hara, Tomomasa, 839
Harada, 492, 601
Harada, Ryuhei Miyazato; Shunsuke Kitada; Kei,
Harada, Tatsuya, 601
Harary, 173
Harary, Sapir, 173
Hardalov, 99
Hardalov, Momchil, 99
Hardy, 90, 891
Hardy, Michael, 90, 891
Hariharan, 150
Hariharan, Bharath, 150
Haring, 496
Haring, Jordan Sinclair; Yousra Shleibik; Kerstin,
Hariri, 294, 472
Hariri, Mohsen, 294, 472
Harish, 798
Harish, Avyukth, 798
Harka, 331
Harka, Leslie, 331
Harman, 195
Harman, Mark, 195
Harris, 486
Harris, Ian, 486
Harry, 547
Harry, Tamunotonye, 547
Hartley, 160
Hartley, John, 160
Hartmann, 568
Hartmann, David, 568
Hartvigsen, 553, 801, 884, 891
Hartvigsen, Thomas, 553, 801, 884, 891
Harvey, 822
Harvey, George, 822
Harwath, 177
Harwath, David, 177
Hasabnis, 794
Hasabnis, Niranjan, 794
HASAN, 613
Hasan, 105, 471, 556
Hasan, Bashar Awwad Shiekh, 105
Hasan, Md Arid, 471, 556
HASAN, SYED MHAMUDUL, 613
Hasani, 611, 644
Hasani, Hosein, 644
Hasani, Moein, 611
Hasanov, 532
Hasanov, Eshgin, 532
Hashemi, 784
Hashemi, Diba, 784

Hashim, 769
Hashim, Mohammed Qusay, 769
Hashimoto, 463
Hashimoto, Ryuji, 463
Hashish, 818
Hashish, Ahmed, 818
Hashmi, 253
Hashmi, Sarim, 253
Hasibi, 736
Hasibi, Faegheh, 736
Haslam, 173
Haslam, Nick, 173
Haslum;,
Haslum;, Patrik,
Hasnat, 232
Hasnat, Abul, 232
Hassan, 178, 260, 532, 589, 763
Hassan, Besher, 179, 589
Hassan, Md. Mahadi, 260, 532
Hassan, Samiah, 763
Hassani, 151
Hassani, Hamed, 151
Hassany, 547
Hassany, Mohammad, 547
Hassidim, 585
Hassidim, Avinatan, 585
Hasson, 595
Hasson, Hilaf, 595
Hatzel, 818
Hatzel, Hans Ole, 818
Hautli-Janisz, 510
Hautli-Janisz, Bayram Ayadi; Annette,
Hawasly, 145, 769
Hawasly, Majd, 145, 769
Hawkins, 306, 764
Hawkins, Robert, 306, 764
Hays, 464
Hays, James, 464
Haznitrama, 590
Haznitrama, Faiz Ghifari, 590
HE, 757
He, 87, 95, 97, 100, 114, 121, 122, 134, 145, 155, 157, 162, 163, 164, 165, 166, 181, 186, 213, 214, 227, 232, 233, 239, 245, 249, 250, 251, 252, 261, 265, 269, 271, 273, 274, 278, 282, 284, 285, 290, 296, 307, 311, 321, 323, 325, 328, 332, 333, 335, 350, 356, 363, 368, 369, 377, 378, 390, 392, 395, 396, 397, 403, 404, 410, 411, 412, 413, 417, 422, 423, 424, 435, 443, 444, 446, 448, 450, 452, 457, 460, 469, 476, 479, 484, 486, 488, 493, 496, 507, 527, 530, 554, 556, 560, 573, 581, 583, 603, 624, 626, 629, 644, 645, 651, 655, 658, 661, 666, 668, 672, 685, 689, 691, 692, 695, 703, 708, 714, 716, 719, 720, 724, 727, 734, 742, 749, 753, 755, 756, 757, 758, 759, 762, 768, 774, 779, 789, 797, 800, 804, 805, 807, 808, 826, 829, 830, 834, 843, 846, 853, 856, 859, 861, 877, 881, 882, 884, 894, 898, 901, 902, 906
He, Allen, 282
He, Ben, 369, 378, 834
He, Bin, 898
He, Bing, 269
He, Bingsheng, 155, 181
He, Bingxiang, 573, 603, 808
He, Conghui, 239, 274, 789, 829, 859
He, Dengwu, 368
He, Di, 165
He, Gaole, 278
He, Gaolei, 350
He, Gaoqi, 734
He, Guannan, 881
He, Guoxiu, 410, 411
He, Haiying, 214
He, Hao, 762, 768
He, Haocong, 666
He, Haodong, 493
He, Haolin, 390
He, Huan, 87, 166, 692
He, Huang, 724
He, Huasen, 774, 853
He, Jie, 469
He, Jingrui, 446, 457, 469, 626, 826
He, Jingwen, 100
He, Junxian, 95
He, Ka Sing, 450
He, Kai, 261
He, Kaiyu, 333
He, Kang, 114, 134
He, Kejun, 162
He, Keyu, 232, 843
He, Kuang, 624
He, Kun, 290, 689
He, Langzhou, 307, 488
He, Lewei, 356, 412
He, Liang, 245, 877
He, Lihuo, 507
He, Linhan, 603
He, Lipeng, 556
He, Liqi, 452
He, Longzhu, 403
He, MaoLin, 114
He, Maoxin, 708

He, Min, 413, 805
HE, Minggui, 757
He, Minggui, 756
He, Minghua, 122
He, Muyu, 284
He, Pengcheng, 486
He, Pengfei, 164, 265, 269, 296, 484
He, Ping, 121
He, Pinjia, 249, 882
He, Qi, 443, 603, 861, 881
He, Qiang, 797
He, Qianyu, 249, 581, 644
He, Qiaozhi, 626
He, Qingfeng, 651
He, Ran, 363
He, Renqing, 328, 395, 884
He, Shaokai, 629
He, Shibo, 583
He, Shiqi, 273
He, Shiyu, 727
He, Shizhu, 97, 396, 424, 703
He, Shuibing, 753
He, Shuyi, 856
He, Shwai, 163
He, Sijun, 377
He, Siyang, 695
He, Tao, 395
He, Tianxing, 672
He, Tianyao, 239
He, Tieke, 448, 554
He, Tingting, 323
He, Wei, 411
He, Weiqing, 227, 894
He, Xiangnan, 311, 325, 645
He, Xiaofei, 685
He, Xiaofeng, 714, 755
He, Xiaoshan, 493
He, Xin, 716
He, Xindong, 404
He, Xinran, 278
He, Xinyang, 460, 800
He, Xiuqiang, 335
He, Xu, 435, 758
He, Xuanli, 742
He, Yancheng, 397, 759
He, Yao, 392, 719, 727
He, Ye, 476, 804
He, Yi, 145
He, Yifei, 661
He, Yifeng, 807
He, Yike, 906
He, Ying, 666, 757
He, Yixiao, 251
He, Yongbo, 749
He, Yongchao, 321, 332
He, Yu, 412
He, Yuanpeng, 155
He, Yueru, 166, 233, 527
He, Yufei, 181, 213, 830, 901
He, Yulan, 157, 668
He, Yun, 444
He, Yunjie, 422
He, Yunzhong, 530, 560, 779
He, Yuxiang, 658
He, Yuxiong, 252, 496
He, Zexue, 479
He, Zhankui, 250
He, Zhaofeng, 285, 422
He, Zhenyu, 165
He, Zheqi, 186
He, Zhitao, 691
He, Zhiyang, 417
He, Zihao, 846
He, Zijian, 720
He, Zirui, 902
He, Ziyi, 271, 423
He, Zoe Wanying, 655
He;,
He;, Peiyang,
Heakl, 95, 156, 253
Heakl, Ahmed, 95, 156, 253
Healey, 156
Healey, Jennifer, 156
Heck, 89
Heck, Michael, 89
Hedderich, 166, 583, 812, 858
Hedderich, Michael A., 166, 583, 812, 858
Hefetz, 816
Hefetz, Ofri, 816
Heffernan, 287
Heffernan, Bláithín, 287
Hei, 316, 552, 787
Hei, Ou Ocean Kun, 552
Hei, Xusen, 316
Hei, Yiming, 787
Heilmann, 207
Heilmann, Xenia, 207
Heinz, 109
Heinz, Michael, 109
Heinzerling, 273
Heinzerling, Benjamin, 273
Hejabi, 226, 539
Hejabi, Parsa, 226, 539
Helcl;,
Helcl;, Jan Bronec; Jindich,
Held, 827

Held, William Barr, 827
Helff, 298
Helff, Lukas, 298
Helm, 226, 800
Helm, Hayden, 226, 800
Henao, 483, 819
Henao, Ricardo, 483, 819
Hendy, 886
Hendy, Mohamed, 886
Henry, 798
Henry, Felix, 798
Heo, 147, 217, 230
Heo, Chaewon, 147
Heo, Inbum, 217
Heo, Ryang, 230
Herbold, 772
Herbold, Steffen, 772
Herel, 625
Herel, David, 625
Hernandez-Orallo, 227
Hernandez-Orallo, Jose, 227
Hernández-Cano, 784
Hernández-Cano, Alejandro, 784
Herraoui, 864
Herraoui, Omar El, 864
Hershcovich, 854
Hershcovich, Daniel, 854
Herwanto, 715
Herwanto, Guntur Budi, 715
Herzig, 116
Herzig, Roei, 116
Herzinger, 744
Herzinger, Andreas, 744
Heumann, 778
Heumann, Christian, 778
Hewavitharana, 841
Hewavitharana, Sanjika, 841
Hidayatullah, 589
Hidayatullah, Ahmad Fathan, 589
Hieber, 542
Hieber, Felix, 542
Higgins, 232
Higgins, Helen, 232
Higuchi, 243
Higuchi, Yosuke, 243
Hill, 495, 892
Hill, Heather, 495
Hill, Shawndra, 892
Hin, 347
Hin, Tang Pok, 347
Hincapie-Monsalve, 862
Hincapie-Monsalve, Sara, 862
Hinczewski, 472
Hinczewski, Michael, 472
Hindriks, 113, 191
Hindriks, Koen, 113, 191
Hingmire, 292
Hingmire, Swapnil, 292
Hinojosa, 291
Hinojosa, Carlos, 291
Hinostroza, 110
Hinostroza, Carlos Rosas, 110
Hira, 589
Hira, Medha, 589
Hiraoka, 838, 841
Hiraoka, Tatsuya, 838, 841
Hirota, 237, 529
Hirota, Wataru, 237
Hirota, Yusuke, 529
Hirsch, 173
Hirsch, Eran, 173
Hisada, 307
Hisada, Shohei, 307
Ho, 107, 154, 198, 208, 219, 391, 448, 609, 615, 789
Ho, Chien-Ju, 107
Ho, Derek, 609
Ho, Edmond S. L., 198, 448
Ho, Johnny K. W., 391
Ho, Joyce C., 219
Ho, Sin Yu Bonnie, 615
Ho, Tsung-Yi, 154, 208, 789
HOANG, 192
Hoang, 121, 320, 639, 701
Hoang, Anh Dung, 639
Hoang, Anh Nguyen, 121
Hoang, Anh T., 701
Hoang, Cong Duy Vu, 320
HOANG, PHAN NHAT, 192
Hochman, 578
Hochman, Hail, 578
Hockenmaier, 248, 474, 807
Hockenmaier, Julia, 248, 474, 807
Hockenmaier;,
Hockenmaier;, Julia,
Hoefler, 784
Hoefler, Torsten, 784
Hoeken, 218
Hoeken, Sanne, 218
Hoeve, 580
Hoeve, Maartje Ter, 580
Hofmann-Coyle, 694
Hofmann-Coyle, Ella, 694
Hoi, 410
Hoi, Steven, 410
Hojae_Han1,

Hojae_Han1, Hojae_Han1,
Holdcroft, 744
Holdcroft, Jamie, 744
Holden, 309
Holden, Eun-Jung, 309
Homan, 522
Homan, Deepak Pandita; Christopher M,
Honda, 272, 810
Honda, Ukyo, 272, 810
Hong, 129, 161, 162, 164, 182, 212, 256, 267, 275, 309, 348, 381, 398, 416, 427, 444, 457, 465, 470, 600, 637, 669, 679, 683, 702, 708, 714, 730, 743, 751, 755, 770, 794, 806, 877
Hong, Chang, 416
Hong, Cheng, 770
Hong, Eunjin, 256
Hong, Geng, 348
Hong, Geonwoo, 129
Hong, Giwon, 275
Hong, Haiwen, 398, 708, 751
Hong, Hanhua, 637, 669
Hong, Haokai, 743
Hong, Haoran, 161
Hong, Ilgee, 794
Hong, Jiajie, 730
Hong, Jiale, 309
Hong, Liang, 877
Hong, Pingjun, 457
Hong, Qichen, 381
Hong, Richang, 714, 755
Hong, S. K., 600
Hong, Sangwoo, 182
Hong, Seongtae, 465, 806
Hong, Seunghyeok, 267
Hong, Shiwei, 470
Hong, Wangyang, 702
Hong, Wenyi, 164
Hong, Yu, 162, 212
Hong, Yuecheng, 679
Hong, Yuri, 444
Hong, Zhang-Wei, 427
Hong, Zhuojian, 683
Hongler, 875
Hongler, Peter, 875
Hoofd, 439
Hoofd, François, 439
Hooi, 213, 830, 901
Hooi, Bryan, 213, 830, 901
Hooker, 826
Hooker, Sara, 826
Hoorn, 391
Hoorn, Johan F., 391
Hope, 89, 96, 109, 176
Hope, Tom, 89, 96, 109, 176
Hoque, 822
Hoque, Enamul, 822
Horio, 485
Horio, Kaito, 485
Horne, 523
Horne, Jacob, 523
Hortacsu, 607
Hortacsu, Ali, 607
Horwood, 192
Horwood, Graham, 192
Hoseinzade, 655
Hoseinzade, Ehsan, 655
Hoshino, 810
Hoshino, Sho, 810
Hossain, 249, 556
Hossain, Md Fahad, 556
Hossain, Md Khalid, 249
Hosseini, 909
Hosseini, Parsa, 909
Hotta, 862
Hotta, Hajime, 862
Hou, 102, 155, 169, 194, 202, 228, 236, 249, 250, 254, 286, 303, 322, 327, 357, 373, 415, 420, 422, 432, 455, 463, 473, 476, 488, 493, 543, 545, 554, 562, 571, 604, 606, 619, 624, 631, 636, 640, 658, 703, 722, 728, 759, 774, 782, 784, 787, 810, 815, 853, 857, 874, 880, 885
hou, 179
Hou, Bairu, 604
Hou, Bingxuan, 874
Hou, Bo WangYang XiangYihong TangDongming ZhaoZifei YuYuexian, 420
Hou, Bojing, 857
Hou, Dylan Xinming, 202
Hou, Guiyang, 631, 722
Hou, Haotian, 432
Hou, Haowen, 228, 543
Hou, Jinchang, 357, 422
Hou, Jinyu, 787
Hou, Junyi, 155
Hou, Lei, 169, 476, 488, 554, 571, 782, 880
Hou, Liang, 286
hou, massimiliano pronestianya belzyufang, 179
Hou, Nana, 455
Hou, Peng, 624
Hou, Qianzi, 554
Hou, Qibin, 810
Hou, Ruihui, 254, 327, 463
Hou, Shizheng, 562
Hou, Shuyang, 476

Hou, TianQi, 759
Hou, Wen, 249
Hou, Wenqing, 373
Hou, Xiaofeng, 619
Hou, Xinmeng, 640
Hou, Xinrui, 703
Hou, Yixuan, 322
Hou, Yu, 102, 473
Hou, Yubo, 728
Hou, Yueling, 636
Hou, Yufang, 493, 545
Hou, Yujie, 874
Hou, Yunpeng, 194, 774, 853
Hou, Yupeng, 236, 250
Hou, Yusen, 815
Hou, Yutai, 658
Hou, Zhendong, 606
Hou, Zhengxu, 303, 885
Hou, Zhiyan, 415
Hou, Zhongni, 784
Hou;,
Hou;, Yufang,
Hougen, 894
Hougen, Dean F., 894
Hovhannisyan, 576
Hovhannisyan, Vahagn, 576
Hovy, 83, 458, 529, 681
Hovy, Dirk, 529
Hovy, Eduard, 83, 458, 681
Hoyle, 784
Hoyle, Alexander Miserlis, 784
Hrinchuk, 542
Hrinchuk, Oleksii, 542
Hruschka, 579, 840
Hruschka, Estevam, 579, 840
Hsiao, 174, 262, 312, 496
Hsiao, Chi-Hsiang, 262
Hsiao, Chun-En, 174
Hsiao, Hsien-Fu, 312
Hsiao, Janet Hui-wen, 496
Hsin, 285
Hsin, Eric, 285
Hsiung, 154, 208
Hsiung, Lei, 154, 208
Hsu, 110, 142, 153, 168, 196, 428, 467, 484, 513, 547, 881
Hsu, Benjamin, 142
Hsu, Chan-Jan, 110
Hsu, Chao-Chun, 484, 881
Hsu, Chih-Hao, 513
Hsu, Chun-Chia, 467, 547
Hsu, Chun-Yu, 547
Hsu, Winston H., 153, 196
Hsu, Wynne, 428
Hsu, Yi-Li, 168, 547
Htet, 471
Htet, Aung Kyaw, 471
HU, 560
Hu, 84, 87, 92, 97, 104, 129, 134, 137, 143, 144, 146, 153, 161, 170, 171, 172, 180, 187, 197, 202, 205, 207, 209, 212, 219, 222, 227, 271, 278, 279, 283, 285, 288, 290, 292, 298, 303, 304, 305, 309, 321, 322, 323, 329, 331, 332, 338, 339, 342, 343, 348, 354, 361, 363, 370, 377, 378, 379, 384, 389, 395, 397, 403, 404, 405, 412, 414, 416, 417, 418, 422, 428, 431, 439, 440, 445, 450, 459, 462, 463, 467, 469, 476, 482, 487, 496, 501, 513, 515, 561, 562, 568, 572, 585, 592, 616, 619, 623, 624, 630, 635, 642, 643, 646, 647, 651, 652, 656, 666, 668, 669, 670, 672, 673, 675, 676, 679, 680, 683, 684, 691, 693, 694, 697, 701, 707, 727, 728, 730, 733, 734, 735, 738, 749, 752, 755, 759, 760, 762, 765, 770, 772, 774, 776, 786, 792, 793, 796, 798, 801, 803, 809, 819, 820, 824, 836, 845, 847, 854, 859, 868, 879, 882, 886, 899, 901, 905, 909
hu, 605
Hu, Baotian, 338, 379, 397, 403, 404, 630, 668, 759, 798
Hu, Beizhe, 170
Hu, Bin, 97, 727
Hu, Boren, 675
Hu, Chao, 222, 283
Hu, Chaowen, 329, 691
Hu, Chen, 180
Hu, Cheng, 304
Hu, Chengyin, 354
Hu, Chenkai, 348
Hu, Chuanrui, 332, 680
Hu, Chunming, 619
Hu, Dou, 693
Hu, Fang, 209
Hu, Fei, 348
Hu, Feng, 635
Hu, Gang, 428
Hu, Guoping, 395, 697
Hu, Guosheng, 137
Hu, Haibo, 207, 482
Hu, Haimin, 513
Hu, Haixiang, 679
Hu, Han, 187, 669
Hu, Hang, 463, 568
Hu, Hangkai, 752

Hu, Haoqi, 868
Hu, Haoyi, 271
Hu, Haoyuan, 905
Hu, Jennifer, 153, 765
Hu, Jianwei, 377
Hu, Jiaxi, 647
Hu, Jiaxue, 417
Hu, Jiayi, 624
Hu, Jie, 412
Hu, Jijin, 707
Hu, Jinbo, 271
Hu, Jingbin, 363
Hu, Jingcheng, 292
Hu, Jingyang, 487
Hu, Jingyu, 803
Hu, Jinpeng, 84, 561, 592
Hu, Jinwei, 170
Hu, Junhao, 501
Hu, Junjie, 683
Hu, Junxing, 439
Hu, Kai, 285
Hu, Kairui, 361
Hu, Lijie, 171, 487, 886
Hu, Lingxiang, 87, 616
Hu, Litao, 845
Hu, Lulu, 290
Hu, Luoming, 431
Hu, Mengting, 733
Hu, Mengxuan, 219
Hu, Minghao, 384
Hu, Nan, 469
Hu, Paul Jen-Hwa, 459
Hu, Peizhao, 207
Hu, Qi, 161
Hu, Qian, 752
Hu, Qiangqiang, 774
Hu, Qigen, 414
Hu, Qin, 170
Hu, Qingguo, 397
Hu, Qirui, 676
Hu, Qisheng, 322
Hu, Ranjuexiao, 735
Hu, Rui, 760, 909
Hu, Sen, 129, 278, 467, 786, 899
Hu, Shasha, 730
Hu, Shiqi, 414
Hu, Shiyu, 642
Hu, Shujie, 643
Hu, Shuming, 562
Hu, Shuyue, 309
Hu, Songlin, 143, 389, 462, 693, 792, 793
Hu, Tiancheng, 104, 137, 227, 501, 572, 845
Hu, Tiansheng, 623
Hu, Tianyi, 854
Hu, Wei, 450, 670
Hu, Weiming, 134
Hu, Wenbin, 496, 647, 882
Hu, Wenpeng, 733
Hu, Wenping, 672
Hu, Wentao, 370
Hu, Wenxing, 422
Hu, Xiang, 92
Hu, Xiangdong, 170
Hu, Xianming, 212
Hu, Xiaohui, 370
Hu, Xiaolin, 879
Hu, Xiaomeng, 728
Hu, Xiaoyang, 772
Hu, Xing, 418, 515
Hu, Xingyu, 824
Hu, Xinshuo, 403
Hu, Xinyan, 303
Hu, Xinyi, 796
Hu, Xinyu, 445, 684
Hu, Xiyang, 271, 288, 836
Hu, Xuecai, 776
Hu, Xueyu, 323
Hu, Xuming, 144, 205, 666, 694, 707
Hu, Yan, 331, 416
Hu, Yangen, 796
Hu, Yao, 197, 298, 404, 405, 652, 701
Hu, Yazhe, 819
Hu, Yi, 476, 859
Hu, Yibo, 749
Hu, Yifan, 656
Hu, Yifei, 414
Hu, YiKun, 329
Hu, Yiming, 378
Hu, Yiqun ZhangHao LiZihan WangShi FengXiaocui YangDaling WangBo ZhangLei BaiShuyue, 342
Hu, Yiqun ZhangPeng YeXiaocui YangShi FengShufei ZhangLei BaiWanli OuyangShuyue, 343
HU, Yiran, 560
Hu, Yue, 202, 673, 765
Hu, Yuepeng, 770, 820
Hu, Yujia, 469
Hu, Yujun, 734
Hu, Yulan, 459, 793, 801
Hu, Yulin, 755, 762
Hu, Yunhai, 279
Hu, Yupeng, 651, 755
Hu, Yusong, 440
Hu, Yuwei, 738
Hu, Yuxin, 384

Hu, Yuxuan, 146, 339, 646, 647
Hu, Zefa, 321
Hu, Zhecheng, 666
Hu, Zheng, 669
Hu, Zhengyu, 172
Hu, Zhenlin, 585
Hu, Zhenyu, 847
hu, zhenyu, 605
Hu, Zhiting, 809
Hu, Zhiyuan, 901
Hu, Zirui, 321
Hu, Zongye, 305
Hua, 155, 201, 230, 259, 413, 418, 501, 597, 598, 767, 776, 790, 797, 808, 870
Hua, Cheng, 870
Hua, Ermo, 413
Hua, Hang, 259
Hua, Harper, 797
Hua, Jiacheng, 776
Hua, Peichun, 767
Hua, Renjie, 155
Hua, Ting, 201, 790, 808
Hua, Wen, 418
Hua, Yilun, 598
Hua, Yining, 230
Hua, Yuncheng, 597
Hua, Ziyue, 501
Huai, 254, 827
Huai, Mengdi, 827
Huai, Ziyue, 254
Huan, 349
Huan, Zhaoxin, 349
HUANG, 164
Huang, 82, 84, 87, 89, 93, 94, 95, 96, 97, 99, 100, 106, 108, 111, 116, 119, 120, 124, 125, 126, 128, 129, 130, 132, 135, 136, 138, 142, 145, 149, 151, 152, 153, 155, 156, 161, 162, 164, 165, 166, 167, 170, 175, 178, 180, 181, 186, 187, 201, 203, 205, 206, 213, 214, 220, 221, 226, 228, 233, 237, 241, 244, 245, 251, 253, 255, 258, 259, 261, 264, 266, 271, 272, 273, 278, 279, 281, 283, 284, 289, 291, 292, 300, 305, 307, 308, 311, 312, 315, 319, 321, 322, 323, 324, 325, 326, 328, 331, 334, 336, 337, 338, 339, 343, 344, 350, 351, 353, 355, 358, 368, 371, 372, 375, 376, 377, 378, 379, 382, 384, 385, 387, 388, 390, 391, 392, 398, 399, 400, 401, 403, 404, 407, 409, 412, 413, 417, 418, 421, 425, 428, 429, 430, 431, 433, 434, 437, 442, 444, 446, 449, 452, 455, 458, 460, 462, 471, 473, 475, 478, 488, 490, 492, 497, 499, 501, 506, 513, 514, 515, 520, 527, 528, 531, 532, 533, 536, 537, 538, 539, 540, 541, 543, 545, 546, 555, 558, 559, 560, 562, 564, 569, 570, 582, 583, 588, 598, 600, 601, 607, 609, 616, 620, 625, 628, 630, 631, 632, 636, 638, 639, 643, 644, 645, 652, 653, 656, 661, 662, 670, 671, 676, 678, 680, 683, 684, 686, 687, 688, 689, 690, 692, 696, 697, 700, 706, 708, 713, 714, 717, 723, 725, 737, 738, 739, 740, 741, 745, 747, 749, 750, 751, 759, 762, 768, 770, 774, 776, 781, 784, 786, 788, 789, 790, 791, 792, 794, 796, 797, 799, 802, 807, 808, 814, 820, 821, 822, 823, 824, 826, 828, 838, 839, 841, 844, 849, 851, 854, 856, 861, 867, 868, 870, 874, 875, 882, 894, 895, 896, 897, 903, 904, 906, 907, 909
huang, 353, 800
Huang, Ailin, 292
Huang, Allen Hao, 784
Huang, Baichuan, 371
Huang, Baixiang, 264
Huang, Baizhou, 392, 697
Huang, Baodai, 545
Huang, Beichen, 130
Huang, Binxuan, 616
Huang, Biwei, 130, 273
Huang, Bo, 271
Huang, Brandon, 116
Huang, Caishuang, 838
Huang, Cassie, 151, 790
Huang, Chao, 138, 251, 271, 490, 537, 588, 907
Huang, Chao-Ming, 241
Huang, Chen, 135, 531, 692
Huang, Chengkai, 514
Huang, Chengrui, 324
Huang, Chenhao, 838
Huang, Chenyang, 203
Huang, Chien-yu, 111
Huang, Chiyue, 839
Huang, Chongxuan, 433
Huang, Chu-Ren, 385, 506
Huang, Chung-ju, 155
Huang, Chuyue, 129
Huang, Dan, 206
Huang, Degen, 351, 741
Huang, Di, 418
Huang, Dong, 478
Huang, Dongqi, 358
Huang, Edward W, 607
Huang, Fan, 153
Huang, Fanding, 145

Huang, Fei, 93, 120, 126, 214, 368, 601, 706, 739, 794, 796, 797, 897
Huang, Furong, 226, 492, 781
Huang, Guan, 125
Huang, Guanbo, 145
Huang, Guanhua, 540
Huang, Guolin, 261
Huang, Haiduo, 400
Huang, Haizhen, 431, 896
Huang, Hanbo, 152
Huang, Hanxian, 444
Huang, Hao, 108, 279, 372, 452, 814
Huang, Haoliang, 284
Huang, Haoran, 344
Huang, Haoyu, 161, 253
Huang, Hen-Hsen, 291, 300, 559
Huang, Heqing, 662
huang, Hexiang, 800
Huang, Heyan, 89, 136, 187, 541, 631
Huang, Heyuan, 558
Huang, Hong, 375, 774
Huang, Hongsen, 155
Huang, Hsiu-Yuan, 149, 539
Huang, Hu, 434
Huang, Hua, 237, 874
Huang, Hui, 272, 598, 802
Huang, Huihui, 515
Huang, James Y., 792
Huang, Jen-tse, 425, 462, 582, 687, 882
Huang, Jerry, 166, 555, 788, 807
Huang, Jiahao, 823
Huang, Jiajia, 428
Huang, Jiajian, 747
Huang, Jiameng, 165
Huang, Jiani, 473
huang, Jianjie, 353
Huang, Jianzhao, 745
Huang, Jiatan, 106, 221
Huang, Jie, 311, 749
Huang, Jimin, 95, 166, 226, 233, 442, 527
Huang, Jimmy, 822
Huang, Jin, 307
Huang, Jin-Xia, 904
Huang, Jincai, 514
Huang, Jing, 531, 824, 826, 861, 875
Huang, Jingang, 437
Huang, Jingyuan, 644, 725
Huang, Jixuan, 545
Huang, Jizhou, 713, 849
Huang, Jun, 87, 562, 896
Huang, Junming, 533
Huang, Junzhou, 283
Huang, Kaibo, 714
Huang, Kaihong, 625
Huang, Kaiyu, 96, 351, 583, 741
Huang, Kaizhu, 538
Huang, Ke-Wei, 84
Huang, Kecheng, 791
Huang, Keduan, 407
Huang, Keman, 630
Huang, Kexin, 228, 696
Huang, Kezhao, 379
Huang, Kung-Hsiang, 609, 789
Huang, Langchu, 399
Huang, Lei, 328, 353, 697
Huang, Leo, 895
Huang, Li, 628
Huang, Lifu, 281
Huang, Linjiang, 339
Huang, Longtao, 398, 708, 751
Huang, Min, 460
Huang, Mingxuan, 738
Huang, Minlie, 132, 164, 398, 413, 497, 520, 652, 786, 851, 909
Huang, Mitt, 560
Huang, Muye, 821
Huang, Nannan, 768
Huang, Peiwen, 513
Huang, Pengcheng, 377, 387
Huang, Pengxiang, 828
Huang, Ping, 471
Huang, Pingyang, 326
Huang, Qiang, 178, 312, 407
Huang, Qinghua, 325
Huang, Qingjia, 499
Huang, Qipeng, 632, 678
Huang, Rongjie, 339
Huang, Ruihong, 180, 895
Huang, Shaohan, 93, 145, 167, 907
Huang, Sharon, 458
Huang, Shen, 388
Huang, Shijue, 808, 903
Huang, Shiting, 372
Huang, Shizhou, 569
Huang, Shouzheng, 403
Huang, Shuai, 687
Huang, Shuangping, 690
Huang, Shujian, 94, 142, 867
Huang, Shurui, 384
Huang, Siming, 350
Huang, Sukai, 620
Huang, Tang Da, 737
Huang, Tenghao, 789
Huang, Tiansheng, 723
Huang, Tianyu, 841
Huang, Ting-Hao 'Kenneth', 165

Huang, Tinglin, 205
Huang, Tingxuan, 377
Huang, Tongwen, 378
Huang, Tzu-Hung, 513
Huang, Victor Shea-Jay, 398, 520
Huang, Wei, 391, 684
Huang, Wei-Chieh, 488
Huang, Weijing, 244
Huang, Weiqing, 429
Huang, Weiran, 656
Huang, Weitao, 680, 689
Huang, Weiyang, 653
Huang, Wen, 308
Huang, Wenhao, 676
Huang, Wenjun, 136
Huang, Wenkang, 688
Huang, Wenrui, 501
Huang, Wenxin, 186
Huang, Wenxuan, 645, 759
Huang, X. Angelo, 156
Huang, Xiang, 849
Huang, Xianwen, 358
Huang, Xiao, 245, 336, 449, 854
Huang, Xiaobao, 808
Huang, Xiaojiang, 802
Huang, Xiaolei, 132, 213, 536
Huang, Xiaowei, 170, 376
Huang, Xiaoxi, 390
Huang, Xin, 331
Huang, Xingyue, 162, 446
Huang, Xinju, 532
Huang, Xinmiao, 170, 376
Huang, Xinyi, 475
Huang, Xinyue, 636
Huang, Xiuqi, 687
Huang, Xiusheng, 355, 638
Huang, Xuanao, 799
Huang, Xuanjing, 201, 382, 545, 546, 661, 717, 740, 820, 821, 838, 839, 868, 897
Huang, Xudong, 412
Huang, Xuyao, 433
Huang, Yangmin, 417
Huang, Yangyu, 278
HUANG, Yanyi, 164
Huang, Yaokai, 323
Huang, Yeqi, 527
Huang, Yi, 321, 430
Huang, Yi Ting, 906
Huang, Yifei, 776, 844
Huang, Yijie, 671
Huang, Yiling, 545
Huang, Yiming, 97, 391
Huang, Ying, 258
Huang, Ying-Hsiang, 331
Huang, Yinya, 600
Huang, Yiyang, 124
Huang, Yongbo, 762
Huang, Yongfeng, 401
Huang, Yu, 266, 531, 670
Huang, Yu-Shiang, 870
Huang, Yuanhao, 119
Huang, Yuanhong, 421
Huang, Yuchen, 319
Huang, Yucong, 135
Huang, Yue, 82, 259, 311, 471, 684
Huang, Yuehao, 894
Huang, Yuekai, 311, 639
Huang, Yueyuan, 820
Huang, Yujie, 856
Huang, Yukun, 99
Huang, Yunchong, 437
Huang, Yupin, 175
Huang, Yushi, 128
Huang, Yutong, 255, 528
Huang, Yuwei, 186
Huang, Yuxiang, 543
Huang, Zeying, 643
Huang, Zhaoheng, 337, 700
Huang, Zhe, 334
Huang, Zhen, 322, 401, 723, 725, 750
Huang, Zhengxing, 220
Huang, Zhenhan, 343
Huang, Zhenya, 321, 338, 353, 433
Huang, Zhenzhen, 404
Huang, Zhewei, 292
Huang, Zhicong, 770
Huang, Zhiqi, 738
Huang, Zhisheng, 570
Huang, Zhongjing, 311
Huang, Zhongzhen, 255, 528
Huang, Zhuochen, 906
Huang, Zi, 564
Huang, Zihe, 319, 417
Huang, Zijian, 181
Huang, Zijie, 111
Huang, Zikang, 455
Huang, Ziqi, 100, 564
Huang, Zisu, 321
Huang, Ziwei, 128, 683
Huang, Zixuan, 315, 409
Huang, Ziyang, 325, 652
Huang, Ziyi, 305
Huang, Zongmou, 186
Huang, Zuchao, 289
Huang;,
Huang;, Jizhou,

Huang;, Minlie,
Huangyu, 721
Huangyu, Huangyu, 721
Huaping, 569
Huaping, Zhang, 569
Huber, 444, 527
Huber, Patrick, 444
Huber, Thomas, 527
Hudelot, 169, 439
Hudelot, Céline, 169, 439
Hudi, 816
Hudi, Frederikus, 816
Hudspeth, 168
Hudspeth, Marisa, 168
Hueber, 447
Hueber, Thomas, 447
Huh, 234
Huh, Jisu, 234
Hui, 383, 462, 510, 540, 634, 783, 845, 856
Hui, Binyuan, 634
Hui, Haojia, 540
Hui, Julie, 462
Hui, Sun, 783
Hui, Tingfeng, 383
Hui, Zheng, 510, 845
Hui, Zhiwen, 856
HuiKai, 155
HuiKai, Andre Lin, 155
Hulden, 188
Hulden, Mans, 188
Hung, 496, 782
Hung, Ki Sen, 496, 782
Hunter, 811
Hunter, Julie, 811
Huo, 143, 144, 205, 470, 508, 657, 687, 690, 708, 780
Huo, Chang, 508
Huo, Feiye, 708
Huo, Fushuo, 470
Huo, Jiahao, 144, 205
Huo, Jing, 657
Huo, Liangyu, 690
Huo, Mingyue, 143, 780
Huo, Yintong, 687
Hur, 182, 465, 515
Hur, Chan, 515
Hur, Yuna, 182, 465
Hurych, 502
Hurych, Luká Eigler; Jindich Libovický; David,
Hussain, 222, 556
Hussain, Aafiya, 222
Hussain, Syed Rafiul, 556
Huy, 192, 645
Huy, Phan Nht, 192
Huy, Phung Gia, 645
Hwang, 217, 240, 266, 309, 317, 439, 461, 466, 496, 515, 569, 604, 605, 690, 811, 833, 846, 855, 891
Hwang, Angel Hsing-Chi, 439
Hwang, Eui Jun, 515, 833
Hwang, EunJeong, 891
Hwang, Gyurim, 846
Hwang, Jenq-Neng, 569
Hwang, Kyomin, 466
Hwang, Seojin, 811
Hwang, Seokju, 690
Hwang, SeokJun, 266
Hwang, Seonjeong, 309, 317, 461, 605
Hwang, Seung-won, 240, 496
Hwang, Sung Ju, 604, 855
Hwang, Taewook, 217
Hyman, 486
Hyman, Jayden Kana, 486
Hyun, 844
Hyun, Lee, 844
Hägele, 784
Hägele, Alexander, 784
Højer, 891
Højer, Bertram, 891

Iacovides, 538
Iacovides, Giorgos, 538
Ibrahim, 86, 862
Ibrahim, Lujain, 86
Ibrahim, Nuhu, 862
Ibtehaz, 696
Ibtehaz, Nabil, 696
Ichikawa, 445
Ichikawa, Yuma, 445
Idris, 862
Idris, Inshirah, 862
Ieiri, 884
Ieiri, Yuya, 884
Igarashi, 563
Igarashi, Takeo, 563
Iglesias, 803
Iglesias, Gonzalo, 803
II, 127, 629
II, Ahmed Haj Ahmed; Ruochen Zhang; Alvin Grissom,
II, Andrew Ressler, 127
III, 103, 473, 609, 610, 833
III, Fenno F. Heath, 610
III, Hal Daumé, 103, 473, 833
III, Vincil C. Bishop, 609
Ilasariya, 862

Ilasariya, Fenal Ashokbhai, 862
Ilic, 784
Ilic, Alexander, 784
Ili, 898
Ili, David, 898
Illina, 104
Illina, Irina, 104
Im, 612, 824
Im, Chaemoon, 612
Im, Jinbae, 824
Imaizumi, 839
Imaizumi, Masaaki, 839
Imperial, 862
Imperial, Joseph Marvin, 862
Imteaj, 613
Imteaj, Ahmed, 613
In, 142
In, Yeonjun, 142
Inan, 123
Inan, Mert, 123
Inciarte, 102
Inciarte, Alcides Alcoba, 102
Indurthi, 118, 486
Indurthi, Sathish Reddy, 118, 486
Ingle, 770
Ingle, Digvijay Anil, 770
Inostroza, 167
Inostroza, Demian, 167
Inoue, 469
Inoue, Koji, 469
Inui, 273, 839
Inui, Kentaro, 273, 839
Ioannidis, 452, 797, 880
Ioannidis, Vassilis, 797, 880
Ioannidis, Vassilis N., 452
Ioannou, 419
Ioannou, Yani, 419
Iolov, 238
Iolov, Alexandre, 238
Ionescu, 121
Ionescu, Radu Tudor, 121
Iraola, 308
Iraola, David Martinez, 308
Isch, 177
Isch, Calvin, 177
Ishibashi, 812
Ishibashi, Yoichi, 812
Ishigaki, 237, 452
Ishigaki, Tatsuya, 237, 452
Ishihara, 328
Ishihara, Shotaro, 328
Ishii, 158, 598
Ishii, Etsuko, 158, 598
Ishmam, 222, 371
Ishmam, Alvi, 222
Ishmam, Md Farhan, 371
Ishtiaq, 556
Ishtiaq, Abdullah Al, 556
Isingizwe, 815
Isingizwe, Ganza Belise, 815
Islah, 258
Islah, Nizar, 258
Islam, 82, 100, 135, 229, 268, 453, 511, 609, 822
Islam, Md Asiful, 268
Islam, Md Mofijul, 82, 609
Islam, Md. Tanjeed, 100
Islam, Mohammed Saidul, 822
Islam, Mst Rafia, 135
Islam, Sekh Mainul, 453
Islam, Tunazzina, 229, 511
Ismayilzada, 784
Ismayilzada, Mete, 784
Isono, 102, 113
Isono, Shinnosuke, 102, 113
Israel, 148
Israel, Daniel Mingyi, 148
Israeli, 472
Israeli, Abraham, 472
Issaka, 538
Issaka, Sheriff, 538
Ito, 307, 416, 863
Ito, Hitoshi, 863
Ito, Kazuhiro, 307
Ito, Takayuki, 416
Ittichaiwong, 823
Ittichaiwong, Piyalitt, 823
IV, 560
IV, David Randolph, 560
Ivanov, 491, 614
Ivanov, Todor, 614
Ivanov, Vladimir V., 491
Ivetta, 174, 566
Ivetta, Guido, 174, 566
Ivey, 472
Ivey, Jonathan, 472
Iwakura, 838
Iwakura, Tomoya, 838
Iwasawa, 608, 660
Iwasawa, Yusuke, 608, 660
Iyatomi;,
Iyatomi;, Hitoshi,
Iyengar, 526
Iyengar, Anirudh Iyengar Kaniyar Narayana, 526
Iyer, 238, 764
Iyer, Lina, 764
Iyer, Vijayasri, 238

Iyyer, 233, 257, 784
Iyyer, Mohit, 233, 257, 784
Izadi, 644
Izadi, Amirmohammad, 644
Izumi, 463
Izumi, Kiyoshi, 463

Jabr, 471
Jabr, Aisha, 471
Jacobson, 109
Jacobson, Nicholas, 109
Jafari, 597
Jafari, Mehdi, 597
Jagatap, 226
Jagatap, Gauri, 226
Jaggi, 784
Jaggi, Martin, 784
Jahara, 257
Jahara, Fatima, 257
Jain, 122, 161, 265, 285, 428, 512, 557, 562, 608, 609, 632, 784
Jain, Akriti, 122, 557, 784
Jain, Ankit, 285, 562
Jain, Arin, 161
Jain, Ayush, 608
Jain, Hemang, 428
Jain, Niharika, 609
Jain, Nishita, 512
Jain, Siddhartha, 265
Jain, Soumya, 632
Jalilian, 631
Jalilian, Samaneh, 631
Jamalabadi, 529
Jamalabadi, Hamidreza, 529
Jameel, 854
Jameel, Shoaib, 854
James, 464, 830
James, Jim, 464
James, Joseph, 830
Jamil, 238
Jamil, Umar, 238
Jamshidi, 634
Jamshidi, Neda, 634
Jana, 91, 532, 875
Jana, Abhik, 875
Jana, Suman, 91, 532
Jang, 121, 140, 147, 182, 195, 259, 275, 282, 292, 301, 517, 550, 553, 557, 564, 594, 679, 779, 806, 823, 865, 903
Jang, Beakcheol, 140, 550
Jang, Brian, 779
Jang, Geonhui, 517
Jang, Haeun, 182, 557
Jang, Han, 823
Jang, Hwanseok, 594
Jang, Hyeju, 259
Jang, Ijun, 147, 553
Jang, Jaeeun, 195
Jang, Jaehyeok, 301
Jang, Jaehyuk, 282
Jang, Jinhee, 865
Jang, Jinkwan, 553
Jang, Joonwon, 121
Jang, Min Ho, 903
Jang, Seokgyu, 147, 292
Jang, Wonjune, 275, 564
Jang, Yoonna, 594
Jang, Youngjoon, 806, 903
Jang, Youngsoo, 679
Janghorbani, 127
Janghorbani, Sepehr, 127
Janiak, 494
Janiak, Denis, 494
Jannesari, 118
Jannesari, Ali, 118
Jansen, 763, 889
Jansen, Peter, 763, 889
Jara, 454
Jara, Ignacio Meza De la, 454
Jarrar, 494, 859, 860
Jarrar, Mustafa, 494, 859, 860
Jasin, 471
Jasin, Stefanus, 471
Jason, 506
Jason, Lev Sukherman; Artem Frenk; Nina Klimenkova; Connor,
Jatowt, 622, 639, 697, 744, 873
Jatowt, Adam, 622, 639, 697, 744, 873
Jauhar, 479
Jauhar, Sujay Kumar, 479
Javed, 808
Javed, Hamza A, 808
Jayakumar, 654
Jayakumar, Thanmay, 654
Jayannavar, 168
Jayannavar, Prashant, 168
Jena, 420
Jena, Sai Satyam, 420
Jennings, 177
Jennings, Grace, 177
Jeon, 217, 461, 552, 592, 596, 769, 833, 871
Jeon, Donghyeon, 217, 592, 769
Jeon, Eojin, 871
Jeon, Jaehyun, 552
Jeon, Mingyu, 596
Jeon, Myeongho, 217

Jeon, Woosang, 833
Jeon, Yejin, 461
Jeong, 240, 550, 560, 595, 604, 612, 855, 888
Jeong, Jaehwan, 560
Jeong, Jinhong, 595
Jeong, Minchan, 888
Jeong, Seogyeong, 612
Jeong, Soyeong, 604, 855
Jeong, Yeonseok, 240
Jeong, Yujin, 550
Jeoung, 797, 880
Jeoung, Sullam, 797, 880
Jerzak, 188
Jerzak, Connor T., 188
Jerzak, Ethan, 188
Jha, 263, 317, 637, 831
Jha, Aman Vaibhav, 637
Jha, Samyak, 317
Jha, Somesh, 831
Jha, Sumit Kumar, 263
Jha, Susmit, 263
Jhang, 269, 765
Jhang, Minha, 269, 765
Jhaveri, 613
Jhaveri, Parin Rajesh, 613
Ji, 85, 95, 97, 114, 121, 125, 151, 191, 210, 221, 230, 234, 262, 280, 292, 335, 361, 373, 428, 431, 434, 440, 456, 470, 475, 513, 514, 527, 535, 543, 545, 550, 566, 573, 597, 603, 606, 607, 626, 647, 678, 703, 725, 733, 748, 761, 780, 802, 818, 856, 877, 897, 901
ji, 97
Ji, Bin, 733
Ji, Bokai, 335
Ji, Daeun, 550
Ji, Deyi, 607
Ji, Donghong, 114, 647, 856
Ji, Fanlin, 818
Ji, Fengxian, 475, 527
Ji, Heng, 151, 262, 292, 440, 573, 603, 626, 802
Ji, Hongru, 703
Ji, Huawei, 361
Ji, Jiaming, 85, 470, 513, 543, 545, 748
Ji, Jiarui, 234
Ji, Jiyuan, 95
Ji, Lei, 210, 280, 431, 573
Ji, Luo, 191
Ji, Luyao, 97
Ji, Mengyuan, 550
Ji, Shaoxiong, 230
Ji, Shengpeng, 114, 780
Ji, Shouling, 121, 566, 703, 877
Ji, Shuiwang, 901
Ji, Tao, 897
Ji, TianTian, 597
Ji, Wei, 514
Ji, Xiangbo, 373
Ji, Yang, 535
Ji, Yicheng, 125, 221
Ji, Yifan, 434
Ji, Yixin, 678, 761
ji, Yu, 97
Ji, Yupeng, 725
Ji, Zhuolin, 606
Ji, Zibo, 428
Ji, Zihao, 456
Jia, 91, 98, 114, 115, 120, 137, 170, 189, 214, 225, 249, 341, 343, 348, 349, 374, 381, 423, 433, 437, 441, 464, 499, 505, 512, 559, 562, 587, 639, 644, 671, 681, 698, 701, 707, 708, 744, 777, 790, 791, 796, 798, 879
Jia, Bing, 644
Jia, Caijun, 777
Jia, Change, 437
Jia, Changguo, 698
Jia, Guoli, 790
Jia, Jiaya, 423, 433
Jia, Jie, 671
Jia, Jinyuan, 559
Jia, Junhao, 91, 374
Jia, Linwei, 796
Jia, Lu, 98
Jia, Mengzhao, 505, 879
Jia, Minghui, 249
Jia, Pengyue, 214, 464, 587, 791
Jia, Qi, 559
Jia, Qianghuai, 441, 798
Jia, Ran, 349
Jia, Robin, 225
Jia, Runsong, 681
Jia, Sihang, 707
Jia, Wei, 701
Jia, Weijia, 137, 189, 512
Jia, Weile, 708
Jia, Xiaofeng, 341
Jia, Xiaohua, 343
Jia, Xiaojun, 170, 639
Jia, Xiaoqi, 499
Jia, Xiyuan, 120
Jia, Yaning, 115
Jia, Yuhang, 744
Jia, Yuheng, 381
Jia, Zheng, 562
Jia, Zhiyang, 114

Jia, Zijian, 348
Jian, 345, 399, 453, 636, 665, 733, 839
Jian, Ai, 345, 636
Jian, Ping, 453
Jian, Songlei, 733
Jian, Tong, 839
Jian, Xiangru, 399
Jian, Zhongquan, 665
Jiang, 82, 84, 87, 88, 97, 117, 120, 125, 135, 138, 142, 145, 146, 149, 156, 158, 163, 165, 166, 167, 170, 173, 181, 184, 185, 187, 188, 194, 201, 227, 228, 230, 233, 235, 258, 260, 264, 275, 279, 280, 288, 290, 295, 303, 311, 313, 323, 324, 337, 341, 344, 350, 352, 353, 361, 378, 380, 384, 397, 399, 407, 409, 413, 418, 423, 424, 429, 441, 445, 450, 454, 455, 459, 467, 471, 474, 475, 485, 488, 490, 491, 504, 505, 515, 531, 540, 542, 543, 545, 549, 550, 555, 560, 563, 569, 583, 588, 592, 598, 599, 600, 605, 607, 608, 611, 616, 621, 623, 628, 631, 639, 647, 650, 653, 654, 662, 664, 665, 666, 679, 683, 685, 688, 689, 690, 694, 701, 703, 716, 717, 723, 733, 741, 749, 750, 757, 758, 759, 769, 770, 777, 779, 784, 786, 788, 789, 793, 796, 801, 804, 805, 809, 813, 817, 820, 823, 844, 847, 848, 850, 852, 853, 857, 864, 877, 879, 881, 893, 897, 901, 903, 907, 908
jiang, 643
Jiang, Angqing, 569
Jiang, Aoyuan, 877
Jiang, Bo, 485
Jiang, Bohan, 185
Jiang, Botian, 228
Jiang, Changhao, 717, 820, 897
Jiang, Changjun, 459
Jiang, Chaoyi, 488
Jiang, Che, 413
Jiang, Cheng, 540
Jiang, Chenhao, 344
Jiang, Chunyang, 543
Jiang, Cong, 549
Jiang, Daxin, 560
Jiang, Dazhi, 786
Jiang, Dongfu, 901
Jiang, Dongmei, 350
Jiang, Dongwei, 598
Jiang, Eric Hanchen, 280, 504, 758
Jiang, Fei, 384
Jiang, Feng, 142, 474
Jiang, Fengqing, 599, 879
Jiang, Guangxin, 679
Jiang, Guanxuan, 857
Jiang, Guanyu, 769
Jiang, Haiyang, 418
Jiang, Haiyun, 88
Jiang, Hanqi, 599
Jiang, Hanyun, 184
Jiang, Hao, 290
Jiang, Hengle, 908
Jiang, Hongchao, 142, 275
Jiang, Hui, 441
Jiang, Jane Yi, 471
Jiang, Jiaming, 784
Jiang, Jian, 685
Jiang, Jianguo, 429
Jiang, Jiawei, 664
Jiang, Jiazhen, 653
Jiang, Jie, 87, 605, 607, 616, 847
Jiang, Jinfeng, 531
Jiang, Jing, 187
Jiang, Jingchi, 397
Jiang, Jingyan, 145
Jiang, Jingyi, 361
Jiang, Jiuchuan, 654
Jiang, Jiuyun, 679
Jiang, Junfeng, 823
Jiang, Kehan, 445, 788
Jiang, Lei, 165, 230, 295, 588
Jiang, Li, 805
Jiang, Lingjie, 167, 793, 907
Jiang, Lu, 290
Jiang, Lutao, 666
Jiang, Meng, 181, 188, 505, 583, 817, 879
Jiang, Menghua, 324
Jiang, Ming, 683
Jiang, Mingyang, 166
Jiang, Mohan, 665
Jiang, Owen, 623
Jiang, Peijie, 694
Jiang, Ping, 881
Jiang, Qian, 117
Jiang, Qingxuan, 149, 716
Jiang, Qinting, 145
Jiang, Qiuping, 407
Jiang, Qizhi, 608
Jiang, Renhe, 488, 563
Jiang, Rongqing, 323
Jiang, Runze, 337
Jiang, Ruomeng, 550
Jiang, Shanshan, 97, 741
Jiang, Shengpei, 621
Jiang, Shenyuan, 397, 759
Jiang, Shixin, 690

Jiang, Shutong, 158
Jiang, Shuyang, 409
Jiang, Sihang, 757, 848, 908
Jiang, Siyu, 689
Jiang, Songtao, 450
Jiang, Tanqiu, 117
Jiang, Tianyi, 399
Jiang, Tianyu, 173, 454, 809, 813, 864
Jiang, Tingsong, 135
Jiang, Wei, 258, 341, 491, 733
Jiang, Weijie, 350
Jiang, Weilong, 689
Jiang, Weipeng, 844
Jiang, Wenhao, 303, 515, 650
Jiang, Wenjia, 146
Jiang, Wenxi, 662
Jiang, Wenyuan, 156, 475, 600
Jiang, Xiaofeng, 194
Jiang, Xiaoqian, 288
jiang, xiaoxi, 643
Jiang, Xiaoyan, 852
Jiang, Xinke, 264, 303, 850
Jiang, Xinqi, 703
Jiang, Xirui, 459
Jiang, Xu, 701
Jiang, Xue, 97, 260, 531, 796
Jiang, Xuejun, 893
Jiang, Xuhui, 723
Jiang, Yanbei, 201, 545
Jiang, Yangqin, 138
Jiang, Yangyang, 170
Jiang, Ye, 235
Jiang, Yi, 353, 378
Jiang, Yikun, 804
Jiang, Yilei, 280
Jiang, Yilun, 125
Jiang, Yishun, 84
Jiang, Yiwei, 631
Jiang, Yiwen, 592
Jiang, Yixuan, 227
Jiang, Yiyang, 542
Jiang, Yong, 120, 380, 853
Jiang, Yu, 82, 455
Jiang, Yu-Gang, 352
Jiang, Yucheng, 789
Jiang, Yuechen, 166, 233
Jiang, Yueqi, 789
Jiang, Yuhan, 786
Jiang, Yuhao, 540
Jiang, Yuhua, 413, 423, 750
Jiang, Yukun, 628, 647
Jiang, Yuning, 341, 733
Jiang, Yuxian, 777
Jiang, Yuxuan, 163, 490, 683, 779, 857
Jiang, Zefang Zong; Dingwei Chen; Yang Li; Qi Yi; Bo Zhou; Chengming Li; Bo Qian; Peng Chen; Jie,
Jiang, Zheng, 688
Jiang, Zhengping, 555
Jiang, Zhengyuan, 611, 770
Jiang, Zhihua, 313
Jiang, Zhiwei, 97, 801
Jiang, Zhiyun, 380
Jiang, Zhonglin, 685
Jiang, Zhongtao, 424
Jiang, Zhuoren, 467, 749
Jiang, Zifan, 903
Jiang, Ziyou, 311, 639, 749
jianyang, 774
jianyang, jianyang, 774
Jiao, 134, 144, 155, 260, 352, 388, 466, 487, 531, 584, 604, 635, 724, 739, 791, 882, 892
Jiao, Difan, 487, 584
Jiao, Hui, 724
Jiao, Huitian, 791
Jiao, Jian, 466
Jiao, Jianbin, 134
Jiao, Kechen, 739
Jiao, Long, 144
Jiao, Pengkun, 352
Jiao, Wenpin, 155, 260, 531
Jiao, Wenxiang, 882, 892
Jiao, Yihan, 604
Jiao, Ying, 635
Jiao, Ziyun, 388
JiaQingCaiWang, 809
JiaQingCaiWang, JiaQingCaiWang, 809
Jiawen, 391
Jiawen, Duan, 391
Jiayang, 243
Jiayang, Cheng, 243
Jie, 246
Jie, Shan, 246
Jimenez, 538
Jimenez, Nicte Aguillon, 538
Jin, 88, 90, 94, 97, 112, 117, 132, 135, 138, 141, 147, 155, 156, 162, 194, 195, 204, 216, 220, 239, 260, 293, 303, 306, 313, 314, 324, 351, 353, 355, 358, 361, 375, 377, 384, 408, 418, 425, 432, 443, 450, 468, 482, 491, 497, 526, 531, 538, 553, 554, 584, 590, 591, 600, 606, 613, 631, 642, 650, 652, 659, 678, 689, 692, 706, 708, 715, 721, 725, 728, 749, 752, 754, 760, 783, 784, 787, 795, 796, 801, 809, 821,

836, 838, 848, 853, 856, 861, 881, 904, 905
Jin, Bo, 138, 351
Jin, Bojun, 715
Jin, Can, 141
Jin, Changhong, 468
Jin, Cheyon, 147
Jin, Chunxiang, 220
Jin, Dong, 194, 853
Jin, Fangzhou, 754
Jin, Fengmei, 418
Jin, Guanghao, 689
Jin, Hai, 375
Jin, Haibo, 584
Jin, Han, 905
Jin, Haojian, 809
Jin, Hengzhe, 135
Jin, Hongbo, 760
Jin, Huiyan, 303
Jin, Hyunbin, 553
Jin, Jiahe, 314
Jin, Jiahui, 88
Jin, Jiajie, 728
Jin, Jihang, 836
Jin, Jing, 117, 659
Jin, Kaiming, 591
Jin, Keyan, 783
Jin, Lei, 652
Jin, Li, 162
Jin, Lianwen, 324
Jin, Luozhijie, 204, 708
Jin, Ming, 443
Jin, Peiquan, 353
Jin, Qin, 90, 195, 293, 497, 861
Jin, Renren, 94, 306, 590
Jin, Rong, 749, 795
Jin, Ruihan, 313, 706
Jin, Ruochun, 721
Jin, Senjie, 821, 838
Jin, Song, 384
Jin, Tao, 355, 650, 678, 749
Jin, Wei, 443, 881
Jin, Xiaobo, 725
Jin, Xiaokang, 132, 216
Jin, Xiaolong, 784
Jin, Xudong, 554
Jin, Yaochu, 752
Jin, Yaohui, 377
Jin, Yeying, 450
Jin, Yichao, 606
Jin, Yichen, 760
Jin, Yilin, 408, 425
Jin, Yingbin, 482
Jin, Yiqiao, 112, 538, 613, 904
Jin, Yongkang, 162
Jin, Yuan, 361
Jin, Zehao, 425
Jin, Zengrui, 432, 848
Jin, Zhanpeng, 491, 801, 856
Jin, Zhengda, 692
Jin, Zhenghui, 787
Jin, Zhenjiang, 239
Jin, Zhi, 97, 155, 260, 531, 796
Jin, Zhijing, 156, 526, 600
Jin, Zhuoran, 358, 631, 642
Jinan, 96, 583
Jinan, Xu, 96, 583
Jindal, 161
Jindal, Divay, 161
JING, 496, 647, 882
Jing, 286, 333, 347, 388, 417, 463, 465, 576, 634, 660, 785, 790, 874, 883, 886
Jing, Baoyu, 347
Jing, Bohao, 286
Jing, Chenchen, 576, 785
Jing, Dong, 333
JING, Huihao, 496, 647, 882
Jing, Huiyun, 660
Jing, Liqiang, 463
Jing, Peiyuan, 790
Jing, Shaoling, 417
Jing, Yan, 874
Jing, Yi, 465
Jing, Yongcheng, 388, 883
Jing, Yuheng, 634
Jing, Zhang, 886
Jingwen, 560
Jingwen, Zhang, 560
Jinji, 823
Jinji, Michinori, 823
Jinyuan, 536
Jinyuan, Fu, 536
Jo, 133, 149, 206, 223, 280, 441, 447, 787
Jo, Dongwon, 206
Jo, Hwiyeol, 280
Jo, Sungmin, 149
Jo, Yohan, 133, 149, 223, 441, 447, 787
Jo;,
Jo;, Yohan,
Johansson, 572
Johansson, Richard, 572
John_Guerrerio,
John_Guerrerio, John_Guerrerio,
Johnson, 500, 553, 831
Johnson, Jack A., 500
Johnson, Kristina, 831

Johnson, Steven L., 553
Jones, 108
Jones, Cameron, 108
Jong, 364
Jong, Matei Benescu; Ivo Pascal de,
Jongyoon_Kim1,
Jongyoon_Kim1, Jongyoon_Kim1,
Joo, 780
Joo, Hyunjung, 780
Joseph, 104, 156
Joseph, Sebastian, 104, 156
Joshi, 207, 236, 282, 843
Joshi, Abhinav, 282
Joshi, Brihi, 843
Joshi, Shruti, 207
Joshi, Smriti, 236
Josyula, 730
Josyula, Yasasvi, 730
Joty, 123, 285, 822, 866
Joty, Shafiq, 123, 285, 822, 866
Joung, 775
Joung, Ji Hoon, 775
Jourdan;,
Jourdan;, Fanny,
Joy, 375
Joy, Saman Sarker, 375
Jr., 588
Jr., Wagner Meira, 588
Ju, 162, 371, 375, 401, 446, 644, 741, 779, 884
Ju, Hyeonjong, 741
Ju, Mingxuan, 162, 446, 644
Ju, Tong, 401
Ju, Wei, 375
Ju, Xiaozhu, 884
Ju, Ying, 371
Ju, Zeqian, 779
Jubair, 815
Jubair, Fahed, 815
Juefei-Xu, 562
Juefei-Xu, Felix, 562
Jumelet, 114
Jumelet, Jaap, 114
Jun, 285
Jun, Hyunmin, 285
Juneja, 235
Juneja, Prerna, 235
Jung, 91, 115, 118, 177, 182, 214, 217, 269, 275, 283, 296, 301, 462, 469, 477, 580, 594, 604, 616, 753, 765, 785, 831, 834, 859, 900, 905
Jung, Chani, 831
Jung, Dahyun, 182
Jung, Donghoon, 905
Jung, Dongwon, 594
Jung, Hayeong, 118
Jung, Hayoung, 115
Jung, Hojung, 214
Jung, Hyunji, 91
Jung, Jeesu, 462, 859
Jung, Jimin, 177, 469
Jung, Junha, 275
Jung, Kyomin, 269, 283, 296, 765, 900
Jung, Kyudan, 301, 616
Jung, MinJae, 753
Jung, Sangkeun, 217, 462, 859
Jung, Sangwon, 580
Jung, Seohyon, 905
Jung, Sungmok, 477
Jung, Taehee, 604
Jung, Young Cheol, 834
Jung, Yuchul, 785
Jung-Mok, 844
Jung-Mok, Lee, 844
Junhao, 405
Junhao, Ren, 405
Junias, 468
Junias, Obed, 468
Junjie, 329
Junjie, Hu, 329
Jurafsky, 458, 764
Jurafsky, Dan, 458, 764
Juraska, 342
Juraska, Juraj, 342
Jurgens, 108, 472, 764, 844
Jurgens, David, 108, 472, 764, 844
Jyothi, 486
Jyothi, Sangeetha Abdu, 486
Jäger, 396
Jäger, Gerhard, 396

K, 818
K, Thennal D, 818
Ka, 259
Ka, Keummin, 259
Kabbara, 250, 808
Kabbara, Jad, 250, 808
Kabir, 117
Kabir, Tasnim, 117
Kachwala, 560
Kachwala, Zoher, 560
Kadhe, 224
Kadhe, Swanand Ravindra, 224
Kadiyala;,
Kadiyala;, Bharadwaj,
Kadlík, 891
Kadlík, Marek, 891

Kaesberg, 174
Kaesberg, Lars Benedikt, 174
Kaestner, 306
Kaestner, Christian, 306
Kafai, 612
Kafai, Mehran, 612
Kagitha, 151
Kagitha, Prabhu Prakash, 151
Kahng, 121
Kahng, Minsuk, 121
Kahou, 258
Kahou, Samira, 258
Kai, 903
Kai, Jushi, 903
Kajare, 649
Kajare, Prajwal Vijay, 649
Kajdanowicz, 483, 494, 495
Kajdanowicz, Mateusz Bystroski; Doheon Han; Nitesh V. Chawla; Tomasz Jan,
Kajdanowicz, Monika, 494
Kajdanowicz, Tomasz Jan, 483, 494
Kajiwara, 489
Kajiwara, Kanade Nonomura; Keita Fukushima; Risa Kondo; Tomoyuki,
KAKADE, 293
KAKADE, ADITYA NAMDEV, 293
Kakizaki, 489
Kakizaki, Kazuya, 489
Kalagi, 184
Kalagi, Vikas, 184
Kaliosis, 533
Kaliosis, Panagiotis, 533
Kallala, 613
Kallala, Aymen, 613
Kallini, 591
Kallini, Julie, 591
Kalohia, 612
Kalohia, Ankur, 612
Kalyanpur, 385
Kalyanpur, Aditya, 385
Kamarshi, 226
Kamarshi, Vijay, 226
Kambadur, 139
Kambadur, Melanie, 139
Kambhampati, 185
Kambhampati, Subbarao, 185
Kambhatla, 209
Kambhatla, Gauri, 209
Kamigaito, 99, 492, 512, 587, 605, 832, 838
Kamigaito, Hidetaka, 99, 492, 512, 587, 605, 832, 838
Kamsin, 857
Kamsin, Amirrudin, 857
Kan, 241, 248, 363, 625, 846
Kan, Min-Yen, 241, 248, 363, 625
Kan, Xuan, 846
Kanade, 484, 895
Kanade, Aditya Sanjiv, 484, 895
Kanagarajan, 846
Kanagarajan, Abinesh, 846
Kanametov, 330
Kanametov, Azamat, 330
Kanayama, 240
Kanayama, Hiroshi, 240
Kanazawa, 445
Kanazawa, Yuzi, 445
Kancheti, 484, 895
Kancheti, Sai Srinivas, 484, 895
Kaneko, 236, 463, 565, 813
Kaneko, Masahiro, 236, 463, 565, 813
Kanemoto, 416
Kanemoto, Yuya, 416
Kang, 162, 182, 185, 186, 216, 217, 230, 249, 252, 262, 267, 275, 285, 308, 372, 391, 424, 445, 461, 465, 474, 486, 507, 509, 516, 535, 563, 564, 592, 601, 604, 605, 669, 686, 691, 762, 769, 783, 788, 818, 834, 894
Kang, Byeongcheol, 486
Kang, Chaean, 834
Kang, Chen, 372
Kang, Choongwon, 285
Kang, Daniel, 262
Kang, Deokhyung, 605
Kang, Donghun, 267
Kang, Dongyeop, 563, 604, 783
Kang, Hong, 509
Kang, Inho, 217, 592, 769
Kang, Jaewoo, 275, 564
Kang, Jikun, 507, 509
Kang, Joeun, 834
Kang, Kunpeng, 762
Kang, Kyeongpil, 186
Kang, Liu, 216
Kang, Myeong-Cheol, 691
Kang, Myunghoon, 182, 465
Kang, Peng, 474
Kang, Sangwoo, 252, 535
Kang, Seokhoon, 461
Kang, SeongKu, 230
Kang, Suho, 601
Kang, U, 516
Kang, Xiaomian, 162, 391
Kang, Xiaoxi, 249
Kang, Yipeng, 424, 818
Kang, Yu, 185

Kang, Yueping, 687
Kang, Zhanhui, 669
Kang, Zhaolu, 186, 445, 788, 894
Kang, Zhongfeng, 308
Kanganis, 216
Kanganis, Alisa, 216
Kani, 823
Kani, Nickvash, 823
Kanithi, 808
Kanithi, Praveenkumar, 808
Kannan, 250
Kannan, Vidhyakshaya, 250
Kanoulas, 254
Kanoulas, Evangelos, 254
Kao, 150, 159, 466, 514
Kao, Ben, 159
Kao, Chia Hsiang, 150
Kao, Hsien-Te, 466
Kao, Hung-Yu, 514
Kapadia, 443
Kapadia, Shashank, 443
Kapadnis, 189
Kapadnis, Manav, 189
Kapdi, 81
Kapdi, Mohsin, 81
Kaplan, 168, 603, 794, 887
Kaplan, Elizabeth, 168
Kaplan, Erel, 794
Kaplan, Guy, 603, 887
Kapoor, 609, 616
Kapoor, Hitesh, 616
Kapoor, Vishaal, 609
Kapur, 306
Kapur, Rhea, 306
Karande, 293
Karande, Shirish, 293
Karasik, 284
Karasik, Michael, 284
Kardinata, 481
Kardinata, Eunike Andriani, 481
Kargupta, 789
Kargupta, Priyanka, 789
Karimi, 589
Karimi, Akbar, 589
Karimireddy, 83
Karimireddy, Sai Praneeth, 83
Karivelil, 611
Karivelil, Mathew, 611
Karkar, 513
Karkar, Ravi, 513
Karlina, 469
Karlina, Wiwik, 469
Karlinsky, 116
Karlinsky, Leonid, 116
Karmakar, 875
Karmakar, Priyanshu, 875
Karmaker, 260, 532
Karmaker, Santu, 260, 532
Karpenko, 658
Karpenko, Volodymyr, 658
Karpinska, 233
Karpinska, Marzena, 233
Karray, 88
Karray, Fakhri, 88
Karthika, 188
Karthika, NJ, 188
Kartono, 121
Kartono, Brigitta Jesica, 121
Kartá, 872
Kartá, Ivan, 872
Kashima, 823
Kashima, Hisashi, 823
Kashiwagi, 111
Kashiwagi, Yosuke, 111
Kashyap, 458, 683, 845
Kashyap, Gautam Siddharth, 458, 845
Kashyap, Omkar Mahesh Kashyap; Padegal Amit; Ashwini M Joshi; Shylaja S S; Madhav,
Kasneci, 495, 497, 540, 762, 878
Kasneci, Enkelejda, 495
Kasneci, Gjergji, 497, 540, 762, 878
Katerenchuk, 238
Katerenchuk, Denys, 238
Kathuria, 892
Kathuria, Tarun, 892
Katsis, 558
Katsis, Yannis, 558
Katyal, 227
Katyal, Yatin, 227
Katz, 600, 771
Katz, Dmitriy A., 771
Katz, Shahar, 600
Katz, Yoav, 771
Katz-Samuels, 444
Katz-Samuels, Julian, 444
Kaur, 115, 180, 189, 613, 904
Kaur, Kirandeep, 189
Kaur, Navreet, 115
Kaur, Rachneet, 180, 613, 904
Kaushik, 844
Kaushik, Roshni, 844
Kautsar, 864
Kautsar, Muhammad Dehan Al, 864
Kavathekar, 428
Kavathekar, Ishan, 428

Kawabata, 92
Kawabata, Akira, 92
Kawada, 480
Kawada, Takuro, 480
Kawaguchi, 625
Kawaguchi, Kenji, 625
Kawahara, 113, 255, 469, 485
Kawahara, Daisuke, 255, 485
Kawahara, Tatsuya, 469
Kawahara, Yoshinobu, 113
Kawahara;,
Kawahara;, Daisuke,
Kawai, 863
Kawai, Yoshihiko, 863
Kawamura, 156
Kawamura, Kazuki, 156
Kawsar, 549
Kawsar, Fahim, 549
Kazienko, 518
Kazienko, Jacek Duszenko;Jan Koco;Przemysaw,
Kazistova, 342
Kazistova, Kristina, 342
Ke, 392, 400, 418, 497, 608, 649, 719, 727, 851
Ke, Ao, 400
Ke, Changxin, 418
Ke, Guolin, 649
Ke, Pei, 497, 608, 851
Ke, Wenjun, 392, 719, 727
Kechadi, 468
Kechadi, Tahar, 468
keerthi, 471
keerthi, Satya, 471
Keith, 84, 216
Keith, Katherine, 84
Keith, Katherine A., 216
Keleg, 201, 862, 864
Keleg, Amr, 201, 862, 864
Kele, 302
Kele, Onur, 302
Keller, 302, 825
Keller, Frank, 302
Keller, Lennart, 825
Kemp, 167
Kemp, Charles, 167
Kennington, 902
Kennington, Casey, 902
Keramati, 523
Keramati, Ali, 523
Kerdabadi, 220, 530
Kerdabadi, Mohsen Nayebi, 220, 530
Kerdthaisong, 862
Kerdthaisong, Kun, 862
Kersting, 289, 298
Kersting, Kristian, 289, 298
Kertkeidkachorn, 306
Kertkeidkachorn, Natthawut, 306
Kesen, 862
Kesen, Ilker, 862
Keshari, 160
Keshari, Arsh, 160
Kessaci, 598
Kessaci, Yacine, 598
Keswani, 575
Keswani, Sumit, 575
Ketharaman, 574
Ketharaman, Harshita, 574
Kettimuthu, 872
Kettimuthu, Rajkumar, 872
Kevin, 862
Kevin, Jun, 862
Keyaki, 485
Keyaki, Atsushi, 485
Keyaki, Ribeka, 485
Keydar, 82
Keydar, Renana, 82
Khabsa, 118
Khabsa, Madian, 118
Khakwani, 526
Khakwani, Kaneez Zahra Rubab, 526
Khalifa, 133
Khalifa, Muhammad, 133
khalil, 95
khalil, Marwa Elsaid, 95
Khan, 99, 139, 145, 247, 253, 293, 328, 355, 394, 408, 526, 562
Khan, Ahmad Raza, 562
Khan, Fahad Shahbaz, 139, 394
Khan, Faisal Nadeem, 145
Khan, Hamza, 355
Khan, Kris Prasad; Md Abdullah Al Hafiz,
Khan, Mohammad Aflah, 99
Khan, Muhammad Ali, 526
Khan, Muhammad Haris, 394, 408
Khan, Salman, 139, 253, 394
Khan, Zaid, 293
Khan, Zohaib, 247
Khandelwal, 337
Khandelwal, Dinesh, 337
Khang, 536
Khang, Le Nguyen, 536
Khanina, 308
Khanina, Anna, 308
Khanna, 107
Khanna, Rajiv, 107
Khanuja, 112, 246
Khanuja, Harneet Singh, 112

Khanuja, Simran, 246
Kharlamov, 422
Kharlamov, Evgeny, 422
Kharlamova, 253
Kharlamova, Arina, 253
Khashabi, 313, 423
Khashabi, Daniel, 313, 423
Khatri, 167
Khatri, Chandra, 167
Khatun, 480
Khatun, Ayesha, 480
Khatwani, 822
Khatwani, Saksham, 822
Kheir, 341
Kheir, Yassine El, 341
Khentout, 447
Khentout, Manel, 447
Khodabandeh, 285
Khodabandeh, Mehran, 285
Khodorchenko, 655
Khodorchenko, Maria, 655
Khor, 87
Khor, Huai Yuen, 87
Khorramrouz, 199
Khorramrouz, Adel, 199
Khosla, 444
Khosla, Sopan, 444
Khot, 452
Khot, Tushar, 452
Khoury, 146
Khoury, Richard, 146
Khraishi, 160
Khraishi, Raad, 160
KhudaBukhsh, 839
KhudaBukhsh, Ashiqur R., 839
Ki, 473
Ki, Dayeon, 473
Kiefer, 107
Kiefer, Lotta, 107
Kiegeland, 110, 243
Kiegeland, Samuel, 110, 243
Kieu, 281
Kieu, Tung, 281
Kihara, 696
Kihara, Daisuke, 696
Kilicoglu, 240
Kilicoglu, Halil, 240
Kim, 90, 93, 101, 104, 108, 113, 116, 118, 123, 129, 139, 140, 142, 147, 151, 157, 160, 161, 166, 171, 175, 177, 179, 183, 185, 186, 188, 189, 195, 197, 202, 206, 209, 210, 213, 214, 217, 223, 227, 229, 240, 242, 251, 252, 255, 256, 259, 262, 266, 267, 274, 279, 280, 282, 283, 285, 288, 290, 291, 292, 296, 297, 299, 305, 309, 317, 385, 390, 395, 441, 444, 450, 460, 461, 466, 469, 472, 474, 477, 478, 479, 480, 516, 528, 532, 533, 535, 550, 552, 553, 564, 568, 572, 593, 595, 596, 600, 601, 604, 605, 612, 616, 637, 639, 651, 679, 690, 741, 753, 769, 775, 796, 803, 806, 811, 823, 828, 830, 831, 833, 834, 840, 846, 855, 858, 865, 870, 881, 885, 887, 889, 890, 891, 893, 898, 900, 903, 908
Kim, A Young, 123
Kim, Beomsoo, 550
Kim, Bugeun, 118
Kim, Byeongjeong, 811
Kim, Byeongjin, 898
Kim, Byung-Hoon, 197
Kim, Byungjun, 395
Kim, Chaeyun, 753
Kim, Changhyeon, 292
Kim, Changhyun, 461
Kim, Changick, 282
Kim, Chanhwi, 564
Kim, Dae Hyun, 552
Kim, Daehui, 605
Kim, Dahye, 461
Kim, Dohoon, 450
Kim, Dohyoung, 305
Kim, Dong-Kyum, 283
Kim, Dong-Yun, 834
Kim, Donghoon, 291
Kim, Donghyun, 690, 741
Kim, Dongjun, 297, 469, 858
Kim, Dongwhi, 188
Kim, Doyoung, 828
Kim, Eunsu, 197, 612, 831
Kim, Euntae, 890
Kim, Eunwon, 116
Kim, Gabeen, 186
Kim, Gun Il, 140
Kim, Gunhee, 908
Kim, Gyuhwan, 274
Kim, Gyunyeop, 252, 535
Kim, Gyuwan, 898
Kim, HanBeul, 564
Kim, Hanbin, 552
Kim, Hannah, 840
Kim, Hansaem, 593, 834
Kim, Heedou, 161
Kim, Heeju, 118
Kim, Heeseung, 113
Kim, Hong-Gee, 460, 535

Kim, Hye Hyeon, 870
Kim, Hyemi, 870
Kim, Hyeongseob, 157
Kim, Hyeonjin, 466
Kim, Hyeonsang, 157
Kim, Hyo Jin, 480
Kim, Hyounghun, 213, 309, 317, 605, 803
Kim, Hyunjae, 823
Kim, Hyunjong, 299
Kim, Iseo, 256
Kim, Jae-Joon, 206
Kim, Jaehee, 593
Kim, Jaeho, 612
Kim, JaeHong, 601
Kim, Jaehong, 893
Kim, Jaehyung, 166, 528, 595, 885, 889
Kim, Jaekwang, 279
Kim, Jaemin, 223
Kim, Jaeyeon, 908
Kim, Jaeyoung, 240
Kim, Jang Hyun, 285
Kim, Janie, 532
Kim, Jeonghoon, 280, 616
Kim, Jeonghwan, 262, 292
Kim, Jieun, 185, 550
Kim, JiHak, 593
Kim, Jihwan, 595, 616
Kim, Jihyuk, 240
Kim, Jimyeong, 600
Kim, Jin Young, 903
Kim, Jin-Dong, 251
Kim, Jina, 217
Kim, Jinmo, 157
Kim, Jiseon, 893
Kim, Jisung, 775
Kim, Jiyeon, 179
Kim, Joeun, 828
Kim, Jong Wook, 140
Kim, Jongho, 240
Kim, Jongjin, 516
Kim, Jongwoo, 390
Kim, Joo-Kyung, 604
Kim, Joongheon, 612
Kim, Juae, 256
Kim, Juhyeong, 252
Kim, Jun Seo, 870
Kim, Jung-jae, 637, 651
Kim, JungBae, 305
Kim, Junghwan, 108, 472, 753
Kim, JungHyoun, 266
Kim, Jungwhan, 267
Kim, Junhyeok, 775
Kim, Junmo, 887
Kim, Junseok, 296
Kim, Junu, 210
Kim, Junwan, 900
Kim, Junyeong, 596
Kim, Junyoung, 568
Kim, Juyeon, 846
Kim, Kang-Min, 255
Kim, Kangsan, 855
Kim, Kihyun, 753
Kim, Kitae, 104
Kim, Kon Woo, 251
Kim, Kwang In, 139
Kim, Kyeonghyun, 116
Kim, Kyumin, 139
Kim, Kyungho, 846
Kim, Kyusik, 104, 229, 267, 533
Kim, Laerdon, 478
Kim, Mansu, 255
Kim, Mi-Young, 227
Kim, MiHyeon, 803
Kim, Min Soo, 552
Kim, Min-Jae, 189
Kim, Minhyuk, 267, 806
Kim, Minjae, 160
Kim, Minju, 741
Kim, Minjun, 516
Kim, Minkyoung, 550
Kim, Minsung, 283, 900
Kim, Minwoo, 753
Kim, Moonkwon, 769
Kim, Myeongsoo, 881
Kim, MyoungJin, 177
Kim, Najoung, 197, 796
Kim, Nalim, 441
Kim, San, 171
Kim, Sangho, 477
Kim, Sangmin, 195
Kim, Sejoon, 830
Kim, Seongwan, 444
Kim, Seung-Bin, 101, 209
Kim, Seunghee, 292
Kim, Seungone, 474
Kim, Seungyeon, 93
Kim, Shubin, 259
Kim, Siun, 385
Kim, Soohyeong, 266
Kim, Soomin, 533
Kim, Sooyong, 639
Kim, Soyeon, 202
Kim, Soyoon, 616
Kim, Sumin, 160
Kim, Sungkyun, 223
Kim, Sunwoo, 846

Kim, Taehee, 595
Kim, Taehwan, 129
Kim, Taehyeon, 214, 290, 679
Kim, Taehyeong, 833
Kim, Taesup, 147, 553, 885
Kim, Taeuk, 90, 147, 292
Kim, Takyoung, 183, 441
Kim, To Eun, 151
Kim, Unsol, 593
Kim, Wonjoong, 142
Kim, Yeachan, 479
Kim, Yejin, 288
Kim, Yeongho, 846
Kim, Yeonjea, 890
Kim, Yoon, 139, 160, 242
Kim, Yoonjae, 217
Kim, Yoonji, 550
Kim, Youna, 572
Kim, YoungBin, 116, 803, 865, 891
Kim, Youngdong, 887
Kim, Yu Jin, 240
Kim, Yulhwa, 206
Kim, Yunsu, 175
Kim, Yunsung, 90, 891
Kim;,
Kim;, Junho,
Kimani, 471
Kimani, Jane Wanjiru, 471
Kimhi, 543
Kimhi, Moshe, 543
Kindratenko, 261
Kindratenko, Volodymyr, 261
King, 82
King, Joe, 82
Kini, 95
Kini, Prajakta, 95
Kinugawa, 863
Kinugawa, Kazutaka, 863
Kirchhoff, 320
Kirchhoff, Katrin, 320
Kirmayr, 440
Kirmayr, Johannes, 440
Kirsten, 571
Kirsten, Elisabeth, 571
Kirtani, 209
Kirtani, Chhavi, 209
Kiselev, 233
Kiselev, Igor, 233
Kishinami, 512
Kishinami, Yosuke, 512
Kishino, 885
Kishino, Ryo, 885
Kishore, 467
Kishore, Varsha, 467
Kitada, 480
Kitada, Shunsuke, 480
Kiyomaru, 485
Kiyomaru, Hirokazu, 485
Kjell, 207
Kjell, Oscar, 207
Klakow, 120
Klakow, Dietrich, 120
Klampanos;,
Klampanos;, Liu Zai; Iraklis,
Klein, 331, 570, 784
Klein, Mark, 784
Klein, Michel, 570
Klein, Samuel, 331
Klejch, 126
Klejch, Ondrej, 126
Klimaszewski, 499
Klimaszewski, Mateusz, 499
Klimovic, 784
Klimovic, Ana, 784
Klug, 785
Klug, Katrin, 785
Klumpp, 765
Klumpp, Jan-Felix, 765
Kmainasi, 232
Kmainasi, Mohamed Bayan, 232
Kneevi, 812
Kneevi, Marta, 812
Knoll, 881
Knoll, Alois, 881
Ko, 267, 282, 347, 510, 717, 815, 831, 833, 888
Ko, Changgeon, 717, 831, 833
Ko, Ho Tin, 347
Ko, Hyeonmok, 815
Ko, Hyunwoo, 267
Ko, Jongwoo, 888
Ko, Jungmin, 510
Ko, Kangwook, 282
Koa, 84
Koa, Kelvin J.L., 84
Kobashi, 608
Kobashi, Yohei, 608
Kobayashi, 823
Kobayashi, Kazuma, 823
Kobler, 534
Kobler, Sebastian, 534
Kobyzev, 273
Kobyzev, Ivan, 273
Kochedykov, 613
Kochedykov, Denis, 613
Kodali, 837
Kodali, Prashant, 837

KODATI, 471
KODATI, DHEERAJ, 471
Kodavanti, 305
Kodavanti, Sravanth, 305
Koeppl, 88
Koeppl, Heinz, 88
Kogundi, 129
Kogundi, Jayavibhav, 129
Koh, 639, 765, 887
Koh, Hyukhun, 765
Koh, Jungin, 639
Koh, Junyoung, 639
Koh, Seunghee, 887
Kohail, 471
Kohail, Sarah, 471
Koike, 236, 813
Koike, Ryuto, 236, 813
Kojima, 608, 660
Kojima, Takeshi, 608, 660
Kola, 615
Kola, Sneha, 615
Kolahdouzi, 696
Kolahdouzi, Mojtaba, 696
Kolasani, 615
Kolasani, Lakshman, 615
Kolodiazhnyi, 438
Kolodiazhnyi, Maksim, 438
Komachi, 101
Komachi, Mio Ohashi; Hajime Kiyama; Zhidong Ling; Mamoru,
Kona, 606, 617
Kona, Uma, 606, 617
Konev, 655
Konev, Georgii, 655
Kong, 96, 99, 136, 159, 209, 282, 292, 349, 352, 353, 359, 390, 407, 410, 422, 449, 454, 475, 564, 638, 646, 699, 744, 804, 834, 860, 878, 888, 889
Kong, Aobo, 744
Kong, Beibei, 638
Kong, Chuyi, 699
Kong, Cunliang, 282, 449
Kong, Dechang, 359
Kong, Fang, 407
Kong, Huanjun, 209
Kong, Jiawei, 349
Kong, Keyi, 136, 860
Kong, Lingdi, 834
Kong, Lingjing, 454, 878
Kong, Lingpeng, 99, 159, 353
Kong, Ming, 352, 564
Kong, Mingze, 646
Kong, Qiuqiang, 390
Kong, Quan, 96
Kong, Quyu, 410
Kong, Rui, 804
Kong, Xiangwen, 292, 475
Kong, Xiangzheng, 888, 889
Kong, Zicheng, 422
Konishi, 416
Konishi, Yuta, 416
Konomi, 865
Konomi, Mersin, 865
Konushin, 438
Konushin, Anton, 438
Koo, 166, 182, 445
Koo, Hamin, 166
Koo, Myoung-Wan, 445
Koo, Seonmin, 182
Kopilov, 551
Kopilov, Adi, 551
Koppisetti, 398
Koppisetti, Surya, 398
Kordjamshidi, 774
Kordjamshidi, Parisa, 774
Kordjazi, 609
Kordjazi, Neda, 609
Koreeda, 833
Koreeda, Yuta, 833
Korhonen, 104
Korhonen, Anna, 104
Korikov, 568
Korikov, Anton, 568
Kornmann, 98
Kornmann, Marcus, 98
Kostromina, 342
Kostromina, Alina, 342
Kotevska, 909
Kotevska, Olivera, 909
Kothapalli, 124, 356
Kothapalli, Vignesh, 124, 356
Kothiyal, 572
Kothiyal, Akshat, 572
Kothur, 694
Kothur, Sachith Sri Ram, 694
Koto, 201, 442, 555, 562, 589, 695, 864, 869, 874
Koto, Fajri, 201, 442, 555, 562, 589, 695, 864, 869, 874
Kotoge, 198, 779
Kotoge, Rikuto, 198, 779
Kotov, 207
Kotov, Romon, 207
Koul, 221, 299
Koul, Anurag, 299
Koul, Deeksha, 221
Kour, 610

Kour, George, 610
Kouris, 223
Kouris, Alexandros, 223
Kouzy, 104, 106
Kouzy, Ramez, 104, 106
Kovacevic, 715
Kovacevic, Filip, 715
Kowalczyk, 606
Kowalczyk, Witold, 606
Koychev, 442
Koychev, Ivan, 442
Kozaki, 854
Kozaki, Koiji, 854
Koos, 548
Koos, Anna, 548
Kpei, 538
Kpei, Sylvester, 538
Kranti, 862
Kranti, Chalamalasetti, 862
Krasnodbska, 740
Krasnodbska, Aleksandra, 740
Krause, 784
Krause, Andreas, 784
Kreiss, 306
Kreiss, Elisa, 306
Krimmel, 678
Krimmel, Markus, 678
Krishna, 128, 253, 473, 879
Krishna, Amrith, 473
Krishna, Satyapriya, 128, 253, 879
Krishnamoorthi, 444
Krishnamoorthi, Raghuraman, 444
Krishnamurthy, 589
Krishnamurthy, Parameswari, 589
Krishnan, 515
Krishnan, Dhriti, 515
Kriz, 154
Kriz, Reno, 154
Kropko, 884
Kropko, Jonathan, 884
Krsteski, 838
Krsteski, Stefan, 838
Krumdick, 871
Krumdick, Michael, 871
Kruse, 822
Kruse, Maya, 822
Kryvosheieva, 447
Kryvosheieva, Daria, 447
Ku, 547, 840
Ku, Lun-Wei, 547, 840
Kuan, 455
Kuan, Shih-Yun Shan, 455
Kuang, 377, 408, 435, 467, 727, 754, 907
Kuang, Jun, 408
Kuang, Kun, 467, 754
Kuang, Li, 435
Kuang, Minchi, 727
Kuang, Zexi, 907
Kuang, Zhaoqi, 377
Kucharavy, 784
Kucharavy, Andrei, 784
Kucha, 891
Kucha, Josef, 891
Kucherenko, 784
Kucherenko, Anastasiia, 784
Kuempel, 771
Kuempel, Clara, 771
Kuhar, 299
Kuhar, Sachit, 299
Kuishibekova, 874
Kuishibekova, Kamila, 874
Kukreja, 569
Kukreja, Dikshant, 569
Kulkarni, 167, 219, 240, 245, 562, 606, 617
Kulkarni, Ashish, 167
Kulkarni, Chaitanya, 606, 617
Kulkarni, Kalmit, 562
Kulkarni, Rohan, 240
Kulkarni, Vivek, 245
Kulkarni, Yogesh, 219
Kulmizev, 173, 246
Kulmizev, Artur, 173, 246
Kuma, 622
Kuma, Raghvendra, 622
Kumar, 127, 139, 140, 142, 221, 232, 239, 266, 284, 295, 299, 305, 374, 394, 406, 411, 472, 473, 484, 526, 538, 558, 587, 588, 632, 644, 672, 700, 782, 787, 881, 892, 904
kumar, 558
Kumar, Aayush, 239
Kumar, Abhishek, 142, 588
Kumar, Anand, 284
Kumar, Anuj, 700
Kumar, Aounon, 394
Kumar, Ayush, 632
Kumar, Deepak, 644
Kumar, Dhruv, 782
Kumar, Gaurav, 221
Kumar, Komal, 139
Kumar, Nalin, 411
Kumar, Nitin, 374
Kumar, Prince, 558
Kumar, Rishabh, 473
Kumar, Sachin, 127, 295, 587, 892
Kumar, Sandeep, 374

Kumar, Satyam, 406
Kumar, Sawan, 266
Kumar, Shivani, 472
Kumar, Sonal, 140
Kumar, Srijan, 538, 904
Kumar, Tampu Ravi, 526
Kumar, Tarun, 672
Kumar, Uttam, 305
Kumar, Varun, 299, 484, 787, 881
Kumar, Vishal, 232
kumar, vishwajeet, 558
Kumarage, 879
Kumarage, Tharindu, 879
Kumaraguru, 428
Kumaraguru, Ponnurangam, 428
Kummerfeld, 534
Kummerfeld, Jonathan K., 534
Kumon, 206
Kumon, Ryoma, 206
Kundapur, 292
Kundapur, Niran, 292
Kundu, 269, 498
Kundu, Souvik, 269, 498
Kung, 850
Kung, Po-Nien, 850
Kuo, 110, 312
Kuo, Heng-Cheng, 312
Kuo, Yen-Chun, 110
Kuratov, 178
Kuratov, Yuri, 179
Kurdydyk, 117
Kurdydyk, Dmytro, 117
Kuribayashi, 102, 105, 110, 589
Kuribayashi, Tatsuki, 102, 105, 110, 589
Kurita, 839
Kurita, Hiroto, 839
Kurohashi, 213
Kurohashi, Sadao, 213
Kurup, 233
Kurup, Devi G., 233
Kurz, 247
Kurz, Simon, 247
Kurzawe, 620
Kurzawe, Daniel, 620
Kusa, 740
Kusa, Wojciech, 740
Kushner, 705
Kushner, Jason, 705
Kusuma, 898
Kusuma, Christopher Adrian, 898
Kutralingam, 288
Kutralingam, T Pranav, 288
Kuzkina, 330
Kuzkina, Yulia, 330
Kuznetsov, 633
Kuznetsov, Andrey, 633
KV, 226
KV, Gouthaman, 226
Kwak, 93, 153, 466, 560, 889
Kwak, Haewoon, 153, 560
Kwak, Minseo, 889
Kwak, Nojun, 93, 466
Kwan, 742
Kwan, Wai-Chung, 742
Kwok, 144, 782
Kwok, James, 144
Kwok, Tsun On, 782
Kwon, 116, 118, 121, 160, 182, 223, 274, 283, 295, 438, 492, 587, 600, 741, 769, 803, 893
Kwon, Haeyoung, 182
Kwon, Hyeokhyen, 438
Kwon, Jea, 283, 893
Kwon, Jeongyoun, 118
Kwon, Jihoon, 160
Kwon, Jingun, 492, 587
Kwon, JuneHyoung, 116, 803
Kwon, Soonhwan, 769
Kwon, Taekyoung, 274
Kwon, Taeyoon, 741
Kwon, Taeyoun, 121
Kwon, Yeong-Dae, 600
Kwon, Yongchan, 295
Kwon, Young D., 223
Kyomuhendo, 862
Kyomuhendo, Bruhan, 862
Käser, 812
Käser, Tanja, 812
Körner, 868
Körner, Felicia, 868

L, 260
L, Leo, 260
Laban, 81, 299
Laban, Guy, 81, 299
Labat, 477
Labat, Léo, 477
Labatut, 559
Labatut, Vincent, 559
Labeau, 176
Labeau, Matthieu, 176
Lacera, 588
Lacera, Anisio, 588
Lachenmaier, 766
Lachenmaier, Clara, 766
Lade, 250
Lade, Ravij, 250

Lahaussois, 228
Lahaussois, Aimée, 228
Lahib, 477
Lahib, Ali El, 477
Lahlou, 95, 442
Lahlou, Salem, 95, 442
Lai, 330, 333, 347, 377, 417, 463, 471, 482, 560, 584, 714, 751, 839, 875, 895
Lai, Changhao, 751
Lai, Fuming, 895
Lai, Jin, 875
Lai, Jinshan, 377
Lai, Junhong, 330
Lai, Kee-hung, 333
Lai, Lixian, 417
Lai, Mingyoung, 839
Lai, Shuzhong, 330
Lai, Viet, 463, 584
Lai, Yanlin, 560
Lai, Yingchuan, 714
Lai, Yunghwei, 482
Lai, Zhengfeng, 471
Lai, Zhengzhao, 347
Laiyk, 236, 874
Laiyk, Nurkhan, 236, 874
Lakkaraju, 164, 394
Lakkaraju, Himabindu, 164, 394
Lalor, 806
Lalor, John, 806
Lam, 110, 126, 150, 458, 540, 541, 702, 875
Lam, Cheryl, 150
Lam, Kwok-Yan, 702
Lam, Kyle, 110
Lam, Michelle S., 458
Lam, Tsz Kin, 126, 875
Lam, Wai, 540, 541
Lambert, 168
Lambert, Charlotte, 168
Lamprier, 346
Lamprier, Sylvain, 346
LAN, 376
Lan, 89, 97, 136, 141, 169, 186, 231, 240, 278, 318, 380, 401, 441, 467, 547, 586, 596, 628, 635, 656, 687, 719, 723, 762, 798, 832, 850
Lan, Andrew, 186, 547, 832
Lan, Cuiling, 723
Lan, Fangping, 586, 596
Lan, Ge, 656
Lan, Junwei, 89
Lan, Man, 635, 850
Lan, Mengfei, 240
Lan, Michael, 586
Lan, Qizhen, 467
LAN, Su, 376
Lan, Tian, 97, 136, 141, 231, 318, 401, 441, 762, 798
Lan, Tianwei, 169
Lan, Tingfeng, 628, 687
Lan, Yu, 719
Lan, Yunshi, 278
Lan, Zhenzhong, 380
Land, 194
Land, Sander, 194
Lang, 342, 401, 444, 446, 675, 706, 821
Lang, Fei, 342
Lang, Hao, 675, 706
Lang, Hunter, 444, 446
Lang, Jianjun, 821
Lang, Yong, 401
Langbridge, 610
Langbridge, Abigail, 610
Langford, 492
Langford, John, 492
Langlais, 110
Langlais, Pierre-Carl, 110
Langlois, 232
Langlois, Lucía, 232
Lango, 138, 872
Lango, Mateusz, 138, 872
Lao, 399, 687
Lao, Jiale, 687
Lao, Leyi, 399
Lapata, 176, 302
Lapata, Mirella, 176, 302
Lapuschkin, 858
Lapuschkin, Sebastian, 858
Larraz, 881
Larraz, Irene, 881
Larsen, 263
Larsen, Carter, 263
Larsson, 615
Larsson, Erik, 615
Laskar, 822
Laskar, Md Tahmid Rahman, 822
Latecki, 596
Latecki, Longin Jan, 596
Latimer, 637
Latimer, Christopher, 637
Lau, 201, 571, 583, 881
Lau, Jey Han, 201, 571
Lau, Max SY, 881
Lau, Stephen, 583
lavelli, 521
lavelli, pietro ferrazzu; bernardo magnini; alberto,
Lavergne, 712

Lavergne, Thomas, 712
Lavrukhin, 542
Lavrukhin, Vitaly, 542
Lawrence, 275, 474
Lawrence, Carolin, 275, 474
Le, 121, 150, 192, 221, 245, 277, 351, 389, 405, 411, 426, 427, 432, 534, 558, 563, 639, 645, 686, 701, 727, 787, 828, 862, 882
Le, Bach, 121
Le, Cuong Chi, 192
Le, Hoang-Quynh, 686
Le, Hung, 351, 787
Le, Huu-Loi, 862
Le, Khanh Chi, 563
Le, Lihong, 558
Le, Long Bao, 277
Le, Phuc-Lu, 701
Le, Qi, 534
Le, Thai, 245
Le, Thang, 432
Le, Thao Minh, 882
Le, Thu, 150
Le, Trung, 405, 411, 426, 427, 639, 645
Le, Xiaoying, 389
Le, Xinyi, 221
Le, Xuan-Bach, 828
Le, Yankun, 727
Le-Anh, 121
Le-Anh, Minh, 121
Leach, 266, 531
Leach, Kevin, 266, 531
Leang, 742
Leang, Joshua Ong Jun, 742
Lease, 209
Lease, Matthew, 209
Lebbah, 219
Lebbah, Mustapha, 219
LeBlond, 884
LeBlond, Beau, 884
LeCun, 243
LeCun, Yann, 243
Lee, 85, 91, 93, 101, 106, 107, 110, 112, 113, 116, 121, 122, 133, 139, 142, 149, 151, 160, 171, 175, 179, 182, 183, 190, 195, 207, 209, 212, 213, 215, 217, 218, 222, 223, 225, 226, 230, 240, 244, 248, 259, 267, 275, 280, 282, 287, 290, 293, 294, 305, 309, 312, 317, 331, 351, 428, 441, 444, 445, 446, 447, 455, 456, 458, 460, 461, 469, 472, 474, 477, 479, 480, 486, 491, 496, 500, 508, 510, 513, 515, 516, 528, 534, 549, 550, 551, 557, 559, 563, 564, 569, 572, 575, 580, 592, 600, 605, 612, 626, 639, 651, 679, 690, 691, 705, 717, 735, 766, 769, 775, 780, 787, 792, 796, 797, 800, 803, 806, 811, 815, 823, 824, 828, 830, 832, 833, 846, 854, 858, 865, 866, 870, 871, 883, 892, 904
lee, 797
Lee, Allen, 866
Lee, Andrew, 207
Lee, Ayoung, 133
Lee, Benjamin, 331
Lee, Bruce W., 766
Lee, Byung-Jun, 351
Lee, Chanhee, 690
Lee, Che Hyun, 113
Lee, Chen-Yu, 446
Lee, Daeun, 222
Lee, Daeyeop, 151
Lee, Dohyeon, 240
Lee, Dokyun, 858
Lee, DongGeon, 121
Lee, Dongha, 230, 559
Lee, Dongjin, 865
Lee, Dongwon, 85, 456, 458, 626, 854
Lee, Dongwook, 113
Lee, Dosung, 580
Lee, Eun-Ju, 447
Lee, Eunju, 803
Lee, Euntaik, 305
Lee, Gary, 171, 213, 309, 317
Lee, Gary Geunbae, 175, 461, 605, 803
Lee, GyeongTaek, 780
Lee, Gyun, 575
Lee, Hansle, 195
Lee, Hanwool, 267
Lee, Harim, 461
Lee, Hochul, 870
Lee, Hong kyu, 883
Lee, Hongseok, 846
Lee, Huije, 717, 833
Lee, Hung-yi, 110, 112, 142, 218, 225, 275, 312, 455
Lee, Hwanhee, 182, 557, 811
Lee, Hwaran, 116
Lee, HyeonSu, 305
Lee, Hyunji, 179
Lee, Hyunsoo, 679
Lee, I-Ta, 800
Lee, Insung, 445
Lee, Jae-Gil, 828
Lee, Jaeho, 212, 223, 444
Lee, Jaehoon, 183, 828
Lee, Jaejin, 477
Lee, Jaeri, 516

Lee, Jaewoo, 293
Lee, Jaewook, 691, 832
Lee, Jaeyoung, 259
Lee, Jaeyun, 639
Lee, Jay-Yoon, 472
Lee, Jea Sung Kim ; Suan,
Lee, Jee-Hyong, 183, 564, 828
Lee, Jeongah, 513
Lee, Jeonghwan, 830
Lee, Jihun, 139
Lee, Jihyung, 183
Lee, Jin-Seop, 183
Lee, Jinkyu, 106
Lee, Jinu, 248, 474
Lee, Jonggeun, 223, 787
Lee, Jongha, 846
Lee, Jonghoon, 612
Lee, Jongwuk, 769
Lee, Joonhak, 477
Lee, Joosung, 280
Lee, Juhyeon, 797
Lee, Jungmin, 294
Lee, Jungwoo, 182
Lee, Jungyu, 575
Lee, Junhoo, 93
Lee, Junhyeok, 823
Lee, Junhyun, 815
Lee, Junmyeong, 515
Lee, JunSeung, 600
Lee, Juntae, 107
Lee, Junyeol, 223
Lee, Justin, 892
Lee, Kunhui, 575, 592
Lee, Kwonjoon, 792
Lee, Kyenghun, 815
Lee, Kyong-Ho, 690
Lee, Kyungsu, 217
Lee, Luke Zhang; Justin Vasselli; Aditya Khan; York Hay Ng; En-Shiun Annie,
Lee, Lung-Hao, 551
Lee, Meng-Chieh, 797
Lee, Mingyu, 479
Lee, Minhwa, 563
Lee, Minjae, 215
Lee, Minwoo, 534
Lee, Mong-Li, 428
Lee, Moontae, 240, 679
Lee, Nakyung, 182
Lee, Namhoon, 91
Lee, Nayeon, 486
Lee, Nicholas, 244
Lee, Owen, 705
Lee, Paige, 226
Lee, Roy Ka-Wei, 460, 469, 651
Lee, Sang-goo, 572
Lee, Sangam, 230
Lee, Sangeyl, 775
Lee, SangKeun, 479, 871
Lee, Seohyeong, 116
Lee, Seokhan, 122
Lee, Seong-Whan, 101, 209
Lee, Seongmin, 612
Lee, Seongyun, 474
Lee, Seungbeen, 259
Lee, Seungeon, 287
Lee, Seungho, 217
lee, Seunghyun, 797
Lee, Seungkyu, 149, 441
Lee, Seungmin, 830
Lee, Seungmin OhEunseok, 500
Lee, Seungyoon, 806
Lee, Sua, 824
Lee, Suhyun, 290
Lee, Sunbowen, 800
Lee, Sung Jun, 217
Lee, Sung-Ju, 549
Lee, Sungjae, 139, 508
Lee, Sunkyung, 769
Lee, Wenke, 569
Lee, Wonjin, 139
Lee, Wonjun, 213, 282
Lee, Wonkee, 183
Lee, Wonseok, 510
Lee, Woojin, 904
Lee, Yejin, 190
Lee, Yeongheon, 766
Lee, Yeonjin, 639
Lee, Yeonjoon, 294
Lee, Yohan, 550
Lee, Yong Jae, 480
Lee, Yongjae, 160
Lee, Yoonhyung, 106
Lee, Youngwon, 496
Lee, Yueh Z, 528
Lee, Yukyung, 796
Lee, Yumin, 828
Lee, Yun-Yu, 870
Lee, Zhemg, 735
Leeb, 156
Leeb, Felix, 156
Lehmann, 542
Lehmann, Jens, 542
Lei, 93, 97, 118, 135, 159, 163, 212, 256, 259, 261, 405, 450, 481, 491, 611, 692, 693, 734
Lei, Bin, 118
Lei, Bohan, 450

Lei, Deren, 118
Lei, Fangyu, 97
Lei, Jingdi, 405
Lei, Minglong, 734
Lei, Qingyuan, 261
Lei, Shuo, 611
Lei, Wenqiang, 135, 692, 693
Lei, Xiaoxuan, 212
Lei, Xuanyu, 93
Lei, Xuejing, 481
Lei, Yibin, 163
Lei, Yu, 159
Lei, Yuanting, 491
Lei, Yuanyuan, 256, 259
Leippold, 777
Leippold, Markus, 777
Leiter, 107
Leiter, Christoph, 107
Lejeune, 477
Lejeune, Gaël, 477
Lemeshko, 438
Lemeshko, Andrey, 438
Leng, 119, 138, 358, 433, 487, 519, 796
Leng, Fangling, 358
Leng, Jia, 796
Leng, Jiaming, 433
Leng, Jingwen, 138, 487
Leng, Jinman, 519
Leng, Yongqi, 119
Lent, 246
Lent, Heather, 246
Leong, 738
Leong, Chak Tou, 738
Lepori, 500
Lepori, Michael, 500
Lequeu, 477
Lequeu, Pierre-Antoine, 477
Leto, 800
Leto, Alexandria, 800
Levchenko, 831
Levchenko, Kirill, 831
Lever, 198
Lever, Jake, 198
Levy, 82, 153, 160, 199, 242, 257, 615
Levy, Joshua H., 615
Levy, Oscar, 160
Levy, Roger, 153
Levy, Roger P., 242
Levy, Shahar, 82
Levy, Sharon, 199, 257
Lewis, 194
Lewis, Ashley, 194
Lhoneux, 246
Lhoneux, Miryam de, 246
LI, 185, 208, 260, 347, 416, 425, 485, 540, 577, 588, 649, 687, 846
Li, 81, 84, 85, 88, 89, 91, 92, 93, 94, 96, 97, 104, 105, 106, 108, 110, 111, 112, 114, 117, 119, 120, 121, 123, 124, 125, 126, 128, 129, 130, 131, 132, 134, 136, 137, 138, 141, 144, 145, 147, 148, 149, 150, 151, 155, 156, 158, 159, 161, 162, 163, 165, 166, 167, 169, 170, 172, 176, 178, 181, 184, 185, 186, 187, 191, 192, 194, 197, 198, 199, 200, 201, 202, 203, 204, 205, 206, 207, 208, 209, 210, 212, 214, 216, 217, 219, 220, 221, 223, 224, 227, 228, 229, 234, 237, 238, 239, 240, 243, 248, 249, 250, 251, 252, 253, 255, 256, 258, 259, 260, 261, 264, 266, 267, 270, 271, 272, 273, 274, 276, 278, 279, 280, 281, 282, 283, 284, 286, 288, 289, 291, 292, 294, 296, 298, 300, 301, 303, 304, 307, 308, 310, 311, 312, 313, 315, 316, 318, 320, 321, 322, 323, 324, 325, 326, 327, 330, 332, 335, 336, 337, 338, 339, 340, 341, 344, 345, 346, 347, 348, 350, 351, 352, 353, 354, 355, 356, 358, 359, 360, 361, 362, 363, 364, 368, 370, 372, 373, 374, 376, 377, 378, 379, 380, 381, 384, 385, 386, 388, 390, 391, 392, 394, 395, 396, 397, 400, 401, 403, 405, 406, 407, 408, 412, 413, 414, 415, 416, 417, 418, 419, 420, 421, 422, 423, 425, 427, 428, 429, 431, 432, 435, 437, 438, 439, 440, 442, 443, 444, 446, 449, 452, 453, 455, 456, 458, 459, 460, 461, 462, 464, 465, 466, 469, 470, 471, 473, 474, 475, 476, 477, 480, 482, 483, 484, 486, 488, 489, 490, 491, 496, 497, 499, 501, 503, 504, 505, 506, 509, 513, 514, 515, 516, 527, 528, 531, 533, 537, 540, 541, 542, 543, 544, 545, 546, 550, 551, 552, 554, 557, 558, 561, 562, 563, 565, 566, 569, 571, 572, 573, 575, 576, 577, 578, 579, 581, 583, 584, 585, 587, 588, 589, 590, 591, 593, 596, 597, 599, 600, 604, 605, 606, 607, 609, 611, 617, 619, 620, 621, 624, 626, 627, 628, 629, 630, 631, 632, 634, 636, 637, 638, 639, 642, 643, 647, 649, 650, 651, 653, 654, 659, 660, 662, 663, 664, 665, 666, 667, 668, 671, 672, 673, 675, 676, 677, 678, 680, 681, 684, 685, 686, 687, 688, 689, 690, 691, 692, 693, 694, 695, 696, 697, 698, 699, 700, 702, 703, 704, 705, 706, 708, 709, 713, 714,

715, 716, 717, 718, 719, 722, 723, 724, 725, 726, 728, 730, 731, 732, 733, 735, 738, 739, 740, 741, 742, 743, 744, 745, 746, 748, 749, 750, 751, 752, 754, 755, 756, 758, 759, 760, 761, 762, 765, 767, 768, 769, 770, 773, 775, 777, 778, 779, 780, 781, 782, 783, 784, 785, 786, 787, 788, 790, 792, 793, 794, 795, 796, 797, 798, 801, 802, 803, 804, 807, 808, 809, 810, 812, 814, 816, 817, 818, 819, 821, 823, 824, 825, 826, 827, 829, 830, 833, 835, 837, 838, 840, 843, 844, 845, 848, 849, 850, 851, 852, 853, 855, 856, 857, 859, 861, 862, 865, 866, 867, 869, 871, 875, 876, 877, 878, 879, 880, 881, 882, 883, 887, 888, 889, 892, 893, 895, 896, 897, 900, 901, 902, 903, 904, 905, 908
li, 87, 379, 435, 577, 616, 674, 733
Li*, 725
Li*, Piji, 725
Li, Alan, 783
Li, Ang, 163, 252
Li, Aoxue, 301
Li, Baishi, 84
Li, Baixuan, 853
Li, Bangze Pan;Yang,
Li, Baochun, 421, 773
Li, Baohang, 381
Li, Bei, 335, 413, 456, 739, 748
Li, Beibin, 444
Li, Bin, 284, 723, 726
Li, Bing, 134
Li, Binghang, 418
Li, Bingxuan, 292, 573
Li, Bingzhe, 503
Li, Binhua, 796
Li, Binyang, 184, 726
Li, Bo, 134, 259, 361, 551, 626, 651, 688, 690, 699, 775, 887
Li, Bobo, 428
Li, Bohao, 743
Li, Botao, 255
Li, Bowen, 695
Li, Boyan, 205, 552, 562, 853
Li, Cangyuan, 460
Li, Changcheng, 664
Li, ChangHao, 385
Li, Changjiang, 121, 566, 877
Li, Changjin, 186
Li, Changye, 165
Li, Chao, 256, 272
Li, Chaofan, 89
Li, Chaoyu, 219
Li, Chaozhuo, 128, 787, 880
Li, Chengkai, 861
Li, Chenglin, 220, 228, 543
Li, Chengxi, 788
Li, Chengzhengxu, 719, 807
Li, Chengzhi, 453
Li, Chengzu, 503, 572
Li, Chenhao, 261
Li, Chenjie, 260
Li, Chenliang, 93, 214
Li, Chentao, 629
Li, Chenyang, 735
Li, Chenyi, 725
Li, Chin-Jou, 217, 456
Li, Chongxuan, 501
Li, Chuan, 901
Li, Chuanhao, 134, 844
Li, Chuanyi, 336
Li, Chuntao, 833
Li, Chunyang, 161, 733
LI, CIXING, 425
Li, Da, 379
Li, Dailin, 345
Li, Dawei, 185, 221, 300, 490, 563
Li, Dayang, 322, 363
Li, Derek, 112
Li, Deyu, 676
Li, Dong, 355, 505, 628, 837
Li, Dongfang, 338, 630, 798
Li, Dongsheng, 401, 665, 702, 723, 725
Li, Dongyang, 159, 714
Li, Dongyuan, 488, 563
Li, Dongze, 804
Li, Fan, 358
Li, Fang, 775
Li, Fangyuan, 437, 893
Li, Fangze, 501
Li, Fanxiao, 248, 322, 363
Li, Fei, 114, 597, 856
Li, Feng, 749
Li, Fenghua, 416
Li, Fengnan, 313
Li, Franklin Mingzhe, 414
Li, Gang, 209, 429, 528
Li, Gaotang, 575
Li, Ge, 97, 260, 531, 577, 760, 796
Li, Gen, 634
Li, Gengwang, 429
Li, Gonghui, 374
Li, Guangyan, 385
Li, Guanlin, 745
Li, Guannan, 818
Li, Guanpeng, 503

Li, Guocong, 676
Li, Guohao, 545, 667
Li, Guoqi, 491
Li, Guozheng, 740
Li, Haifeng, 303, 330
Li, Haijun, 798
Li, Haiqing, 283
Li, Haitao, 220, 228, 373
Li, Haitian, 136
Li, Haizhou, 397, 474, 740, 759
Li, Han, 777
Li, Hanchen, 465
Li, Hang, 124, 205
Li, Hanhao, 558
Li, Hao, 201, 278, 323, 488, 573, 621, 643, 767, 856, 883
Li, Haohang, 166
Li, Haoliang, 204, 704
Li, Haoming, 496, 769
Li, Haonan, 88, 442, 589
Li, Haoran, 191, 496, 513, 647, 691, 713, 782, 830, 882
Li, Haowen, 812
Li, Haoxi, 148, 359
Li, Haoxuan, 81, 476, 554, 687, 804
Li, HaoYang, 856
Li, Haoyuan, 880
Li, Haoze, 728
Li, Hengyu, 626
Li, Hongliang, 583
Li, Hongming, 251
Li, Hongsheng, 339, 432
Li, Hongxing, 751
Li, Hongyan, 191
Li, Hongyu, 344
Li, Hongzhi, 264
Li, Houqiang, 315, 461
Li, Houyi, 350
Li, Hua, 466
Li, Huajie, 723
Li, Huan, 96, 125, 221, 497
Li, Huatao, 617
Li, Huawei, 460
Li, Huayang, 378
Li, Huayu, 204, 795
LI, HUI, 416
Li, Hui, 380, 397, 897
Li, Huicong, 373
Li, Huijia, 561, 677
Li, Irene, 823
Li, Jerry, 271
Li, Ji-Fu, 372
Li, Jia, 84, 237, 260, 458, 531, 577, 790
Li, Jia-Nan, 865
Li, Jiacheng, 250, 708
Li, Jiafan, 403
Li, Jiahao, 284, 606, 689, 709
Li, Jiahua, 654
Li, Jiahuan, 151
Li, Jiajia, 234
Li, Jialin, 471
Li, Jiaming, 790
Li, Jian, 276, 332, 724, 788
Li, Jianbin, 413
Li, Jianbo, 378
Li, Jiang, 147, 318, 401, 762, 814
Li, Jiangnan, 425
Li, Jiangtong, 459
Li, Jianguo, 92, 380
Li, Jianhua, 768
li, Jianhui, 577
Li, Jianjun, 394, 407, 790
LI, JIANSHE, 846
Li, Jiaoda, 848
Li, Jiaqi, 145, 313, 395, 544, 818
Li, Jiashan, 300
Li, Jiatong, 151
Li, Jiawei, 509, 705, 830
Li, Jiaxuan, 607
Li, Jiayang, 533, 798
Li, Jiayu, 192, 880, 892
Li, Jiayun, 418
Li, Jiazheng, 150
Li, Jie, 696
Li, Jierui, 636
Li, Jieyu, 266
Li, Jiguo, 395
Li, Jinbao, 121, 877
Li, Jindong, 732
Li, Jing, 354, 362, 692, 694
Li, Jingguang, 449
LI, JINGJING, 577
Li, Jingjing, 514
Li, Jingming, 403
Li, Jingru, 423
Li, Jingyuan, 205
Li, Jinming, 578
Li, Jinpeng, 130, 718, 730, 905
Li, Jinyu, 199, 352
Li, Jinyuan, 676
Li, Jinze, 159
Li, Jinzhe, 209
Li, Jiuyi, 351
Li, Jiwei, 237, 425
Li, Juan, 408
Li, Juanzi, 169, 476, 488, 554, 571, 782, 880

Li, Jun, 406
Li, Junan, 848
Li, Jundong, 585, 824
Li, Junlin, 362
Li, Juntao, 632, 678, 716, 722, 733, 761
Li, Junyan, 224
Li, Junyi, 85, 253, 587, 717
Li, Junyi Jessy, 104, 106, 156, 209
Li, Junzhuo, 694
Li, Kaiyue, 184
Li, Kan, 405
Li, kang, 783
Li, Ke, 588, 802
Li, Kejiao, 540
Li, Kejie, 480
Li, Kenli, 284, 286
Li, Kenneth, 238
Li, Kevin Chenhao, 650
Li, Keyu, 665
Li, Kezhi, 298
Li, Krinos, 790
Li, Kuan, 120, 853
Li, Kun, 540, 792, 848
Li, Lei, 248, 354, 537, 542, 569, 604, 867, 869, 904
Li, Levina, 280
Li, Li, 315, 449, 488, 905
Li, Liangyue, 819
Li, Lihong, 181
Li, Lijian, 155
Li, Lijun, 685, 777, 902
Li, Lin, 110
Li, Linghao, 296
Li, Lingyao, 470
Li, Linjie, 112
Li, Linjing, 249
Li, Linyang, 159, 579, 877
Li, Linyi, 695
Li, Long, 93, 185, 469, 506
Li, Lujun, 197, 272, 347
Li, Luming, 676
Li, Manling, 151, 875
Li, Marissa, 261
Li, Marvin, 785
Li, Meimingwei, 778
Li, Meng, 904
Li, Mengfan, 446
Li, Menghan, 653
Li, Mengjie, 795
Li, Mengyang, 662
Li, Mengyuan, 770
Li, Mengze, 754
Li, Mengzhang, 527
Li, Miao, 176
Li, Miaomiao, 767
Li, Michael, 291
Li, Min, 888
LI, MING, 185
Li, Ming, 91, 131, 187, 208, 861
Li, Mingchen, 106, 221
Li, Minghan, 731, 748
Li, Minghao, 638
Li, Mingjie, 417
Li, Mingqi, 583
Li, Mingxin, 473
Li, Mingxuan, 312, 465
Li, Mingyang, 311, 639, 749
Li, Mingye, 543
Li, Mingzhe, 117, 129, 876
Li, Minxi, 566
Li, Moxin, 303, 852
Li, Moyan, 258, 835
Li, Mukai, 159, 576, 838
Li, Muyang, 797
Li, Nan, 700
Li, Nanjie, 108
Li, Ning, 321, 338
Li, Ninghui, 178
Li, Ouxiang, 303
Li, Pei Pei, 422
Li, Peifeng, 123, 240, 857
Li, Peiji, 159
Li, Peiming, 420
Li, Peiyu, 808
Li, Peize, 326
Li, Peng, 93, 214, 284, 462, 631, 684, 797
Li, Pengfei, 437, 893
Li, Pengfeng, 692
Li, Pengna, 775
Li, Pengqi, 323
Li, Pengzhen, 514
Li, Piji, 243, 296
Li, Ping, 346
Li, Qi, 378, 460
Li, Qian, 643
Li, Qiang, 341
Li, Qicheng, 744
Li, Qien, 435
Li, Qiming, 353, 360, 739
Li, Qin, 578
Li, Qing, 88, 380, 662
Li, Qingbin, 698
Li, Qinghua, 489
Li, Qingyun, 591
Li, Qiongxiu, 810
Li, Qiudan, 650
Li, Qiwei, 184, 234, 814

Li, Renhao, 158
Li, Rong-Hua, 458
Li, Rongjin, 550, 796
Li, Ru, 403, 427
Li, Rui, 223, 321, 338, 475, 738, 889
Li, Ruihui, 286
Li, Ruilin, 228
Li, Ruiqing, 850
Li, Ruixuan, 660
Li, Ruizhe, 316
Li, Rumei, 151
Li, Rumeng, 105
Li, Runan, 833
Li, Runchao, 131
Li, Runkai, 672
Li, Ruoran, 289, 739
Li, Ruosen, 709
Li, Ruoxiang, 396
Li, Scarlett, 278
Li, Serena, 588
Li, Sha, 845
Li, Shanghao, 307
Li, Shangyang, 125
Li, Sharon, 132, 151, 229, 237, 464, 896
Li, Shasha, 733
Li, Sheng, 219, 786
Li, Shengjie, 627
LI, Shengzhi, 649
Li, Shichen, 506, 840, 857
Li, Shike, 427
Li, Shimin, 228, 708
Li, Shiyu, 420, 471
Li, Shoubin, 749
LI, SHUAI, 540
Li, Shuaimin, 751, 762
Li, Shuanglong, 368
Li, Shuangyin, 483
Li, Shuyu, 597
Li, Si, 348
Li, Sihang, 649
Li, Simiao, 412
Li, Siqi, 255
Li, Siqing, 739
Li, Siran, 557
Li, Siyou, 871
Li, Siyuan, 326
Li, Songlin, 444
Li, Songze, 354, 624, 715
Li, Steven, 779
Li, Sujian, 678
Li, Sunzhu, 421
Li, Tao, 678
Li, Tengchao, 680
Li, Tengju, 320
Li, Tianchun, 611
Li, Tianhao, 94
Li, Tianjiao, 200
Li, Tianlin, 414
Li, Tianlong, 439
Li, Toby Jia-Jun, 531, 826
Li, Tong, 680, 723, 803
Li, Tsz Ho, 882
Li, Wanchun, 167
Li, Wanling, 714
Li, Wei, 239, 322, 388, 412, 584, 659, 775, 778, 901
Li, Weibin, 620
Li, Weijia, 239
Li, Weikang, 849
Li, Weiling, 533
Li, Weiping, 267, 348, 558
Li, Weishi, 673
Li, Weitao, 482, 855
LI, WEIWEI, 588
Li, Weiyuan, 687
Li, Wenbing, 849
Li, Wenbo, 826, 905
Li, Wenda, 742
Li, Wendi, 132, 151
Li, Wenhao, 138, 282, 449
Li, Wenjie, 149, 391, 415, 663, 665, 692, 738
Li, Wenkai, 844
li, wenwei, 87, 616
Li, Wenxi, 243
Li, Wenyao, 843
Li, Wenyi, 619, 671
Li, Wenzhe, 444
Li, Wenzhong, 600
Li, Wu, 354, 374
Li, Wuchao, 414
Li, Xia, 743
Li, Xian, 111
Li, Xiang, 162, 200, 227, 289, 599, 697, 713, 802
Li, Xiang Lorraine, 449
Li, Xiangang, 114, 865
Li, Xiangci, 484
Li, Xianghua, 703
Li, Xiangpeng, 551
Li, Xianyong, 176
Li, Xiao, 879
Li, Xiaodong, 573
Li, Xiaohan, 221
Li, Xiaojian, 85
Li, Xiaoli, 637, 676
Li, Xiaolong, 149, 276, 425, 716
Li, Xiaomeng, 782

Li, Xiaopeng, 214, 733
Li, Xiaoting, 581
Li, Xiaoxiao, 206
Li, Xiaoxue, 540
Li, Xiaoyan, 169
Li, Xiaoyu, 311, 408, 425
Li, Xiaoyuan, 852
Li, Xiaozhe, 159, 579, 794
Li, Xin, 339, 405, 459, 631, 793, 801
Li, Xinfeng, 280, 482, 504
Li, Xing, 759
Li, Xingyang, 758
li, xingyang, 435
Li, Xinlu, 308
Li, Xinpeng, 569, 895
Li, Xinran, 288
Li, Xintong, 514, 680, 866
Li, Xinwei, 724
Li, Xinyan, 664
Li, Xinye, 321, 630
Li, Xinyu, 212, 443
Li, Xinze, 377, 406
Li, Xinzhe, 318, 337
Li, Xiping, 270
Li, Xiquan, 390, 691, 748
Li, Xiyun, 289
Li, Xu, 456
Li, Xuanchen, 455
Li, Xuancheng, 373
Li, Xuanwu, 327
Li, Xuanyu Zhang Yunxin, 397
Li, Xuchen, 340, 642
Li, Xue, 418
Li, Xuelong, 370, 756, 761
Li, Xueting, 654
Li, Xuhui, 876
Li, Xujun, 786
Li, Xunkai, 458
Li, Xuzhao, 340, 642
Li, Yafu, 550, 654, 756
Li, Yahui, 159
Li, Yakai, 499
Li, Yaliang, 178, 273, 296, 619
Li, Yan, 681
Li, Yancui, 407
Li, Yang, 96, 134, 170, 181, 202, 540, 565, 579, 593, 849
Li, Yanghao, 596
Li, Yangning, 488
Li, Yangyi, 827
Li, Yangzhuo, 114, 780
Li, Yanhao, 210
Li, Yanshu, 137, 692
Li, Yanxi, 161
Li, Yanyan, 144
Li, Yao, 692
Li, Yaolin, 631
Li, Yaqian, 540
Li, Yaxi, 486
Li, Yaxuan, 488
Li, Ye, 315
Li, Yi, 300, 503, 693
Li, Yichen, 660
Li, Yidi, 629
LI, YIFAN, 347
Li, Yifan, 119
Li, Yifei, 821, 851
Li, Yihang, 125, 829
Li, Yijiang, 908
Li, Yijie, 809
Li, Yikun, 515, 531
Li, Yilin, 386
Li, Yiming, 609
Li, Ying, 261, 403, 781
Li, Yinghui, 460, 488, 668
Li, Yingji, 325, 431
Li, Yingjian, 335
Li, Yingjie, 802
Li, Yining, 221
Li, Yinxi, 266
Li, Yishen, 339
Li, Yishun, 759
Li, Yiwei, 405
Li, Yixi, 606, 617
Li, Yixia, 93, 167, 440, 746, 793
Li, Yiyuan, 862
Li, Yizhe, 676
LI, Yizhi, 260, 485
Li, Yizhi, 637
Li, Yizhuo, 693
Li, Yong, 255, 258, 420, 491, 673
Li, Yongbin, 126, 289, 675, 706, 739, 796, 849
Li, Yongmin, 97
Li, Yongqi, 149, 415, 663, 675, 706, 738, 755, 903
Li, Yongyuan, 356
Li, Youru, 506, 848
Li, Yu, 141, 274, 381, 609, 829, 859
Li, Yu-Kun, 379
Li, Yuan, 134, 204, 896
Li, Yuan-Fang, 340
Li, Yuanchun, 680, 804
Li, Yuang, 619
Li, Yuankai, 483
Li, Yuanze, 545, 550
Li, Yuchen, 304, 310, 699, 713, 752
Li, Yucheng, 830

Li, Yudong, 136, 286
Li, Yudu, 261
Li, Yuechan, 392
Li, Yuetai, 599, 879
Li, Yuhang, 900
Li, Yuhao, 327
Li, Yuhua, 660
Li, Yuliang, 480
Li, Yulin, 384
Li, Yun, 528
Li, Yung-Hui, 840
Li, Yunpeng, 279, 593, 765, 793
Li, Yuntao, 672
Li, Yunxin, 759
li, yunxin, 379
Li, Yunzhu, 573
Li, Yupeng, 418, 419
Li, Yuqiang, 437, 893
Li, Yuqing, 425
Li, Yuxi, 738
Li, Yuxiao, 878
Li, Yuxin, 691
Li, Yuxuan, 665, 716
Li, Yuying, 698
Li, Yuyuan, 877
LI, ZAITANG, 208, 687
Li, Zang, 638, 904
Li, Ze Shi, 292
Li, Zechao, 93
Li, Zechen, 237, 443
Li, Zechuan, 545
Li, Zehan, 438, 477
Li, Zehao, 159, 591
Li, Zehuan, 92
Li, Zejian, 281, 798
Li, Zeju, 298
Li, Zejun, 516, 546, 562
Li, Zekun, 194, 476, 804
Li, Zelin, 796, 896
Li, Zelong, 892
Li, Zeping, 440, 716, 746
Li, Zexin, 740, 790
Li, Zeyuan, 356
Li, Zhang, 756
Li, Zhao, 750
Li, Zhaochen, 605
Li, Zhaolin, 825
Li, Zhaoyang, 311, 639, 881
Li, Zhaoye, 733
Li, Zhaoyi, 338
Li, Zhe, 355
Li, Zheng, 264, 647, 817
Li, Zhenghua, 149
Li, Zhengnan, 346
Li, Zhengyi, 138
Li, Zhenhao, 515, 790
Li, Zhenxiang, 239
Li, Zhenyang, 162, 629, 725
Li, Zhenyu, 362, 400
Li, Zheyuan, 577
Li, Zhicong, 793
Li, Zhigen, 590
Li, Zhiheng, 134
Li, Zhijiang, 364
Li, Zhijing, 158
Li, Zhijun, 422
Li, Zhiqiang, 336
Li, Zhixing, 460
Li, Zhixu, 606, 666
Li, Zhixun, 280, 698, 848
Li, Zhiyang, 400
li, Zhiyu, 674, 733
Li, Zhiyuan, 706
Li, Zhong-Zhi, 431, 758
Li, ZhongDong, 830
Li, Zhonghang, 537
Li, zhongwen, 702
Li, Zhongyang, 726
Li, Zhongzhi, 797
Li, Zhoujun, 260
Li, Zhuang, 203, 816, 856
Li, Zhuo, 561, 643
Li, Zhuofeng, 901
Li, Zhuohang, 281
Li, Zhuolin, 460
Li, Zhuoqun, 741
Li, Zhuoran, 216
Li, Zhuoyu, 540
Li, Zhuoyun, 172, 376
Li, Zihan, 324
Li, Zihao, 621
Li, Ziheng, 216
Li, Zijian, 388
Li, Ziming, 401
Li, Zitao, 178, 296
Li, Zitong, 207
Li, Zixu, 698, 755
Li, Zixuan, 541, 784
Li, Ziyang, 151
Li, Ziyu, 267, 348
Li, Zizhen, 134, 844
Li, Zonggen, 289
Li, Zongmin, 198
Li, Zongxia, 819
Li, Zuchao, 184, 234, 292, 294, 362, 452, 814, 888
Li, Zuolong, 358

Li1;,
Li1;, Fu,
Li;,
Li;, Weiping,
Liakata, 183, 529, 763
Liakata, Maria, 183, 529, 763
Lian, 89, 187, 298, 330, 341, 375, 393, 400, 405, 435, 569, 668, 700, 726, 819
Lian, Defu, 89, 298, 569, 726, 819
Lian, Heng, 405
Lian, Jianxun, 375
Lian, Jingchun, 330
Lian, Niu, 393
Lian, Sheng, 187
Lian, Shiguo, 341
Lian, Wangwang, 435
Lian, Xiaoying, 700
Lian, Ying, 400
Lian, Zheng, 668
Liang, 92, 96, 104, 114, 117, 123, 125, 130, 132, 142, 144, 145, 146, 152, 157, 163, 165, 172, 196, 207, 218, 230, 239, 244, 249, 258, 261, 264, 269, 273, 276, 279, 295, 298, 313, 321, 334, 340, 342, 352, 354, 357, 363, 364, 377, 379, 391, 395, 402, 431, 437, 456, 459, 475, 478, 491, 507, 513, 564, 568, 573, 576, 577, 581, 583, 593, 599, 631, 632, 636, 644, 645, 651, 666, 678, 680, 684, 688, 690, 691, 692, 693, 700, 701, 707, 708, 709, 720, 726, 738, 743, 748, 758, 764, 780, 783, 795, 800, 808, 819, 821, 832, 836, 838, 843, 847, 848, 849, 864, 879, 894, 907
Liang, Bin, 157, 692, 701, 726
Liang, Chao, 843
Liang, Chen, 459
Liang, Di, 165, 230, 295, 848
Liang, Dongyang, 364
Liang, Du, 726
Liang, Guang, 239
Liang, Guoyan, 632
Liang, Hanzhong, 847
Liang, Hao, 340
Liang, Hongru, 693
Liang, James Chenhao, 507
Liang, Jiacheng, 117, 879
Liang, Jiaen, 391, 684
Liang, Jiafeng, 395
Liang, Jiahui, 849
Liang, Jian, 363
Liang, Jiaqing, 249, 581, 644, 666, 848
Liang, Jiazhou, 125, 568, 764
Liang, Jingcheng, 783
Liang, Jinglin, 690
Liang, Jingxi, 577
Liang, Junhong, 475, 864
Liang, Kaiqu, 513
Liang, Kun, 738
Liang, Lei, 146, 334, 354, 720
Liang, Lili, 342
Liang, Ming, 258, 491
Liang, Mingfu, 795
Liang, Paul Pu, 104
Liang, Qiao, 402, 707, 708
Liang, Qifan, 354
Liang, Qiliang, 172
Liang, Qingyuan, 478
Liang, Ruigang, 593
Liang, Ruiqi, 92
Liang, Shining, 264
Liang, Shiyu, 152
Liang, Shuang, 165, 230, 273, 295
Liang, Sichu, 437
Liang, Sirui, 819
Liang, Siyuan, 321
Liang, Tian, 352, 564
Liang, Tianle, 114, 780, 832
Liang, Tianyuan, 651
Liang, Weijie, 261
Liang, Wenfeng, 379
Liang, Xiao, 145, 163, 280, 431, 573, 636, 758, 821
Liang, Xiaobo, 632, 678
Liang, Xin, 688
Liang, Xinyue, 631
Liang, Xucheng, 377
Liang, Yan, 680
Liang, Yangyang, 244
Liang, Yaozhen, 583, 894
Liang, Yidan, 132, 144
Liang, Yiming, 491, 599
Liang, Yingyu, 123
Liang, Yiwei, 298
Liang, Yong-Ching, 196
Liang, Yu, 142, 276
Liang, Yu-Hsuan Li, 218
Liang, Yuan, 836
Liang, Yunhao, 800
Liang, Yunlong, 96
Liang, Yupu, 645
Liang, Yushi, 700
Liang, Yuxuan, 313, 907
Liang, Yuzhe, 456, 691, 748
Liang, Yuzhi, 743
Liang, Zhenwen, 269, 357, 576, 709, 838
Liang, Zhixiang, 130
Liang, Zhiyuan, 636

Liang, Zi, 207
Liang, Zihan, 402, 808
Liao, 93, 118, 136, 186, 196, 264, 289, 292, 327, 328, 350, 361, 388, 396, 424, 434, 645, 656, 659, 666, 668, 674, 693, 703, 739, 749, 759, 777, 788, 810, 815, 819, 842
Liao, Chenfei, 666
Liao, Chonghua, 289, 656, 668, 739
Liao, Chunhua, 118
Liao, Dingshi, 186
Liao, Disen, 842
Liao, Hao, 659
Liao, Huanxuan, 396, 424, 703
Liao, Jianxin, 327, 388
Liao, Jingyun, 136
Liao, Junjie, 815
Liao, Lizi, 196, 328, 350
Liao, Manwen, 292
Liao, Minpeng, 788
Liao, Qing, 434, 645, 674
Liao, Weibin, 264
Liao, Wenrui, 693
Liao, Xiangwen, 396, 703, 819
Liao, Xin, 749
Liao, Yanghaoyu, 777
Liao, Yusheng, 93
Liao, Zhaohe, 361
Liao, Zhenyi, 759
Liao, Zhiyuan, 810
Libbrecht, 695
Libbrecht, Maxwell W., 695
Libovický, 157
Libovický, Duan Vari; Abishek Stephen; Jindich,
Liew, 327, 376
Liew, Alan Wee-Chung, 327, 376
Lifan, 674
Lifan, Chen, 674
Lijing, 754
Lijing, Ren, 754
LIM, 267
Lim, 90, 116, 133, 147, 177, 182, 184, 185, 186, 187, 197, 202, 217, 290, 447, 465, 476, 515, 541, 554, 594, 601, 622, 741, 753, 800, 803, 804, 806, 830, 842, 858, 903, 908
Lim, Byeonggeuk, 116, 803
Lim, Chae-Gyun, 217
Lim, Dahyun JungJaewook LeeHeuiseok, 541
Lim, Ee-Peng, 187, 290
Lim, Heuiseok, 177, 182, 465, 594, 601, 806, 842, 858, 908
Lim, Hyeongsoo, 903
Lim, Hyunggu, 147
Lim, Hyunkuk, 830
Lim, Jongmin, 217
Lim, Jongwon, 133
Lim, Joy Jia Yin, 476, 554, 804
Lim, Jungwoo, 465
Lim, Kahee, 202
Lim, Kyungho, 197
Lim, KyungTae, 515
Lim, Lawrence, 184
Lim, Sehun, 90
Lim, Seoha, 185
Lim, Ser-Nam, 622
Lim, Seungwon, 741
Lim, Sungjib, 447
LIM, TEABIN, 267
Lim, YongTaek, 753
Lim, Younghun, 800
Lim, Zi Pong, 186
Lima, 839
Lima, Victor De, 839
Limkonchotiwat, 169, 902
Limkonchotiwat, Peerat, 169, 902
LIN, 166
Lin, 89, 98, 105, 109, 110, 112, 116, 118, 134, 136, 137, 140, 147, 152, 158, 159, 161, 164, 165, 166, 172, 182, 184, 186, 191, 210, 211, 218, 222, 225, 239, 243, 245, 248, 251, 252, 254, 262, 264, 268, 274, 296, 300, 302, 305, 309, 311, 312, 319, 324, 329, 338, 340, 344, 347, 349, 355, 357, 369, 372, 374, 378, 380, 384, 393, 413, 417, 432, 434, 445, 455, 467, 469, 474, 479, 487, 490, 491, 494, 501, 513, 514, 530, 538, 547, 549, 550, 551, 560, 562, 568, 573, 575, 594, 599, 601, 621, 631, 632, 634, 646, 653, 663, 664, 669, 674, 680, 684, 688, 691, 699, 703, 714, 720, 724, 730, 736, 737, 740, 741, 742, 743, 744, 745, 746, 748, 749, 750, 754, 758, 759, 760, 762, 765, 774, 775, 776, 784, 794, 796, 798, 810, 826, 830, 831, 840, 843, 849, 851, 856, 859, 870, 877, 880, 887, 896, 897
lin, 338, 436, 857
Lin, Abraham, 538
Lin, Ailiang, 172
Lin, Allen, 551
Lin, Bill Yuchen, 599
Lin, Biyuan, 758
Lin, Bohan, 714
Lin, Can, 530
Lin, Che, 870
Lin, Chenchen, 305

Lin, Chenghua, 268, 669, 830
Lin, Chenxi, 749
Lin, Chenyang, 191
Lin, Chi-Heng, 245
Lin, Chung-Ching, 112
Lin, Dahua, 159, 239, 859
Lin, Dechen, 239
Lin, Dianqing, 762
Lin, DingYang, 413, 748
Lin, Dongding, 663
Lin, Dongyan, 243
Lin, Dongying, 688
Lin, Erliang, 98
Lin, Fan, 674
Lin, Fulin, 393
Lin, Gang, 338
lin, Ge, 857
Lin, Guan-Ting, 455
Lin, Hangyu, 826
Lin, Haokun, 491
Lin, Haotian, 222
Lin, Haowen, 831
Lin, Hongfei, 378
Lin, Honglin, 274, 750, 859
Lin, Hongyu, 252, 369, 653, 741
Lin, Hongzhan, 105, 699
Lin, Hsien-Chin, 89
Lin, I-Fan, 736
Lin, Jessica, 765
Lin, Jiahang, 740
Lin, Jianghao, 248, 601, 794, 877
Lin, Jiaye, 329, 417, 691
Lin, Jifan, 134
Lin, Jimmy, 161, 594
Lin, Jiuheng, 549, 810
Lin, Jun, 467
lin, Junda, 436
Lin, Junyang, 158, 434, 560, 634, 897
Lin, Kangheng, 251
Lin, Kevin, 112
Lin, Kuanwei, 760
Lin, Lanbo, 344
Lin, Li, 445, 684, 724
Lin, Liang, 147, 621, 741, 775, 776
Lin, Lu, 730
Lin, Mingbao, 774
LIN, MINGQUAN, 166
Lin, Nianyi, 880
Lin, Pei-Hung, 118
Lin, Peiqin, 663
Lin, Peng, 843
Lin, Qiang, 745, 746
Lin, Qika, 851
Lin, Qiliang, 743
Lin, Qingwei, 211, 349, 684, 720
Lin, Rexar, 136
Lin, Shengyuan, 166
Lin, Shou-De, 467, 513, 547
Lin, Shuxia, 374
Lin, Sida, 347
Lin, Siying, 550
Lin, Tao, 264
Lin, Ting-En, 849
Lin, Tingting, 184
Lin, Tzu-Mi, 551
Lin, Tzu-Quan, 218, 312
Lin, Tzung-Sheng, 262
Lin, Wang, 355, 754
Lin, Wei, 384, 568, 737, 784
Lin, Weicheng, 302
Lin, Weihao, 309
Lin, Weiyao, 380
Lin, Weizhe, 594, 759
Lin, Wenjie, 494
Lin, Wenjun, 300
Lin, Xiao, 469, 575
Lin, Xiaowen, 562
Lin, Xingyu, 357
Lin, Xixun, 319, 324, 490
Lin, Xuanxu, 736
Lin, Xubo, 182
Lin, Xuemin, 680
Lin, Xueyuan, 796
Lin, Yang, 340
Lin, Yankai, 432, 501, 573, 798, 887
Lin, Yehao, 311
Lin, Yi, 109
Lin, Yi-Cheng, 110, 218
Lin, Yiguan, 631
Lin, Yiming, 744
Lin, Ying-Chun, 479
Lin, Ying-Jia, 514
Lin, Yonghua, 186, 742
Lin, Youfang, 703
Lin, Youyuan, 896
Lin, Yu-Xiang, 225
Lin, Yu-Zheng, 856
Lin, Yue, 254
Lin, Yueqian, 140
Lin, Yuhao, 664
Lin, Yun, 487
Lin, Yupian, 632
Lin, Yuping, 164, 296
Lin, Zekai, 165
lin, zekun wangjingjie zengyingxu liliang yanghongfei, 338

Lin, Zhanteng, 137
Lin, Zhen, 840
Lin, Zheng, 152
Lin, Zhenghao, 210, 758
Lin, Zheyu, 646
Lin, Zhihao, 372
Lin, Zhiqiu, 116
Lin, Zhiyu, 474
Lin;,
Lin;, Chenghua,
Lin;, Jimmy,
LING, 818
Ling, 114, 250, 272, 288, 497, 600, 715, 786, 849, 851, 865
ling, 674
LING, BIN, 818
Ling, Fenghua, 250
Ling, Kan, 715
Ling, Mingyang, 786
ling, Rui, 674
Ling, Xiaoying, 497, 851
Ling, Yaobin, 288
Ling, Zhan, 272
Ling, Zhen-Hua, 114, 865
Ling, Zhenqing, 849
Ling, Ziyang, 600
linsheng, 538
linsheng, linsheng, 538
Linzen, 500
Linzen, Tal, 500
Lioma, 496
Lioma, Christina, 496
Lipani, 451, 740
Lipani, Aldo, 451, 740
Lipka, 102, 557, 598, 784
Lipka, Nedim, 102, 557, 598, 784
Lipka;,
Lipka;, Abhinav Sharma; Brian Zhang; Deepti Guntur; Zhiyang Zuo; Shreyas Chaudhari; Wenlong Zhao; Franck Dernoncourt; Puneet Mathur; Ryan A. Rossi; Nedim,
Litchiowong;,
Litchiowong;, Napassorn,
Litman, 385
Litman, Ron, 385
Litschko, 104
Litschko, Robert, 104
LIU, 112, 199, 250, 255, 554
Liu, 82, 84, 86, 87, 88, 89, 93, 95, 96, 97, 100, 107, 108, 109, 110, 111, 112, 113, 114, 118, 119, 122, 125, 129, 132, 133, 134, 135, 137, 138, 140, 142, 144, 145, 147, 148, 149, 152, 156, 157, 158, 159, 161, 162, 164, 165, 166, 169, 172, 175, 180, 181, 184, 185, 187, 188, 189, 191, 194, 197, 200, 202, 204, 208, 209, 210, 212, 213, 214, 215, 216, 219, 225, 227, 230, 233, 235, 239, 248, 251, 253, 255, 256, 258, 260, 261, 262, 265, 269, 272, 273, 274, 276, 278, 279, 280, 282, 284, 286, 295, 296, 297, 301, 303, 304, 306, 310, 311, 312, 313, 314, 315, 316, 321, 322, 323, 324, 325, 326, 327, 328, 329, 330, 332, 333, 334, 335, 336, 338, 339, 340, 341, 342, 343, 344, 345, 346, 347, 348, 350, 351, 352, 354, 355, 356, 357, 358, 359, 360, 361, 364, 368, 369, 370, 372, 373, 377, 378, 380, 382, 383, 384, 386, 387, 389, 391, 392, 393, 394, 395, 396, 397, 400, 401, 404, 405, 406, 407, 410, 411, 412, 413, 414, 415, 417, 418, 419, 420, 421, 422, 423, 424, 425, 427, 429, 430, 431, 433, 434, 435, 436, 437, 438, 439, 441, 442, 444, 446, 448, 449, 450, 455, 456, 457, 458, 459, 460, 462, 463, 464, 465, 467, 469, 472, 474, 475, 476, 477, 481, 482, 483, 485, 486, 487, 490, 491, 493, 500, 501, 504, 505, 506, 507, 508, 509, 513, 514, 524, 529, 530, 531, 532, 537, 538, 539, 540, 541, 543, 544, 545, 548, 550, 551, 552, 554, 555, 558, 560, 561, 562, 563, 564, 566, 568, 569, 571, 573, 575, 576, 577, 578, 579, 583, 584, 586, 587, 588, 589, 590, 591, 592, 593, 594, 596, 599, 602, 603, 605, 607, 611, 612, 616, 619, 620, 621, 622, 624, 625, 626, 628, 630, 631, 633, 634, 635, 636, 638, 640, 641, 642, 643, 645, 646, 647, 648, 649, 650, 651, 652, 654, 656, 657, 658, 660, 663, 664, 665, 669, 673, 675, 676, 677, 678, 679, 680, 682, 683, 684, 685, 686, 687, 688, 689, 690, 691, 694, 695, 697, 698, 699, 701, 702, 703, 704, 705, 706, 707, 708, 712, 713, 715, 716, 717, 718, 719, 720, 721, 723, 724, 725, 726, 727, 729, 732, 735, 736, 737, 739, 740, 741, 742, 743, 744, 745, 746, 747, 748, 750, 751, 753, 754, 756, 758, 759, 760, 764, 765, 770, 771, 773, 774, 775, 776, 777, 779, 780, 781, 782, 783, 788, 789, 790, 791, 792, 793, 794, 795, 796, 797, 799, 802, 803, 804, 805, 807, 808, 809, 812, 814, 817, 819, 820, 821, 825, 826, 827, 828, 830, 831, 832, 836, 837, 838, 843, 844, 845, 847, 848, 849, 851,

852, 853, 854, 855, 858, 859, 860, 861, 863, 864, 865, 869, 871, 875, 876, 877, 879, 881, 882, 883, 884, 885, 886, 887, 892, 893, 894, 896, 897, 898, 900, 901, 902, 904, 905, 906
liu, 224, 265, 271, 344, 357, 475, 654, 691
Liu, Ahao, 312
Liu, Aishan, 260
Liu, Aiwei, 122, 701, 893
Liu, Akide, 675
Liu, Alexander, 209
Liu, Alisa, 446
Liu, An, 394, 830
Liu, Andrew, 161
Liu, Anjin, 524
Liu, Anmin, 501
Liu, Ao, 900
Liu, Baihui, 665
Liu, Bao-di, 387
Liu, Bei, 197, 272, 347, 619
Liu, Beidan, 335
Liu, Bin, 100
Liu, Binbin, 657
Liu, Bing, 530, 560, 779
Liu, Binglin, 476
Liu, Bingquan, 680, 689
Liu, Bingyu, 456
LIU, Bo, 250
Liu, Bo, 88, 169
Liu, Bowen, 458
Liu, Boyang, 303
Liu, Cao, 475
Liu, Chang, 417, 469, 571, 622, 718, 782
Liu, Changyu, 507
Liu, ChaoChun, 505
Liu, Chaoran, 213
Liu, Che, 260, 630, 884
Liu, Chen, 137, 278, 347, 410, 756
Liu, Cheng-Lin, 187, 431, 636
Liu, Chengao, 134
Liu, Chengkai, 551
Liu, Chengwen, 405
Liu, Chengwu, 255
Liu, Chenqing, 397
Liu, Chenxin, 756
Liu, Chenxu, 306, 501
Liu, Chenyu, 545
Liu, Chi, 112
Liu, Chi Harold, 748
Liu, Chong, 765
Liu, Chonghan, 630
Liu, Chuanyi, 391, 776
Liu, Chunhua, 620
Liu, Chunli, 418, 419
Liu, Ciyi, 718
Liu, Congying, 427
Liu, Daizong, 640
Liu, DanDan, 328
Liu, Daohuan, 682
Liu, Dawei, 423, 750
Liu, Dayiheng, 560, 852, 858
Liu, Di, 386
Liu, Dianbo, 383
Liu, Dingming, 113
Liu, Dongfang, 507
Liu, Donghan, 578
Liu, Dongnan, 352
Liu, Dongqi, 724
Liu, Dongrui, 368, 419, 545
Liu, Dugang, 329, 335
Liu, Dunqiang, 87
Liu, Fan, 145, 149, 356
Liu, Fangxin, 805
Liu, Fei, 588
Liu, Feiran, 278, 475
Liu, Feng, 723
Liu, Fengyuan, 321
Liu, Fengze, 657
Liu, Fujun, 845
Liu, Furui, 310
Liu, Fuyuan, 391
Liu, Gangao, 631
Liu, Gaowen, 548, 633
Liu, Gongshen, 902
Liu, Guangyi, 304, 583, 894
Liu, Guisong, 343
Liu, Guoyang, 689
Liu, Haijiang, 258
Liu, Haijun, 333
Liu, Han, 457, 530, 775, 793, 837, 884
Liu, Hanbing, 593, 773, 836
Liu, Hanmeng, 348
Liu, Hao, 147, 612
Liu, Haochen, 611
Liu, Haodong, 725
Liu, Haoliang, 569
Liu, Haoran, 132
Liu, Haoxuan, 877
Liu, Haoyang, 746
Liu, Haoyue, 566
Liu, Heng, 406, 729, 732
Liu, Hengwei, 404, 621
Liu, Hexin, 455, 736
Liu, Hong, 89
Liu, Hongbin, 820
Liu, Hongbo, 100

Liu, Hongcheng, 322, 364
Liu, Hongye, 483, 819
Liu, Hongyu, 748
Liu, Houde, 142
Liu, Huadai, 339
Liu, Huan, 185, 563, 741, 773
Liu, Huan-Yu, 418
Liu, Huang, 716
Liu, Huanghai, 560
Liu, Huanyu, 577, 796
liu, huaxiao, 265
Liu, Hui, 175, 204, 269, 420, 481, 575, 704, 821
LIU, Huiqin, 554
Liu, Huiqin, 476
Liu, Huiyao, 469
Liu, Jason Klein, 591
Liu, Jeng-Yue, 743
Liu, Ji, 514
Liu, Jia, 694, 747
Liu, Jiacheng, 197, 272, 338, 347, 545, 550, 619, 680, 796
Liu, Jiahao, 333, 355
Liu, Jiahe, 592
Liu, Jiaheng, 657, 660, 676, 707, 825, 900
Liu, Jiahong, 732
LIU, JIAJI, 255
Liu, Jiaji, 135
Liu, Jiajin, 783
Liu, Jiajun, 392, 401, 421, 631, 719, 727, 740
Liu, Jiale, 906
liu, Jiale, 344
Liu, Jialin, 303, 885
Liu, Jiaming, 184
Liu, Jiamou, 181
Liu, Jiang, 348, 490
Liu, Jianliang, 324
Liu, Jianxing, 437
Liu, Jiaqi, 545, 832
Liu, Jiaqing, 865
Liu, Jiarui, 156
Liu, Jiateng, 802
Liu, Jiawei, 683, 726
Liu, Jiayao, 344
Liu, Jiayi, 588
Liu, Jiayu, 472, 691, 808
Liu, Jie, 119, 253, 314, 352, 397, 654, 704, 861
Liu, Jieyuan, 809
Liu, Jin, 688
Liu, Jinbo, 836
Liu, Jincheng, 377
Liu, Jinfei, 369
Liu, Jing, 111, 736, 746
Liu, Jingguo, 111
Liu, Jingjiang, 132, 144, 216
Liu, Jingping, 341, 463, 568, 836
Liu, Jingxuan, 605
Liu, Jingyao, 256
Liu, Jinhan, 648
Liu, Jinliang, 657
Liu, Jinshuo, 647
Liu, Jinyan, 368
Liu, JiYan, 235
Liu, Jiyuan, 852
Liu, Jun, 248, 436, 464, 821, 851, 871
Liu, Junfei, 435
liu, Junfeng, 475
Liu, Junhan, 439
Liu, Junhao, 771
Liu, Junle, 324
Liu, Junlin, 311
Liu, Junpeng, 351
liu, junxi, 691
Liu, Junxu, 369
Liu, Kai, 540
Liu, Kaifeng, 389, 893
Liu, Kaikui, 500, 892
Liu, Kaili, 191
Liu, Kaiming, 93
Liu, Kainan, 803
Liu, Kang, 97, 147, 272, 325, 355, 358, 396, 424, 631, 638, 642, 652, 685, 703, 814, 819, 853
Liu, Kangjun, 400
Liu, Kyle, 576
Liu, Langming, 675, 695, 698
Liu, Lei, 742
Liu, Lemao, 296, 425, 605
Liu, Liang, 583, 894
Liu, Liangxin, 142, 276
Liu, Liangzu, 713
Liu, Lieqi, 538
Liu, Lijuan, 118
Liu, Lin, 368, 380
Liu, Linfeng, 864
Liu, Lingyu, 330
Liu, Lingyuan, 336
Liu, Liqun, 87, 605, 607, 616, 847
Liu, Liu, 84
Liu, Lu, 109
Liu, Luhui, 705
Liu, Marianne Menglin, 477, 819
Liu, Meihan, 262
Liu, Meizhu, 552
Liu, Meng, 415
Liu, Mengjie, 789
Liu, Mengqiao, 230, 295

Liu, Mengyang, 531
Liu, Mengyao, 272
Liu, Miao, 776
Liu, Michael, 560
Liu, Michael Xieyang, 306
Liu, Ming, 342, 345, 395, 429, 690
Liu, Mingchao, 611
Liu, Minghao, 676
Liu, Mingqing, 626
Liu, Mingwei, 620
Liu, Mingyu, 590, 881, 900
Liu, Minqian, 781
Liu, Naiming, 586
Liu, Nayu, 391, 621
Liu, Ning, 138
Liu, Ningyuan, 219
Liu, Ollie, 225
Liu, Pei, 548
Liu, Peipei, 427
Liu, Peiyang, 165, 230, 295, 506, 848
Liu, Peiyu, 532, 634, 753
Liu, Pengfei, 134, 145, 255, 665, 750
Liu, Pengwei, 643
Liu, Pengyuan, 86, 372
Liu, Qi, 159, 273, 321, 338, 576, 634, 702, 838
liu, Qi, 654
Liu, Qian, 95, 389, 561, 641, 677
Liu, Qiang, 634, 698
Liu, Qiankun, 251
Liu, Qianying, 213
Liu, Qiao, 588
Liu, Qibin, 540
Liu, Qin, 225, 594, 792, 886
Liu, Qing, 369
Liu, Qingbin, 630
Liu, Qingwen, 579, 794
Liu, Qingxiang, 313
Liu, Qingyang, 361
Liu, Qingzhi, 191
Liu, Quan, 726
Liu, Qun, 875
Liu, Renyang, 322
Liu, Rui, 339, 469, 656
Liu, Ruishan, 439
Liu, Ruitong, 326
Liu, Ruixuan, 883
Liu, Run, 504
Liu, Runtao, 736
Liu, Runxuan, 395
Liu, Ryan, 513
Liu, Sen, 395
Liu, Shanshan, 854
Liu, Shaofan, 382
Liu, Shaonan, 253
Liu, Sheng, 441, 544
Liu, Shengchao, 719
Liu, Shenghua, 635, 686
Liu, Shengping, 391, 684, 760
Liu, Shichun, 382, 740, 820
Liu, Shihao, 831, 906
Liu, Shilei, 675, 695, 698
Liu, Shixuan, 897
Liu, Shiyan, 420
Liu, Shuang, 892
Liu, Shudong, 129, 274
Liu, Shujian, 486
LIU, Shujie, 112, 199
Liu, Shujie, 253, 352
Liu, Shukai, 260, 485
Liu, Shuliang, 144, 707
Liu, Shunyu, 421
Liu, Shuochen, 436, 791
Liu, Shuxin, 718
Liu, Si, 339
Liu, Sijia, 360, 795
Liu, Siming, 631
Liu, Siran, 297, 321, 332
Liu, Sirui, 727
Liu, Siyuan, 706
Liu, Songwei, 538
Liu, Tang, 596
Liu, Tao, 140, 773
Liu, Tengxiao, 194
Liu, Tianchi, 455
Liu, Tianci, 611, 794
Liu, Tianjian, 695
Liu, Tianle, 723, 725
Liu, Tianlun, 723, 725
Liu, Tianming, 387, 599
Liu, Tianyang, 809
Liu, Tianyu, 96, 825
Liu, Timothy, 209
Liu, Ting, 328, 389, 626, 643, 658, 697
liu, ting, 357
Liu, Tingwen, 304, 332, 384, 591, 715
Liu, Tongliang, 340, 797
Liu, Tongxin, 303
Liu, Tzu-Chi, 265
Liu, Wei, 129, 152, 256, 387, 545, 558, 702, 770, 849, 876
liu, wei, 224
Liu, Weicong, 697
Liu, Weifeng, 387
Liu, Weijie, 149, 539
Liu, Weisi, 132
Liu, Weiwen, 157, 721, 794, 863, 877, 892

Liu, Wenbin, 357
Liu, Wenhan, 310, 699
Liu, Wenxi, 310, 345
Liu, Wenyuan, 665
Liu, Wu, 282
Liu, Wulong, 312
Liu, Xi, 795
Liu, Xian, 301
Liu, Xiang, 341, 561
Liu, Xianglong, 260, 485
Liu, Xiangyan, 301
Liu, Xiao, 89, 97, 164, 210, 280, 286, 573, 675
liu, xiao, 271
Liu, Xiao-Yang, 166
Liu, Xiaoguang, 539
Liu, xiaoguang, 783
Liu, Xiaole, 189
Liu, Xiaolong, 315
Liu, Xiaoming, 719, 807
Liu, Xiaoou, 602
Liu, Xiaoqian, 456
Liu, Xiaoqun, 493
Liu, Xiaorui, 901
Liu, Xiaoxiao, 686
Liu, Xiaoyang, 407
Liu, Xiaoyu, 412
Liu, Xiaoyuan, 882
Liu, Xiaozhong, 467, 529, 821, 843
Liu, Xihui, 301
Liu, Xin, 133, 486, 633, 678, 721, 758, 790
Liu, Xinchen, 108
Liu, Xinggao, 788
Liu, Xingjia, 799
Liu, Xingyu, 506, 543
Liu, Xinyu, 342, 739
Liu, Xinyuan, 407
Liu, Xinyun, 554
Liu, Xiong, 849
Liu, Xu, 239, 401, 715
Liu, Xuanyi, 607
Liu, Xudong, 110
Liu, Xue, 370, 393, 474, 507, 509, 712, 774
Liu, Xuebo, 323
Liu, Xueding, 687
Liu, Xuepeng, 702
Liu, Xuwei, 828
Liu, Xuye, 161
Liu, Xuyuan, 771
Liu, Yafei, 87
Liu, Yahui, 777
Liu, Yaluo, 462
Liu, Yan, 378, 836, 884
Liu, Yanbing, 304, 765
Liu, Yanchi, 771, 837
Liu, Yang, 23, 82, 93, 172, 200, 214, 215, 227, 251, 284, 329, 339, 414, 462, 482, 550, 640, 651, 684, 713, 797, 844, 855, 898
Liu, Yanguang, 442, 902
Liu, Yangyang, 406, 729, 732
Liu, Yanming, 107
Liu, Yanshan, 152
Liu, Yantao, 897
Liu, Yanzhi, 372
Liu, Yao, 707, 708
Liu, Yaowenqi, 555
Liu, Yefeng, 677
Liu, Yepeng, 592
Liu, Yi, 404, 405
Liu, Yichen, 550
Liu, Yicheng, 437
Liu, Yiding, 775
Liu, Yidong, 562
Liu, Yifan, 284, 589
Liu, Yifan Simon, 125, 568, 764
Liu, Yifeng, 869
Liu, Yihao, 413, 750
Liu, Yihe, 212
Liu, Yihong, 583, 812, 826
Liu, Yijin, 701
Liu, Yikang, 346
Liu, Yilong, 490
Liu, Yilun, 564, 584, 756
Liu, Yiming, 114, 134
Liu, Yinan, 688
Liu, Ying, 86
Liu, Yingen, 286
Liu, Yingxue, 664
Liu, Yiping, 548
Liu, Yiqi, 669
Liu, Yiqun, 373
Liu, Yiran, 788
Liu, Yisheng, 420
Liu, Yixin, 327, 717
Liu, Yiyang, 474, 507
Liu, Yong, 148, 384, 419, 459, 583, 587, 793, 894
Liu, Yongbin, 625
Liu, Yongjiang, 148, 359
Liu, YongKang, 422, 474, 643, 663
Liu, Yongmei, 346
Liu, Youzheng, 235
Liu, Yozen, 162, 446
Liu, Yu, 145, 304, 508
Liu, Yu-An, 721, 754
Liu, Yuan, 545
Liu, Yuanjun, 830
Liu, Yuankun, 904

Liu, Yuansen, 354, 744
Liu, Yuanxiang, 354, 624, 715
Liu, Yuanxin, 537
Liu, Yuanxing, 328, 389
Liu, Yuanyang, 505
Liu, Yubo, 368
Liu, Yuchen, 736, 777, 845
Liu, Yue, 830
Liu, Yuewen, 134
Liu, Yuexiao, 685
Liu, Yufei, 658
Liu, Yuhan, 129, 202, 679, 887
Liu, Yuhang, 316
Liu, Yujie, 723, 726
Liu, Yuling, 279, 593, 793
Liu, Yun, 735
Liu, Yunhao, 651
Liu, Yunhui, 448, 554
Liu, Yunpeng, 688
Liu, Yunxin, 680, 804
Liu, Yusen, 164
Liu, Yuxing, 555
Liu, Yuxuan, 897
liu, yuxuan, 224
Liu, Yuyang, 158
Liu, Yuying, 373
Liu, Yuyu, 108
Liu, Yuzhe, 433
Liu, Zechun, 444
Liu, Zedi, 339
Liu, Zeming, 169, 541, 737
Liu, Zenghao, 751
Liu, Zequn, 180, 191
Liu, Zeshuo, 361
Liu, Zewen, 881
Liu, Zhanyang, 350
Liu, Zhanyu, 397, 414
Liu, Zhaoxiang, 341
Liu, Zhaoyang, 122, 591
Liu, Zhe, 475
Liu, Zheli, 723
Liu, Zhen, 180
Liu, Zhendong, 660
Liu, Zheng, 89, 135, 239, 274, 361, 482, 491, 508, 789, 853, 859
Liu, Zhenghao, 358, 377, 387, 406, 434, 673, 735
Liu, Zhengqing, 780
Liu, Zhengwu, 821
Liu, Zhengyuan, 460, 469, 649
Liu, Zhengzhong, 809
Liu, Zhenhua, 400, 777
Liu, Zhenke, 439
Liu, Zhenyu, 397, 759
Liu, Zheyuan, 181, 188, 569, 817, 860, 879, 896
Liu, Zhi, 311
Liu, Zhichen, 356, 603
Liu, Zhidan, 410, 646
Liu, Zhihan, 384
Liu, Zhijun, 438
Liu, Zhining, 457, 469, 575
Liu, Zhiqiang, 354, 715, 720
Liu, Zhiwei, 166, 233, 420, 573
Liu, Zhiyu, 261, 611
Liu, Zhiyuan, 405, 476, 543, 716, 735, 905
Liu, Zhizhen, 334, 354
Liu, zhonghao zhang; NA LIU; Jiajia Ma; Nier Wu; Guiping,
Liu, Zhongliang, 456
Liu, Zhongxin, 389, 628, 745
Liu, Zhou, 434, 750
Liu, Zhoufu, 652
Liu, Zhoumianze, 159, 591
Liu, Zhu, 449
Liu, Zhuang, 827
Liu, Zhuo, 303
Liu, Zichen, 301, 316
Liu, Zicheng, 348, 490
Liu, Zihan, 425, 487, 577, 887
Liu, Zihang, 208, 420
Liu, Zihao, 669
Liu, Zijun, 797
Liu, Zikang, 119, 430
Liu, Zining, 279
Liu, Ziniu, 626
Liu, Ziqi, 744
Liu, Ziqian, 684
Liu, Ziqiang, 849
Liu, Ziqiong, 505
Liu, Ziquan, 465, 807
Liu, Zirui, 164, 418, 419
Liu, Ziwei, 100, 361, 564, 775
Liu, Zixuan, 338
Liu, Ziyan, 341, 463, 836
Liu, Ziyi, 561, 650, 736
Liu, Zuozhu, 450, 645
Liu;,
Liu;, Zhang Zhang; Shuqi Lu; Hongjin Qian; Di He; Zheng,
Liu;, Zhiyuan,
Livescu, 863
Livescu, Karen, 863
Livyatan, 816
Livyatan, Lior, 816
Liò, 519
Liò, Pietro, 519
Lo, 429, 515, 531, 587

Lo, Chun Hei, 587
Lo, David, 515, 531
Lo, Kyle, 429
Locatelli, 826
Locatelli, Acyr, 826
Loiseau, 713
Loiseau, Gabriel, 713
Loison, 169
Loison, António, 169
Lokam, 577
Lokam, Satya, 577
Lombart, 173
Lombart, Erika, 173
Lombrozo, Tania, 22
Lomidze, 235
Lomidze, Lika, 235
Lonca, 633
Lonca, Emmanuel, 633
Long, 128, 131, 227, 319, 322, 435, 462, 469, 473, 540, 604, 637, 649, 679, 762, 815, 871, 899
Long, Dingkun, 473
Long, Eileen Margaret Peters, 899
Long, Jiangxuan, 637
Long, Kaiwen, 540
Long, Lingkun, 128
Long, Luo, 462
Long, Meixiu, 604
Long, Qi, 227
Long, Qingqing, 319
Long, Quanyu, 128, 322, 649
Long, Wanqiu, 469
Long, Weicai, 815
Long, Xiang, 435
Long, Xianxuan, 131
Long, Yunbo, 679
Long, Yunfei, 871
Long, Zimo, 762
Longépé, 238
Longépé, Nicolas, 238
Loo, 512
Loo, Nicholas, 512
Lopardo, 798
Lopardo, Antonio, 798
Loparo, 131
Loparo, Kenneth, 131
Lopez, 238
Lopez, Antonio, 238
Lopez-Lira, 160, 166, 233
Lopez-Lira, Alejandro, 160, 166, 233
Lothritz, 868
Lothritz, Cedric, 868
Lou, 86, 140, 214, 297, 317, 392, 403, 585, 677, 721, 817, 882, 906
Lou, Chanhou, 882
Lou, Fanghao, 677
Lou, Haoye, 86
Lou, Jungang, 585
Lou, Qian, 882
Lou, Shengfeng, 214
Lou, Siwei, 817
Lou, Siyu, 297
Lou, Xingyu, 721
Lou, Xingzhou, 317
Lou, Yibin, 140
Lou, Yihang, 403
Lou, Yiling, 906
Lou, Yuhang, 392
Loukachevitch, 551
Loukachevitch, Natalia V, 551
Lovering, 871
Lovering, Charles, 871
Lovichit, 388
Lovichit, Pak, 388
Low, 294, 811
Low, Bryan Kian Hsiang, 294, 811
Loye, 398
Loye, Xian Qi, 398
LU, 432
Lu, 89, 114, 126, 143, 145, 165, 166, 174, 180, 197, 199, 201, 202, 205, 206, 219, 230, 239, 244, 249, 252, 263, 265, 272, 284, 285, 295, 298, 304, 316, 320, 325, 328, 332, 333, 339, 350, 352, 353, 354, 357, 360, 363, 368, 369, 374, 381, 382, 390, 400, 404, 409, 413, 431, 432, 435, 441, 443, 453, 455, 470, 471, 494, 514, 519, 520, 524, 535, 541, 544, 545, 572, 578, 583, 587, 596, 599, 600, 607, 608, 618, 621, 627, 631, 636, 644, 647, 648, 653, 656, 659, 662, 673, 676, 681, 684, 692, 698, 717, 722, 723, 738, 739, 741, 745, 751, 752, 762, 765, 767, 775, 788, 798, 824, 826, 832, 840, 842, 848, 852, 861, 870, 875, 886, 892, 894, 901
Lu, Alex, 244
Lu, Chengqiang, 197, 298
Lu, Chenyang, 201
Lu, Chenyao, 230, 295
Lu, Chonggang, 745
Lu, Fengyu, 435
Lu, Hailing, 249
Lu, Haiqi, 328
Lu, Hanqing, 875
Lu, Haonan, 627

Lu, Haoran, 775
Lu, Haotian, 382
Lu, Heng, 199
Lu, Hong, 443
Lu, Hongyu, 717
Lu, Hongyuan, 541
Lu, Huchuan, 514
Lu, Huilin, 608
Lu, Jiacheng, 202, 870
Lu, Jiahua, 265
Lu, Jiajie, 767
LU, Jialiang, 432
Lu, Jianfeng, 738
Lu, Jianglin, 143, 618
Lu, Jiasen, 219
Lu, Jiayi, 409
Lu, Jie, 524, 681
Lu, Jinghui, 431
Lu, Jingyu, 114, 832
Lu, Junting, 644, 684
Lu, Junyu, 607
Lu, Liangsi, 738
Lu, Lin, 350
Lu, Lindong, 239
Lu, Miao, 272
Lu, Mingfei, 852
Lu, Mingzhe, 765, 886
Lu, Nan, 354
Lu, Pan, 441
Lu, Panzhong, 369, 840
Lu, Peiqing, 205
Lu, Peng, 166, 587, 788
Lu, Pengqian, 524
Lu, Pengrui, 470
Lu, Pinyan, 752
Lu, Qingyu, 656
Lu, Ruifang, 325
Lu, Ruiying, 648
Lu, Sanglu, 600
Lu, Shijia, 390
Lu, Shuai, 636
Lu, Shuo, 363
Lu, Sidi, 357
Lu, Siyu, 432
Lu, Tingwei, 535, 698
Lu, Wei-Chung, 263
Lu, Weihai, 692
Lu, Weiming, 126, 304, 368, 404, 621, 631, 722, 751
Lu, Wenhao, 673
Lu, Wenhuan, 332
Lu, Wenjian, 316
Lu, WenJie, 676
Lu, Wenpeng, 374, 400, 453, 762
Lu, Wensheng, 842
Lu, Xiaoman, 544
Lu, Xiaoyue, 333, 353
Lu, Xingyu, 145
Lu, Xu, 316
Lu, Xuesong, 180
Lu, Yan, 199, 284, 352, 494, 723
Lu, Yang, 471
Lu, Yao, 519
Lu, Yaojie, 252, 369, 653, 741
Lu, Yi, 599, 901
Lu, Yi-Fan, 89
Lu, Yida, 413, 520
Lu, Yifan, 647
Lu, Yifei, 659
Lu, Yu, 165, 230, 295, 320
Lu, Yu-An, 174
Lu, Yuan, 892
Lu, Yuheng, 455, 572
Lu, Yunfei, 360, 381, 739
Lu, Yutong, 206
Lu, Yuxiang, 404, 596
Lu, Yuxing, 824
Lu, Yuxuan, 824, 826, 861, 875
Lu, Ze, 578
Lu, Zeyi, 662
Lu, Zhengxi, 126, 304, 583, 894
Lu, Zhengxuan, 798
Lu, Zhichao, 201
Lu, Zhicong, 470
Lu, Zhifeng, 285
Lu, Zhiyu, 848
Lu, Zhonghang, 545
Lu, Zimu, 339, 432
Luan, 124, 152, 256, 508, 558, 693, 788, 849
luan, 224
Luan, Chengyu, 693
Luan, Hao, 508
Luan, Jian, 124, 152, 256, 558, 849
luan, jian, 224
Luan, Sitao, 788
Lubana, 207
Lubana, Ekdeep Singh, 207
Lubis, 89
Lubis, Nurul, 89
Lucas, 85
Lucas, Jason S, 85
Luft, 207
Luft, Benjamin J., 207
Luger, 862
Luger, Sarah, 862
Luhtaru, 820

Luhtaru, Agnes, 820
Lui, 734
Lui, John C.S., 734
Lukas, 202, 887
Lukas, Nils, 202, 887
Lukashevskiy, 424
Lukashevskiy, Dmitriy, 424
Lukasiewicz, 283, 786
Lukasiewicz, Thomas, 283, 786
Luo, 81, 89, 97, 105, 117, 124, 145, 146, 161, 162, 180, 185, 186, 191, 210, 214, 246, 249, 252, 253, 261, 279, 284, 288, 294, 299, 305, 316, 317, 321, 327, 338, 347, 350, 375, 379, 384, 387, 406, 412, 413, 414, 422, 436, 441, 446, 448, 449, 450, 452, 462, 471, 481, 482, 545, 552, 562, 566, 573, 591, 597, 614, 621, 627, 632, 656, 659, 664, 665, 668, 669, 671, 676, 677, 680, 683, 685, 687, 688, 689, 701, 705, 729, 732, 733, 738, 745, 746, 747, 748, 758, 791, 795, 798, 810, 849, 851, 853, 856, 875, 879, 886, 888, 889, 895, 901, 908, 909
Luo, Alan, 294
Luo, Ali, 249
Luo, Andy, 566
Luo, Chao, 705
Luo, Chen, 446, 481, 875
Luo, Chong, 145
Luo, Chu Fei, 180
Luo, Dezhi, 908
Luo, Fuwen, 214, 462
Luo, Guangchun, 671
Luo, Guanran, 665
Luo, Guanting, 252
Luo, Guibo, 210, 452
Luo, Guoqing, 214
Luo, Han, 81, 299
Luo, Hanjun, 471, 482
Luo, Hao, 387, 664, 895
Luo, Haochen, 347
Luo, Haohao, 849
Luo, Haoming, 591
Luo, Haoran, 656, 668, 851
Luo, Haotian, 676
Luo, Hongchen, 185
Luo, Hongyu, 161
Luo, Hui, 327, 632
Luo, Jiaming, 246
Luo, Jian, 745, 746
Luo, Jianwen, 97, 162
Luo, JiaTang, 279, 685
Luo, Jiaxin, 747
Luo, Junyu, 191, 448
Luo, Kairong, 901
Luo, Kangcheng, 180
Luo, Kangyang, 449
Luo, Kun, 853
Luo, Liang, 795
Luo, Linhao, 449
Luo, Minnan, 888, 889
Luo, Pengfei, 791
Luo, Ping, 621
Luo, Qinyu, 573
Luo, Qiqi, 680
Luo, Ruilin, 450
Luo, Run, 316
Luo, Shirui, 261
Luo, Sichun, 321
Luo, Sigang, 688
Luo, Sijia, 146
Luo, Siwen, 545
Luo, Tao, 614
Luo, Tiancheng, 683
Luo, Tianyu, 186
Luo, Tongxu, 627
Luo, Weihua, 379, 406, 441, 677, 701, 729, 732, 738, 798
Luo, Weiqing, 305
Luo, Wen, 117, 659
Luo, Xianzhen, 350
Luo, Xiao, 191, 284, 448, 810
Luo, Xiaohao, 384, 422
Luo, Xiaoling, 253
Luo, Xiaoyu, 810
Luo, Xinchen, 414
Luo, Yang, 597
Luo, Yanyi, 856
Luo, Yawei, 669
Luo, Yaxin, 338
Luo, Yifan, 448, 591
Luo, Yilun, 665
Luo, Yinbo, 338
Luo, Yingfeng, 413, 748
Luo, Yu, 909
Luo, Yujie, 317, 436
Luo, Yuping, 89
Luo, Yuxiao, 379
Luo, Yuyu, 552, 562, 687, 886
Luo, Zhen, 758
Luo, Zhenbo, 124
Luo, Zheng, 288
Luo, Zhensheng, 347
Luo, Zhihao, 412
Luo, Zhiwen, 689
Luo, Zhunchen, 384, 733

Luo, Zihan, 375
Luo, Ziqin, 879
Luo, Ziyang, 105
Luong, 469, 799
Luong, Hoang-Chau, 799
Luong, Tuan, 469
Lupart, 254
Lupart, Simon, 254
Lutgen, 868
Lutgen, Anne-Marie, 868
Luthra, 243, 250
Luthra, Aditya, 250
Luthra, Mahi, 243
Lutz, 529
Lutz, Wolfgang, 529
Luu, 210, 432, 851, 901
Luu, Anh Tuan, 210, 432, 851, 901
Lv, 97, 134, 135, 158, 260, 286, 294, 298, 300, 311, 321, 344, 357, 422, 487, 560, 571, 593, 621, 661, 665, 672, 695, 698, 699, 731, 748, 798, 819, 839, 874, 900, 905
Lv, Ang, 487, 798, 905
Lv, Bo, 621
Lv, Bowen, 286
Lv, Changze, 321, 661
Lv, ChenXu, 158
Lv, Chunli, 300
Lv, Dongyi, 593
Lv, Haijun, 839
Lv, Hang, 298, 819
Lv, Jianwei, 621
Lv, Kangtao, 695
Lv, Liuzhenghao, 158
Lv, Minxuan, 672
Lv, Peizhuo, 294, 593
Lv, Qingsong, 698
Lv, Ruilin, 900
Lv, Tianpeng, 344
Lv, Tianrui, 731
Lv, Weifeng, 260
Lv, Weijiang, 874
Lv, Wenbo, 560
Lv, Wenhan, 665
Lv, Xiaolei, 134
Lv, Xin, 571
Lv, Xinxuan, 311, 748
Lv, Zhao, 699
Lv, Zhenyu, 97
Lv, Zheqi, 135
Lv, Zhonghou, 357, 422
Lyng, 127
Lyng, Gregory D, 127
LYU, 416
Lyu, 85, 87, 88, 147, 164, 264, 296, 345, 404, 415, 474, 507, 515, 531, 584, 599, 621, 634, 653, 654, 666, 687, 694, 742, 788, 826, 896
Lyu, Chenyang, 88, 296
Lyu, Fuyuan, 474, 507
Lyu, Guangtao, 634, 654
Lyu, Jiayang, 264
Lyu, Mengsi, 742
Lyu, Michael R., 531, 687
Lyu, Nuoyan, 584
Lyu, Pin, 653
Lyu, Qiner, 85
Lyu, Qingqing, 404, 621
Lyu, Qinhan, 147
Lyu, Ruiliang, 164
Lyu, Sicheng, 788
Lyu, Weimin, 826
Lyu, Xinglin, 345
LYU, YANG, 416
Lyu, Yibo, 694
Lyu, Yuanhuiyi, 666
Lyu, Yuanjie, 87, 896
Lyu, Yunbo, 515
Lyu, Zhiheng, 599
Lyu, Ziyu, 415
Lübeck, 784
Lübeck, Frederike, 784

MA, 630
Ma, 87, 92, 95, 97, 99, 105, 108, 112, 115, 117, 132, 133, 134, 136, 137, 144, 145, 148, 149, 159, 161, 174, 178, 185, 191, 195, 198, 202, 204, 208, 210, 212, 214, 222, 231, 232, 239, 257, 259, 260, 264, 268, 270, 273, 278, 283, 285, 289, 293, 294, 299, 303, 304, 306, 307, 310, 311, 314, 320, 326, 328, 332, 338, 341, 345, 353, 355, 360, 361, 362, 377, 378, 390, 394, 395, 396, 404, 405, 407, 409, 412, 415, 422, 425, 432, 436, 440, 451, 455, 456, 457, 464, 470, 472, 474, 476, 482, 486, 487, 524, 528, 539, 545, 548, 551, 560, 561, 565, 581, 584, 587, 611, 622, 623, 626, 634, 638, 640, 651, 653, 662, 665, 667, 674, 676, 678, 680, 681, 692, 694, 699, 707, 709, 712, 716, 719, 721, 732, 735, 736, 737, 741, 748, 757, 759, 766, 773, 774, 775, 783, 790, 795, 796, 806, 814, 822, 825, 830, 831, 834, 843, 844, 852, 855, 861, 870, 874, 879, 884, 886, 888, 892, 903, 907, 908
ma, 357

Ma, Bolei, 95, 97, 174, 806, 886
Ma, Chang, 95
Ma, Chen, 474, 712
Ma, Chenrui, 137
Ma, Chiyu, 115, 136
Ma, Chuan, 332
Ma, Congbo, 212
Ma, Dan, 311
Ma, De, 843
Ma, Dehua, 422
Ma, Dongsheng, 239, 674
Ma, Duohe, 303
Ma, Fengbo, 581
Ma, Fenglong, 117, 528, 611, 766
Ma, Gengchen, 852
Ma, George, 299
Ma, Guoqing, 132, 144
Ma, Guoxin, 719
Ma, Haokai, 415
Ma, Haolun, 830
Ma, Haoyuan, 404
Ma, Haozhe, 830
Ma, Hehuan, 283
Ma, Hongli, 680
Ma, Hongxia, 757
Ma, Huan, 326
Ma, Huidong, 539, 783
Ma, Jackie, 822
Ma, Jianfei, 259
Ma, Jianghong, 270
Ma, Jianzhe, 195, 293
Ma, Jiaqi W., 307, 626
Ma, Jiatong, 736
Ma, Jiaxin, 198
Ma, Jin, 304
Ma, Jing, 105, 412, 472, 699
Ma, Jingkun, 565
Ma, Jizheng, 831
Ma, Jun-Yu, 436
Ma, Junpeng, 707
Ma, Junwei, 551
Ma, Kaijing, 676
Ma, Ke, 87, 651
Ma, Kexin, 721
Ma, Liantao, 264
Ma, Linrui, 587
Ma, Lintao, 145
Ma, Lu, 210
Ma, Maolin, 640
Ma, Marci Chi, 407
Ma, Mengyang, 214
MA, Ming, 630
ma, ming, 357
Ma, MingHua, 345
Ma, Minghua, 584
Ma, Pingchuan, 185
Ma, Qian, 208, 294
Ma, Qiang, 377
Ma, Qianli, 361, 732, 741
Ma, Qianou, 232, 306
Ma, Qingsen, 285
Ma, Qingzhi, 394
Ma, Rui, 134
Ma, Runmin, 440
Ma, Ruotian, 149, 425, 716
Ma, Ruoyun, 892
Ma, Shaoxuan, 464
Ma, Shiqing, 844
Ma, Shixuan, 709
Ma, Shuai, 257
Ma, Shujuan, 674
Ma, Shuoyoucheng, 735
Ma, Simin, 486
Ma, Siwei, 204
Ma, SiYuan, 622
Ma, Teng, 456
Ma, Tengfei, 108, 548
Ma, Tianyi, 178, 202, 396, 774, 870
Ma, Tzu-Yen, 545
Ma, Wei-Ying, 268
Ma, Weitao, 99, 328, 353
Ma, Weiyu, 667
Ma, Weizhi, 482, 855
Ma, Wenping, 222
Ma, Wenquan, 707
Ma, Xiangyao, 314
Ma, Xiangyu, 888
Ma, Xiaojun, 773
Ma, Xiaoqian, 638
Ma, Xiaosong, 133, 148
Ma, Xiaoteng, 440, 903
Ma, Xiaowen, 191, 289
Ma, Xiaoxiao, 320, 662
Ma, Xiaoxu, 834
Ma, Xilai, 362, 694
Ma, Xinbei, 425, 470
Ma, Xinchen, 278
Ma, Xindian, 665
Ma, Xing, 345, 487
Ma, Xingjun, 451
Ma, Xinhang, 92
Ma, Xinhua, 355
Ma, Xinyan, 395
Ma, Xinyu, 273, 310, 699, 795
Ma, Xu, 524
Ma, Xudong, 260

Ma, Xueguang, 161
Ma, Xueqi, 681
Ma, Xuetao, 874
Ma, Yanbiao, 692
Ma, Yanxiang, 409
Ma, Yao, 208, 879, 907
Ma, Yichuan, 159
Ma, Yihan, 457
Ma, Yiling, 623
Ma, Yinchao, 707
Ma, Yingwei, 796, 825
Ma, Yiran, 405
Ma, Yixuan, 360
Ma, Yong, 678
Ma, Youmi, 814
Ma, Yubo, 560, 852
Ma, Yuchen, 759
Ma, Yue, 736
Ma, Yunpu, 191, 289, 737
Ma, Yunshan, 338
Ma, Zehan, 561
Ma, Zejun, 775
Ma, Zexin, 476
Ma, Zeyao, 634
Ma, Zhanyu, 328, 884
Ma, Zhaoyu, 97
Ma, Zhengzhao, 653
Ma, Zhicheng, 341
Ma, Zhiming, 422
Ma, Zhiyuan, 353, 790
Ma, Zi-Ao, 231
Ma, Zicheng, 259
Ma, Zihan, 888
Ma, Ziqiao, 908
Ma, Zixian, 112
Ma, Ziyang, 390, 432, 455, 748
Ma, Ziyu, 378
Ma, Zizhan, 861
Ma, Zongyang, 134
Maab, 768
Maab, Iffat, 768
Macdonald, 426
Macdonald, Craig, 426
Machacek, 784
Machacek, Roman, 784
Macken, 622
Macken, Lieve, 622
Mackenzie, 439
Mackenzie, Harry, 439
Mackert, 104
Mackert, Michael, 104
Mackin, 109
Mackin, Daniel, 109
Macko, 667
Macko, Dominik, 667
MacLellan, 866
MacLellan, Christopher J., 866
Macé, 169
Macé, Quentin, 169
Madala, 807
Madala, Siddarth, 807
Madani, 681
Madani, Mohammad Reza Ghasemi, 681
Madhusudan, 250
Madhusudan, Sangmitra, 250
Madhusudhan, 441
Madhusudhan, Sathwik Tejaswi, 441
Madhwal, 181
Madhwal, Dhruv, 181
Madkoor, 802
Madkoor, Anas, 802
Madmoun, 442
Madmoun, Hachem, 442
Maestroni, 415
Maestroni, Michele, 415
Magdy, 468, 860
Magdy, Samar M., 468, 860
Mager, 207, 286, 894
Mager, Manuel, 207, 286, 894
Magesh, 509
Magesh, Akshayaa, 509
Maggini, 415, 634
Maggini, Marco, 634
Maggini, Michele Joshua, 415
Magu, 839
Magu, Rijul, 839
Mahato, 642
Mahato, Priyanshu, 642
Mahawar, 305
Mahawar, Utkarsh Kumar, 305
Mahbub, 579
Mahbub, Khairul, 579
Mahjoub, 792
Mahjoub, Hossein Nourkhiz, 792
Mahmoud, 84, 236, 253
Mahmoud, Abdulrahman, 253
Mahmoud, Tarek, 84, 236
Mahongxia, 756
Mahongxia, Mahongxia, 756
Mahowald, 153, 894
Mahowald, Kyle, 153, 894
Mai, 102, 146, 299, 402, 484, 527, 686, 702
Mai, Gengchen, 146, 299
Mai, Hung, 102
Mai, Luo, 527
Mai, Peihua, 702

Mai, Quan, 484
Mai, Sijie, 402
Mai, Tu-Phuong, 686
Maillard, 862
Maillard, Jean, 862
Maimon, 113
Maimon, Gallil, 113
Maistro, 496, 854
Maistro, Maria, 496, 854
Majumdar, 265
Majumdar, Somshubra, 265
Majumder, 156, 257
Majumder, Bodhisattwa Prasad, 257
Majumder, Suvajit, 156
Makaaka, 862
Makaaka, Kamohelo, 862
Makinae, 832
Makinae, Mana, 832
Malamut, 232
Malamut, Jim, 232
Malanga, 611
Malanga, Christopher, 611
Malaviya, 829
Malaviya, Chaitanya, 829
Malay, 441
Malay, Shiva Krishna Reddy, 441
Malekzadeh, 549
Malekzadeh, Mohammad, 549
Malema, 98
Malema, Gabofetswe, 98
Maliha, 263, 894
Maliha, Maisha, 263, 894
Malin, 281
Malin, Bradley A., 281
Mall, 150
Mall, Utkarsh, 150
Mallat, 212
Mallat, Jihad, 212
Mallick, 875
Mallick, Shubhojit, 875
Malmasi, 835
Malmasi, Shervin, 835
Mamta, 193
Mamta, Mamta, 193
Mamun, 199
Mamun, Nursadul, 199
Man, 142, 553
Man, Hieu, 142, 553
Manacop, 232
Manacop, John Allen, 232
Manchanda, 227
Manchanda, Sahil, 227
Manco, 770
Manco, Maxim, 770
Mandic, 538
Mandic, Danilo, 538
Mandyam, 444
Mandyam, Karishma, 444
Mang, 882
Mang, Qiuyang, 882
Mangal, 604
Mangal, Ravi, 604
Manitaras, 784
Manitaras, Theofilos Ioannis, 784
Manjunath, 105
Manjunath, Sagar, 105
Manjunatha, 784
Manjunatha, Varun, 784
Mannor, 816
Mannor, Alon, 816
Mannu, 81
Mannu, Gurdeep Singh, 81
Manocha, 127, 140, 526, 598
Manocha, Dinesh, 127, 140, 526, 598
Manoharan, 687
Manoharan, Sathiamoorthy, 687
Mansimov, 385
Mansimov, Elman, 385
Mansurov, 236
Mansurov, Jonibek, 236
Mantri, 160
Mantri, Yoages Kumar, 160
Manzoor, 545
Manzoor, Muhammad Arslan, 545
MAO, 272
Mao, 89, 136, 231, 308, 316, 317, 320, 358, 396, 427, 429, 491, 522, 572, 659, 689, 709, 729, 757, 796, 888
Mao, Chenhui, 491
Mao, Dongxing, 888
MAO, KEYU, 272
Mao, Rui, 396, 659
Mao, Shunqi, 757
Mao, Wenji, 316
Mao, Xian-Ling, 89, 136, 231
Mao, Yanxu, 427
Mao, Yuchen, 308
Mao, Yueheng, 429
Mao, Yujiao, 358
Mao, Yujun (Audrey), 796
Mao, Yuren, 317
Mao, Zhendong, 320, 689, 709, 729
Mao, Zhenjiang, 522, 572
Maracani, 677
Maracani, Andrea, 677
Maraia, 426

Maraia, Gabriele, 426
Marashian, 246
Marashian, Ali, 246
Marconi, 311
Marconi, Gian Maria, 311
Mardani, 899
Mardani, Morteza, 899
Marfurt, 784
Marfurt, Andreas, 784
Margatina, 158
Margatina, Katerina, 158
Marinas, 784
Marinas, Inés Altemir, 784
Marinescu, 493
Marinescu, Radu, 493
Marivate, 862
Marivate, Vukosi, 862
Markert, 663
Markert, Katja, 663
Marneffe, 173, 457
Marneffe, Marie-Catherine de, 173, 457
Marquis, 633, 722
Marquis, Pierre, 633, 722
Marr, 189
Marr, Clayton, 189
Marrese-Taylor, 823
Marrese-Taylor, Edison, 823
Marshall, 104, 615
Marshall, Eric, 615
Marshall, Iain J., 104
Marteau, 602
Marteau, Benoit, 602
Martin, 154, 545
Martin, Alexander, 154
Martin, Geoffrey, 545
Martinelli, 174, 566, 581
Martinelli, Giuliano, 581
Martinelli, Sofía, 174, 566
Martinez, 291, 610
Martinez, Brais, 291
Martinez, Natalia, 610
Martino, 232, 548
Martino, Giovanni Da San, 232, 548
Martins, 114, 705
Martins, André F. T., 705
Martins, Jonas Mayer, 114
Martínez, 862
Martínez, Juan Pablo, 862
Marzollo, 517
Marzollo, Michele, 517
Masano, 485
Masano, Miwa, 485
Maskey, 652
Maskey, Utsav, 652
Mass, 230, 553
Mass, Yosi, 230, 553
Masud, 487
Masud, Sarah, 487
Matari, 570
Matari, Maja, 570
Mathew, 574
Mathew, Ruben, 574
Mathur, 189, 526, 584, 598, 784
Mathur, Puneet, 526, 584, 598, 784
Mathur, Yash, 189
Matias, 585
Matias, Yossi, 585
Matoba, 784
Matoba, Kyle, 784
Matrenok, 784
Matrenok, Simon, 784
Matsumoto, 252, 854
Matsumoto, Yuji, 252, 854
Matsuo, 608, 660
Matsuo, Yutaka, 608, 660
Matthes, 194, 907
Matthes, Florian, 194, 907
Matuszek, 820
Matuszek, Cynthia, 820
Maurya, 320, 633
Maurya, Chandresh Kumar, 320
Maurya, Sunil Kumar, 633
Mausam, 94
Mausam, Mausam, 94
Maveli, 805
Maveli, Nickil, 805
Maximov, 330
Maximov, Egor, 330
May, 258, 478, 555, 789
May, Chandler, 555
May, Jonathan, 478, 789
May, Victor, 258
Mayer, 232
Mayer, Brian, 232
Maynard, 235
Maynard, Diana, 235
Mazanov, 233
Mazanov, Viktor, 233
Mazein, 655
Mazein, Nikita, 655
Mazhar, 236
Mazhar, Abdullah, 236
Mazor, 109
Mazor, Nir, 109
Mazumder, 371
Mazumder, Akm Moshiur Rahman, 371

Mazzone, 560
Mazzone, Maria, 560
Mañacop, 232
Mañacop, John Andrew, 232
MC, 399
MC, Manjunatha Naik, 399
McAuley, 250, 257, 483, 514
McAuley, Julian, 23, 250, 257, 483, 514
McCurdy, 448
McCurdy, Kate, 448
McGuinness, 226
McGuinness, Harvey, 226
McKeown, 199
McKeown, Kathleen, 199
McLeman, 497
McLeman, Micah Helzerman; Steven R Wilson; Cam,
McMillan, 531
McMillan, Collin, 531
Meadows, 682
Meadows, Jordan, 682
Meagher, 813
Meagher, Margaret, 813
Medya, 200, 889
Medya, Sourav, 200, 889
Meek, 586
Meek, Austin, 586
Meghwani, 552
Meghwani, Hansa, 552
Mehnaz, 556
Mehnaz, Shagufta, 556
Mehrabi, 851
Mehrabi, Ninareh, 851
Mehri, 441
Mehri, Shikib, 441
Mehri, Shuhaib, 441
Mehrotra, 239, 606, 617, 621
Mehrotra, Nikita, 239
Mehrotra, Prakhar, 606, 617, 621
Mei, 148, 168, 303, 468, 594, 617, 635, 636, 686, 746, 752
Mei, Congyao, 746
Mei, Jianbiao, 148, 303
Mei, Jingbiao, 594
Mei, Katelyn Xiaoying, 168
Mei, Lijun, 752
Mei, Lingrui, 635, 686
Mei, Yongsheng, 617
Mei, Yuke, 468
Mei, Zihan, 636
Meij, 119
Meij, Edgar, 119
Meister, 247
Meister, Clara, 247
Mejia, 199
Mejia, Ivan Ernesto Perez, 199
MEKKI, 468
Mekki, 860
MEKKI, Abdellah EL, 468
Mekki, Abdellah El, 860
Meltzer-Asscher, 862
Meltzer-Asscher, Aya, 862
Men, 631, 852
Men, Rui, 852
Men, Tianyi, 631
Menczer;,
Menczer;, Filippo,
Mendel, 664
Mendel, Shahar, 664
Mendelson, 220
Mendelson, Avi, 220
Mendes, 464
Mendes, Ethan, 464
Mendonça, 784
Mendonça, Henrique, 784
Mendozza, 641
Mendozza, Alessandro Corona, 641
Menezes, 167
Menezes, Arul, 167
Meng, 93, 96, 97, 113, 161, 178, 191, 198, 259, 281, 296, 304, 339, 390, 397, 426, 435, 480, 501, 527, 537, 543, 555, 583, 625, 626, 644, 657, 665, 670, 678, 701, 729, 745, 756, 759, 786, 848, 859, 893, 908
Meng, BingXu, 555
Meng, Changhua, 93
Meng, Chong, 304
Meng, Fanbo, 397, 759
Meng, Fandong, 96, 281, 296, 537, 543, 626, 701
Meng, Fanyu, 657
Meng, Feifan, 501
Meng, Haoqian, 665
Meng, Helen, 435
Meng, Helen M., 339
Meng, Huanran, 259
Meng, Jinxiang, 97
Meng, Linghui, 625
Meng, Linjian, 657
Meng, Liqiu, 786
Meng, Liyuan, 527
Meng, Long, 729
Meng, Qi, 339
Meng, Qinglin, 859
Meng, Qingyu, 113, 191
Meng, Rui, 678
Meng, Ruoxi, 161

Meng, Shiao, 178
Meng, Shichang, 670
Meng, Wenchao, 583, 893
Meng, Xiangfeng, 625
Meng, Xiaojun, 644, 756, 848, 908
Meng, Yu, 480
Meng, Zaiqiao, 198, 390, 426
Meng, Zijie, 745
Meo, 216
Meo, Pasquale De, 216
Merdjanovska, 864
Merdjanovska, Elena, 864
Merrill, 891
Merrill, Mike A, 891
Mesgar, 97
Mesgar, Mohsen, 97
Messmer, 784
Messmer, Bettina, 784
Metaxas, 141, 291, 534
Metaxas, Dimitris N., 141, 534
Metaxas, oannis Maniadis, 291
Metcalf, 580
Metcalf, Katherine, 580
Meyer, 713
Meyer, Maxime, 713
Meymand, 560
Meymand, Mohammad Mahmoudi, 560
Meyoyan, 115
Meyoyan, Gonzalo Ariel, 115
Mgonzo, 837
Mgonzo, Macton, 837
Mhaskar, 314
Mhaskar, Shivam Ratnakant, 314
Mi, 130, 170, 242, 264, 289, 357, 398, 436, 576, 676, 687, 700, 709, 722, 750, 838
Mi, Fei, 130, 264, 398
Mi, Haitao, 289, 357, 436, 576, 687, 709, 722, 838
Mi, Hao, 170
Mi, Hongze, 676
Mi, Maggie, 242, 700
Mi, Tiantian, 750
Miao, 176, 239, 244, 262, 272, 315, 327, 361, 407, 488, 542, 622, 657, 717, 756
Miao, Chunyu, 262, 488
Miao, Dezhuang, 272
Miao, Duoqian, 176
Miao, Fen, 315
Miao, Guoyi, 407
Miao, Lingqi, 756
Miao, Rui, 327, 717
Miao, Runqing, 657
Miao, Yibo, 361
Miao, Yongliang, 244
Miao, Yuyang, 622
Miao, Zhongtao, 542
Miao, Ziyang, 239
Michaelov, 108
Michaelov, James, 108
Michail, 817
Michail, Andrianos, 817
Michieli, 537
Michieli, Umberto, 537
Michielon, 567
Michielon, Edoardo, 567
Mickus, 891
Mickus, Timothee, 891
Miculescu, 81
Miculescu, Angelo, 81
Miguez, 881
Miguez, Ruben, 881
Mihalcea, 626, 902
Mihalcea, Rada, 626, 902
Mikhailov, 236
Mikhailov, Vladislav, 236
Mikros, 494
Mikros, George, 494
Mikulova, 597
Mikulova, Marie, 597
Milano, 868
Milano, Emilia, 868
Miletic, 591
Miletic, Milan, 591
Mille, 287
Mille, Simon, 287
Miller, 106, 189, 214, 822
Miller, Lee M., 106, 214
Miller, Moritz, 189
Miller, Timothy A., 822
Milliken, 678
Milliken, Louis Mark, 678
Mills, 773
Mills, Keith G., 773
Min, 118, 119, 190, 217, 296, 335, 339, 555, 580, 847
Min, Erxue, 335
Min, Haewon, 217
Min, Hyunji, 580
Min, Kyungkoo, 190
Min, Kyungmin, 296
Min, Sehee, 118
Min, Xiongkuo, 339
Min, Yingqian, 119, 847
Min, Yuecong, 555
Minamoto, 255, 485
Minamoto, Rei, 255, 485
Minegishi, 660

Minegishi, Gouki, 660
Minervini, 742
Minervini, Pasquale, 742
Ming, 123, 474, 866
Ming, Yidong, 474
Ming, Yifei, 123, 866
Minggad, 809
Minggad, Quulgan, 809
Minh, 347, 536
Minh, An Binh, 347
Minh, Nguyen Le, 536
Minixhofer, 137
Minixhofer, Benjamin, 137
Mino, 863
Mino, Hideya, 863
Mintz, 82, 230
Mintz, Reshef, 82
Mintz, Toben, 230
Mira, 648
Mira, Antonietta, 648
Mira;,
Mira;, Antonietta,
Mire, 476
Mire, Joel, 476
Miriyala, 305
Miriyala, Srinivas Soumitri, 305
Mirovsky, 597
Mirovsky, Jiri, 597
Mirsaidova, 459
Mirsaidova, Aziza, 459
Mishra, 261, 357, 442, 443, 615, 633, 642, 672, 742
Mishra, Aniket Santosh, 642
Mishra, Ashish, 672
Mishra, Bamdev, 615
Mishra, Deep Naryan, 443
Mishra, Deepali, 357
Mishra, Shresth, 742
Mishra, Soumya Smruti, 442
Mishra, Venkatesh, 633
Mishra, Vishala, 261
Misra, 258, 863, 878
Misra, Amita, 878
Misra, Diganta, 258
Misra, Kanishka, 863
Mistica, 167
Mistica, Meladel, 167
Mita, 639
Mita, Masato, 639
Mitani, 873
Mitani, Ryosuke, 873
Mitchell, 298, 360
Mitchell, Cassie S., 360
Mitchell, Rupert, 298
Mithsara, 613
Mithsara, Wanniarachchi Kankanamge Malithi, 613
Mitra, 115, 116, 483, 523, 906
Mitra, Chancharik, 116
Mitra, Ekata, 523
Mitra, Saayan, 906
Mitra, Subrata, 483
Mitra, Tanu, 115
Mittal, 115, 131, 558, 562, 613
Mittal, Anshul, 131
Mittal, Ashish, 558
Mittal, Prateek, 562
Mittal, Rohan, 613
Mittal, Shravika, 115
Miwa, 747
Miwa, Makoto, 747
Miyagawa, 489
Miyagawa, Taiki, 489
Miyao, 174
Miyao, Yusuke, 174
Mkrtchyan, 504
Mkrtchyan, Arya Bhardia; Julian Ramirez; Siddhanta Verma; Karen,
Mo, 106, 113, 188, 191, 267, 348, 352, 369, 378, 404, 488, 558, 564, 565, 583, 594, 817, 904
Mo, Fan, 404
Mo, Fengran, 188, 565, 583, 817
Mo, Guozhao, 369, 378
Mo, Kaijie, 106
Mo, Linjian, 352, 564
Mo, Tong, 267, 348, 558, 904
Mo, Wenjie Jacky, 594
Mo, Yifan, 113, 191
Mo, Zhiwen, 488
Moakhar, 909
Moakhar, Arshia Soltani, 909
Moalla, 784
Moalla, Skander, 784
Mochihashi, 101
Mochihashi, Mao Shimada; Hajime Kiyama; Zhidong Ling; Mamoru Komachi; Toshinobu Ogiso; Hiroya Takamura; Daichi,
Modi, 186, 282, 375
Modi, Ashutosh, 282
Modi, Divyaksh Shukla; Ashutosh,
Modi, Sarthak Ketanbhai, 186
Modzelewski, 203, 548
Modzelewski, Arkadiusz, 203, 548
Moens, 218, 624
Moens, Marie-Francine, 218, 624
Moghaddam, 220, 530

Moghaddam, Arya Hadizadeh, 220, 530
Moghimifar, 320
Moghimifar, Farhad, 320
Mohamed, 98, 280, 784, 822, 860
Mohamed, Amr, 280, 822
Mohamed, Fawzi Roberto, 784
Mohamed, Muhidin A, 98
Mohamed, Yehdih, 860
Mohammad, 545, 551, 845
Mohammad, Abdullah, 845
Mohammad, Saif M., 545, 551
Mohammadi, 813
Mohammadi, Seyedali, 813
Mohammadian, 644
Mohammadian, Sadegh, 644
Mohammadpour, 765
Mohammadpour, Amir, 765
Mohammed;,
Mohammed;, Nabeel,
Mohan, 790, 846
Mohan, Shyam, 846
Mohan, Stuti, 790
Mohananey, 575
Mohananey, Anhad, 575
Mohapatra, 728
Mohapatra, Biswesh, 728
Mohor, 667
Mohor, Banerjee, 667
Mok, 244, 267
Mok, Jisoo, 267
Mok, Micah, 244
Mokhiamar, 247
Mokhiamar, Mahmoud O., 247
Molchanov, 899
Molchanov, Pavlo, 899
Molfese, 873
Molfese, Francesco Maria, 873
Molfetta, 835
Molfetta, Lorenzo, 835
Molinari, 614
Molinari, Andrea, 614
Molitz, 884
Molitz, Penelope, 884
Momeni, 903
Momeni, Liliane, 903
Monazzah, 704
Monazzah, Erfan Moosavi, 704
Mondal, 522, 556, 574, 819
Mondal, Aishee, 574
Mondal, Ishani, 522, 819
Mondal, Mainack, 556
Mondorf, 891
Mondorf, Philipp, 891
Moniz, 201
Moniz, Nuno, 201
Monshizadeh, 821
Monshizadeh, Mahsa, 821
Montagnani, 648, 658
Montagnani, Daniele, 648, 658
Montalan, 169
Montalan, Jann Railey, 169
Montariol, 784
Montariol, Syrielle, 784
Monterola, 232, 525
Monterola, Christopher, 232, 525
Montgomery, 127
Montgomery, Kyle, 127
Montoya, 560
Montoya, Ernesto Gabriel Hernández, 560
Monz, 873
Monz, Christof, 873
Moon, 189, 207, 287, 450, 474, 587, 601
Moon, Hyeonseok, 601
Moon, Hyosik, 287
Moon, Jun-Yeong, 189
Moon, Minkyeong, 474
Moon, Sangjun, 587
Moon, Suhong, 207
Moon, Taesup, 450
Mooney, 448
Mooney, Michael, 448
Moore, 547, 820
Moore, Juston, 820
Moore, Steven James, 547
Moosavi, 700, 830
Moosavi, Nafise Sadat, 700, 830
Mor-Lan, 585
Mor-Lan, Guy, 585
Moradi, 329
Moradi, Mohammad Mahdi, 329
Moradizeyveh, 471
Moradizeyveh, Sahar, 471
Moraffah, 805
Moraffah, Raha, 805
Morais, 175
Morais, Michael J., 175
Morales, 234
Morales, Gianmarco De Francisci, 234
Morales-Garzón, 854
Morales-Garzón, Andrea, 854
Morand, 601
Morand, Victor, 601
Moratelli, 803
Moratelli, Nicholas, 803
More, 250
More, Trush Shashank, 250

Moreira, 169
Moreira, Gabriel de Souza P., 169
Mori, 489, 778, 800
Mori, Junki, 489
Mori, Paolo, 778
Mori, Tatsunori, 800
Morishita, 176, 512
Morishita, Makoto, 512
Morishita, Terufumi, 176
Moritz, 243
Moritz, Rashel, 243
Moro, 468, 667, 835
Moro, Gianluca, 468, 835
Moro, Robert, 667
Moroni, 873
Moroni, Luca, 873
Mortensen, 177, 189, 217, 456, 821, 841
Mortensen, David, 841
Mortensen, David R., 177, 189, 217, 456
Mortensen, Genevieve, 821
Mosbach, 278, 806
Mosbach, Marius, 278, 806
Moshtaghi, 878
Moshtaghi, Masud, 878
Moskvoretskii, 315
Moskvoretskii, Maria Marina; Daniil Moskovskiy; Sergey Pletenev; Mikhail Salnikov; Alexander Panchenko; Viktor,
Mosquera-Gómez, 862
Mosquera-Gómez, Rafael, 862
Mostafavi, 551
Mostafavi, Ali, 551
Mothe, 601
Mothe, Josiane, 601
Mothukuri, 358
Mothukuri, Viraaji, 358
Mou, 214, 349, 382, 448, 796
Mou, Lili, 214, 796
Mou, Yuchen, 382, 448
Mou, Yueqi, 349
Mou;,
Mou;, Lili,
Moubayed, 268
Moubayed, Noura Al, 268
Mouchel, 784
Mouchel, Luca, 784
Moukafih, 806
Moukafih, Youness, 806
Mourad, 859
Mourad, Aya, 859
Mousavi, 459, 463
Mousavi, Reza, 459, 463
Mousi, 897
Mousi, Basel, 897
Mozafari, 639, 697, 744
Mozafari, Jamshid, 639, 697, 744
Mozharovskyi, 176
Mozharovskyi, Pavlo, 176
Mraikhat, 590
Mraikhat, Osama Al, 590
Mu, 83, 255, 335, 345, 353, 363, 395, 424, 475, 626, 628, 736
Mu, Bingshen, 363, 736
Mu, Fangwen, 475
Mu, Jian, 424
Mu, Lin, 353
Mu, Linjie, 255
Mu, Nankun, 628
Mu, Rongchuan, 345, 395
Mu, Ruotong, 83
Mu, Yongyu, 335, 626
Mubarak, 494, 769
Mubarak, Hamdy, 494, 769
Muchemi, 862
Muchemi, Carol, 862
Mudur, 329
Mudur, Sudhir, 329
Mueller, 90, 207, 277, 858
Mueller, Aaron, 90, 207, 277, 858
Muhamed, 454
Muhamed, Aashiq, 454
Muhammad, 380, 471, 551, 862, 871
Muhammad, Abubakar, 380
Muhammad, Ahmed, 871
Muhammad, Shamsuddeen Hassan, 471, 551, 862
Muhan_Zhang1,
Muhan_Zhang1, Yiding_Wang1 Yuxuan_Chen11 Fauxu_Meng1 Xifan_Chen1 Xiaolei_Yang1,
Mujahid, 517, 545
Mujahid, Zain Muhammad, 517, 545
Mujral, 156
Mujral, Kushaldeep, 156
Mukherjee, 270, 504, 587, 633, 722
Mukherjee, Anjishnu, 504
Mukherjee, Koyel, 270, 587
Mukherjee, Tanmoy, 633, 722
Mukhituly, 589
Mukhituly, Nurdaulet, 589
Mukhopadhyay, 633
Mukhopadhyay, Souradeep, 633
Mukkamla, 731
Mukkamla, Aneesh, 731
Mulay, 122
Mulay, Anish, 122
Mule, 192

Mule, Srujan, 192
Mulian, 517
Mulian, Hadar, 517
Muller, 258, 517
Muller, Eilif, 258
Muller, Lorenz K., 517
Mullins, 488
Mullins, Robert, 488
Mullor, 439
Mullor, Thomas, 439
Mumin, 538
Mumin, Sadick Abdul, 538
Mundada, 257, 273
Mundada, Gagan, 257
Mundada, Rishika, 273
Munigala, 558
Munigala, Vitobha, 558
Munkhdalai, 575
Munkhdalai, Tsendsuren, 575
Munne, 854
Munne, Rumana Ferdous, 854
Murai, 529
Murai, Fabricio, 529
Muraki, 416
Muraki, Yusuke, 416
Muralidharan, 560
Muralidharan, Rasika, 560
Murata, 272
Murata, Eiki, 272
Murawaki, 178, 884
Murawaki, Yugo, 178, 884
Murray, 862
Murray, Kenton, 862
Murtaza, 204
Murtaza, Sahil, 204
Murugesan, 178, 201, 215, 536
Murugesan, Keerthiram, 178, 201, 215, 536
Mushtaq, 128
Mushtaq, Erum, 128
Mussabayeva, 874
Mussabayeva, Ayana, 874
Mustafid, 862
Mustafid, Ahmad, 862
Muszyski, 763
Muszyski, Jakub, 763
Mutlu, 677
Mutlu, Sinan, 677
Muttram, 226
Muttram, Harry G., 226
Mutyala, 305
Mutyala, Narendra, 305
Muziri, 862
Muziri, Casper, 862
Möller, 826
Möller, Sebastian, 826
Møller, 472
Møller, Anders Giovanni, 472
Müller, 111
Müller, Markus, 111

Na, 183, 217, 230, 564, 575, 592, 828
Na, ChelWon, 183
Na, CheolWon, 564, 828
Na, Hongbin, 230
Na, Seung-Hoon, 217, 575, 592
Naaman, 800
Naaman, Zohar, 800
Nacar, 860
Nacar, Omer, 860
Nadeem, 682
Nadeem, Afrozah, 682
Nadel, 160
Nadel, Ryan, 160
Nafea, 860
Nafea, Youssef, 860
Nafisi, 644
Nafisi, Ali, 644
Nagata, 542
Nagata, Masaaki, 542
Nagireddy, 781
Nagireddy, Manish, 781
Naidu, 188
Naidu, Nagasai Saketh, 188
Naik, 189, 467
Naik, Aakanksha, 467
Naik, Atharva, 189
Naim, 147, 301
Naim, Omar, 147, 301
Nair, 107, 461, 862
Nair, Inderjeet, 461
Nair, Sathvik, 107, 862
Najem-Meyer, 784
Najem-Meyer, Sven, 784
Najjar, 579
Najjar, Nadia, 579
Najjuuko, 201
Najjuuko, Claire, 201
Nakadai, 346
Nakadai, Kazuhiro, 346
Nakagawa, 445
Nakagawa, Takumi, 445
Nakamura, 873
Nakamura, Satoshi, 873
Nakano, 190
Nakano, Yukiko, 190
Nakash, 610

Nakash, Itay, 610
Nakashole, 95
Nakashole, Ndapa, 95
Nakayama, 485
Nakayama, Kouta, 485
Nakov, 84, 95, 126, 178, 233, 236, 248, 442, 527, 545, 555, 813, 864, 874, 876, 881, 886
Nakov, Preslav, 84, 95, 126, 179, 233, 236, 248, 442, 527, 545, 555, 813, 864, 874, 876, 881, 886
Nakshatri, 766
Nakshatri, Nishanth_Sridhar,
Nam, 232, 478, 828
Nam, HyunJi, 232
Nam, Yoonsoo, 478
Nam, Youngeun, 828
Nambiar, 156
Nambiar, Karthik, 156
Nan, 273, 324, 707
Nan, Fang, 273
Nan, Jiayan, 707
Nan, Yixuan, 324
Nandi, 469
Nandi, Palash, 469
Naous, 247
Naous, Tarek, 247
Naphade, 511
Naphade, Atharv, 511
Narasimha, 763
Narasimha, Pragnya, 763
Narayan-Chen, 168
Narayan-Chen, Anjali, 168
Narayanam, 392
Narayanam, Ramasuri, 392
Narayanan, 226, 529, 547, 821
Narayanan, Arun Balajiee Lekshmi, 547
Narayanan, Arvind, 821
Narayanan, Shrikanth, 226, 529
Narenthiran, 265
Narenthiran, Sean, 265
Nasar, 864
Nasar, Mohammad Rustom Al, 864
Naseem, 368, 458, 462, 471, 556, 652, 682, 845
Naseem, Mustafa, 462
Naseem, Usman, 368, 458, 471, 556, 652, 682, 845
Nasim, 721
Nasim, Mehwish, 721
Nasonov, 655
Nasonov, Denis, 655
Nasrallah, 841
Nasrallah, Hadi, 841
Nassar, 340
Nassar, Islam, 340
Natarajan, 791
Natarajan, Nagarajan, 791
Nathani, 194
Nathani, Deepak, 194
Nauck, 195
Nauck, Detlef, 195
Navasartian, 524
Navasartian, Karine, 524
Navigli, 264, 581, 873
Navigli, Roberto, 264, 581, 873
Nayak, 441
Nayak, Shravan, 441
Nayeem, 822
Nayeem, Mir Tafseer, 822
Nayel, 862
Nayel, Hamada, 862
Nayim, 273
Nayim, Meshal, 273
Nazari, 341
Nazari, Ali, 341
Naziri, 696
Naziri, Amirreza, 696
Nchifor, 538
Nchifor, Faith, 538
Near, 547
Near, Joseph, 547
Nedoluzhko, 597
Nedoluzhko, Anja, 597
Neel, 785
Neel, Seth, 785
Neeser, 637
Neeser, Andrew, 637
Neiswanger, 225
Neiswanger, Willie, 225
Nemcovsky, 220
Nemcovsky, Yaniv, 220
Nenadic, 856
Nenadic, Goran, 856
Nenkova, 156
Nenkova, Ani, 156
Neo, 702
Neo, Clement, 702
Nepal, 109
Nepal, Subigya, 109
Neplenbroek, 263, 713
Neplenbroek, Vera, 263, 713
Neubig, 246, 474, 539
Neubig, Graham, 246, 474, 539
Neville, 479
Neville, Jennifer, 479
Ng, 130, 135, 169, 209, 336, 369, 569, 627, 692, 777, 898, 901
Ng, Aik Beng, 209, 569

Ng, Hwee Tou, 898
Ng, Michael K., 369
Ng, Raymond, 169
Ng, See-Kiong, 130, 135, 692, 777, 901
Ng, Vincent, 336, 627
Ngai, 159, 305
Ngai, Edith Cheuk-Han, 159, 305
Nghiem, 103, 833
Nghiem, Huy, 103, 833
Ngo, 507, 515, 553
Ngo, Chris, 515
Ngo, Linh, 507
Ngo, Nghia Trung, 553
Ngong, 547
Ngong, Ivoline, 547
Ngui, 169, 902
Ngui, Jian Gang, 169, 902
Nguyen, 102, 123, 192, 210, 233, 285, 327, 351, 405, 408, 411, 426, 427, 432, 466, 478, 485, 507, 510, 553, 574, 584, 585, 600, 639, 645, 701, 745, 779, 816, 826, 851, 862, 866, 882, 890
Nguyen, 'Hoang D., 574
Nguyen, Cam-Tu, 408, 585, 600
Nguyen, Chieu, 233
Nguyen, Cong Huy, 745
Nguyen, Cong-Duy, 210
Nguyen, Cuong C., 639
Nguyen, Daniel, 466
Nguyen, Dung, 485, 882
Nguyen, Duy, 890
Nguyen, Hoang Son, 426, 427
Nguyen, Huy, 584
Nguyen, Huy Huu, 432
Nguyen, Khang, 862
Nguyen, Khanh, 102
Nguyen, Khoa, 510
Nguyen, Manh, 351
Nguyen, Minh-Tien, 862
Nguyen, Minh-Vuong, 816
Nguyen, My Chiffon, 862
Nguyen, Nam V., 701
Nguyen, Phi Le, 745
Nguyen, Quoc Viet Hung, 327
Nguyen, Son Dinh, 745
Nguyen, Thai Binh, 779
Nguyen, Thanh Hong, 411, 645
Nguyen, Thanh Trung, 745
Nguyen, Thien, 507
Nguyen, Thien Huu, 427, 432, 553, 584, 851
Nguyen, Thuat, 411
Nguyen, Tien N., 192
Nguyen, Tien-Phat, 405
Nguyen, Tuan Dung, 745
Nguyen, Tung, 426
Nguyen, Van Bach, 826
Nguyen, Van Chien, 584
Nguyen, Van-Chien, 507
Nguyen, Vivian, 478
Nguyen, Xuan-Phi, 123, 285, 866
Nhiem, 840
Nhiem, Tran Van, 840
Ni, 192, 233, 296, 301, 310, 353, 401, 421, 443, 488, 544, 561, 599, 607, 638, 653, 784, 800, 835, 874, 881, 884, 901
Ni, Anqi, 607
Ni, Bo, 901
Ni, Hanchu, 884
Ni, Jingwei, 233, 784, 835
Ni, Jinjie, 301
Ni, Juntong, 443, 881
Ni, Kangqi, 784
Ni, Li, 353
Ni, Minzhe, 544
Ni, Shiwen, 800
Ni, Shiyu, 638
Ni, Wanchun, 401, 421
Ni, Wei, 310
Ni, Xuanfan, 296
Ni, Yiming, 192
Ni, Yuansheng, 599
Ni, Yuchen, 561
Ni, Zhanpeng, 874
Ni, Zhiyu, 488
Ni, Ziyi, 653
Nian, 757
Nian, Junjie, 757
Nickleach, 148
Nickleach, Scott, 148
Nie, 161, 166, 243, 255, 266, 289, 326, 415, 474, 475, 501, 583, 599, 643, 694, 715, 723, 735, 755, 791, 821, 853
Nie, Ercong, 243, 289, 474, 643, 821
Nie, Jiahao, 791
Nie, Jian-Yun, 166, 583
Nie, Lihai, 723
Nie, Liqiang, 326, 415, 694, 755
Nie, Pengyu, 266
Nie, Ping, 161, 475, 599, 853
Nie, Shen, 501
Nie, Shuaiyi, 715
Nie, Wei, 255
Nie, Zhijie, 735
Niehues, 126, 298, 779, 825
Niehues, Jan, 126, 298, 779, 825
Niekerk, 89

Niekerk, Carel van, 89
Nihal, 346
Nihal, Ragib Amin, 346
Nikishina, 589
Nikishina, Irina, 589
Niklaus, 247
Niklaus, Joel, 247
Niklaus;,
Niklaus;, Christina,
Nikolentzos, 865
Nikolentzos, Giannis, 865
Ning, 110, 275, 285, 299, 457, 469, 473, 658, 662, 826, 840
Ning, Lechen, 285
Ning, Liang-Bo, 473
Ning, Liangbo, 662
Ning, Ruiyu, 299
Ning, Ruoxi, 110
Ning, Wu, 658
Ning, Xuying, 457, 469, 826
Ning, Yue, 275
Ning, Zhiyuan, 840
Ningtyas, 715
Ningtyas, Annisa Maulida, 715
Nishida, 252, 512, 781, 854, 896
Nishida, Kosuke, 781
Nishida, Kyosuke, 781
Nishida, Noriki, 252, 854
Nishida, Shin'ya, 896
Nishida, Yuto, 512
Nishimura, 198
Nishimura, Mai, 198
Nishizaki, 509
Nishizaki, Futo Kajita;Nobuyori Nishimura; Takehito Utsuro;Naoki Muto;Chee Siang Leow;Hiromitsu,
Nisioi, 716
Nisioi, Sergiu, 716
Niu, 132, 144, 239, 304, 361, 372, 386, 407, 409, 416, 421, 434, 455, 468, 497, 541, 559, 578, 599, 627, 662, 689, 691, 718, 748, 773, 786, 807, 851, 853, 879
Niu, Ben, 416
Niu, Boyu, 239
Niu, Cheng, 559, 807
Niu, Conghui, 372
Niu, Di, 421, 773
Niu, Fang, 541
Niu, Fuqiang, 434
Niu, Jianwei, 386
Niu, Junbo, 239
Niu, Li, 361
Niu, Luyao, 599, 879
Niu, Peizhi, 407
Niu, Qingyu, 132, 144
Niu, Simin, 304
Niu, Tianhao, 627, 662
Niu, Tianyi, 853
Niu, Xi, 578
Niu, Yilin, 497, 786, 851
Niu, Yingjie, 468
Niu, Yiyi, 689
Niu, Yuhang, 718
Niu, Zhikang, 455, 691, 748
Niu, Ziwei, 409
Niwa, 813
Niwa, Ayana, 813
Niyigaba, 466
Niyigaba, Manzi Fabrice, 466
Niyogi, 139
Niyogi, Mitodru, 139
Nkhata, 484
Nkhata, Gibson, 484
Nnamdi, 602, 824
Nnamdi, Micky, 602
Nnamdi, Micky C., 824
Nobakhtian, 862
Nobakhtian, Melika, 862
Noel, 811
Noel, Charlotte, 811
Nok, 498
Nok, Poon Tsz, 498
NOONEY, 82
NOONEY, AKHIL, 82
Noori, 667
Noori, Ayush, 667
Nopparatbundit, 728
Nopparatbundit, Thitiwat, 728
Noppen, 195
Noppen, Joost, 195
Noroozi, 265
Noroozi, Vahid, 265
Noughabi, 277
Noughabi, Havva Alizadeh, 277
Nouri, 802
Nouri, Marzia, 802
Nouwen, 229
Nouwen, Rick, 229
Novak, 597
Novak, Michal, 597
Nuriel, 385
Nuriel, Oren, 385
Nursal, 360
Nursal, Batuhan, 360
Nus, 770
Nus, Alexander, 770

Nutanong, 169, 205, 388, 430, 728
Nutanong, Sarana, 169, 205, 388, 430, 728
Nuti, 252
Nuti, Gaurav, 252
Nweke, 547
Nweke, Chima, 547
Nyah, 247
Nyah, Sergius Justus, 247
Nyati, 889
Nyati, Pranav, 889
Nzeyimana, 869
Nzeyimana, Antoine, 869

O'Connell, 287
O'Connell, Ornait, 287
O'Donncha, 610
O'Donncha, Fearghal, 610
O'Sullivan", 574
O'Sullivan", "Barry, 574
Oba, 447
Oba, Miyu, 447
Obaida, 590
Obaida, Qusai Abu, 590
Obiefuna, 98
Obiefuna, Abigail OppongNnaemeka Casmir, 98
Obreiter, 600
Obreiter, Max, 600
Ochs, 277
Ochs, Sebastian, 277
Oco, 520
Oco, Juliana Isabelle A. Guillermo; Jasper Kyle Catapang; Nathaniel,
Oda, 255
Oda, Yusuke, 255
Odumakinde, 468
Odumakinde, Ayomide, 468
Oduwole, 837
Oduwole, Mardiyyah, 837
Oermann, 694
Oermann, Yanbing Chen; Ruilin Wang; Zihao Yang; Lavender Yao Jiang; Eric Karl,
Ofek, 664
Ofek, Hila, 664
Ogawa, 485
Ogawa, Hayato, 485
Ogaz, 560
Ogaz, Lucy, 560
Oguz, 810
Oguz, Cennet, 810
Oh, 101, 149, 151, 156, 197, 237, 259, 290, 509, 533, 547, 590, 806, 831, 834, 844, 862, 870, 888
Oh, Alice, 197, 259, 590, 831
Oh, Byung-Doh, 862
Oh, Changdae, 151
Oh, Changhoon, 533
Oh, Dongsuk, 806
Oh, Hanseok, 834
Oh, Hyung-Seok, 101
Oh, Hyunjin Jane, 547
Oh, Jihwan, 290, 888
Oh, Juhyun, 197, 590, 831
Oh, Minjae, 149
Oh, Minsik, 237
Oh, Myeong Seok, 834
Oh, Seong Joon, 156, 509
Oh, Tae-Hyun, 844
Oh, Woo Joo, 870
Oh;,
Oh;, Alice,
Ohashi, 480
Ohashi, Takumi, 480
Ohkuma, 237
Ohkuma, Tomoko, 237
Ojewale, 869
Ojewale, Victor, 869
Ok, 175, 212, 223, 508
Ok, Hyunjong, 212, 223
Ok, Jungseul, 175, 508
Okabe, 862
Okabe, Shu, 862
Okada, 872
Okada, Kensuke, 872
Okazaki, 813, 814
Okazaki, Naoaki, 813, 814
Okewunmi, 837
Okewunmi, Paul, 837
Okoani, 288
Okoani, Ogochukwu N., 288
Okocha, 196
Okocha, Chibuzor, 196
Okoh, 247
Okoh, Ifeoma, 247
Okumura, 172, 492, 587
Okumura, Manabu, 172, 492, 587
Okuno, 608
Okuno, Yasushi, 608
Olabisi, 523
Olabisi, Olubusayo, 523
Oliva, 784
Oliva, Gennaro, 784
Ollagnier, 83
Ollagnier, Anaïs, 83
Olson, 118, 813
Olson, Elizabeth A., 813
Olson, Matthew Lyle, 118

Omar, 298
Omar, Ahmad, 298
Omi, 237
Omi, Takahiro, 237
On, 248
On, Kyoung-Woon, 248
Onderková, 521
Onderková, Kristýna, 521
Ondrejová, 563
Ondrejová, Marek SuppaAndrej RidzikDaniel HládekNatália KaekováViktória, 563
ONeil, 763
ONeil, Mike, 763
Ong, 702
Ong, Luke, 702
Onizuka, 480
Onizuka, Makoto, 480
Onoue, 844
Onoue, Koichi, 844
Onyame, 202
Onyame, Eric, 202
Onyango, 471
Onyango, Nelson Odhiambo, 471
Oostdijk, 765
Oostdijk, Nelleke, 765
Oozeer, 586
Oozeer, Narmeen Fatimah, 586
Opedal, 842
Opedal, Andreas, 842
Opitz, 817
Opitz, Juri, 817
Orel, 178, 442, 545, 874
Orel, Daniil, 179, 442, 545, 874
Oren, 794
Oren, Gal, 794
Ornik, 274
Ornik, Melkior, 274
Oroojlooy, 192
Oroojlooy, Afshin, 192
Ortega, 302
Ortega, Gerardo, 302
Ortman, 466
Ortman, Tracy, 466
Ortu, 526
Ortu, Francesco, 526
Oruche, 484
Oruche, Roland, 484
Orvieto, 258
Orvieto, Antonio, 258
Osebe, 253
Osebe, Samuel, 253
Osei, 538
Osei, Jemimah, 538
Oseki, 102, 105, 639
Oseki, Yohei, 102, 105, 639
Osmelak, 448
Osmelak, Doreen, 448
Ostarek, 650
Ostarek, Julian, 650
Ostendorf, 112
Ostendorf, Mari, 112
Ostendorff, 862
Ostendorff, Malte, 862
Ostermann, 306, 806, 810, 858
Ostermann, Simon, 306, 806, 810, 858
Osuji, 287
Osuji, Chinonso Cynthia, 287
Otiende, 862
Otiende, Verrah, 862
Otmakhova, 83, 458, 470
Otmakhova, Yulia, 83, 458, 470
Ottoborgo, 288
Ottoborgo, Mattia, 288
OU, 867
Ou, 120, 144, 205, 248, 313, 364, 395, 504, 619, 671
Ou, Jie, 619, 671
Ou, Jiefu, 313
Ou, Justice, 205
Ou, Litu, 120
Ou, Mingdong, 144
Ou, Siqu, 364
Ou, Weixuan, 504
Ou, Xianhao, 395
OU, Yulin, 867
Ou, Zhuoying, 248
Ouahidi, 238
Ouahidi, Yassine El, 238
Ouh, 531
Ouh, Eng Lieh, 531
Oukid, 219
Oukid, Ilyes, 219
Ousidhoum, 545
Ousidhoum, Nedjma, 545
Ouyang, 115, 117, 136, 153, 221, 239, 309, 413, 418, 419, 496, 520, 534, 542, 575, 626, 643, 869
Ouyang, Feiyun, 221
Ouyang, Guanxian, 117
Ouyang, Jessica, 496
Ouyang, Jie, 418, 419
Ouyang, Linke, 239
Ouyang, Penghao, 575
Ouyang, Siqi, 542, 869
Ouyang, Wanli, 153, 309
Ouyang, Xuan, 520

Ouyang, Yangou, 626, 643
Ouyang, Yuxuan, 413
Ouyang, Zhongyu, 115, 136, 534
Ouzzani, 769
Ouzzani, Mourad, 769
Overwiening, 603
Overwiening, Jorin, 603
Owens, 331
Owens, Trevor, 331
Owodunni, 127, 837
Owodunni, Abraham, 837
Owodunni, Abraham Toluwase, 127
Owusu, 878
Owusu, Hillary N, 878
Oyama, 885
Oyama, Momose, 885
Oyamada, 812
Oyamada, Masafumi, 812
Oyarhoseini, 161
Oyarhoseini, Hosna, 161
Oyshi, 484
Oyshi, Uttamasha Anjally, 484
Oza, 337
Oza, Neeva, 337
Ozay, 537, 677
Ozay, Mete, 537, 677
Ozkan, 677
Ozkan, Savas, 677
Ozyurek, 448
Ozyurek, Asli, 448

P, 374
P, Prathosh A, 374
Pacheco, 95, 468, 800, 843, 856
Pacheco, Jesus Horacio, 856
Pacheco, Maria, 843
Pacheco, Maria Leonor, 95, 468, 800
padala, 609
padala, Vaibhavi, 609
Padhi, 781
Padhi, Inkit, 781
Padmanabhan, 829
Padmanabhan, Veena, 829
Padte, 226
Padte, Akshay Kiran, 226
Padó, 263, 548
Padó, Sebastian, 263, 548
Pagliardini, 784
Pagliardini, Matteo, 784
Pagès, 862
Pagès, Quentin, 862
Pahwa, 598
Pahwa, Khushbu, 598
Paige, 533
Paige, Amie, 533
Painuli, 473
Painuli, Abhinav, 473
Pal, 227, 392
Pal, Harsh, 227
Pal, Soumyabrata, 392
Paladugu, 314
Paladugu, Abhijay Sai, 314
Palan, 728
Palan, Mayank, 728
Paldhe, 705
Paldhe, Manas, 705
Paliwal, 866
Paliwal, Bhawna, 866
Palme, 784
Palme, Elia, 784
Palmer, 188, 246
Palmer, Alexis, 188, 246
Palnitkar, 117
Palnitkar, Aadi, 117
Palo, 608
Palo, Flavio Di, 608
Palombini, 174, 566
Palombini, Pietro, 174, 566
Palta, 567
Palta, Shramay, 567
Pan, 94, 111, 122, 128, 134, 201, 225, 232, 274, 278, 286, 289, 312, 319, 327, 328, 329, 356, 369, 374, 398, 401, 402, 405, 412, 419, 422, 439, 445, 451, 459, 469, 481, 520, 538, 546, 555, 599, 603, 606, 607, 617, 627, 636, 641, 647, 672, 676, 677, 687, 691, 717, 722, 730, 753, 787, 793, 801, 808, 829, 832, 843, 846, 857, 859, 903
Pan, Binqiang, 627
Pan, Bo, 607, 687, 846
Pan, Boyuan, 405
Pan, Changhao, 832
Pan, Chengjun, 546
Pan, Chengwei, 520
Pan, Deng, 201
Pan, Gang, 676, 843
Pan, Guanzhong, 232
Pan, Haonan, 439
Pan, Jeff Z., 289, 374, 398, 451, 469, 647, 903
Pan, Jiahui, 356, 412
Pan, Jinhao, 225
Pan, Junjun, 327
Pan, Junshu, 369
Pan, Leiyu, 94, 672
Pan, Leyi, 122

Pan, Liangming, 445, 859
Pan, Lihu, 328
Pan, Lu, 329, 691
Pan, Menghai, 538, 787
Pan, Mingjun, 422
Pan, Qiutong, 278
Pan, Rui, 555
Pan, Shirui, 319, 327, 717
Pan, Sinno Jialin, 312
Pan, Teng, 722
Pan, Tianjun, 128
Pan, Tianyu, 636
Pan, Xiao, 481
Pan, Xiaoman, 111
Pan, Xu, 603
Pan, Xuyan, 286
Pan, Yi, 599
Pan, Yihong, 641, 677
Pan, Yudai, 730
Pan, Yulan, 419
Pan, Yuntong, 134
Pan, Zhe, 753
Pan, Zheng, 459, 793, 801
Pan, Zhuangzhuang, 857
Pan, Zhuoshi, 274, 829, 859
Pan, Zijie, 606, 617
Pan, Ziwen, 402, 808
Pan, Ziyi, 401
Panchenko, 551, 576, 714
Panchenko, Alexander, 551, 576, 714
Panda, 576, 862
Panda, Ashwinee, 576
Panda, Srikant, 862
Pandey, 122, 156, 457, 600, 793
Pandey, Abhishek, 457, 793
Pandey, Aishani, 122
Pandey, Punya Syon, 156, 600
Pandit, 866
Pandit, Shrey, 866
Panferov, 784
Panferov, Andrei, 784
Pang, 154, 169, 208, 355, 374, 407, 417, 444, 504, 535, 606, 617, 621, 664, 702, 753, 833, 852
Pang, Hao, 664
Pang, Honglin, 833
Pang, Liang, 355, 417, 753
Pang, Linsey, 606, 617, 621
Pang, Richard Yuanzhe, 444
Pang, Shengyuan, 504
Pang, Tianyu, 154, 208, 535
Pang, Xianghe, 407
Pang, Xiaotao, 374
Pang, Yan, 702
Pang, Yunhe, 169, 852
Panov, 126, 424
Panov, Aleksandr, 424
Panov, Maxim, 126
Paoletti, 784
Paoletti, Léo, 784
Papadopoulos, 166
Papadopoulos, Triantafillos, 166
Papalampidi, 575
Papalampidi, Pinelopi, 575
Papalexakis, 492, 602
Papalexakis, Evangelos E., 492, 602
Papotti, 216
Papotti, Paolo, 216
Papoudakis, 302
Papoudakis, Argyrios, 302
Pappas, 271, 894
Pappas, Nikolaos, 271, 894
Parchiev, 685
Parchiev, Rauf, 685
Pardeshi, 575, 611
Pardeshi, Vishwa, 575, 611
Parekh, 694
Parekh, Tanmay, 694
Pari, 792
Pari, Ehsan Moradi, 792
Parida, 471
Parida, Shantipriya, 471
Parimi, 243
Parimi, Surya, 243
Parizi, 358
Parizi, Reza M., 358
Park, 106, 120, 121, 133, 142, 151, 161, 179, 189, 190, 197, 259, 266, 267, 269, 274, 279, 280, 295, 351, 395, 426, 444, 450, 476, 533, 553, 557, 564, 592, 593, 595, 601, 612, 616, 717, 741, 765, 775, 799, 805, 811, 824, 833, 834, 842, 858, 885, 890, 898, 901
Park, Bumjin, 890
Park, Chan Young, 295
Park, Changjun, 279
Park, Chanhee, 601
Park, Chanjun, 858
Park, Chanyoung, 142
Park, Cheonbok, 280, 616
Park, Cheoneum, 592
Park, Gyeong-Moon, 189
Park, Hae Won, 901
Park, Haon, 121
Park, Hyejin, 799
Park, Hyunjung, 834

Park, Hyunsin, 106
Park, Jaewoo, 601, 775
Park, Jeongbae, 842
Park, JeongHyeon, 266
Park, Jeonghyun, 557, 811
Park, Jeongjae, 553, 885
Park, Jinhwan, 106
Park, Jong C., 717, 833
Park, Joonhyung, 120
Park, Joonsuk, 280
Park, Jueon, 564
Park, Juneyoung, 444
Park, Junghun, 595
Park, Jungyang, 601
Park, Junha, 476
Park, Junhee, 775
Park, Junho, 450
Park, Junyeong, 197, 259
Park, Kiwoong, 197, 612
Park, Kyeongman, 269, 765
Park, Kyungwon, 593
Park, Leo Hyun, 274
Park, Minjun, 190, 741
Park, Noseong, 805
Park, Sanghee, 824
Park, Sanghyun, 426
Park, Sangkwon, 267
Park, Sangwu, 142
Park, Sejun, 133
Park, Seo Yeon, 898
Park, Seojeong, 592
Park, Seongheon, 151
Park, Seonghyeon, 885
Park, Seonjoon, 601
Park, Seonyeong, 395
Park, Seoyoon, 593
Park, Shinwoo, 799
Park, Sihwan, 120
Park, Sue Hyun, 179
Park, Sungmi, 161
Park, Sungwon, 351
Park, Yangryeol, 592
Park, Yein, 564
Park, Yeji, 533
Park, Yoonah, 133
Parllaku, 477
Parllaku, Fjona, 477
Parmar, 172
Parmar, Mihir, 172
Partha, 100
Partha, Pankaj Chowdhury, 100
Parvez, 135, 145, 549
Parvez, Md Rizwan, 135, 145, 549
Pasad, 529
Pasad, Ankita, 529
Pasareanu, 604
Pasareanu, Corina, 604
Pascale, 173, 493, 863
Pascale, Alessandra, 493, 863
Pascale, Stefano De, 173
Passerini, 784
Passerini, Marco, 784
Passonneau, 479
Passonneau, Rebecca J., 479
Pasti, 842
Pasti, Clemente, 842
Pastor, 765
Pastor, Martial, 765
Patel, 161, 192, 224, 238, 240, 552, 856, 862, 873
patel, 282
Patel, Dhaval, 238
Patel, Hima, 224, 240
Patel, Hitesh Laxmichand, 192, 552, 862, 873
Patel, Karan, 856
Patel, Kshama, 161
patel, Ujjaval, 282
Patella, 835
Patella, Marco, 835
Pathmanathan, 781
Pathmanathan, Pankayaraj, 781
Patil, 85, 108, 166, 439, 444, 616, 896
Patil, Kailash, 85
Patil, Mahavir, 166
Patil, Namratha, 439
Patil, Nilesh, 616
Patil, Sarang Rajendra, 108
Patil, Shishir G, 444
Patil, Vaidehi, 896
Patki, 175
Patki, Rohit, 175
Patras, 465, 807
Patras, Ioannis, 465, 807
Pattanayak, 263
Pattanayak, Ronit, 263
Pattnaik, 770
Pattnaik, Anup, 770
Pattnaik;,
Pattnaik;, Anup,
Pattnayak, 473
Pattnayak, Priyaranjan, 473
Patwardhan, 192, 623
Patwardhan, Manasi, 192, 623
Paul, 211, 247, 905
Paul, Bikash Kumar, 211
Paul, Debjit, 247
Paul, Sudipta, 905

Pavlov, 784, 805
Pavlov, Ivan, 784
Pavlov, Oleg, 805
Pawar, 770
Pawar, Rushikesh, 770
Pawelczyk, 394
Pawelczyk, Martin, 394
Payani, 633
Payani, Ali, 633
Payoungkhamdee, 205
Payoungkhamdee, Patomporn, 205
Pedapati, 215, 536
Pedapati, Tejaswini, 215, 536
Pedram, 269
Pedram, Massoud, 269
Pei, 113, 155, 191, 266, 268, 274, 312, 487, 559, 562, 577, 829, 859
Pei, Changhua, 577
Pei, Dun, 559
Pei, Jiahuan, 113, 191
Pei, Kexin, 266
Pei, Qizhi, 268, 274, 487, 829, 859
Pei, Shichao, 155
Pei, Wenqi, 562
Pei, Zehua, 312
PeiguangLi, 335
PeiguangLi, PeiguangLi, 335
PeixuanChen, 155
PeixuanChen, PeixuanChen, 155
Peled-Cohen, 196
Peled-Cohen, Lotem, 196
Penaloza, 507
Penaloza, Emiliano, 507
PENG, 609, 669
Peng, 94, 95, 100, 103, 109, 112, 139, 141, 161, 165, 166, 179, 198, 230, 236, 243, 262, 273, 280, 292, 295, 299, 307, 337, 341, 343, 344, 348, 349, 354, 359, 368, 418, 421, 428, 442, 444, 455, 457, 465, 467, 476, 483, 498, 501, 518, 527, 545, 547, 548, 581, 597, 599, 601, 604, 637, 654, 659, 669, 675, 676, 698, 707, 713, 773, 782, 791, 794, 796, 799, 811, 820, 824, 832, 850, 859, 906
peng, 110
Peng, Andi, 112
Peng, Baolin, 599, 791
Peng, Bei, 100
Peng, Boci, 675
Peng, Boya, 139
Peng, Cheng, 637
Peng, Dehua, 354
Peng, Guangyue, 659
Peng, Guowei, 343
Peng, Hao, 476, 713, 773, 782, 859
Peng, Haoxuan, 368
Peng, Hongwu, 141
Peng, Houwen, 669
peng, jiachuan, 110
Peng, Jianxiang, 94
Peng, Jie, 348
Peng, Jinghong, 299
Peng, Jingyu, 243
Peng, Khizar QureshiGeoffrey MartinYifan, 198
Peng, Letian, 236, 273, 548
Peng, Li, 179
Peng, Lingyun, 597
Peng, Min, 428
Peng, Minlong, 165, 230, 295, 337
Peng, Nanyun, 262, 467, 483, 547, 850
PENG, QIMING, 669
Peng, Qiyuan, 820
Peng, Rui, 824
Peng, Ruoshui, 707
Peng, Shaowen, 307
Peng, Shiyao, 796
Peng, Siyao, 457, 811
Peng, Tianshuo, 280
Peng, Ting, 501
Peng, Wanli, 581, 782
Peng, Weihan, 518
Peng, Wen-Chih, 103
Peng, Wenjie, 455
Peng, Wenjun, 359
Peng, Xi, 161
PENG, XIANGYU, 609
Peng, Xiao, 344
Peng, Xiaoliang, 444
Peng, Xin, 601, 794, 906
Peng, Xingchen, 341
Peng, Xiong, 418
Peng, Xueqing, 95, 166, 442, 527
Peng, Yibo, 604, 832
Peng, Yifan, 109, 498, 545
Peng, Yihang, 465
Peng, Yiheng, 421
Peng, Yizhou, 455
Peng, Yue, 292
Peng, Yun, 343
Peng, Yuxuan, 601
Peng, Yuyang, 799
Peng, Zhiqi, 698
Peng, Zhiyuan, 349
Peng, Zhongyuan, 654, 676
Perdekamp, 571
Perdekamp, Jost GroSSe, 571

Perez, 565
Perez, Natália da Silva, 565
Pergola, 761
Pergola, Gabriele, 761
Peris, 488, 851, 879
Peris, Charith, 488, 851, 879
Periti, 173
Periti, Francesco, 173
Perlitz, 517
Perlitz, Yotam, 517
Perrault, 194
Perrault, Andrew, 194
Perroy, 110
Perroy, Bastien, 110
Pesiakhovsky, 553
Pesiakhovsky, Yehonatan, 553
Petrangeli, 845
Petrangeli, Stefano, 845
Petruzzi, 567
Petruzzi, Simone, 567
Petullo, 305
Petullo, James, 305
Peyrard, 604
Peyrard, Maxime, 604
Pezeshkpour, 579
Pezeshkpour, Pouya, 579
Pezzelle, 437, 890
Pezzelle, Sandro, 437, 890
Pfister, 446, 678
Pfister, Tomas, 446, 678
Pham, 192, 203, 211, 245, 257, 553, 639, 784
Pham, Chau Minh, 257, 784
Pham, Cuong, 639
Pham, Dzung, 257
Pham, Minh V.T., 192
Pham, Van-Cuong, 553
Pham, Viet, 245
Pham, Viet-Thanh, 203, 211
Phan, 118, 510, 738, 862
Phan, Hai, 510
Phan, Hoang, 118
Phan, Manh-Cuong, 862
Phan, Vu Minh Hieu, 738
Phi, 747
Phi, Hieu, 747
Philippy, 868
Philippy, Fred, 868
Phillips, 331, 639
Phillips, Jordan, 639
Phillips, Mark, 331
Phoncharoen, 728
Phoncharoen, Kasidit, 728
Phoo, 150
Phoo, Cheng Perng, 150
Phukan, 587
Phukan, Anirudh, 587
Phung, 211, 478, 485
Phung, Dinh, 211
Phung, Huy, 485
Phung, Myles, 478
Pi, 477, 642, 736
Pi, Renjie, 642, 736
Pi, Xinyu, 477
Pianzola, 395, 463
Pianzola, Federico, 395, 463
Piao, 112, 244, 426, 514, 756
Piao, Hongming, 112, 244
Piao, Mengyao, 756
Piao, Shengmin, 426
Piao, Yongri, 514
Piatrashyn, 126
Piatrashyn, Dzianis, 126
Piccardi, 293
Piccardi, Massimo, 293
Piech, 891
Piech, Christopher J, 891
Pierri, 234
Pierri, Francesco, 234
Pilehvar, 602, 704
Pilehvar, Mohammad Taher, 602, 704
Pillai, 109, 613
Pillai, Arvind, 109
Pillai, Siddharth, 613
Pimentel, 278, 808, 876
Pimentel, Marco AF, 808
Pimentel, Tiago, 278, 876
Pineda-Castañeda, 624
Pineda-Castañeda, Nataly, 624
Ping, 383, 616, 887
Ping, Bowen, 383
Ping, Qing, 616
Ping, Wei, 887
Piniera, 232
Piniera, Keith, 232
Pino, 243
Pino, Juan, 243
Pinter, 194, 799
Pinter, Yuval, 194, 799
Piotrowski, 483
Piotrowski, Grzegorz, 483
Piper, 463, 476
Piper, Andrew, 463, 476
Piroi, 715
Piroi, Florina, 715
Piryani, 622, 639, 697, 744
Piryani, Bhawna, 622, 639, 697, 744

Pisano, 264
Pisano, Raffaele, 264
Piwowarski, 477, 601
Piwowarski, Benjamin, 477, 601
Pikos, 648
Pikos, Piotr, 648
Plank, 104, 167, 416, 457, 811, 835, 868
Plank, Barbara, 19, 104, 167, 416, 457, 811, 835, 868
Plaza-del-Arco, 529
Plaza-del-Arco, Flor Miriam, 529
Ploeger, 246
Ploeger, Esther, 246
Plum, 868
Plum, Alistair, 868
Po, 259
Po, Yip Tin, 259
Poché, 439
Poché, Antonin, 439
Podasca, 245
Podasca, Andreia, 245
Poelman, 246
Poelman, Wessel, 246
Pohlmann, 568
Pohlmann, Lena, 568
Pohovey, 274
Pohovey, John, 274
Poiroux, 784
Poiroux, Auguste, 784
Polev, 685
Polev, Konstantin, 685
Poli, 243
Poli, Maxime, 243
Politi, 238
Politi, Marcello, 238
Polonuer, 667
Polonuer, Joaquín, 667
Pombal, 705
Pombal, José, 705
Ponkshe, 784
Ponkshe, Kaustubh, 784
Ponwitayarat, 169
Ponwitayarat, Wuttikorn, 169
Ponzetto, 622
Ponzetto, Simone Paolo, 622
Poon, 792
Poon, Hoifung, 792
Poovendran, 599, 879
Poovendran, Radha, 599, 879
Popa, 558
Popa, Lucian, 558
Popovi, 659
Popovi, Nicholas, 659
Porcaro, 873
Porcaro, Ciro, 873
Portet, 581, 604
Portet, Felix Herron;Solange Rossato;Alexandre Allauzen;François,
Portet, François, 604
Post, 871
Post, Matt, 871
Potapov, 640
Potapov, Anatolii, 640
Pothugunta, 806
Pothugunta, Krishna, 806
Potraghloo, 269
Potraghloo, Erfan Baghaei, 269
Pouly, 829
Pouly, Marc, 829
Poznanski, 429
Poznanski, Jake, 429
Pozorski, 763
Pozorski, Pawe, 763
Prabhakaran, 174, 566, 846
Prabhakaran, Vinodkumar, 174, 566
Prabhakaran, Vishnu, 846
Prakash, 356
Prakash, Chandan, 356
Prakrankamanant, 388
Prakrankamanant, Patawee, 388
Pralomram, 356
Pralomram, Surawat, 356
Pramanik, 263
Pramanik, Vishal, 263
Pranida, 589
Pranida, Salsabila Zahirah, 589
Prasad, 268, 293, 694, 700, 890
Prasad, Archiki, 293, 890
Prasad, Shreyas, 268
Prasad, Srivas, 694
Prasad, Yamuna, 700
Prenassi, 154
Prenassi, Gabriel, 154
Prenkaj, 540, 878
Prenkaj, Bardh, 540, 878
Pretschner, 88
Pretschner, Alexander, 88
Preum, 557, 815
Preum, Sarah Masud, 557, 815
Priebe, 226
Priebe, Carey, 226
Priesemann, 114
Priesemann, Viola, 114
Priya, 649
Priya, Priyanshu, 649
Proietti, 871

Proietti, Lorenzo, 871
Pronesti, 81, 493
Pronesti, Massimiliano, 81, 493
Providencia, 81
Providencia, Rui, 81
Pruthi, 812
Pruthi, Danish, 812
Prutko, 330
Prutko, Aleksandr, 330
Pryor, 820
Pryor, Mitch, 820
PU, 798
Pu, 114, 128, 261, 335, 361, 383, 623, 671, 672, 676, 753, 832, 880
Pu, Bin, 261
Pu, Fanyi, 361
Pu, Hongli, 671
Pu, Jiaxing, 753
Pu, Jingwen, 383, 671
PU, Qihang, 798
Pu, Rui, 128, 880
Pu, Ruili, 676
Pu, Tianle, 335
Pu, Xiao, 623, 672
Pu, Xueyi, 114, 832
Pu, Yuan, 261
Puccetti, 236
Puccetti, Giovanni, 236
Pujari, 766
Pujari, Rajkumar, 766
Puranik, 606
Puranik, Vinayak S, 606
Puri;,
Puri;, Joseph Matveyenko; James Liu; John David Parsons; Ryan A. Brown; Alina I. Palimaru; Vipul Gupta; Prateek,
Purohit, 392
Purohit, Kiran, 392
Purschke, 868
Purschke, Christoph, 868
Purushothama, 875
Purushothama, Abhishek, 875
Putra, 862
Putra, Vallerie Alexandra, 862
Putri, 715
Putri, Rifki Afina, 715
Putta, 861
Putta, Akshith Reddy, 861
Pyayt, 614
Pyayt, Anna, 614
Pyo, 223
Pyo, Junseong, 223
Pyun, 787
Pyun, Haesung, 787
Pásztor, 784
Pásztor, Barna, 784

Qawasmeh, 860
Qawasmeh, Baraah, 860
Qazi, 471
Qazi, Ihsan Ayyub, 471
Qendro, 549
Qendro, Lorena, 549
Qi, 145, 251, 271, 272, 286, 314, 325, 327, 335, 352, 375, 394, 401, 413, 415, 421, 423, 433, 437, 444, 470, 514, 533, 560, 564, 572, 583, 632, 637, 696, 750, 772, 782, 801, 836, 877, 879, 884, 893, 894
qi, 357
Qi, Biqing, 413, 423, 437, 750, 877, 893
Qi, Guilin, 145
Qi, Heli, 572, 772
Qi, Huachen, 696
Qi, Huiyan, 352
Qi, Ji, 632
Qi, Jiayin, 514
Qi, Junbo, 884
Qi, Peng, 470, 879
Qi, Qi, 251, 327, 444, 836
Qi, Qingfeng, 801
Qi, Qingguo, 750
Qi, Qiuyi, 352, 564
Qi, Rui, 583
Qi, Ruihua, 325
qi, shuren, 357
Qi, Siyuan, 271
Qi, Tao, 375, 401, 421, 533
Qi, Wei, 335
Qi, Wenqi, 394
Qi, Xiaojuan, 433
Qi, Yang, 314
Qi, Yirui, 272
Qi, Yuan, 637
Qi, Yuemeng, 560
Qi, Yunjia, 782
Qi, Yupeng, 415
Qi, Zehan, 286
Qi, Zheng, 271, 894
Qi, Zhenting, 394
Qian, 97, 118, 127, 130, 135, 151, 166, 239, 260, 278, 292, 308, 311, 312, 321, 339, 362, 374, 389, 419, 440, 442, 444, 482, 494, 567, 573, 587, 596, 603, 617, 626, 675, 706, 728, 745, 746, 775, 778, 808, 845, 853
Qian, Angela, 118

Qian, Chen, 339, 419, 596, 617
Qian, Cheng, 151, 292, 440, 573, 603, 626, 808
Qian, Chengxuan, 775
Qian, Fulan, 746
Qian, Haofu, 745
Qian, Hong, 311
Qian, Hongjin, 135, 482, 728, 853
Qian, Huijun, 845
Qian, Jiachen, 362
Qian, Jiaru, 97, 260
Qian, Lijun, 567
Qian, Lingfei, 166, 442
Qian, Lingyun, 97
Qian, Livia, 778
Qian, Pengfei, 389
Qian, Qi, 321
Qian, Qian, 127
Qian, Rui, 312
Qian, Siyi, 239
Qian, Tieyun, 675, 706
Qian, Weining, 278
Qian, Xiaowei, 587
Qian, Xinying, 374
Qian, Yanmin, 308
Qian, Yining, 130
Qian, Yundi, 444
Qian, Zhuzhong, 494
Qiang, 227, 455, 456, 673
Qiang, Chunyu, 455, 456
Qiang, Evan, 227
Qiang, Jipeng, 673
Qiao, 100, 107, 120, 211, 303, 326, 334, 375, 412, 427, 451, 462, 591, 665, 702, 747, 761, 842
Qiao, Bo, 211
Qiao, Dan, 427, 761
Qiao, Fengcai, 747
Qiao, Kyle, 107
Qiao, Linbo, 665
Qiao, Peng, 702
Qiao, Ruizhi, 842
Qiao, Shuofei, 334, 451
Qiao, Xuanjia, 462
Qiao, Yi, 412
Qiao, Yu, 100, 303, 591
Qiao, Zile, 120
Qiao, Ziyue, 326, 375
QIN, 161
Qin, 99, 103, 107, 142, 163, 186, 251, 273, 328, 329, 341, 345, 346, 352, 353, 360, 375, 378, 381, 395, 396, 404, 417, 424, 433, 460, 528, 540, 555, 603, 608, 626, 630, 643, 659, 691, 692, 693, 697, 708, 718, 720, 725, 728, 729, 730, 739, 744, 750, 755, 761, 762, 765, 796, 867
qin, 357
Qin, Bing, 328, 345, 353, 360, 378, 381, 395, 626, 643, 697, 739, 755, 761, 762
qin, bing, 357
Qin, Bowen, 186
Qin, Bruce, 555
Qin, Chengwei, 424
Qin, Chonghan, 99
Qin, Chuan, 163, 433, 718, 730
Qin, Cong, 329, 417, 691
QIN, Feng, 161
Qin, Guangming, 725
Qin, Haiming, 659
Qin, Jianbin, 396
Qin, Jiayu, 708
Qin, Kai, 142
Qin, Kaitong, 603
Qin, Kang, 107
Qin, Ke, 608
Qin, Lianhui, 273
Qin, Libo, 353, 360, 404, 693
Qin, Melissa Xiaohui, 251
Qin, Peijia, 103
Qin, Ruiyang, 341
Qin, Ruize, 765
Qin, Shang, 460
Qin, Shengchao, 329
Qin, Shengqian, 528
Qin, Shitong, 692
Qin, Sijing, 867
Qin, Xiaolin, 540
Qin, Xiaoting, 720
Qin, Yiwei, 750
Qin, Yong, 352, 744
Qin, Yu, 643
Qin, Yulu, 796
Qin, Yunxiao, 346
Qin, Zengchang, 728
Qin, Zhanyue, 630
Qin, Zhenxin, 341
Qin, Zhiheng, 729
Qin, Zichen, 375
Qing, 473, 542, 573
Qing, Li, 473, 542, 573
Qingyun, 181
Qingyun, Zou, 181
QishengSu, 372
QishengSu, QishengSu, 372
Qiu, 92, 110, 119, 120, 130, 132, 156, 166, 204, 228, 242, 271, 347, 391, 397, 413, 424, 432, 440, 446, 457, 465, 507, 520, 533,

546, 564, 583, 640, 646, 665, 672, 685, 708, 760, 773, 838, 848, 850, 852, 863, 880
Qiu, Delai, 391, 397, 760
Qiu, Haibo, 507
Qiu, Han, 92, 132, 413, 432, 520, 583
Qiu, Hanjia, 640
Qiu, Haonan, 564
Qiu, Jiahao, 440
Qiu, Jianing, 110
Qiu, Junyuan, 533
Qiu, Lanlan, 672
Qiu, Lin, 271, 863
Qiu, Linlu, 242
Qiu, Lirong, 880
Qiu, Meikang, 166
Qiu, Minghui, 119
Qiu, Peijie, 204
Qiu, Qibo, 685
Qiu, Rihong, 850
Qiu, Ruihong, 564
Qiu, Ruizhong, 446, 457
Qiu, Shuang, 424
Qiu, Weiyun, 465
Qiu, Wentao, 665
Qiu, Xihe, 852
Qiu, Xipeng, 130, 204, 228, 546, 708, 838
Qiu, Yuan, 347
Qiu, Yue, 773
Qiu, Yuhan, 120
Qiu, Yukui, 646
Qiu, Zeju, 156
Qiu, Zexuan, 848
Qorib, 898
Qorib, Muhammad Reza, 898
Qu, 159, 168, 203, 211, 215, 236, 239, 249, 289, 392, 418, 550, 605, 640, 654
Qu, Bohao, 640
Qu, Changle, 289
Qu, Jiayi, 159
Qu, Lizhen, 203, 211, 249, 550
Qu, Renyi, 236
Qu, Songxin, 418
Qu, Tingyu, 215
Qu, Xiaoyang, 392
Qu, Xiaoye, 654
Qu, Yuan, 239
Qu, Zhan, 168
Qu, Zhi, 605
Qu;,
Qu;, Rui,
Quan, 364, 695, 879
Quan, Tho, 364
Quan, Xiaojun, 695
Quan, Yihao, 879
Quaremba, 246
Quaremba, Gerrit, 246
Quek, 731
Quek, Tony Q.S., 731
Quensel, 102
Quensel, Carlotta, 102
Quirke, 586
Quirke, Philip, 586

R, 492
R, Govindarajan, 492
Raballo, 648, 658
Raballo, Andrea, 648, 658
Raban, 288
Raban, Ofek, 288
Rabbani, 560, 574, 871
Rabbani, Parisa, 574
Rabbani, Saad, 871
Rabbani, Tahseen, 560
Rabinovich, 660
Rabinovich, Ella, 660
Rabinowitz, 781
Rabinowitz, David, 781
Radevski, 275
Radevski, Gorjan, 275
Radhakrishna, 239
Radhakrishna, Arjun, 239
Radharapu, 238
Radharapu, Bhaktipriya, 238
Radwan, 419
Radwan, Ahmed Y., 419
Ragazzi, 468, 835
Ragazzi, Luca, 468, 835
Raghavendra, 492, 530
Raghavendra, Mohit, 530
Raghavendra, Prakash, 492
Raha, 808
Raha, Tathagata, 808
Rahamim, 230
Rahamim, Adir, 230
Rahimi, 737
Rahimi, Zahra, 737
Rahimirad, 118
Rahimirad, Shahab, 118
Rahimzadeh, 704
Rahimzadeh, Vahid, 704
RAHMAN, 371
Rahman, 260, 567, 664, 822
RAHMAN, AKMMAHBUBUR, 371
Rahman, Akond Ashfaque Ur, 260
Rahman, Fuad, 664

Rahman, Mafizur, 567
Rahman, Mizanur, 822
Rahmani, 847
Rahmani, Hossein A., 847
Rahmati, 226, 539
Rahmati, Elnaz, 226, 539
Rai, 192, 814
Rai, Arushi, 192
Rai, Shahid Iqbal, 814
Raimundo, 85
Raimundo, Rodolfo, 85
Rair, 804
Rair, Nisrine, 804
Raj, 86, 225, 250, 622, 637
Raj, Arnav, 250
Raj, Chahat, 86, 225
Raj, Devankar, 622
Raj, Subham, 637
Rajabzadeh, 329
Rajabzadeh, Hossein, 329
Rajan, 808
Rajan, Ronnie, 808
Rajani, 284
Rajani, Nazneen, 284
Rajaram, 492
Rajaram, Arjun, 492
Rajasekaran, 547
Rajasekaran, Bhavya, 547
Rajendiran, 305
Rajendiran, Vikram Nelvoy, 305
Rajeswar, 441
Rajeswar, Sai, 441
Raji, 869
Raji, Inioluwa Deborah, 869
Rajman, 784
Rajman, Martin, 784
Rajmohan, 211, 349, 599, 684, 720, 791
Rajmohan, Saravan, 211, 349, 599, 684, 720, 791
Rajtmajer, 294
Rajtmajer, Sarah, 294
Raju, 472, 655
Raju, Anandharaju Durai, 655
Raju, Rohan, 472
Rakesh, 538
Rakesh, Vineeth, 538
Rakshit, 472
Rakshit, Sushrita, 472
Ram, 705
Ram, Vishal Sankar, 705
Ramachandran, 492, 538
Ramachandran, Arun, 492
Ramachandran, Rohan, 538
Ramakrishna, 128, 851
Ramakrishna, Anil, 128, 851
Ramakrishnan, 188, 458, 473, 637
Ramakrishnan, Aashish Anantha, 458
Ramakrishnan, Ganesh, 188, 473
Ramakrishnan, Naren, 637
Ramamoorthy, 246
Ramamoorthy, Sathyanarayanan, 246
Raman, 458
Raman, Vyoma, 458
Ramanan, 116
Ramanan, Deva, 116
Ramanathan, 299, 881
Ramanathan, Murali Krishna, 299, 881
Ramasubramanian, 599
Ramasubramanian, Bhaskar, 599
Rambow, 533
Rambow, Owen, 533
Ramirez, 276, 862
Ramirez, Guillem, 276
Ramirez, Ingrid Gabriela Franco, 862
Rammouz, 477, 524
Rammouz, Veronica, 477, 524
Ramu, 122, 270
Ramu, Pritika, 122, 270
ramzi, 883
ramzi, elias, 883
Ran, 278, 636, 794
Ran, Dekang, 636
Ran, Jiaxin, 278
Ran, Maolin, 794
Ran-Milo, 664, 666
Ran-Milo, Yuval, 664, 666
Rana, 167, 826
Rana, Simmi, 826
Rana, Souvik, 167
Ranaldi, 193, 269, 426, 727
Ranaldi, Federico, 193, 269
Ranaldi, Leonardo, 193, 269, 426, 727
Ranasinghe, 868
Ranasinghe, Tharindu, 868
Ranchin, 784
Ranchin, Nathan, 784
Randhawa, 462
Randhawa, Shan, 462
Rando, 784
Rando, Javier, 784
Rangwala, 452, 797, 880
Rangwala, Huzefa, 452, 797, 880
Ranjan, 253
Ranjan, Mukul, 253
Ranjbar, 556
Ranjbar, Ali, 556
Ranjit, 478

Ranjit, Jaspreet, 478
Rankel, 567
Rankel, Peter, 567
Rao, 213, 313, 581, 673, 825, 852, 890, 895
Rao, Delip, 825
Rao, Guozheng, 313
Rao, Hanshu, 213
Rao, Junreng, 673
Rao, Sukruth, 890
Rao, Yanghui, 852, 895
Rao, Zixin, 581
Raquib, 371
Raquib, Faisal Hossain, 371
Rashid, 556, 807
Rashid, Farzana, 807
Rashid, Syed Md Mukit, 556
Rathod, 428
Rathod, Ameya, 428
Rathore, 94
Rathore, Vipul Kumar, 94
Ratnakar, 611
Ratnakar, Shivam, 611
Ratnam, 462
Ratnam, Anjali, 462
Ratnaparkhi, 613
Ratnaparkhi, Adwait, 613
Rattanavipanon, 205
Rattanavipanon, Norrathep, 205
Rauby, 258
Rauby, Brice, 258
Rauniyar, 471
Rauniyar, Kritesh, 471
Raval, 621
Raval, Jay, 621
Ravan, 218
Ravan, Seyedehmoniba, 218
Raveh, 82
Raveh, Barak, 82
Ravenda, 648, 658
Ravenda, Federico, 648, 658
Ravenda, Sofia Irene, 658
Ravi, 131, 136, 552, 819
Ravi, Sujith, 131, 136, 552, 819
Ray, 815, 845
Ray, Rohan, 815
Ray, Sushant Kumar, 845
Rayfield, 610
Rayfield, James T, 610
Raza, 145, 419, 462
Raza, Agha Ali, 462
Raza, Mohammad, 145
Raza, Shaina, 419
Razi, 204
Razi, Abolfazl, 204
Razumenko, 818
Razumenko, Itay, 818
Razzak, 854
Razzak, Imran, 854
Re, 243
Re, Francesco Ignazio, 243
Rechkemmer, 246
Rechkemmer, Amy, 246
Recknor, 154
Recknor, Hannah, 154
Reddy, 148, 197, 505, 799, 871
Reddy, Chandan K., 148, 505
Reddy, Saket, 197
Reddy, Varshini, 799, 871
Redmond, 81, 591
Redmond, Amélie, 591
Redmond, Oisín, 81
Rege, 480
Rege, Aniket, 480
Regev, 876
Regev, Tamar, 876
Rehlinger, 868
Rehlinger, Nils, 868
Rei, 705, 847
Rei, Marek, 847
Rei, Ricardo, 705
Rei_Taniguchi,
Rei_Taniguchi, Rei_Taniguchi,
Reichart, 196, 286, 553
Reichart, Roi, 196, 286, 553
Reif, 603, 887
Reif, Yuval, 603, 887
Reiter, 735
Reiter, Ziyue Liu; Nils,
Reizinger, 207
Reizinger, Patrik, 207
Ren, 94, 119, 125, 128, 155, 166, 168, 184, 209, 239, 251, 255, 301, 313, 327, 334, 336, 339, 349, 360, 369, 373, 383, 386, 388, 389, 391, 394, 400, 404, 409, 414, 421, 423, 425, 432, 433, 471, 482, 497, 501, 516, 540, 548, 560, 575, 581, 583, 588, 590, 618, 629, 635, 645, 666, 681, 720, 738, 739, 742, 746, 748, 773, 786, 795, 798, 843, 854, 855, 879, 884
Ren, Baochang, 334
Ren, Bin, 666
Ren, Can, 209, 501
Ren, Chuan, 155
Ren, Feiyang, 125, 497
Ren, Fuji, 373, 388
Ren, Guanghui, 404

Ren, Haojie, 433
Ren, Haoyu, 360, 739
Ren, He, 391
Ren, Hong, 618
Ren, Houxing, 339, 432, 738
Ren, Huimin, 421
Ren, Jiahan, 645
Ren, Jie, 336, 720
Ren, Jingyi, 482, 855
Ren, Juan, 471
Ren, Junyang, 798
Ren, Kui, 184, 369, 742
Ren, Lei, 540
Ren, Lejian, 414
Ren, Liliang, 168
Ren, Lingfei, 854
Ren, Lu, 414
Ren, Pengfei, 251
Ren, Pengjie, 516
Ren, Pengju, 400, 409
Ren, Qibing, 879
Ren, Qingyu, 128, 581
Ren, Ruiyang, 746
Ren, Shenyuan, 394
Ren, Shuai, 583
Ren, Sijian, 255
Ren, Tao, 386
Ren, Wei, 423
Ren, Wenqi, 635
Ren, Xiancong, 884
Ren, Xiang, 575, 843
Ren, Xiaoqi, 786
Ren, Xiaoxiao, 560
Ren, Xiaoxue, 301, 389
Ren, Xinrui, 383
Ren, Xuanbai, 548
Ren, Xubin, 588
Ren, Yang, 166
Ren, Yinglian, 629
Ren, Yuanyi, 773
Ren, Yubing, 327
Ren, Yupei, 635
Ren, Yuqi, 94, 119, 386, 590, 681
Ren, Zhaochun, 349, 425, 795
Ren, Zhaoxing, 748
Ren, Zhifei, 239
Ren, Zhonghao, 313
Renduchintala, 344
Renduchintala, H S V N S Kowndinya, 344
Resnik, Philip, 20
Retkowski, 779
Retkowski, Fabian, 779
Revanasiddappa, 869
Revanasiddappa, Yatish Hosmane, 869
Revankar, 150
Revankar, Shreelekha, 150
Rezaei, 807
Rezaei, MohammadHossein, 807
Rezaeian, 459
Rezaeian, Amir H., 459
Rezapour, 210
Rezapour, Rezvaneh, 210
Rhee, 510
Rhee, Wonjong, 510
Riaz, 526
Riaz, Irbaz Bin, 526
Ribeiro, 99, 302, 542, 803
Ribeiro, Leonardo, 542
Ribeiro, Leonardo F. R., 99, 803
Ribeiro, Manoel, 302
Ricci, 555
Ricci, Kathryn, 555
Rice, 246, 862
Rice, Benjamin, 862
Rice, Enora, 246
Riedhammer, 812
Riedhammer, Korbinian, 812
Rifat, 100
Rifat, S. M Golam, 100
Rikters, 452
Rikters, Matss, 452
Riley, 342
Riley, Parker, 342
Rim, 182
Rim, Daniel, 182
Ringwald, 207
Ringwald, Whitney, 207
Rios, 477, 596
Rios, Anthony, 477
Rios, Walcy, 596
Rios;,
Rios;, Anthony,
Riquet, 713
Riquet, Damien, 713
Rish, 258
Rish, Irina, 258
Ritter, 464
Ritter, Alan, 464
Riviere, 834
Riviere, Pamela, 834
Rivière, 108
Rivière, Pamela, 108
Rizzi, 238
Rizzi, Giovanni, 238
Ro, 498
Ro, Yeonju, 498

Robinson, 553
Robinson, Jo, 553
Rocci, 624
Rocci, Andrea, 624
Rocha, 154
Rocha, Leonardo, 154
Rocktäschel, 376
Rocktäschel, Tim, 376
Rodjananant, 388
Rodjananant, Narabodee, 388
Rodkin, 178
Rodkin, Ivan, 179
Rodriguez, 513, 560
Rodriguez, Paula, 560
Rodriguez, Violeta J., 513
Roemer, 517
Roemer, Niklas, 517
Rogers, 438
Rogers, Anna, 438
Rogov, 471, 576
Rogov, Oleg, 471, 576
Roh, 123, 275, 885
Roh, Daeyoung, 123
Roh, Taeyun, 275
Roh, Youngji, 885
Rohit, 238
Rohit, Jino, 238
Rohrbach, 98
Rohrbach, Anna, 98
Rohrbach, Marcus, 98
Rohweder, 216
Rohweder, Jonas, 216
Roit, 821
Roit, Paul, 821
Romanou, 784
Romanou, Angelika, 784
Romary, 728
Romary, Laurent, 728
Romeo, 598
Romeo, Salvatore, 598
Romero, 118
Romero, Lucia, 118
Romo, 82, 609
Romo, Spencer, 82, 609
Rong, 434, 470, 499, 669
Rong, Ethan Z., 470
Rong, Wenge, 669
Rong, Yankai, 499
Rong, Yu, 434
Rongata, 161
Rongata, Vinit, 161
Rosa, 770
Rosa, Emanuele Di, 770
Rosati, 642
Rosati, Domenic, 642
Rose, 189
Rose, Carolyn, 189
Rosen, 104, 267
Rosen, Gail, 104, 267
Rosenberg, 781
Rosenberg, David S, 781
Rosenthal, 558, 784
Rosenthal, David, 784
Rosenthal, Sara, 558
Rosero, 217
Rosero, Karen, 217
Ross, 112, 220
Ross, Alexis, 112
Ross, Drew, 220
Ross, Jillian, 112
Rossi, 483, 507, 584, 598, 901
Rossi, Ryan, 483
Rossi, Ryan A., 507, 584, 598, 901
Rossolini, 778
Rossolini, Giulio, 778
Rosá, 882
Rosá, Aiala, 882
Roth, 136, 181, 192, 200, 452, 457, 459, 477, 552, 776, 819, 828
Roth, Benjamin, 200, 457
Roth, Dan, 136, 181, 192, 452, 459, 477, 552, 776, 819, 828
Roth;,
Roth;, Dan,
Rothermel, 98
Rothermel, Mark, 98
Rotman, 551
Rotman, Guy, 551
Rowe, 552
Rowe, Matthew, 552
Roy, 82, 556, 621, 866
Roy, Aryan, 866
Roy, Priyashree, 82
Roy, Soumyadeep, 556
Roy, Souvik, 621
Rozanov, 233
Rozanov, Nikolai, 233
Rozovskaya, 477
Rozovskaya, Alla, 477
Ru, 471
Ru, Yanchi, 471
Ruan, 124, 167, 253, 254, 327, 340, 345, 463, 546, 568, 599, 611, 632, 636, 684, 713, 739, 768, 787, 851, 877
Ruan, Chi, 599
Ruan, Chuanwei, 611

Ruan, Dezhao, 877
Ruan, Jingqing, 340, 345, 636
Ruan, Junhao, 739
Ruan, Kangrui, 787
Ruan, Mengzhe, 713
Ruan, Qian, 546
Ruan, Tong, 254, 327, 463, 568, 632, 768
Ruan, Weitong, 253, 851
Ruan, Wenyu, 124
Ruan, Zhiwen, 167, 684
Ruas, 174, 551, 620, 876
Ruas, Terry, 174, 551, 620, 876
Rubashevskii, 126
Rubashevskii, Aleksandr, 126
Rubinstein, 156
Rubinstein, Alexander, 156
Rubio-Fernandez, 448
Rubio-Fernandez, Paula, 448
Ruchkin, 572
Ruchkin, Ivan, 572
Ruckdeschel, 862
Ruckdeschel, Mattes, 862
Ruder, 139
Ruder, Sebastian, 139
Rudinger, 225, 473, 547, 567, 906
Rudinger, Rachel, 225, 473, 547, 567, 906
Rudman, 894
Rudman, William, 894
Rudolph, 150
Rudolph, Eric, 150
Ruffinelli, 862
Ruffinelli, Daniel, 862
RUI, 471
RUI, SU, 471
Ruichen, 363
Ruichen, Zhang, 363
Rukhovich, 438
Rukhovich, Danila, 438
Rulli, 567
Rulli, Maximo, 567
Ruppik, 89
Ruppik, Benjamin, 89
Russell, 233, 257
Russell, Jenna, 233, 257
Russo, 234, 512, 838
Russo, Alessandra, 512
Russo, Giuseppe, 234, 838
Rust, 243
Rust, Phillip, 243
Rutkowski, 321, 883
Rutkowski, Leszek, 321, 883
Rvanova, 491
Rvanova, Lyudmila, 491
Ryan, 247
Ryan, Michael J., 247
Rykov, 640
Rykov, Elisei, 640
Ryu, 175, 228, 533, 803
Ryu, Jeongwoo, 533
Ryu, Koki, 228
Ryu, Sangwon, 175, 803
Rücklé, 864
Rücklé, Andreas, 864
RückstieSS, 537
RückstieSS, Thomas, 537

Saadatinia, 226
Saadatinia, Mehrshad, 226
Saadi, 808
Saadi, Nada, 808
Saadie, 860
Saadie, Razan, 860
Sabale, 221
Sabale, Yash, 221
Sabharwal, 452
Sabharwal, Ashish, 452
Sabolec, 784
Sabolec, Vinko, 784
SABRI, 328
SABRI, AZNUL QALID MD, 328
Sachan, 126, 156, 233, 600, 777, 835
Sachan, Mrinmaya, 126, 156, 233, 600, 777, 835
Sachdeva, 833
Sachdeva, Advik, 833
Sadallah, 555
Sadallah, Abdelrahman, 555
Sadeghi, 247
Sadeghi, Pouya, 247
Sadhu, 480
Sadhu, Arka, 480
Sadiekh, 836
Sadiekh, Sabrina, 836
Sadqi, 212
Sadqi, Yousra, 212
SadraeiJavaheri, 341
SadraeiJavaheri, MohammadAli, 341
Saeidi, 633
Saeidi, Amir, 633
Saengsukhiran, 388
Saengsukhiran, Teerapol, 388
Saeteng, 728
Saeteng, Treephop, 728
Safarzadeh, 192, 776
Safarzadeh, Mohammadtaher, 192, 776
Safder, 527
Safder, Muhammad Usman, 527

Sagar, 444
Sagar, Adithya, 444
Sagot, 282, 862
Sagot, Benoît, 282, 862
Sah, 569
Sah, Kshitij, 569
Saha, 163, 202, 225, 513, 622, 637, 662
Saha, Anisha, 225
Saha, Koustuv, 513
Saha, Sourajit, 163
Saha, Sriparna, 202, 622, 637, 662
Sahai, 128
Sahai, Sattvik, 128
Sahil, 471
Sahil, P Sam, 471
Sahitaj, 84
Sahitaj, Ariana, 84
Sahitaj, Premtim, 84
Sahnan, 442, 881
Sahnan, Dhruv, 442, 881
Sahoo, 574
Sahoo, Priyam, 574
Sahu, 853
Sahu, Sambit, 853
Sai, 784
Sai, Ananya B., 784
Saidutta, 245, 905
Saidutta, Yashas Malur, 245, 905
Saikia, 667
Saikia, Swaagat Bikash, 667
Saim, 809
Saim, Mohammad, 809
Saint-James, 243
Saint-James, Charles-Eric, 243
Saiyed, 836
Saiyed, Ahson, 836
Saji, 473
Saji, Chriss Philip, 473
Sajjad, 165, 642, 837
Sajjad, Hassan, 165, 642, 837
Sakai, 99, 215, 445, 481, 492, 832
Sakai, Akira, 445
Sakai, Yusuke, 99, 215, 481, 492, 832
Saklad, 805
Saklad, Ryan, 805
Sakshi, 140
Sakshi, S, 140
Sakuma, 346, 489, 565
Sakuma, Jun, 346, 489, 565
Sakunkoo, 523
Sakunkoo, Annabella Sakunkoo; Jonathan,
Saladi, 731
Saladi, Anoop, 731
Saladi;,
Saladi;, Anoop,
Salam, 471
Salam, Hanan, 471
Salamanca, 826
Salamanca, Alejandro R., 826
Saleh, 212, 320
Saleh, Fahimeh, 320
Saleh, Khaled, 212
Salehi, 856
Salehi, Soheil, 856
Salekin, 82, 609
Salekin, Md Sirajus, 82, 609
Salfity, 820
Salfity, Jonathan, 820
Saligrama, 99
Saligrama, Venkatesh, 99
Salim, 230, 295, 443, 597, 840, 854, 862
Salim, Flora, 854
Salim, Flora D., 230, 295, 443, 597
Salim, Luis Frentzen, 840, 862
Sallinger, 820
Sallinger, Emanuel, 820
Salmani-Zarchi, 737
Salmani-Zarchi, Mohammad Mahdi, 737
Saluja, 188
Saluja, Rohit, 188
Salvador, 260
Salvador, John, 260
Samadi, 265
Samadi, Mehrzad, 265
Samanta, 146
Samanta, Kunal, 146
Samaras, 533
Samaras, Dimitris, 533
Sambara, 261
Sambara, Sraavya, 261
Samek, 858
Samek, Wojciech, 858
Samih, 370, 690, 864
Samih, Younes, 370, 690, 864
Samoylov, 510
Samoylov, Elisaveta, 510
Samuel, 862
Samuel, Saron, 862
Samuel-Ipaye, 538
Samuel-Ipaye, Oluwatumininu, 538
Samway, 600
Samway, Keenan, 600
Sanaie, 764
Sanaie, Nicholas, 764
Sanatizadeh, 463
Sanatizadeh, Aida, 463

Sanchez, 291
Sanchez, Karen, 291
Sanders, 126, 154
Sanders, Kate, 154
Sanders, Nicholas, 126
Sandoval-Castaneda, 863
Sandoval-Castaneda, Marcelo, 863
Sang, 124, 353, 363, 396, 405, 716, 723, 760, 861, 875
Sang, Dinh Viet, 405
Sang, Hejian, 124
Sang, Jitao, 363, 396, 723, 760
Sang, Lei, 353
Sang, Yisi, 861, 875
Sang, Zilin, 716
Sanjabi, 356
Sanjabi, Maziar, 356
Sankar, 444
Sankar, Chinnadhurai, 444
Sankaran, 745
Sankaran, Aditya Narayan, 745
Sanner, 125, 568, 764
Sanner, Scott, 125, 568, 764
sanniboina, 262
sanniboina, Saikrishna, 262
Santanam, 464
Santanam, Tejas, 464
Santis, 200
Santis, Antonio De, 200
Santos, 588
Santos, Janderson, 588
Santos-Deonizio, 232
Santos-Deonizio, Liliana, 232
Santos-Rios, 187
Santos-Rios, Roi, 187
Santos-Rodriguez, 428
Santos-Rodriguez, Raul, 428
Santra, 649, 823, 866
Santra, Bikash, 649
Santra, Bishal, 866
Santra, Payel, 823
Sanyal, 671
Sanyal, Debarshi Kumar, 671
Sanz-Guerrero, 286
Sanz-Guerrero, Mario, 286
Sap, 466, 476
Sap, Maarten, 466, 476
Saparina, 176
Saparina, Irina, 176
Saparkhan, 145
Saparkhan, Raman, 145
Saracino, 778
Saracino, Andrea, 778
Sarapat, 582
Sarapat, Pakhapoom, 582
Sarawagi, 221
Sarawagi, Sunita, 221
Sarawgi, 246
Sarawgi, Anjali, 246
Sari, 715
Sari, Yunita, 715
Sarid, 284
Sarid, Amir, 284
Sarikaya, 802
Sarikaya, Ruhi, 802
Sarkar, 498
Sarkar, Gaurav, 498
Sarrafzadeh, 561
Sarrafzadeh, Bahareh, 561
Sarti, 439
Sarti, Gabriele, 439
Sartori, 232
Sartori, Elisa, 232
Sarwar, 320, 896
Sarwar, Nobin, 896
Sarwar, Tabinda, 320
Sasaki, 779
Sasaki, Yuichi, 779
Sastre, 882
Sastre, Ignacio, 882
Satam, 856
Satam, Pratik, 856
Satheesh, 492, 785
Satheesh, Anirudh, 492
Satheesh, Shalaka, 785
Sather, 279
Sather, Eric, 279
Sathyanathan, 812
Sathyanathan, Rose, 812
Sato, 877
Sato, Fumiaki, 877
Sattigeri, 771, 781
Sattigeri, Prasanna, 771, 781
Saubion, 346
Saubion, Frédéric, 346
Sauser, 784
Sauser, Mathieu, 784
Sauter, 713
Sauter, Adrian, 713
Savary, 712
Savary, Agata, 712
Savchenko, 685, 714, 722, 761
Savchenko, Andrey, 685, 714, 761
Savchenko, Danila Lapokin; Andrey,
Savchenko, Maxim, 685
Savelka, 515

Savelka, Jaromir, 515
Sawaf, 841
Sawaf, Hassan, 841
Saxena, 238, 240, 469, 587
Saxena, Apoorv, 587
Saxena, Eshika, 238
Saxena, Ojasva, 469
Saxena, Vaibhav, 240
Saydaliev, 784, 862
Saydaliev, Jakhongir, 784, 862
Sayed, 649
Sayed, Marwan, 649
Sayfiddinov, 784
Sayfiddinov, Mukhammadali, 784
Scardapane, 567
Scardapane, Simone, 567
Scarlatos, 832
Scarlatos, Alexander, 832
Schamschurko, 881
Schamschurko, André, 881
Schiele, 291
Schiele, Bernt, 291
Schijndel, 175
Schijndel, Lucas Y. Li; Zander Lynch; Marten Van,
Schlag, 784
Schlag, Imanol, 784
Schlegel, 179, 856
Schlegel, Viktor, 179, 856
Schluter, 580
Schluter, Natalie, 580
Schmaltz, 832
Schmaltz, Allen, 832
Schmid, 243, 497
Schmid, Helmut, 243
Schmid, Stefan, 497
Schmidt, 107, 218, 799, 876
Schmidt, Craig W, 799
Schmidt, Elena, 107
Schmidt, Fabian, 218
Schmidt, Finn, 876
Schmitt, 84, 826, 858
Schmitt, Vera, 84, 826, 858
Schneider, 599, 784
Schneider, Benjamin, 599
Schneider, Marian, 784
Schoenegger, 200
Schoenegger, Loris, 200
Schopf, 850
Schopf, Tim, 850
Schramowski, 298, 806
Schramowski, Patrick, 298, 806
Schreieder, 850
Schreieder, Tobias, 850
Schuetze, 191, 289, 474, 583, 643, 663, 812, 826
Schuetze, Hinrich, 191, 289, 474, 583, 643, 663, 812, 826
Schuhmacher, 817
Schuhmacher, Elias, 817
Schuller, 780
Schuller, Björn, 780
Schulthess, 784
Schulthess, Thomas C., 784
Schumacher, 477, 524
Schumacher, Dan, 477, 524
Schunn, 449
Schunn, Christian, 449
Schuppli, 784
Schuppli, Stefano, 784
Schuster, 663, 796
Schuster, Jakob, 663
Schuster, Sebastian, 796
Schwandt, 218
Schwandt, Silke, 218
Schwartz, 207, 529, 543, 544, 603, 887
Schwartz, Brian, 529
Schwartz, Eli, 543
Schwartz, H. Andrew, 207, 529
Schwartz, Idan, 544
Schwartz, Roy, 603, 887
Schölkopf, 156, 189, 600
Schölkopf, Bernhard, 156, 189, 600
Schütze, 243, 766, 821
Schütze, Hinrich, 243, 766, 821
Scialanga, 784
Scialanga, Marco, 784
Seddah, 282
Seddah, Djamé, 282
Sedoc, 236
Sedoc, João, 236
See, 209, 243
See, Simon, 209, 243
Seeberger, 812
Seeberger, Philipp, 812
Seegmiller, 557, 815
Seegmiller, Parker, 557, 815
Seifert, 826
Seifert, Christin, 826
Sekii, 877
Sekii, Taiki, 877
Sekiya, 873
Sekiya, Toshiyuki, 873
Seleznyov, 576
Seleznyov, Mikhail, 576
Selialia, 869
Selialia, Khotso, 869
Selvaganapathy, 721

Selvaganapathy, Sanjeevan, 721
Selvam, 240
Selvam, Parameswaran, 240
Semenov, 784
Semenov, Andrei, 784
Semmann, 471
Semmann, Martin, 471
Sen, 308, 558, 667
Sen, Jaydeep, 558
Sen, Procheta, 308
Sen, Tan Min, 667
Sengupta, 299
Sengupta, Aritra, 299
Senichev, 233
Senichev, Sergey, 233
Seniw, 843
Seniw, Maksim, 843
Sennrich, 247, 817
Sennrich, Rico, 247, 817
Seo, 90, 177, 179, 215, 217, 223, 309, 469, 559, 580, 797, 803, 842, 903
Seo, Eun Seo, 903
Seo, Gyuhyeon, 223
Seo, Insuk, 217
Seo, Jaehyung, 177, 469, 842
Seo, Jiwon, 223
Seo, Jun, 309, 803
Seo, KiJung, 90
Seo, Kwangwook, 559
Seo, Minhyuk, 215
Seo, Minjoon, 179
Seo, Paul Hongsuck, 580
Seo, Wonduk, 797
Seong, 190, 461
Seong, Eunseon, 461
Seong, Yongju, 190
Serbina, 95
Serbina, Anna, 95
Sercan_Karakas,
Sercan_Karakas, Sercan_Karakas,
Servan, 712
Servan, Christophe, 712
Seshadri, 468
Seshadri, Preethi, 468
Seth, 127, 140
Seth, Ashish, 127, 140
seung-won_hwang2,
seung-won_hwang2, seung-won_hwang2,
Sevegnani, 519
Sevegnani, Karin, 519
Severini, 634
Severini, Silvia, 634
Seyyed-Kalantari, 696
Seyyed-Kalantari, Laleh, 696
Sgobbi, 278
Sgobbi, Andrea, 278
Sha, 93, 224, 314, 361, 653, 820
Sha, Hangyu, 224
Sha, Huayu, 820
Sha, Jiu, 314, 361
Sha, Lei, 653
Sha, Zeyang, 93
Shaaban, 903
Shaaban, Mohamed, 903
Shabbir, 610, 845
Shabbir, Ebad, 610, 845
Shabtay, 543
Shabtay, Nimrod, 543
Shafran, 276
Shafran, Or, 276
Shah, 131, 162, 189, 209, 236, 246, 250, 357, 446, 477, 529, 544, 549, 558, 569, 615, 644, 672, 831, 841
Shah, Agam, 250
Shah, Arpit, 672
Shah, Arya, 357
Shah, Chirag, 189, 549
Shah, Deep, 615
Shah, Fahad, 131, 477
Shah, Het Riteshkumar, 236
Shah, Mubarak, 544
Shah, Neil, 162, 446, 644
Shah, Raj Sanjay, 529
Shah, Rajiv Ratn, 209, 569
Shah, Siddhant, 831
Shah, Vishwa, 246
Shah, Vraj, 558
Shah, Yash, 841
Shaham, 752
Shaham, Uri, 752
Shahane, 374
Shahane, Aditya Hemant, 374
Shahid, 613
Shahid, Abdur R., 613
Shahriar, 894
Shahriar, Sadat, 894
Shahroor, 232
Shahroor, Ali Ezzat, 232
Shaib, 106
Shaib, Chantal, 106
Shaikh, 84
Shaikh, Hana Fatima, 84
Shaji, 331
Shaji, Shreya, 331
Shakhnarovich, 863
Shakhnarovich, Greg, 863

Shami, 130
Shami, Farzad, 130
Shamout, 212
Shamout, Farah E., 212
Shamsi, 454
Shamsi, Afshar, 454
Shan, 205, 221, 248, 348, 398, 460, 475, 555, 627, 751, 752, 877
Shan, Baocai, 398, 627
Shan, Jinxiao, 460
Shan, Lianlei, 751
Shan, Rong, 248, 877
Shan, Shiguang, 555
Shan, Yi, 475
Shan, Yongxue, 348
Shan, Yubo, 205
Shan, Zifei, 221
Shan, Zikang, 752
Shang, 87, 111, 130, 163, 168, 227, 236, 255, 264, 274, 280, 285, 292, 304, 324, 327, 338, 392, 398, 514, 548, 587, 616, 638, 680, 696, 730, 740, 763, 768, 821, 822, 825, 827, 865
shang, 859
Shang, Ahmad Chamma;Omar El Herraoui; Guokan,
Shang, Changjing, 696
Shang, Chenming, 163, 821
Shang, Guokan, 168, 280, 822, 865
Shang, Jiaming, 827
Shang, Jingbo, 111, 236, 514, 548
Shang, Junyuan, 304
Shang, Lifeng, 130, 255, 264, 398, 587
Shang, Linfang, 87, 616
Shang, Mingxi, 327, 768
Shang, Shijie, 292
Shang, Shouduo, 680
Shang, Shuo, 324, 638
Shang, Tianqi, 227
Shang, Xiaoran, 274
shang, Xiaoran, 859
Shang, Xinyi, 338
Shang, Xuequn, 730
Shang, Yanmin, 324
Shang, Yuhao, 680
Shang, Yuhu, 285
Shang, Zeyu, 825
Shang, Ziyu, 392, 740
Shangguan, 231
Shangguan, Ziyao, 231
Shankar, 501
Shankar, Shiv, 501
Shao, 143, 155, 171, 255, 278, 329, 350, 363, 406, 419, 464, 519, 605, 656, 679, 685, 691, 694, 706, 724, 725, 737, 777, 780, 797, 809, 820, 856, 871, 902
Shao, Chengcheng, 725
Shao, Chenyang, 255, 656
Shao, Hongxin, 605
Shao, Jiandong, 519
Shao, Jing, 419, 685, 777, 902
Shao, Juexi, 871
Shao, Kailai, 679
Shao, Liangying, 406
Shao, Mingchen, 363
Shao, Pengpeng, 706
Shao, Pengyang, 171
Shao, Qian, 155
Shao, Rui, 694
Shao, Run, 278
Shao, Sicong, 856
Shao, Wei, 406, 724
Shao, Yanjun, 797
Shao, Yifei, 809
Shao, Yijia, 464
Shao, Yiwen, 143, 780
Shao, Zedian, 820
Shao, Zekai, 329, 691
Shao, Zhenzhou, 737
Shao, Zongru, 350
ShaobingLian, 895
ShaobingLian, ShaobingLian, 895
Shapira, 578, 835
Shapira, Natalie, 578
Shapira, Ori, 835
Shar, 531
Shar, Lwin Khin, 531
Sharan, 225
Sharan, Vatsal, 225
Shareghi, 358, 510
Shareghi, Ehsan, 358, 510
Sharif, 171
Sharif, Mahmood, 171
Sharifymoghaddam,
Sharifymoghaddam, Sahel,
sharji, 860
sharji, Asila Al, 860
Sharma, 212, 270, 301, 309, 315, 443, 538, 561, 615, 649, 764, 823, 866
Sharma, Abhisheik, 764
Sharma, Amit, 866
Sharma, Anuj, 309, 315
Sharma, Anupriya, 443
Sharma, Ashish, 561
Sharma, Kartik, 538

Sharma, Karun, 649
Sharma, Krish, 301
Sharma, Laasya, 212
Sharma, Lavisha, 823
Sharma, Pankaj Kumar, 615
Sharma, Saransh, 270
Shatabda, 375
Shatabda, Swakkhar, 375
Shatnawi, 860
Shatnawi, Sara, 860
Shayanfar, 180
Shayanfar, Radin, 180
She, 94, 325, 383, 405, 491, 801, 856, 867
She, Chaoyin, 325
She, Dong, 491, 801, 856
She, Kun, 383
She, Rui, 405
She, Shuaijie, 94, 867
Sheha, 516
Sheha, Dr. Mohamed Abu, 516
shehzad, 394
shehzad, Khawar, 394
Sheikh, 246, 394
Sheikh, Muhammad Umer, 394
Sheikh, Zaid, 246
Shejole, 406
Shejole, Kaustubh Shivshankar, 406
Shekhar, 606, 617, 841
Shekhar, Ashish Raj, 841
Shekhar, Shivani, 606, 617
Shelat, 621
Shelat, Shlok, 621
Shelmanov, 126, 179, 233, 236, 491, 777
Shelmanov, Artem, 126, 179, 233, 236, 491, 777
SHEN, 666, 767
Shen, 84, 93, 96, 97, 112, 126, 132, 144, 149, 171, 177, 194, 205, 214, 221, 226, 230, 243, 252, 253, 255, 257, 283, 286, 303, 304, 313, 319, 331, 338, 341, 352, 355, 359, 368, 371, 380, 398, 402, 404, 408, 412, 414, 417, 419, 435, 443, 446, 449, 452, 470, 472, 495, 498, 506, 518, 546, 559, 560, 573, 576, 584, 585, 592, 604, 611, 619, 621, 625, 630, 631, 635, 641, 646, 647, 661, 670, 676, 685, 696, 703, 707, 708, 717, 719, 722, 723, 728, 735, 742, 743, 747, 751, 753, 785, 792, 797, 802, 807, 820, 821, 842, 844, 845, 849, 852, 861, 879, 880, 892, 895, 896, 901, 905
shen, 636, 758
Shen, Beijun, 283, 518
Shen, Chao, 719, 807, 844
Shen, Chen, 670, 879
Shen, Chenxinran, 470
Shen, Chunhua, 576, 785
Shen, Dian, 625, 641
Shen, Fei, 97, 171
Shen, Guangyao, 573
SHEN, HANWEN, 767
SHEN, HAO, 666
Shen, Hao, 743
Shen, Hua, 472, 661
Shen, Huawei, 319, 355, 417, 584, 685, 753, 901
Shen, Hui, 821
Shen, Jiajie, 443
Shen, Jialie, 546
Shen, Jianbing, 625, 646
Shen, Jianhao, 255
Shen, Jiawei, 132, 144, 676
Shen, Jiayi, 243
Shen, Jocelyn, 112
Shen, Jucheng, 498
Shen, Jundong, 371
Shen, Junfeng, 221, 703
Shen, Junyi, 96
Shen, Kehan, 177
Shen, Lei, 896
Shen, Li, 723
Shen, Linlin, 253, 286, 861
Shen, Minghe, 630
Shen, Nanhan, 226
Shen, Qiang, 696
Shen, Qing, 585
Shen, Ruijie, 619
Shen, Shannon, 112, 331
Shen, Shannon Zejiang, 446
Shen, Shiqi, 573
Shen, Tao, 230, 257
Shen, Tiancheng, 470
Shen, Wei, 435
Shen, Weixing, 560, 735
Shen, Weizhou, 93, 214
Shen, Wen, 341, 419
Shen, Xiang, 611
Shen, Xiangqing, 414
Shen, Xiaobin, 506
Shen, Xiaoyu, 412
Shen, Xingyu, 905
Shen, Xinshu, 635
Shen, Xinyue, 647
Shen, Xu, 717, 792
Shen, Xuan, 495
Shen, Xudong, 359
Shen, Yaling, 84, 592
Shen, Ye, 559
shen, yelong, 636, 758

Shen, Yi, 341
Shen, Yibin, 408
Shen, Yifei, 205
Shen, Yikai, 194
Shen, Ying, 149, 402, 707, 708, 849
Shen, Yingli, 449
Shen, Yingtao, 747
Shen, Yiting, 792
Shen, Yixian, 313
Shen, Yongliang, 126, 304, 368, 398, 404, 621, 631, 722, 751
Shen, Yu, 845
Shen, Yue, 604, 742
Shen, Yufan, 303
Shen, Yuhao, 96, 313, 352
Shen, Yujiong, 717, 820
Shen, Zhanming, 728
Shen, Zhanyu, 802
Shen, Zhengyuan, 452, 797, 880
Shen, Zheyu, 252
Shen, Zhibin, 892
Shen, Zhifeng, 842
Shen, Zhihong, 419
Shen, Zhiqiang, 338
Shen, Zhiyu, 852, 895
Shen, Zizhuo, 380
Shenaj, 537
Shenaj, Donald, 537
Sheng, 96, 131, 144, 149, 170, 323, 332, 552, 625, 640, 819, 849, 877
Sheng, Honghui, 323
Sheng, Jiawei, 332
Sheng, Jiayi, 877
Sheng, Junwei, 625
Sheng, Junxi, 849
Sheng, Mingzhi, 144
Sheng, Mu, 131
Sheng, Qiang, 96, 170
Sheng, Tao, 552, 819
Sheng, Zhichao, 149
Sheng, Ziji, 640
Shenguanlin, 239
Shenguanlin, Shenguanlin, 239
Sheth, 166, 238, 582
Sheth, Amit, 238
Sheth, Ivaxi, 582
Sheth, Rajvee, 166
SHI, 212, 864
Shi, 85, 87, 94, 108, 110, 111, 114, 123, 124, 132, 142, 144, 148, 161, 164, 165, 167, 178, 191, 201, 216, 221, 222, 229, 230, 260, 268, 269, 276, 281, 283, 290, 295, 297, 303, 304, 306, 310, 311, 314, 324, 325, 339, 341, 348, 350, 354, 355, 358, 369, 371, 381, 383, 391, 392, 395, 405, 409, 417, 425, 433, 437, 439, 440, 446, 450, 454, 455, 462, 470, 479, 500, 515, 518, 531, 539, 550, 551, 560, 561, 562, 569, 570, 578, 579, 590, 601, 617, 621, 627, 632, 641, 644, 647, 665, 669, 677, 685, 688, 696, 697, 703, 716, 723, 729, 733, 736, 738, 743, 745, 758, 759, 760, 767, 795, 798, 814, 830, 835, 842, 860, 877, 901
shi, 540
Shi, Bei, 371
Shi, Bingkang, 462, 500
Shi, Botian, 148, 303
Shi, Bowen, 381, 696
Shi, Boxin, 348
Shi, Boyan, 221, 703
Shi, Chuan, 339
Shi, Chuanqi, 369
Shi, Dachuan, 569
Shi, Daiting, 142, 276, 304, 310, 795
Shi, Dan, 306, 590
Shi, Duo, 409
Shi, Fangcheng, 745
Shi, Freda, 110, 450, 842
Shi, FuLin, 297
Shi, Ge, 314
Shi, Haibo, 165, 230, 295
Shi, Hailong, 290
Shi, Haochen, 647, 733
Shi, Haolin, 417, 743
Shi, Haoyuan, 85, 759
Shi, Hongchang, 561
Shi, Javen, 454
Shi, Jiaqi, 392
Shi, Jiatong, 111, 455
Shi, Jiawen, 350
Shi, Jiayi, 405
Shi, Jie, 688
Shi, Jieke, 515, 531
Shi, Jieming, 578
Shi, Jingwei, 531
Shi, Jinxin, 440
Shi, Junhao, 665
Shi, Kai, 627
Shi, Kaiwen, 178
Shi, Ke, 324
shi, kun, 540
Shi, Lei, 395
Shi, Liang, 124
Shi, Lida, 325
Shi, Lin, 860

Shi, Ling, 94, 697, 729
Shi, Lingfeng, 551
Shi, Luohe, 814
Shi, Mingjun, 383
Shi, Minjing, 835
Shi, Qi, 191, 716
Shi, Qilong, 87
Shi, Qixu, 281
Shi, Qundong, 758
Shi, Rongyao, 560
Shi, Ruijie, 617
Shi, Ruqian, 87
Shi, Shanghao, 767
Shi, Shenghui, 358
Shi, Shumin, 437
Shi, Shuming, 341
Shi, Taiwei, 479
Shi, Tianqi, 677
SHI, TIANYU, 212, 864
Shi, Wei, 108
Shi, Weijia, 446, 550
Shi, Weijie, 132, 144, 470, 830
Shi, Weikang, 339
Shi, Weizhe, 641
Shi, Wenjie, 355
Shi, Wentao, 303, 311, 325
Shi, Wenxuan, 229
Shi, Xian, 736
Shi, Xianjie, 260
Shi, Xiaodong, 433
Shi, Xiaoming, 877
Shi, Xiaowei, 201
Shi, Xin, 669
Shi, Xinhe, 164
Shi, Xinyan, 539
Shi, Xinyu, 161, 632, 901
Shi, Xuanhua, 446
Shi, Yabing, 216
Shi, Yafei, 369
Shi, Yang, 738
Shi, Yaqing, 167
Shi, Yike, 306
Shi, Yongcheng, 439
Shi, Yuanyuan, 601
Shi, YuFei, 114
Shi, Yukun, 798
Shi, Yuling, 222, 283, 518
Shi, Yuntian, 268
Shi, Yuzhen, 560
Shi, Zesheng, 354
Shi, Zhan, 269
Shi, Zhenbo, 391
Shi, Zhengliang, 425, 716
Shi, Zhengpeng, 760
Shi, Zhenmei, 123
Shi, Zhichao, 417, 562, 723
Shi, Zhiyuan, 685
Shi, Zhonghao, 570
Shi, Zhouxing, 579
Shi, Zijing, 621
Shi, Zixing, 644
Shieh, 625
Shieh, Michael Qizhe, 625
Shieh;,
Shieh;, Michael Qizhe,
Shih, 547, 856
Shih, Bono Po-Jen, 856
Shih, Wei-Kuan, 547
Shihab, 309, 315
Shihab, Ibne Farabi, 309, 315
Shikoda, 512
Shikoda, Naoki, 512
Shim, 107, 217, 552, 592, 842
Shim, Gyuho, 842
Shim, Hyunjung, 592
Shim, Kyuhong, 107
Shim, Ryan Soh-Eun, 217
Shim, Sumin, 552
Shimizu, 252, 416
Shimizu, Seiji, 252
Shimizu, Takafumi, 416
Shimodaira, 213, 885
Shimodaira, Hidetoshi, 213, 885
Shin, 124, 140, 223, 280, 290, 351, 445, 601, 691, 717, 816, 831, 833, 842, 846
Shin, Haebin, 280
Shin, Jaewoo, 601
Shin, Jiho, 223
Shin, Jisu, 124, 717, 831, 833
Shin, Jong-hun, 691
Shin, Joongmin, 842
Shin, Jungkyu, 140
Shin, Juyeon KimGeon LeeDongwon ChoiTaeuk KimKijung, 816
Shin, Kijung, 846
Shin, Sunguk, 351
Shin, Yongseop, 445
Shin, Yooju, 290
Shinde, 566
Shinde, Swapnil, 566
Shindo, 298
Shindo, Hikaru, 298
Shinoda, 781
Shinoda, Kazutoshi, 781
Shirai, 306, 863
Shirai, Kiyoaki, 306

Shirai, Naoto, 863
Shiri, 816
Shiri, Fatemeh, 816
Shivakumar, 261
Shivakumar, Esha, 261
SHIYU, 409
SHIYU, TENG, 409
Shklovski, 577
Shklovski, Irina, 577
Shmueli-Scheuer, 517, 783
Shmueli-Scheuer, Michal, 517, 783
Shmueli-Scheuer;,
Shmueli-Scheuer;, Michal,
Shoeybi, 887
Shoeybi, Mohammad, 887
Shohan, 146
Shohan, Faisal, 146
Shojaie, 484
Shojaie, Mehdi, 484
Shokri, 631
Shokri, Mohammad, 631
Shou, 125, 221, 497, 536, 566
Shou, Haoran, 566
Shou, Lidan, 125, 221, 497, 536
Shrestha, 247, 570, 614, 620
Shrestha, Gyanendra, 614
Shrestha, Kabir Manandhar, 620
Shrestha, Kaleen, 570
Shrestha, Siddhartha, 247
Shridhar, 784
Shridhar, Kumar, 784
Shu, 87, 112, 125, 128, 145, 297, 424, 442, 605, 607, 616, 847
Shu, Dong, 442
Shu, Peng, 87, 605, 607, 616, 847
Shu, Tao, 125
Shu, Xiang, 145
Shu, Yan, 112
Shu, Yao, 424
Shu, Ying, 128
Shu, Yuanchao, 297
Shuai, 174, 640, 774
Shuai, Hong-Han, 174
Shuai, Jingxian, 774
Shuai, Zhihao, 640
Shuai_Ling,
Shuai_Ling, Shuai_Ling,
Shubham, 554
Shubham, Kumar, 554
Shuguang, 666
Shuguang, Ma, 666
Shulga, 685
Shulga, Andrey, 685
Shum, 292
Shum, Heung-Yeung, 292
Shun, 265
Shun, Shiow-Ching, 265
Shuraim, 538
Shuraim, Mohammed, 538
Shutova, 591
Shutova, Ekaterina, 591
Shuvo, 211
Shuvo, Uddip Acharjee, 211
Shuyuan, 892
Shuyuan, Liu, 892
Shvetsov, 330, 342
Shvetsov, Egor, 330, 342
Shwartz, 891
Shwartz, Vered, 891
Shyalika, 238
Shyalika, Chathurangi, 238
Shypula, 502
Shypula, Alexander, 502
Si, 112, 135, 216, 303, 374, 417, 449, 665, 668, 750, 762, 772
Si, Chenglei, 112
Si, Jianfeng, 303
Si, Jiasheng, 374, 668, 762
Si, Qingyi, 216
Si, Shuzheng, 449, 772
Si, Wai Man, 417
Si, Weiye, 665, 750
Si, Yuxuan, 135
Siam, 529
Siam, Shakhrul Iman, 529
Siangliulue, 467
Siangliulue, Pao, 467
Sibaee, 860
Sibaee, Serry, 860
Sicilia, 123
Sicilia, Anthony, 123
Siddique, 545
Siddique, Umer, 545
Siddiqui, 424
Siddiqui, Ayaan, 424
Sidhu, 168
Sidhu, Risham, 168
Sieker, 580
Sieker, Judith, 580
Sierra-Arévalo, 226
Sierra-Arévalo, Michael, 226
Siheng_Li,
Siheng_Li, Siheng_Li,
Sijan, 100
Sijan, Tanvir Ahmed, 100
Sil, 320

Sil, Avirup, 320
Sileo, 713
Sileo, Damien, 713
Silpasuwanchai, 357
Silpasuwanchai, Chaklam, 357
Silva, 408, 596
Silva, Diogo Fernandes Costa, 596
Silva, Sathira, 408
Silvestri, 519, 567
Silvestri, Fabrizio, 519, 567
Sim, 190, 720
Sim, Myoseop, 190
Sim, Robert, 720
Simai, 752
Simai, Idan, 752
Simakov, 685
Simakov, Dmitry, 685
Simhadri, 525
Simhadri, Rishi N., 525
Simon, 795
Simon, Luke, 795
Simonazzi, 614
Simonazzi, Stefano, 614
Simoni, 778
Simoni, Marco, 778
Simperl, 246
Simperl, Elena, 246
Simpson, 428
Simpson, Edwin, 428
Sinan, 769
Sinan, Omar, 769
Sinclair, 427
Sinclair, Tom Utting; Mario Giulianelli; Arabella,
Singh, 88, 131, 136, 141, 160, 166, 218, 240, 250, 305, 441, 442, 467, 468, 503, 552, 560, 575, 606, 613, 628, 700, 731, 827, 837, 845, 894, 895
Singh, Aaryamonvikram, 442
Singh, Aashu, 845
Singh, Achal Pratap, 305
Singh, Adarsh, 88
Singh, Amanpreet, 467
Singh, Apoorva, 606
Singh, Ayushman Kumar, 827
Singh, Divyajot, 503
Singh, Janvijay, 575
Singh, Jyotika, 131, 136, 552
Singh, Mayank, 166, 441
Singh, Mukul, 160
Singh, Muskaan, 218
Singh, Navyansh, 628
Singh, Pavit, 560
Singh, Prateek, 613
Singh, Pratik Rakesh, 731
Singh, Rishabh, 141
Singh, Ritambhara, 837, 894
Singh, Sameer, 468
Singh, Shailender, 131
Singh, Shivdeep, 240
Singh, Varun, 250
Singh, Vikash, 895
Singh, Virendra, 700
Singhal, 160, 161, 784, 828
Singhal, Bhavuk, 160, 161
Singhal, Himanshu, 828
Singhal, Raghav, 784
Singhu, 572
Singhu, Saithej, 572
Singla, 94, 337, 659, 809, 845
Singla, Adish, 659
Singla, Parag, 94, 337
Singla, Somanshu, 809
Singla, Sumedha, 845
Sinha, 166, 754, 891, 895
Sinha, Aman, 891
Sinha, Debu, 754
Sinha, Rohit, 895
Sinha, Samridhi Raj, 166
Sircar, 613
Sircar, Prateek, 613
Siro, 471
Siro, Clemencia, 471
Sirohi, 374
Sirohi, Anuj Kumar, 374
Sisman, 780
Sisman, Berrak, 780
Sitaram, 577, 837
Sitaram, Sunayana, 577, 837
Siu, 784
Siu, Alexa, 784
Sivapiromrat, 510
Sivapiromrat, Sanhanat, 510
SiyuWang, 732, 741
SiyuWang, SiyuWang, 732, 741
Skantze, 399, 778
Skantze, Gabriel, 399, 778
Skianis, 865
Skianis, Konstantinos, 865
Skorupska, 203
Skorupska, Kinga, 203
Skrynnik, 424
Skrynnik, Alexey, 424
Slobodkin, 173
Slobodkin, Aviv, 173
Slusallek, 120
Slusallek, Philipp, 120

Smart, 222
Smart, Andrew J, 222
Smerdon, 478
Smerdon, Claire J., 478
Smirnov, 178, 233, 342
Smirnov, Konstantin, 179
Smirnov, Vladimir, 342
Smirnov, Vladislav, 233
Smith, 141, 166, 454
Smith, Kaleb E, 166
Smith, Noah A., 141
Smith, Virginia, 454
Smithard, 763
Smithard, Alison Q., 763
Smock, 871
Smock, Brandon, 871
Snæbjarnarson, 110
Snæbjarnarson, Vésteinn, 110
So, 147, 163, 477, 821
So, Hayden Kwok-Hay, 163, 821
So, Namyeong, 147
So, Yeonkyoung, 477
Socolinsky, 82, 609
Socolinsky, Diego A., 82, 609
Sodagar, 909
Sodagar, Zahra, 909
Soetedjo, 492
Soetedjo, Riza Setiawan, 492
Sogawa, 833
Sogawa, Yasuhiro, 833
Sogawa;,
Sogawa;, Ryo Mitsuhashi; Gaku Morio; Ayana Niwa; Masahiro Kaneko; Kentaro Inui; Terufumi Morishita; Yuta Koreeda; Yasuhiro,
Sohn, 823
Sohn, Jiwoong, 823
Sok, 885
Sok, Jaewon, 885
Soldaini, 429
Soldaini, Luca, 429
Solergibert, 784
Solergibert, Antoni-Joan, 784
Solissa, 395
Solissa, Noa Visser, 395
Solodkin, 761
Solodkin, Vladimir, 761
Solopova, 84
Solopova, Veronika, 84
Solorio, 98, 432
Solorio, Thamar, 98, 432
Som, 661
Som, Subhojit, 661
Someki, 111
Someki, Masao, 111
Someya, 102, 639
Someya, Taiga, 102, 639
Son, 157, 182, 259, 267, 465, 595, 705, 745
Son, Guijin, 267, 595
Son, Mai Hong, 745
Son, Sang-Woo, 157
Son, Suhyune, 182, 465
Son, Yejin, 259
Son, Youngseo, 705
Sonbol, 526
Sonbol, M Bassam, 526
SONG, 390, 646
Song, 91, 113, 117, 119, 120, 123, 124, 127, 129, 135, 149, 151, 153, 154, 158, 161, 169, 193, 197, 206, 212, 221, 229, 241, 243, 253, 260, 267, 280, 297, 304, 307, 313, 321, 325, 341, 370, 380, 384, 388, 389, 391, 401, 409, 431, 434, 440, 445, 447, 450, 454, 456, 467, 473, 474, 475, 479, 486, 496, 508, 515, 516, 519, 527, 530, 541, 559, 562, 567, 576, 592, 600, 601, 604, 608, 611, 612, 636, 639, 642, 643, 647, 658, 659, 661, 671, 702, 707, 709, 717, 722, 730, 733, 747, 749, 754, 761, 777, 782, 787, 788, 795, 796, 800, 802, 828, 830, 831, 833, 838, 846, 849, 872, 876, 878, 881, 882, 883, 895, 902
Song, Changhao, 642
Song, Chaoda, 895
Song, Dawei, 541, 796
Song, Dawn, 127, 151, 592
Song, Eunwoo, 113
Song, Feifan, 659
Song, Guanchun, 761
Song, Guojie, 380, 445, 788
Song, Hao, 389
Song, Haohua, 158
Song, Haonan, 658
Song, Hongru, 754
Song, Hoyun, 515, 717, 831, 833
Song, Hwanjun, 123, 124, 828
Song, Hyun-Je, 297
Song, Jaeyun, 120
Song, James, 604
Song, Jiajun, 643
Song, Jian, 169
Song, Jiangcheng, 409
Song, Jie, 193, 388, 881
Song, Jingkuan, 883
Song, Jingwei, 212
Song, Jinho, 516

Song, Jiwon, 206
Song, Jiwoo, 486
Song, Jun, 341, 636, 707, 733
Song, Junru, 135
Song, Juntong, 559
Song, Kaiqiang, 486
Song, Kaisong, 474, 671, 749
Song, Kaitao, 304, 321, 391
Song, Kangsong, 467
Song, Linfeng, 576, 709, 722, 838
Song, Lingyun, 730
Song, Linyue, 154
Song, Min, 639
Song, Mingli, 388
Song, Mingyang, 872
Song, Mingzi, 527
Song, Minkyoung, 91
Song, Myeong Ho, 846
Song, Qi, 530
Song, Qingquan, 124
Song, Rui, 325, 431, 881
Song, Ruihua, 777
Song, Sangjun, 149
Song, Seockbean, 508
Song, Seok Hwan, 119
Song, Seok_Hwan,
Song, Seongjong, 592
Song, Seyoung, 197, 612
Song, Shouyou, 384
Song, Shuangyong, 370
Song, Sibo, 450
Song, Siqing, 401
Song, Siyuan, 153
Song, Tingyu, 241, 473, 800
Song, Wei, 380
Song, Weixi, 601
Song, William J., 795
Song, Woojung, 447, 787
Song, Xia, 479, 661
Song, Xiangchen, 454, 878
Song, Xiaoshuai, 702
Song, Xinhao, 902
Song, Xiujie, 559
Song, Xueqing, 129
Song, Xuwen, 761
Song, Yang, 611
Song, Yangqiu, 161, 243, 253, 313, 496, 647, 733, 782, 830, 882
Song, Yeeho, 600
Song, Yi, 280
Song, Yi-Zhe, 802
Song, Yifan, 229, 440
Song, Yiming, 221
Song, Youngsook, 267
Song, Yu, 409
SONG, Yuanfeng, 390, 646
Song, Yuanfeng, 434
Song, Yufei, 519
Song, Yuhan, 117
Song, Yuhang, 608
Song, Yun, 562
Song, Yuyang, 260
Song, Zewen, 456
Song, Zihan, 747
Song, Zihe, 307
Song, Zijia, 313
Song, Zikai, 260, 849
Song, Zirui, 117, 475, 876
Song;,
Song;, Dawn,
Song;, Hwanjun,
Song;, Yunxiang MO; Tianshi Zheng; Qing Zong; Jiayu Liu; Baixuan Xu; Yauwai Yim; Chunkit Chan; Jiaxin Bai; Yangqiu,
Soni, 473, 846
Soni, Devesh, 473
Soni, Manan, 846
Sonkar, 586
Sonkar, Shashank, 586
Sonomoto, 416
Sonomoto, Eisuke, 416
Sontag, 112
Sontag, David, 112
Sonwane, 263
Sonwane, Atharv, 263
Sood, 770
Sood, Prajwal, 770
Soon, 249
Soon, Lay-Ki, 249
Sorokin, 424
Sorokin, Nikita, 424
Soroush_Vosoughi,
Soroush_Vosoughi, Soroush_Vosoughi,
Sosnowski, 203
Sosnowski, Witold, 203
Sotiras, 204
Sotiras, Aris, 204
Sotnikova, 784
Sotnikova, Anna, 784
Soufleri, 166
Soufleri, Efstathia, 166
Soumik, 613
Soumik, Mohd. Farhan Israk, 613
Sourati, 226, 819
Sourati, Zhivar, 226, 819
Souri, 661

Souri, Yaser, 661
Speas, 150
Speas, Samuel, 150
Speicher, 99
Speicher, Till, 99
Spero, 233
Spero, Max, 233
Spiegel, 891
Spiegel, Michal, 891
Spiliopoulou, 271
Spiliopoulou, Evangelia, 271
Spiro, 591
Spiro, Emma, 591
Spliethoff, 218
Spliethoff, Sophie, 218
Splitt, 826
Splitt, Fedor, 826
Sproat, 572
Sproat, Richard, 572
Sra, 183
Sra, Misha, 183
Srba, 667
Srba, Ivan, 667
Srey, 210
Srey, Ponhvoan, 210
Sridhar, 207, 498, 776
Sridhar, Dhanya, 207
Sridhar, Roshan, 776
Sridhar, Sharath Nittur, 498
Srinath, 827
Srinath, Mukund, 827
Srinivas, 843
Srinivas, Dananjay, 843
Srinivasa, 250, 548, 633
Srinivasa, Haricharana, 250
Srinivasa, Jayanth, 548, 633
Srinivasan, 209, 843
Srinivasan, Ravi, 209
Srinivasan, Tejas, 843
Sripada, 772
Sripada, Chandra, 772
Sripheanpol, 388
Sripheanpol, Witthawin, 388
Sriraman, 112
Sriraman, Siddharth, 112
Srishankar, 180
Srishankar, Nishan, 180
Srivastava, 91, 222, 234, 236, 293, 453, 532, 637
srivastava, 249
srivastava, aashish, 249
Srivastava, Aseem, 236
Srivastava, Gaurav, 222, 637
Srivastava, Jaideep, 234
Srivastava, Prashast, 91, 532
Srivastava, Shashank, 453
Srivastava, Vivek, 293
Staab, 422
Staab, Steffen, 422
Stacey, 847
Stacey, Joe, 847
Stadler, 542
Stadler, Bernhard, 542
Stafford, 228, 864
Stafford, Eli, 228
Stafford, Tom, 864
Stahl, 253, 857
Stahl, Gustavo Bertolo, 253
Stahl, Maja, 857
Stammer, 298
Stammer, Wolfgang, 298
Stamou, 865
Stamou, Giorgos, 865
Stanczak, 243
Stanczak, Karolina, 243
Stanojevi, 898
Stanojevi, David, 898
Stanovsky, 82
Stanovsky, Gabriel, 82
Stappen, 440
Stappen, Lukas, 440
Stark, 243, 495
Stark, Philipp, 495
Stark, Vanessa, 243
State, 580
State, Radu, 580
Staub, 659
Staub, Jacqueline, 659
Steele, 481
Steele, Brady, 481
Steigerwald, 150
Steigerwald, Philipp, 150
Stein, 154
Stein, Gregory J., 154
Steinebach, 792
Steinebach, Martin, 792
Steinert, 609
Steinert, Rebecca, 609
Stenetorp, 519, 891
Stenetorp, Pontus, 519, 891
Stengel-Eskin, 293, 853, 890
Stengel-Eskin, Elias, 293, 853, 890
Stepachev, 862
Stepachev, Pavel, 862
Stepanek, 597
Stepanek, Jan, 597
Stepankova, 597

Stepankova, Barbora, 597
Stepputtis, 118
Stepputtis, Simon, 118
Sternfeld, 784
Sternfeld, Alexander, 784
Sternlicht, 96, 176
Sternlicht, Noy, 96, 176
Stewart, 862
Stewart, Damian, 862
Stieger, 527
Stieger, Benjamin, 527
Stillwell, 227
Stillwell, David, 227
Stocker, 497
Stocker, Volker, 497
Storkey, 903
Storkey, Amos, 903
Storks, 105, 772, 902
Storks, Shane, 105, 772, 902
Stowe, 85
Stowe, Kevin, 85
Stoyanov, 95, 442
Stoyanov, Veselin, 95, 442
Strahan, 82, 609
Strahan, Bob, 82, 609
Stranisci, 471
Stranisci, Marco Antonio, 471
Strimel, 111
Strimel, Grant, 111
Strobl, 165
Strobl, Lena, 165
Stroebl, 821
Stroebl, Benedikt, 821
Strohmaier, 835
Strohmaier, Markus, 835
Strubell, 123
Strubell, Emma, 123
Stupachenko, 534
Stupachenko, Evgeny, 534
Sturm, 818
Sturm, Arnon, 818
SU, 190
Su, 86, 113, 128, 146, 153, 158, 161, 164, 180, 191, 196, 198, 205, 218, 226, 227, 236, 284, 296, 311, 318, 329, 336, 357, 371, 381, 383, 384, 395, 397, 401, 403, 414, 419, 422, 427, 442, 505, 527, 531, 565, 627, 629, 638, 645, 659, 660, 672, 675, 677, 680, 682, 684, 695, 698, 702, 705, 707, 720, 721, 722, 751, 753, 762, 775, 776, 777, 784, 808, 815, 821, 826, 836, 892, 895
Su, Binxian, 86
Su, Du, 357, 371, 422, 753
Su, Guanhao, 527
Su, Hang, 180, 296, 892
Su, Hongyang, 638
Su, Houcheng, 815
Su, Hsuan, 218
Su, Hui, 311
Su, Hung-Ting, 153, 196
Su, Jiamin, 205
Su, Jiaming, 505
Su, Jian, 198
Su, Jiaoyang, 682
Su, Jiayuan, 645
Su, Jie, 329
Su, Jinbo, 146, 680
Su, Jinhang, 776
Su, Jinsong, 336, 397, 677, 684, 720, 751
Su, Jinyan, 226, 236, 442
Su, Junyou, 284
Su, Lingfeng, 707
Su, Lixin, 721
Su, Liya, 384
Su, Miao, 784
Su, Minghao, 419
Su, Mingyi, 161
Su, Sen, 128, 383, 403, 775
Su, Shu, 381
Su, Tianhuang, 627
Su, Tianyao, 836
Su, Weijie J., 227
SU, Weiwen, 190
Su, Wenbo, 660, 675, 695, 698, 702
Su, Xi, 311
Su, Xiangdong, 318, 401, 762
Su, Xiaohong, 659
Su, Xuanbo, 895
Su, Ya, 427
Su, Yang, 158
Su, Yanning, 395
Su, Yi, 414, 722
Su, Yu, 826
Su, Yuanhao, 419
Su, Yue, 113, 191
Su, Yueyang, 629
Su, Yuxin, 531
Su, Zhan, 565, 705
Su, Zhaochen, 777, 808
Su, Zhe, 164
Su, Zhenpeng, 672
Su, Zunhai, 821
Suarez, 862
Suarez, Pedro Ortiz, 862
Subramani, 291

Subramani, Nishant, 291
Subramanian, 572, 584, 878
Subramanian, Jayakumar, 584
Subramanian, Shivashankar, 878
Subramanian, Sindhura Kumbakonam, 572
Such, 140
Such, Jose, 140
Suciu, 549
Suciu, Dan, 549
Suda, 444
Suda, Naveen, 444
Sudevan, 227
Sudevan, Saurav, 227
Sudharsan, 837
Sudharsan, Deepthi, 837
Sugawara, 92, 242, 447, 867
Sugawara, Saku, 92, 242, 447, 867
Suh, 104, 207, 229, 267, 533
Suh, Bongwon, 104, 229, 267, 533
Suh, Joseph, 207
Sui, 213, 349, 374, 465, 475, 630, 755, 762, 830, 907
Sui, Dianbo, 630
Sui, Xiaohang, 349
Sui, Xingyu, 755, 762
Sui, Xuhui, 374
Sui, Yuan, 213, 830
Sui, Zhifang, 465, 475, 907
Suleiman, 854
Suleiman, Basem, 854
Sullivan, 102
Sullivan, Peter, 102
Sultana, 371, 556
Sultana, Faria, 371
Sultana, Najrin, 556
Sumida, 469
Sumida, Ryuichi, 469
SUN, 306, 647
Sun, 83, 85, 87, 98, 113, 125, 128, 131, 134, 136, 137, 140, 142, 145, 146, 151, 152, 153, 159, 162, 165, 169, 170, 196, 198, 200, 211, 220, 227, 235, 239, 243, 245, 248, 251, 252, 272, 274, 276, 279, 281, 282, 283, 285, 286, 300, 301, 303, 304, 315, 319, 326, 327, 334, 335, 337, 342, 346, 348, 350, 352, 353, 355, 361, 369, 371, 372, 377, 378, 380, 381, 385, 387, 388, 394, 397, 403, 406, 408, 409, 413, 415, 416, 418, 421, 424, 434, 437, 448, 449, 450, 453, 455, 458, 461, 474, 478, 487, 490, 491, 503, 507, 508, 509, 511, 533, 534, 535, 537, 539, 541, 543, 545, 546, 549, 550, 552, 565, 568, 577, 578, 581, 591, 593, 596, 600, 604, 611, 616, 621, 625, 626, 631, 635, 636, 638, 643, 647, 649, 651, 653, 665, 673, 676, 680, 681, 698, 699, 705, 708, 712, 715, 716, 717, 719, 720, 724, 730, 732, 733, 734, 735, 736, 741, 749, 753, 756, 759, 768, 775, 783, 786, 787, 798, 809, 826, 829, 834, 840, 842, 844, 847, 851, 856, 867, 893, 896, 897, 900, 905
Sun, Aixin, 198, 638
Sun, Angela, 534
Sun, Ao, 381, 543
Sun, Bo, 151, 826
Sun, Bolun, 83
Sun, Di, 565, 676
Sun, Duolin, 604
Sun, Fei, 355, 371, 577, 753
Sun, Feng, 896
Sun, Guanglu, 342
Sun, Guangzhi, 140
Sun, Guoheng, 252
Sun, Gyeonghun, 511
Sun, Haifeng, 251, 327, 388
Sun, Haiyang, 545, 550
Sun, Han, 578
Sun, Hao, 409, 681
Sun, Haoqin, 352
SUN, Haoran, 306
Sun, Haosen, 775
Sun, Haoxiang, 847
Sun, Henan, 458
Sun, Hengyu, 708
Sun, Hongda, 636
Sun, Hongru, 448
Sun, Hongyue, 581
Sun, Huashan, 631, 705
Sun, Hui, 539
Sun, Huyang, 200
Sun, Jiacheng, 165
Sun, Jiahang, 734
Sun, Jiahao, 283, 786
Sun, Jiajun, 545, 717, 897
Sun, Jian, 749
Sun, Jianwen, 134, 844
Sun, Jiarui, 549, 787
Sun, Jie, 134, 145, 665
Sun, Jimeng, 450
Sun, Jingbo, 568, 621
Sun, Jingyi, 453, 577
Sun, Jun, 220
Sun, Kaiyue, 301
Sun, Ke, 840
Sun, Kexuan, 851

Sun, Le, 252, 369, 378, 653, 741, 834
Sun, Li, 128, 403, 775
Sun, Lianxin, 768
Sun, Libo, 503
Sun, Lichao, 350
Sun, Lihao, 211
Sun, Lillian, 394
Sun, Lin, 87, 303, 437
Sun, Lina, 698
Sun, Lingyun, 281, 533, 798
Sun, Luning, 227
Sun, Luoyang, 385
Sun, Maosong, 282, 377, 387, 406, 434, 449, 596, 673, 716, 905
Sun, Meng, 635
Sun, Mengshu, 146, 334, 720
Sun, Min, 153, 196, 736
Sun, Mingming, 337
Sun, Mingzhu, 844
Sun, Minqi, 125
Sun, Nan, 319
Sun, Penglei, 162
Sun, Pengqi, 550
SUN, Pingwei, 647
Sun, Pingwei, 708
Sun, Qiang, 545
Sun, Qiheng, 369
Sun, Qilin, 705
Sun, Qingfeng, 474, 712
Sun, Qingqing, 421
Sun, Qiushi, 159, 248, 546, 591
Sun, Qiyao, 369, 829, 840
Sun, Renliang, 300
Sun, Ronald XuMingzhai, 600
Sun, Ruixiao, 611
Sun, Ruopei, 315, 461
Sun, Ruoyu, 152
Sun, Seungjong, 285
Sun, Shaoning, 140, 724
Sun, Shenghe, 274
Sun, Shixuan, 487
Sun, Siming, 893
Sun, Sirui, 380
Sun, Suping, 169, 625
Sun, Tai-Ping, 418
Sun, Te, 326
Sun, Tianyang, 397
Sun, Tianze, 651
Sun, Wangtao, 424
Sun, Weigao, 413, 647
Sun, Weiwei, 272, 699
Sun, Weiyi, 131, 136
Sun, Weiyu, 552
Sun, Wu, 487
Sun, Xiao, 408
Sun, Xiaobing, 735
Sun, Xiaokun, 137
Sun, Xiaoyan, 508
Sun, Ximeng, 348, 490
Sun, Xin, 113, 867
Sun, Xing, 245, 449, 842
Sun, Xingwu, 798
Sun, XinYu, 759
Sun, Xu, 537
Sun, Xueqiao, 286
Sun, Yanan, 732
Sun, Yandu, 415
Sun, Yang, 715
Sun, Yanhui, 282
Sun, Yanming, 541
Sun, Yiding, 87, 616, 719
Sun, Yifan, 170
Sun, Ying, 335, 346, 535
Sun, Yining, 756
Sun, Yinqian, 98
Sun, Yitao, 85
Sun, Yizheng, 856
Sun, Yizhou, 280
Sun, Youcheng, 170
Sun, Yu, 304, 611, 715
Sun, Yuan, 809
Sun, Yuancheng, 300
Sun, Yuanfu, 783
Sun, Yuanhao, 361
Sun, Yuanyuan, 378
Sun, Yuefeng, 239
Sun, Yueru, 491
Sun, Yufan, 730
Sun, Yutao, 900
Sun, Yuxi, 699
Sun, Yuxiang, 680
Sun, Yuxuan, 353
Sun, Yuyang, 733
Sun, Zechen, 733
Sun, Zexu, 335
Sun, Zeye, 581
Sun, Zeyu, 478
Sun, Zhaoxu, 593
Sun, Zhe, 416
Sun, Zhen, 152
Sun, Zhifan, 243
Sun, Zhiyuan, 142, 276
Sun, Zhongao, 372
Sun, Zhongqian, 455
Sun, Zhongtian, 649
Sun, Zhouhao, 626, 643

Sun, Zhu, 235
Sun, Zipeng, 507, 509
Sundaram, 121, 587
Sundaram, Hari, 121, 587
Sung, 102, 189, 580, 830
Sung, Junyoung, 580
Sung, Mingi, 830
Sung, Mujeen, 189
Sung, Yoo Yeon, 102
Sung-Bin, 844
Sung-Bin, Kim, 844
Sungeun_An,
Sungeun_An, Sungeun_An,
Sunkara, 158, 598, 894
Sunkara, Monica, 158, 598, 894
Suo, 289
Suo, Jinli, 289
Suominen, 486
Suominen, Hanna, 486
Surana, 257
Surana, Rohan, 257
Surdeanu, 241, 268
Surdeanu, Mihai, 241, 268
Sureddy, 479
Sureddy, Namrata, 479
Suresh, 225
Suresh, Varsha, 225
Suri, 484, 526, 598
Suri, Manan, 484, 526, 598
Surnachev, 640
Surnachev, Nikita, 640
Susanj, 111
Susanj, Nathan, 111
Susanto, 169
Susanto, Yosephine, 169
Susilo, 486
Susilo, Richard, 486
Sutawika, 539
Sutawika, Lintang, 539
Suzuki, 390
Suzuki, Yoshimi, 390
Swaminathan, 111, 257
Swaminathan, Nandhini, 257
Swaminathan, Rupak Vignesh, 111
Swamy, 539
Swamy, Gokul, 539
Swayamdipta, 478, 843
Swayamdipta, Swabha, 478, 843
Swoboda, 293
Swoboda, Paul, 293
Sy, 104
Sy, Yaya, 104
Syme, 195
Syme, Don, 195
Synkova, 597
Synkova, Pavlina, 597
Szpektor, 585, 821
Szpektor, Idan, 585, 821
Sá, 173
Sá, Jáder Martins Camboim de, 173
Sójka, 606
Sójka, Stanisaw, 606
Søgaard, 634, 641
Søgaard, Anders, 634, 641

T.Y.S.S, 416
T.Y.S.S, Santosh, 416
Ta, 233, 236, 364
Ta, Minh Ngoc, 233, 236
Ta, Tai Tien, 364
Tachibana, 485
Tachibana, Hideyuki, 485
Tachioka, 574
Tachioka, Yuuki, 574
Tadiparthi, 792
Tadiparthi, Vaishnav, 792
Tagarelli, 667
Tagarelli, Andrea, 667
Tagliavini, 468
Tagliavini, Giuseppe, 468
Tagowski, 494
Tagowski, Kamil, 494
Taguchi, 572
Taguchi, Chihiro, 572
Tahmasebi, 173
Tahmasebi, Nina, 173
Tajer, 771
Tajer, Ali, 771
Tak, 83
Tak, Ala N., 83
Takagi, 445
Takagi, Naoya, 445
Takahashi, 237, 328
Takahashi, Hiromu, 328
Takahashi, Kosuke, 237
Takamura, 452
Takamura, Hiroya, 452
Takase, 885
Takase, Yusuke, 885
Takata, 463
Takata, Ryosuke, 463
Takayanagi, 463
Takayanagi, Takehiro, 463
Takeda, 416
Takeda, Rikuho, 416
Takeshita, 107, 480, 862

Takeshita, Michito, 480
Takeshita, Sotaro, 107, 862
Takeuchi, 823
Takeuchi, Koh, 823
Takezoe, 540
Takezoe, Rinyoichi, 540
Tal, 142
Tal, Ro-ee, 142
Talafha, 231
Talafha, Bashar, 231
Talasila, 611
Talasila, Abhiroop, 611
Talat, 462
Talat, Zeerak, 462
Talreja, 464
Talreja, Mohit, 464
Talwalkar, 112
Talwalkar, Ameet, 112
Tam, 449
Tam, Kahou, 449
Tamboli, 192
Tamboli, Dipesh, 192
Tamilselvam, 558
Tamilselvam, Srikanth G., 558
Tamiti, 199
Tamiti, Tarikul Islam, 199
Tamo, 602, 824
Tamo, J. Ben, 824
Tamo, Junior, 602
Tan, 142, 164, 183, 185, 188, 232, 267, 275, 303, 304, 316, 330, 334, 335, 346, 348, 363, 371, 375, 381, 403, 405, 412, 423, 427, 432, 433, 435, 459, 460, 465, 469, 508, 542, 548, 554, 558, 563, 570, 588, 604, 619, 620, 625, 641, 647, 648, 660, 676, 699, 708, 714, 717, 719, 740, 746, 783, 792, 797, 817, 820, 827, 846, 852, 853, 873, 888, 891, 907
Tan, Bryan Chen Zhengyu, 460, 469
Tan, Cheng, 719
Tan, Chenhao, 465, 620, 648
Tan, Chuyi, 405
Tan, Gang, 827
Tan, Guangming, 708
Tan, Haoru, 423, 433
Tan, Hexiang, 371
Tan, Hongye, 403, 427
Tan, Jianchao, 647, 708
Tan, Jie, 164
Tan, Jin, 714
Tan, Jingru, 888
Tan, Kay Chen, 570
Tan, Kexin, 820
Tan, Lide, 459
Tan, Luoda, 548
Tan, Mei, 232
Tan, Mingkui, 335
Tan, Mingtian, 891
Tan, Naiqiang, 676
Tan, Qian, 588
Tan, Qiaoyu, 783
Tan, Qingquan, 619
Tan, Robby T., 142, 275, 740
Tan, Shanwen, 375
Tan, Shaomu, 542, 873
Tan, Shicheng, 625, 746
Tan, Shilei, 334
Tan, Shunwen, 846
Tan, Tao, 316
Tan, Tieniu, 363
Tan, Wenhui, 699
Tan, Xiao, 641
Tan, Xiaobin, 853
Tan, Xiaoyu, 852
Tan, Xin, 412
Tan, Xingwei, 183
Tan, Xu, 304, 432
Tan, Xue, 508
Tan, Xuwei, 907
Tan, Yingshui, 660
Tan, Yongmei, 330
Tan, Yue, 717
Tan, Yuqiao, 316
Tan, Yuting, 346
Tan, Zelin, 714, 797
Tan, Zhaoxuan, 188, 554, 817
Tan, Zhe, 381
Tan, Zhehao, 604
Tan, Zhen, 185, 563, 792
Tan, Zhewen, 303
Tan, Zhijie, 267, 348, 558
Tan, Zhouxing, 435
Tanaka-Ishii, 795
Tanaka-Ishii, Kumiko, 795
Tandazo, 243, 447
Tandazo, Angelo Ortiz, 243, 447
Tang, 118, 124, 126, 141, 143, 146, 149, 154, 164, 169, 189, 203, 210, 239, 244, 260, 261, 262, 265, 268, 269, 270, 271, 284, 286, 296, 304, 309, 335, 336, 338, 343, 347, 352, 371, 383, 385, 389, 395, 397, 408, 412, 414, 420, 422, 425, 431, 444, 451, 454, 479, 519, 535, 539, 541, 545, 550, 560, 566, 584, 592, 620, 621, 627, 628, 636, 641, 646, 651, 652, 675, 678, 680, 682, 687, 690, 692, 698, 707, 728, 732,

733, 734, 736, 737, 744, 752, 761, 768, 777, 786, 789, 793, 796, 807, 815, 821, 829, 843, 848, 851, 875, 881, 888, 896, 897, 908, 909
Tang, Bo, 733
Tang, Chen, 397, 621
Tang, Cheng, 261, 692
Tang, Chenming, 149, 270, 539
Tang, Dezhao, 262
Tang, Fei, 126, 304
Tang, Feilong, 592
Tang, Guo, 690
Tang, Haichuan, 761
Tang, Haixu, 789, 821
Tang, Haochun, 265
Tang, Haoteng, 268
Tang, Haoyu, 651
Tang, Huacong, 768, 815
Tang, Huajin, 843
Tang, Jiabin, 271
Tang, Jiahao, 888
Tang, Jiaqi, 651, 736
Tang, Jiayu, 343
Tang, Jie, 169, 286, 422, 541, 646, 786, 909
Tang, Jiliang, 164, 269, 296
Tang, Jinghua, 395
Tang, Jingxian, 752
Tang, Jinhui, 338
Tang, Jiwei, 535, 675, 698
Tang, Kai, 680
Tang, Ke, 908
Tang, Lexiang, 210
Tang, Ling, 383
Tang, Lingxiao, 389
Tang, Mingjie, 260
Tang, Pingzhi, 444
Tang, Raphael, 519
Tang, Rui, 371
Tang, Ruiming, 414, 777
Tang, Ruixiang, 141
Tang, Shao, 124
Tang, Shengji, 309
Tang, Siliang, 189
Tang, Sizhe, 141
Tang, Tianyi, 560, 897
Tang, Weidong, 636, 737
Tang, Xianfeng, 269, 875, 881
Tang, Xiangyan, 652
Tang, Xiaocheng, 118
Tang, Xiaoying, 451, 566
Tang, Xiaoyu, 620
Tang, Xing, 335
Tang, Xinyu, 848
Tang, Xiu, 347
Tang, Xiuqi, 641
Tang, Xu, 244
Tang, Xuehai, 143, 389, 793
Tang, Xuri, 682
Tang, Yang, 420
Tang, Yehui, 412
Tang, Yingbo, 479
Tang, Yinghao, 628, 687
Tang, Yinhao, 732
Tang, Yixuan, 728, 734, 744
Tang, Yongqiang, 385
Tang, Yung-Chen, 154
Tang, Yuqi, 707
Tang, Zecheng, 678, 733
Tang, Zee Hen, 807
Tang, Zeyu, 454
Tang, Zhanyong, 336
Tang, Zhe, 408, 414, 425
Tang, Zheng, 851
Tang, Zhengyang, 627
Tang, Zhenheng, 786
Tang, Zhenwei, 584
Tang, Zhexiang, 146
Tang, Zhiwen, 203
Tang, Zhuo, 284, 286
Tang, Zichen, 545, 550, 796
Tang, Zihao, 431, 896
Tang, Zinan, 829
Tang, Zirui, 239
Tang, Ziyao, 352
Tang, Ziye, 807
Taniguchi, 237
Taniguchi, Tomoki, 237
Tanjim, 86
Tanjim, Mehrab, 86
Tanner, 799, 871
Tanner, Chris, 799, 871
Tao, 122, 132, 163, 217, 221, 257, 260, 286, 291, 303, 313, 321, 352, 387, 388, 417, 421, 469, 531, 540, 577, 591, 623, 647, 652, 656, 668, 695, 699, 706, 748, 756, 757, 796, 821, 843, 853, 883, 888
Tao, Chaofan, 163, 821
Tao, Chongyang, 257, 623
Tao, Dacheng, 221, 321, 388, 656, 695, 883
Tao, Jialing, 132
Tao, Jianhua, 313, 352, 656, 668, 699, 706
Tao, Jiawen, 303
Tao, Juncheng, 217
Tao, Ming, 748
Tao, Mingxiang, 652
Tao, Ran, 260, 591

Tao, Shengyu, 531
Tao, Shimin, 756, 757
Tao, Shuchang, 122
Tao, Xiaodong, 417
Tao, Xin, 286
Tao, Xiyan, 469
Tao, Yangyu, 540
Tao, Yifei, 888
Tao, Yongding, 531, 577, 796
Tao, Yufei, 291
Tao, Zeyin, 421
Tao, Zhang, 647
Tao, Zhen, 303
Tao, Zhengwei, 853
Tao, Zhimin, 843
Tao, Zui, 387
Tao;,
Tao;, Yaguang,
Tark, 533
Tark, Jiwon, 533
Tarun, 784
Tarun, Ayush Kumar, 784
Tasawong, 169, 902
Tasawong, Panuthep, 169, 902
Tatarinov, 250
Tatarinov, Nikita, 250
Tatariya, 246
Tatariya, Kushal, 246
Tateno, 156
Tateno, Kei, 156
Tavanapong, 119, 567
Tavanapong, Wallapak, 119, 567
Tavernini, 278
Tavernini, Gabriel, 278
Tavor, 610
Tavor, Ateret Anaby, 610
Taxali, 772
Taxali, Aman, 772
Tay, 405
Tay, Wee Peng, 405
Taylor, 108
Taylor, Samuel, 108
Tchrakian, 493
Tchrakian, Tigran T., 493
Te, 605
Te, Cao, 605
Tegtmeyer, 231
Tegtmeyer, Kyle, 231
Teh, 879
Teh, Kai Jun, 879
Tei, 605
Tei, Jin, 605
Teiletche, 784
Teiletche, Paul, 784
Teixeirinha, 705
Teixeirinha, Pedro, 705
Teleki, 180
Teleki, Maria, 180
Tellex, 790
Tellex, Stefanie, 790
Teng, 114, 374, 396, 397, 660, 680, 689, 696, 703, 759, 819, 856
Teng, Chong, 114, 856
Teng, Hao, 680, 689
Teng, Qingyang, 660
Teng, Qixun, 397, 759
Teng, Wenhao, 396, 703, 819
Teng, Yan, 696
Teng, Yeqing, 374
Tenkanen, 130
Tenkanen, Henrikki, 130
Tenneti, 611
Tenneti, Tejaswi, 611
Teo, 186
Teo, Yon Shin, 186
Terberger, 527
Terberger, Maximilian, 527
Terzi, 99
Terzi, Evimaria, 99
Terzopoulos, 424
Terzopoulos, Demetri, 424
Testoni, 103
Testoni, Alberto, 103
Tewari, 782
Tewari, Tulika, 782
Tey, 891
Tey, Joseph, 891
Tezcan, 622
Tezcan, Arda, 622
Thai, 233, 256, 879, 905
Thai, Katherine, 233
Thai, My T., 256
Thai, Zhen Leng, 879
Thai, Zhen_Leng,
Thakar, 516
Thakar, Krishna, 516
Thakur, 118, 161, 224, 240
Thakur, Nandan, 161
Thakur, Rajeev, 118
Thakur, Shailja, 224, 240
Thangali, 606, 617
Thangali, Avinash, 606, 617
Thangarasa, 279
Thangarasa, Vithursan, 279
Thanh, 701
Thanh, Tran Tan, 701

Thanh-Tung, 281
Thanh-Tung, Hoang, 281
Thapa, 222, 441, 471
Thapa, Rahul, 441
Thapa, Surendrabikram, 222, 471
Thareja, 442
Thareja, Rushil, 442
Thellmann, 542
Thellmann, Klaudia-Doris, 542
Theologitis, 549
Theologitis, Michael, 549
Thiem, 505
Thiem, Gabriel, 505
Thille, 891
Thille, Candace, 891
Thomas, 222
Thomas, Chris, 222
Thomason, 230, 582, 843
Thomason, Jesse, 230, 582, 843
Thompson, 516
Thompson, Dr. Emmanuel, 516
Thong, 745
Thong, Mai Huy, 745
Thrampoulidis, 206
Thrampoulidis, Christos, 206
Thronson, 875
Thronson, Emma, 875
Thu, 471
Thu, Ye Kyaw, 471
Tian, 97, 100, 107, 111, 158, 175, 191, 204, 245, 251, 261, 264, 276, 284, 295, 296, 316, 321, 322, 329, 332, 339, 384, 389, 401, 404, 442, 449, 450, 469, 509, 511, 515, 529, 531, 541, 550, 584, 619, 620, 629, 652, 665, 671, 676, 694, 714, 723, 725, 729, 738, 741, 750, 761, 775, 781, 831, 836, 850, 901
Tian, Changxin, 694
Tian, Changyao, 339
Tian, Chen, 284
Tian, Chris Xing, 204
Tian, Chuan, 175
Tian, Chunlin, 449
Tian, Cong, 316, 329
Tian, Fuze, 97
Tian, Haolin, 550
Tian, Haoye, 515, 531
Tian, Harry Yizhou, 107
Tian, Hui, 714
Tian, Jiahe, 450
Tian, Jiayu, 775
Tian, Jinchuan, 111, 529
Tian, Jingyuan, 831
Tian, Jinjin, 175
Tian, Kaiwen, 332
Tian, Kaiyuan, 665
Tian, Mingkai, 729
Tian, Muzhao, 321
Tian, Qixin, 836
Tian, Shiqiang, 245
Tian, Shuangshuang, 264
Tian, Shulin, 100
Tian, Tengfei, 676
Tian, Wei, 469, 850
Tian, Wenhong, 619, 671
Tian, Xiang, 389
Tian, Xin, 251
Tian, Xinmei, 322
Tian, Xiuqi, 761
Tian, Xueyun, 584, 901
Tian, Yanzhi, 541
Tian, Yijun, 191, 442, 620
Tian, Yizhou, 295
Tian, Yonghong, 158
Tian, Yu, 296, 401, 629, 652
Tian, Yuan, 261, 264, 781
Tian, Yucheng, 511
Tian, Yueying, 738
Tian, Yuxing, 276, 509
Tian, Zeying, 384
Tian, Zhiheng, 741
Tian, Zhiliang, 401, 723, 725
Tian, Zhongbo, 750
Tian, Zhuotao, 404
Tianjiao, 818
Tianjiao, Li, 818
Tie, 350, 640
Tie, Guiyao, 350, 640
Tikhomirova, 663
Tikhomirova, Taisiia, 663
Tillman, 127
Tillman, Robert E., 127
Ting, 244, 585
Ting, Kai Ming, 585
Ting, Yujan, 244
Titiya, 469
Titiya, Prasham, 469
Titov, 276
Titov, Alexandra BirchIvan, 276
Tiwari, 233, 239, 305, 565, 606, 649
Tiwari, Ashish, 239
Tiwari, Prayag, 233, 565, 649
Tiwari, Shivam, 606
Tiwari, Utsav, 305
Tiyyala, 558
Tiyyala, Vijay Murari, 558

Tjalk-Bøggild, 238
Tjalk-Bøggild, Rasmus, 238
Tjhi, 169
Tjhi, William Chandra, 169
Tjiaranata, 862
Tjiaranata, Filbert Aurelian, 862
Tjuatja, 246
Tjuatja, Lindia, 246
Tocchetti, 200
Tocchetti, Andrea, 200
Tokala, 455
Tokala, Santosh, 455
Toker, 286, 887
Toker, Gilat, 286
Toker, Michael, 887
Toksoz, 827
Toksoz, Levent, 827
Tokunaga, 854
Tokunaga, Narumi, 854
Tomar, 236
Tomar, Raj Vardhan, 236
Tomeh, 601
Tomeh, Nadi, 601
Tommasi, 713
Tommasi, Marc, 713
Tonellotto, 519
Tonellotto, Nicola, 519
Tong, 81, 98, 152, 181, 228, 234, 262, 333, 360, 410, 446, 469, 575, 586, 693, 695, 717, 820, 823, 826, 846, 898
Tong, Bin, 234
Tong, Chao, 333
Tong, Haibo, 98
Tong, Hanghang, 81, 446, 469, 575, 826
Tong, Jingqi, 228, 693, 717, 820
Tong, Jizhou, 898
Tong, Panrong, 410
Tong, Sherry T., 823
Tong, Weixi, 586
Tong, Xin, 695
Tong, Yicheng, 181
Tong, Yongqi, 846
Tong, Yulai, 262
Tong, Zhengyang, 152
Tong, Zihe, 360
Tonglet, 218, 817
Tonglet, Jonathan, 218, 817
Tonja, 862
Tonja, Atnafu Lambebo, 862
Toroghi, 764
Toroghi, Armin, 764
Torous, 230
Torous, John, 230
Torr, 716, 797
Torr, Philip, 716, 797
Torres-Fonseca, 902
Torres-Fonseca, Josue, 902
Touchent, 253, 282
Touchent, Rian, 253, 282
Toughrai, 862
Toughrai, Yassine, 862
Toyama, 462
Toyama, Kentaro, 462
Toyin, 589
Toyin, Hawau Olamide, 589
Toyoda, 190
Toyoda, Masashi, 190
Trabelsi, 146
Trabelsi, Amine, 146
Trager, 222, 226
Trager, Jackson, 222, 226
Trakic, 249
Trakic, Adnan, 249
Tramèr, 784
Tramèr, Florian, 784
Tran, 355, 510, 542, 544, 574, 645, 686, 701, 882
Tran', 574
Tran', Vinh-Khanh, 574
Tran, Dien X., 701
Tran, Duc-Luong, 686
Tran, Ke, 542
Tran, Khang, 510
Tran, Khanh-Tung, 574
Tran, Quoc-Huy, 355
Tran, Son, 544
Tran, Thang Duc, 645
Tran, Truyen, 882
Trappolini, 519
Trappolini, Giovanni, 519
Trautwein, 495
Trautwein, Ulrich, 495
Trenous, 542
Trenous, Sony, 542
Tresp, 191, 289
Tresp, Volker, 191, 289
Triantafyllopoulos,
Triantafyllopoulos, Ilias,
Trieu, 747
Trieu, Long, 747
Tripathi, 152
Tripathi, Subarna, 152
Tripodi, 813
Tripodi, Ignacio J., 813
Trivedi, 452, 469
Trivedi, Harsh, 452
Trivedi, Jainil, 469

Trott, 108, 834
Trott, Sean, 108, 834
Truong, 156, 277, 536, 560, 645
Truong, Bao Tran, 560
Truong, Do Dinh, 536
Truong, Minh-Phuc, 645
Truong, Van Q., 156
Truong, Vu Tuan, 277
Tsai, 467, 547
Tsai, Ya An, 467, 547
Tsai, Yun-Da, 467, 547
Tsang, 161, 253, 882
Tsang, Hong Ting, 161, 253, 882
Tsao, 312
Tsao, Yu, 312
tSaoir, 287
tSaoir, Fírinne Nic an, 287
Tsarfaty, 452, 585, 804
Tsarfaty, Reut, 452, 585, 804
Tse, 304
Tse, Timothy Tin-Long, 304
Tseng, 110, 143
Tseng, Liang-Hsuan, 110
Tseng, Wei-Cheng, 143
Tshinu, 98
Tshinu, Kanda Patrick, 98
Tsipidi, 243
Tsipidi, Eleftheria, 243
Tsujii, 166, 233
Tsujii, Junichi, 166, 233
Tsujimura, 452
Tsujimura, Tomoki, 452
Tsunoo, 111
Tsunoo, Emiru, 111
Tsuruoka, 542
Tsuruoka, Yoshimasa, 542
Tsutsukawa, 499
Tsutsukawa, Kazuki, 499
Tsvetkov, 170, 446, 550, 554
Tsvetkov, Yulia, 170, 446, 550, 554
Tsvigun, 236
Tsvigun, Akim, 236
Tsvilodub, 765
Tsvilodub, Polina, 765
Tsymbalov, 105
Tsymbalov, Evgenii, 105
Tsymboi, 640
Tsymboi, Olga, 640
TT, 344
TT, TT, 344
Tu, 127, 131, 132, 136, 149, 158, 209, 228, 239, 263, 360, 373, 381, 425, 476, 488, 501, 556, 568, 630, 638, 652, 658, 716, 722, 725, 739, 804, 841, 856
Tu, Chengying, 209, 501
Tu, Dandan, 360, 381, 658, 739
Tu, Eng-Shen, 263
Tu, Fang, 131, 136
Tu, Haoran, 856
Tu, Hongkui, 373
Tu, Jianhong, 127, 158
Tu, Jinzhe, 132
Tu, Kai, 556
Tu, Kewei, 841
Tu, Minzhu, 638
Tu, Qian, 228
Tu, Shangqing, 476, 488, 804
Tu, Songjun, 568
Tu, Wenxuan, 652
Tu, Xinhui, 725
Tu, Zhaopeng, 149, 425, 716, 722
Tu, Zhengzhong, 630
Tu, Zhiying, 630
Tu, Zhongying, 239
Tuckute, 447
Tuckute, Greta, 447
Tung, 744
Tung, Anthony Kum Hoe, 744
Tuo, 692
Tuo, Jiayi, 692
Tur, 151, 441, 509, 573
Tur, Gokhan, 151, 441, 509, 573
Turmakhan, 874
Turmakhan, Diana, 874
Tutek, 90
Tutek, Martin, 90
Tutubalina, 101, 344, 471, 493, 576, 714
Tutubalina, Anna Borisiuk; Andrey Savchenko; Alexander Panchenko; Elena,
Tutubalina, Elena, 471, 576, 714
Tutubalina, Mikhail Seleznyov; Daniil Korbut; Viktor Moskvoretskii; Oleg Somov; Alexander Panchenko; Elena,
Tutubalina, Pavel Tikhonov; Ivan Oseledets; Elena,
Tuytelaars, 215, 218, 817
Tuytelaars, Tinne, 215, 218, 817
Twist, 195
Twist, Lukas, 195
Tyagi, 779
Tyagi, Utkarsh, 779
Tzimiropoulos, 291
Tzimiropoulos, Georgios, 291
Tài, 701
Tài, Dng Vn, 701

Uchendu, 85, 519, 617
Uchendu, Adaku, 85, 617
Uchendu, Angelina Tsai; Shreya Subramanian; Catherine Liu; Kimberly Lopez; Leif Zinn-Brooks; Alexia Schulz; Adaku,
Uchendu, Uchendu, 85
Uddin, 554
Uddin, Md Nayem, 554
Udomcharoenchaikit,
Udomcharoenchaikit, Can,
Udsa, 205
Udsa, Tinnakit, 205
Uehara, 601
Uehara, Kohei, 601
Ukarapol, 582
Ukarapol, Trapoom, 582
Umap, 150
Umap, Aarush, 150
Umar, 429
Umar, Umar Baba, 429
Unanue, 293
Unanue, Inigo Jauregi, 293
Ungar, 236
Ungar, Lyle, 236
Uniyal, 160
Uniyal, Mansi, 160
Unnikrishnan, 544
Unnikrishnan, Jayakrishnan, 544
Upadhyay, 477, 505, 571, 575
Upadhyay, Mihir, 571
Upadhyay, Ritam, 505
Upadhyay, Shyam, 575
Upadhyay, Vikas, 477
Uppadhyay, 837
Uppadhyay, Manan, 837
Upravitelev, 84
Upravitelev, Max, 84
Urooj, 198
Urooj, Midhat, 198
Ushaw, 407
Ushaw, Gary, 407
Ustun, 129
Ustun, Volkan, 129
Utama, 266
Utama, Prasetya Ajie, 266
Uziel, 783
Uziel, Guy, 783
Uzzi, 867
Uzzi, Brian, 867
Uçar, 779
Uçar, Arda, 779

Vabbilisetty, 443
Vabbilisetty, Saipraveen, 443
Vaculin, 610
Vaculin, Roman, 610
Vadithya, 862
Vadithya, Gouthami, 862
Vaduguru, 598
Vaduguru, Saujas, 598
Vajjala, 862
Vajjala, Sowmya, 862
Vajrala, 305
Vajrala, Sowmya, 305
Valentini, 246
Valentini, Maria, 246
Valentino, 183
Valentino, Marco, 183
Valle, 529
Valle, Rafael, 529
Valvoda, 481
Valvoda, Josef, 481
Vamvas, 247, 784
Vamvas, Jannis, 247, 784
Van, 405, 411, 426, 427, 645
Van, Linh Ngo, 405, 411, 426, 427, 645
VandeVondele, 784
VandeVondele, Joost, 784
Vanmassenhove, 406
Vanmassenhove, Eva, 406
Vanover, 845
Vanover, Brian, 845
Varadarajan, 207
Varadarajan, Vasudha, 207
Varda, 447
Varda, Andrea Gregor de, 447
Vargas, 222
Vargas, Francielle, 222
Varia, 271
Varia, Siddharth, 271
Varma, 860
Varma, Vasudeva, 860
Varol, 903
Varol, Gül, 903
Varshney, 399
Varshney, Vaibhav, 399
Vasan, 469
Vasan, Keertana Arun, 469
Vashurin, 126
Vashurin, Roman, 126
Vasisht, 812
Vasisht, Kinshuk, 812
Vasu, 582
Vasu, Sai Suresh Macharla, 582
Vats, 609, 649
Vats, Vaibhav, 609

Vats, Vidushee, 649
Vaughan, 862
Vaughan, Thom, 862
Vazhentsev, 233
Vazhentsev, Artem, 233
Vazirgiannis, 280, 822, 865
Vazirgiannis, Michalis, 280, 822, 865
Vecchietti, 893
Vecchietti, Luiz Felipe, 893
Vechev, 593
Vechev, Martin, 593
Vedula, 835
Vedula, Nikhita, 835
Veerakanjana, 823
Veerakanjana, Kanyakorn, 823
Veeravalli, 509
Veeravalli, Venugopal, 509
Velasquez, 263
Velasquez, Alvaro, 263
Velcin, 598
Velcin, Julien, 598
Veliz, 471
Veliz, Rudy Alexandro Garrido, 471
Veloso, 766
Veloso, Leonor, 766
Veluri, 235
Veluri, Bandhav, 235
Vemireddy, 614
Vemireddy, Karthik, 614
Venieris, 223
Venieris, Stylianos I., 223
Venkat, 572
Venkat, Anirudhh, 572
Venkatasubramanian,
Venkatasubramanian, Suresh,
Venkatayogi, 106
Venkatayogi, Siddhartha, 106
Venkatesh, 351
Venkatesh, Svetha, 351
Venkit, 609
Venkit, Pranav Narayanan, 609
Veprikov, 761
Veprikov, Andrey, 761
Vera, 291
Vera, Alhim, 291
Veraldi, 784
Veraldi, Livio, 784
Verberne, 349, 631, 736
Verberne, Suzan, 349, 631, 736
Vergari, 805
Vergari, Antonio, 805
Verma, 122, 188, 294, 342, 607, 613, 892
Verma, Aastha A K, 342
Verma, Akshay, 613
Verma, Arun, 294
Verma, Divyansh, 122
Verma, Raeyoung Chang; Dongwook Kwon; Jisoo Lee; Nikhil,
Verma, Rajat, 188
Verma, Saurabh, 892
Verspoor, 308, 816
Verspoor, Karin, 308, 816
Veselovsky, 821
Veselovsky, Veniamin, 821
Vetcha, 383
Vetcha, Nitin, 383
Viaud, 169
Viaud, Gautier, 169
Vidigal, 769
Vidigal, Pedro, 769
Vieira, 110, 842
Vieira, Tim, 110, 842
Vijjini, 105
Vijjini, Anvesh Rao, 105
Villa-Cueva, 98
Villa-Cueva, Emilio, 98
Villar, 243
Villar, Angel, 243
Villavicencio, 176, 700
Villavicencio, Aline, 176, 700
Vinayak, 480
Vinayak, Ramya Korlakai, 480
Virinchi, 731
Virinchi, Srinivas, 731
Vishwakarma;,
Vishwakarma;, Dinesh Kumar,
Vitsakis, 438
Vitsakis, Nikolas, 438
Vittor, 667
Vittor, Lucas, 667
Vlachos, 503, 864
Vlachos, Andreas, 503, 864
Vlassis, 584
Vlassis, Nikos, 584
Vlassopoulos, 713
Vlassopoulos, Georgios, 713
Vlassov;,
Vlassov;, Vladimir,
Vo, 364
Vo, Khang Hoang Nhat, 364
Voinea;,
Voinea;, Andrei,
Volkov, 563
Volkov, Ross, 563
Volkova, 466
Volkova, Svitlana, 466

Volodichev, 685
Volodichev, Andrei, 685
Volovikova, 424
Volovikova, Zoya, 424
Vondrick, 150
Vondrick, Carl, 150
VONG, 620
VONG, Chi Man, 620
Vorobeychik, 90, 92
Vorobeychik, Yevgeniy, 90, 92
Vos, 565
Vos, Ida L., 565
Vosoughi, 115, 136, 466, 534, 569, 860
Vosoughi, Soroush, 115, 136, 466, 534, 569, 860
Vosoughi;,
Vosoughi;, Soroush,
Vozila, 615
Vozila, Paul, 615
Vrabie, 804
Vrabie, Valeriu, 804
Vraga, 234
Vraga, Emily K., 234
Vrandecic, 246
Vrandecic, Denny, 246
Vu, 192, 211, 340, 358, 645, 851
Vu, Hai An, 645
Vu, Thanh, 340
Vu, Thuy, 851
Vu, Thuy-Trang, 211, 358
Vu, Tung D., 192
Vujanic, 537
Vujanic, Robin, 537
Vukovic, 89
Vukovic, Renato, 89
Vuli, 845
Vuli, Ivan, 845
Vuong, 779
Vuong, Tyler, 779
Vutla, 314
Vutla, Sasanka, 314
Vyalla, 209
Vyalla, Rishitej Reddy, 209
Vyas, 82
Vyas, Mayank, 82
Vylomova, 167
Vylomova, Ekaterina, 167
Vázquez, 891
Vázquez, Raúl, 891
Vöckler, 851
Vöckler, Jens-S., 851

W, 431, 764
W, Ruifan, 764
W, Xing, 431
Wachsmuth, 102, 857
Wachsmuth, Henning, 102, 857
Wadhwa, 566
Wadhwa, Sahil, 566
Wagley, 571
Wagley, Benjamin, 571
Wahle, 174, 551, 620, 876
Wahle, Jan Philip, 174, 551, 620, 876
Waibel, 779
Waibel, Alexander, 779
Wairagala, 98
Wairagala, Eric Peter, 98
Wakamiya, 252, 307
Wakamiya, Shoko, 252, 307
Waki, 156
Waki, Satoshi, 156
Walden, 154, 555
Walden, William, 555
Walden, William Gantt, 154
Waldendorf, 105
Waldendorf, Jonas, 105
Wali, 716, 862
Wali, Ahmad Mustapha, 716, 862
Wallace, 104, 106
Wallace, Byron C, 106
Wallace, Byron C., 104
Walter, 505
Walter, Susanne, 505
WAN, 687
Wan, 133, 158, 165, 173, 188, 221, 248, 274, 279, 292, 334, 344, 363, 383, 386, 392, 416, 418, 435, 445, 456, 462, 479, 504, 529, 544, 592, 635, 636, 652, 684, 695, 697, 703, 706, 714, 716, 728, 755, 767, 786, 794, 797, 799, 821, 871, 882, 888, 892
Wan, David, 173, 821
Wan, Dazhen, 418, 652
Wan, Fangi, 292
Wan, Fanqi, 695
Wan, Feng, 274
Wan, Frank, 280
Wan, Guancheng, 504, 716, 797
Wan, Guangya, 786
Wan, Herun, 363, 888
Wan, Hong, 882
Wan, Huaiyu, 221, 703
Wan, Jiayong, 892
Wan, Jikun, 133
Wan, Jiuheng, 714, 755
Wan, Jun, 635
Wan, Mengting, 479
Wan, Pengfei, 456

Wan, Ruyuan, 165
Wan, Tao, 728
Wan, Xiang, 416
Wan, Xiaojun, 334, 383, 386, 392, 445, 652, 684, 697, 706
Wan, Xinyan, 636
Wan, Yang, 888
Wan, Yao, 344, 435, 799
Wan, Yixin, 188, 544, 767
Wan, Yuchen, 248, 871
Wan, Yulong, 435
WAN, Yuxuan, 687
Wan, Zhen, 529
Wan, Zhenglin, 158
Wan, Zhennan, 797
Wan, Zihao, 462
Wan, Zimu, 592
Wan, Ziyu, 794
WANG, 149, 185, 197, 347, 398, 436, 440, 451, 496, 630, 701, 903
Wang, 82, 84, 85, 86, 87, 89, 90, 91, 93, 94, 95, 96, 97, 103, 107, 108, 109, 111, 112, 113, 114, 117, 121, 124, 127, 128, 129, 130, 131, 132, 133, 134, 135, 137, 138, 139, 140, 141, 142, 143, 144, 145, 146, 147, 148, 149, 151, 152, 153, 155, 158, 159, 161, 164, 166, 168, 169, 170, 171, 172, 175, 177, 178, 180, 181, 182, 183, 184, 185, 186, 187, 189, 193, 194, 195, 197, 199, 201, 202, 204, 205, 209, 211, 212, 213, 214, 215, 219, 220, 221, 224, 225, 226, 228, 229, 230, 231, 232, 234, 235, 236, 237, 239, 242, 244, 248, 249, 250, 251, 253, 254, 258, 259, 260, 261, 262, 264, 265, 267, 268, 269, 270, 271, 272, 274, 276, 278, 279, 280, 281, 284, 285, 289, 293, 294, 295, 296, 297, 298, 300, 303, 304, 305, 307, 310, 311, 313, 314, 316, 317, 318, 319, 320, 321, 322, 323, 325, 326, 327, 330, 332, 333, 334, 335, 336, 339, 340, 341, 344, 345, 346, 347, 348, 349, 350, 352, 353, 354, 355, 356, 357, 358, 359, 361, 362, 363, 364, 369, 371, 372, 373, 374, 375, 376, 377, 378, 379, 380, 381, 382, 383, 384, 385, 386, 387, 388, 389, 391, 392, 393, 394, 395, 396, 397, 398, 399, 400, 401, 403, 404, 405, 406, 407, 408, 409, 410, 412, 413, 414, 415, 416, 418, 419, 420, 421, 422, 423, 424, 425, 427, 430, 431, 432, 433, 434, 435, 436, 437, 441, 442, 443, 444, 445, 446, 448, 449, 450, 451, 452, 455, 456, 457, 458, 459, 460, 461, 462, 467, 469, 471, 473, 474, 475, 476, 477, 479, 482, 483, 484, 486, 490, 491, 493, 494, 495, 497, 498, 500, 501, 502, 503, 504, 505, 506, 507, 508, 509, 514, 516, 518, 519, 527, 528, 529, 530, 531, 532, 533, 534, 535, 537, 538, 539, 540, 541, 543, 544, 545, 546, 548, 550, 551, 554, 558, 559, 560, 561, 562, 563, 566, 568, 569, 570, 571, 575, 576, 577, 578, 582, 583, 584, 585, 587, 588, 590, 591, 593, 599, 600, 601, 602, 604, 606, 608, 611, 614, 617, 618, 619, 620, 621, 623, 624, 625, 626, 627, 629, 630, 631, 634, 635, 636, 637, 638, 639, 640, 641, 643, 644, 645, 646, 647, 648, 649, 651, 653, 654, 655, 656, 657, 658, 659, 660, 661, 662, 663, 664, 665, 666, 669, 670, 671, 673, 674, 675, 676, 677, 678, 679, 680, 682, 683, 684, 685, 686, 687, 688, 689, 691, 692, 694, 695, 696, 697, 698, 699, 700, 701, 702, 703, 704, 705, 706, 708, 709, 712, 713, 714, 716, 717, 719, 720, 721, 722, 723, 724, 725, 726, 727, 728, 729, 730, 732, 733, 734, 735, 736, 737, 738, 739, 740, 741, 742, 743, 744, 745, 746, 747, 748, 749, 750, 752, 753, 755, 756, 757, 758, 759, 760, 761, 762, 764, 765, 767, 768, 769, 771, 772, 773, 774, 775, 776, 777, 778, 780, 781, 782, 783, 784, 785, 786, 787, 788, 789, 790, 791, 792, 793, 794, 795, 797, 798, 800, 801, 802, 803, 805, 807, 809, 811, 814, 815, 817, 818, 819, 820, 821, 823, 824, 826, 827, 828, 832, 834, 835, 836, 838, 844, 845, 848, 849, 850, 851, 852, 854, 855, 857, 858, 859, 861, 863, 864, 865, 866, 867, 868, 870, 871, 874, 875, 876, 877, 878, 880, 881, 882, 885, 886, 887, 888, 889, 892, 893, 894, 895, 896, 897, 898, 899, 900, 901, 903, 904, 905, 906, 907, 908, 909
wang, 327, 408, 425
Wang*, 857
Wang*, Yuyang, 857
Wang, Alex, 376
Wang, Alex Jinpeng, 624, 888
Wang, Ante, 397, 482, 855
Wang, Ao, 332, 387
Wang, Bang, 310
Wang, Baojun, 726
Wang, Baosheng, 733, 735
Wang, Baoxin, 395, 697
Wang, Beilun, 625, 641, 798
Wang, Benlu, 231

Wang, Benyou, 416, 474, 627, 686, 705, 743
Wang, Bin, 239, 344, 688
Wang, Bingbing, 692
Wang, Binghai, 897
Wang, Bingxuan, 379
Wang, Bingyang, 908
Wang, Binwu, 313
Wang, Bixuan, 730
Wang, Bo, 89, 204, 678, 906
Wang, Bo-Jyun, 514
Wang, Bochao, 540
Wang, Boran, 187
Wang, Botao, 392
Wang, Boxin, 887
Wang, Boxuan, 376
Wang, Boyan, 657
Wang, Bryan, 835
Wang, Can, 388, 745, 746
WANG, Canran, 630
Wang, Changbo, 734
Wang, Chao, 433, 692, 730, 805
Wang, Chaojun, 469
Wang, Chaokun, 541
Wang, Chaoren, 199
Wang, Chen, 656
Wang, Cheng, 225, 886
Wang, Cheng-Long, 378
Wang, Chengbing, 648
Wang, Chengjie, 724
Wang, Chenglin, 689
Wang, Chenglong, 456, 626, 748
Wang, Chengpeng, 184, 828
Wang, Chenguang, 127, 898
Wang, Chengye, 907
Wang, Chengyou, 363
Wang, Chengyu, 87, 562, 714, 755, 896
Wang, Chentai, 630
Wang, Chenxi, 876
Wang, Chenxu, 349
Wang, Chong, 906
Wang, Chuan-Ju, 870
Wang, Chuang, 401
Wang, Chuhan, 195, 514
Wang, Chunlei, 310
Wang, ChunMing, 254, 327, 768
Wang, Cong, 96, 97, 180, 193, 221, 399, 719, 801
Wang, Cunxiang, 497, 541, 851, 909
Wang, Dakuo, 824, 826, 861, 875
Wang, Daling, 326, 422, 474, 643, 663, 671, 673, 730
Wang, Dan, 369
Wang, Danding, 96, 170
Wang, Dawei, 407
Wang, Dayu, 849
Wang, Dequan, 665
Wang, Di, 171, 268, 378, 413, 540, 602, 798, 803, 886
Wang, Dianyun, 285
Wang, Ding, 148
Wang, Dingcheng, 775
Wang, Dingmin, 881
Wang, Dongjie, 220, 530
Wang, Dongxia, 235, 405
Wang, En, 357
Wang, Enming, 425
Wang, Enze, 735
Wang, Enzhi, 744
Wang, Fanfan, 414
Wang, Fangdong, 239
Wang, Fangyuan, 747
Wang, Fanyu, 249
Wang, Fei, 699, 792
Wang, Feifei, 135
Wang, Feng, 905
Wang, Fengtao, 412
Wang, Fu Lee, 852
Wang, Futing, 264
Wang, Fuyu, 459
Wang, Gang, 539, 783
Wang, Gengyu, 554
Wang, Guan, 143, 234, 793
Wang, Guanbo, 138
Wang, Guanghui, 372
Wang, Guangjing, 177
Wang, Guangming, 544
Wang, Guangrun, 147
Wang, Guangyu, 239, 827
Wang, Guanjun, 713
Wang, Guanyu, 267, 348, 558
Wang, Guoqiu, 422
Wang, Guoren, 323, 725
Wang, Guoxia, 297
Wang, Guoyin, 237
Wang, Hai, 505
Wang, Haicheng, 432
Wang, Haifeng, 169, 297, 304, 724, 746
Wang, Hailing, 618
Wang, Hairui, 601, 794
Wang, Haishuai, 91, 374, 420
Wang, Haiwei, 278
Wang, Haixun, 611
Wang, Haiyang, 353
Wang, Hanbin, 212
Wang, Hanchen, 897
Wang, Handing, 752
Wang, Hang, 109, 791

Wang, Hanli, 319
Wang, Hanpin, 670
Wang, Hansi, 172
Wang, Hao, 87, 197, 204, 205, 244, 272, 298, 314, 347, 406, 519, 583, 653, 664, 729, 732, 756, 788, 819, 894
Wang, Haoan, 443
Wang, Haobo, 680, 728, 758
Wang, Haochang, 323
Wang, Haochun, 378, 761
Wang, Haofen, 224
Wang, Haohan, 584
Wang, Haohua, 896
Wang, Haohui, 232
Wang, Haoliang, 584, 906
Wang, Haonan, 248, 535
Wang, Haosen, 391, 713
Wang, Haotian, 721
Wang, Haoyu, 455, 771, 794, 908
Wang, Haozhao, 660
Wang, Haozhe, 599
Wang, Haozhong, 653
Wang, Heng, 554, 757
Wang, Hepeng, 626, 643
Wang, Heran, 648
Wang, Hong, 745, 746
Wang, Hongning, 132, 164, 398, 497, 786, 851, 909
WANG, Hongru, 149, 398, 436, 440, 451, 903
Wang, Hongru, 737, 746
Wang, Hongshuai, 740
Wang, Hongwei, 393, 611
Wang, Houfeng, 117, 399, 659
Wang, Hua, 430
Wang, Huacan, 278, 341, 405, 467, 653, 786, 836, 899
Wang, Hualei, 896
Wang, Huanjie, 414
Wang, Huazheng, 327, 388
Wang, Hui, 300, 352, 432
Wang, Hui-Po, 582
Wang, Huili, 375, 421
Wang, Huimin, 184
Wang, Huiyang, 310, 345
Wang, Jason, 785
Wang, Jeffrey George, 785
Wang, Ji Hun, 608
Wang, Jiacheng, 591, 619
Wang, Jiachuan, 483
Wang, Jiahao, 334, 392, 773
Wang, Jiahui, 752
Wang, Jiajun, 906
Wang, Jiakang, 701
wang, Jiaming, 408, 425
Wang, Jian, 391, 590, 604, 663, 742, 900
Wang, Jianghui, 505
Wang, Jiangyi, 476
Wang, Jiangyuan, 614
Wang, Jianing, 659, 670
Wang, Jianjin, 456
Wang, Jianlei, 400, 425
Wang, Jianlin, 350
Wang, Jiannan, 818
Wang, Jianwei, 680
Wang, Jianyong, 900
Wang, Jianyu, 845
Wang, Jianzhou, 260
Wang, Jianzong, 392
Wang, Jiapeng, 371
Wang, Jiaqi, 228, 239, 363
Wang, Jiashu, 878
Wang, Jiashuo, 391
Wang, Jiaxin, 213
Wang, Jiaxuan, 801
Wang, Jiayi, 749
Wang, Jiayu, 874
Wang, Jie, 625
Wang, Jiecong, 713
Wang, Jiexin, 340, 654
Wang, Jieyi, 113
Wang, Jihong, 721
Wang, Jin, 551
Wang, Jinbo, 353
Wang, Jindong, 767
Wang, Jing, 251
Wang, Jingang, 151, 209, 335, 355, 371, 501, 739
Wang, Jingchao, 738
Wang, Jinghan, 251
Wang, Jingtan, 811
Wang, Jingwen, 662
Wang, Jingya, 158
Wang, Jingyi, 330, 623
Wang, Jingyu, 251, 327, 388
Wang, Jingyuan, 537, 588
Wang, Jinpeng, 393, 717
Wang, Jinqiao, 415
Wang, Jinrui, 375, 421
Wang, Jinzhuo, 824
Wang, Jiongxiao, 442
Wang, Jiye, 413
Wang, Jiyuan, 395
Wang, Jize, 221
Wang, Juan, 97
Wang, Jun, 121, 161, 385, 721, 877
Wang, Junfeng, 725
Wang, Junjie, 140, 311, 475, 604, 639, 749, 858
Wang, Junjue, 772

Wang, Junli, 874
Wang, Junling, 835
Wang, Junpeng, 787
Wang, Junqi, 424
Wang, Junyu, 455
Wang, Junzhe, 387, 545, 820
Wang, Kai, 310, 341, 354, 680
Wang, Kailong, 697
Wang, Kang, 392
Wang, Kangyu, 380
Wang, Ke, 107, 146, 339, 432, 646, 655, 739, 864
Wang, Kesu, 344
Wang, Kexue, 270
Wang, Keyi, 166, 538, 688, 848
Wang, Keze, 147, 219, 590, 720, 899, 900
Wang, Kuang-Da, 103
Wang, Kun, 205, 319, 375, 582, 621, 749
Wang, Lean, 537
Wang, Lecheng, 577
WANG, LEI, 197, 347
Wang, Lei, 460
Wang, Leye, 155
Wang, Liang, 89, 634, 659, 698, 749
Wang, Liangjia, 545
Wang, Lichao, 748
Wang, Lihao, 268
Wang, Lijuan, 112
Wang, Linghe, 563
Wang, Lingzhi, 645, 674
Wang, Longbiao, 455, 456
Wang, Longyue, 296, 379, 406, 441, 548, 677, 684, 729, 732, 798
Wang, Longzheng, 142, 276
Wang, Lu, 133, 316, 349, 355, 450, 461, 575, 764
Wang, Lucy Lu, 168
Wang, Ludi, 319
Wang, Luhan, 409
Wang, Maolin, 214, 587
Wang, May, 602
Wang, May Dongmei, 824
Wang, Mei, 237, 874
Wang, Meihong, 665
Wang, Meiyang, 354
Wang, Meng, 561, 647
WANG, MengChang, 185
Wang, Mengdi, 460
Wang, Mengna, 631
Wang, Mengru, 149, 425, 708
Wang, Mengxian, 727
Wang, Mengyu, 868
Wang, Min, 412
Wang, Ming, 87, 730
Wang, Mingda, 651, 688
Wang, Minghan, 358
Wang, Mingshuo, 861
Wang, Mingxuan, 692
Wang, Mingyang, 474, 821
Wang, Mingyi, 588
Wang, Mingyu, 316
Wang, Minzheng, 316
Wang, Nanshu, 444
Wang, Ning, 341, 350, 600, 611
Wang, Pei, 702
Wang, Peidong, 422
Wang, Peijie, 431, 636
Wang, Peiran, 261, 781
Wang, Peng, 389, 392, 648, 719, 727, 740
Wang, Pengbo, 653
Wang, Pengwei, 479
Wang, Pengyu, 406
Wang, Ping, 234, 294, 362, 452, 676, 814
Wang, Pingjie, 322, 364
Wang, Po-Ya Angela, 448
Wang, Pu, 534
Wang, Qi, 644
Wang, Qian, 155, 181
Wang, Qianli, 821, 826
Wang, Qianlong, 434
Wang, Qianning, 625
Wang, Qianqian, 414
Wang, Qianyu, 345
Wang, Qidong, 683
Wang, Qifan, 294, 303, 507
Wang, Qifei, 588
Wang, Qifeng, 577
Wang, Qilin, 326
Wang, Qimeng, 197
Wang, Qing, 311, 475, 639, 749, 858
Wang, Qingyan, 392
Wang, Qingyun, 250
Wang, Qingzhuo, 341
Wang, Qiqi, 561, 641, 677
Wang, Qiufeng, 725
Wang, Qixiang, 340
Wang, Quan, 320, 709
Wang, Rongguang, 614
Wang, Rongzhen, 501
Wang, Rui, 161, 314, 436, 451, 483, 701, 753, 877, 903
Wang, Ruiheng, 186
Wang, Rujia, 599
Wang, Runze, 885
Wang, Ruobing, 673
Wang, Ruoyu, 506
Wang, Ruxin, 859
Wang, Sa, 297

Wang, Senmiao, 152
Wang, Shangguang, 375, 401, 421
Wang, Shanshan, 762, 767, 899
Wang, Shanyi, 425
Wang, Shaobo, 527, 655
Wang, Shaofei, 170
Wang, Shaohan, 729
Wang, Shaojiang, 221
Wang, Shaojun, 803
Wang, Shaonan, 184
Wang, Shaoqi, 817
Wang, Shaowei, 265, 484
Wang, Shaowen, 382
Wang, Shasha, 239
Wang, Shen, 205
Wang, Sheng, 451
Wang, Shengling, 570
Wang, Shenran, 304
Wang, Shenzhi, 897
Wang, Shi, 186, 327
Wang, Shibo, 559
Wang, Shihao, 182
Wang, Shijian, 800
Wang, Shijie, 437, 473, 893
Wang, Shijin, 380, 395, 418, 419
Wang, Shipeng, 749
Wang, Shiqi, 139, 204, 530
Wang, Shiting, 401
Wang, Shiyao, 414
Wang, Shiyu, 443
Wang, Shouren, 895
Wang, Shuai, 452, 582, 797, 880
Wang, Shuaiqiang, 289, 713, 795
Wang, Shuhe, 467
Wang, Shujie, 451, 669
Wang, Shuo, 377, 387, 406, 418, 583, 608, 645, 673, 716, 734
Wang, Shuohuan, 304
Wang, Shuyi, 694
Wang, Sijie, 405
Wang, Simon, 471
Wang, Siwei, 508
Wang, Siyuan, 503, 516, 543, 546, 562
Wang, Sizhe, 201, 674
Wang, Song, 486, 792
Wang, Songmiao, 131
Wang, Songyao, 541
Wang, Stephen, 130
Wang, Suge, 676
Wang, Suhang, 268, 269, 854
Wang, Suyuan, 823
Wang, Suyuchen, 166
Wang, Tai, 776
Wang, Taifeng, 340, 657
Wang, Taihang, 235
Wang, Tao, 423, 689
Wang, Taoran, 412
Wang, Taowen, 507
Wang, Teng, 721
Wang, Tian, 531, 826
Wang, Tianfu, 197
Wang, Tianhao, 307
Wang, Tianheng, 313
Wang, Tianlong, 219, 701
Wang, Tianrui, 432, 455, 456
Wang, Tianyang, 137
Wang, Tianyi, 93, 575
Wang, Ting, 117, 528
Wang, Ting-Jun, 153
Wang, Tingyan, 500
Wang, Tongnian, 477
Wang, Tongyao, 159
Wang, Wanfu, 678
Wang, Wanyu, 214
Wang, Wei, 82, 230, 280, 300, 325, 432, 560, 665, 688
Wang, Wei-Yao, 103
Wang, Weida, 774
Wang, Weidong, 350, 640
Wang, Weihan, 164
Wang, Weihua, 318
Wang, Weijie, 675
Wang, Weikang, 86
Wang, Weiping, 152
Wang, Weiqi, 313, 733
Wang, Weiqin, 272, 802
Wang, Weixun, 660, 702
Wang, Wen, 114, 576, 780, 785, 865
Wang, Wendan, 375
Wang, Wenhai, 405, 532, 750
Wang, Wenhan, 332
Wang, WenHao, 583, 894
Wang, Wenhao, 144, 407
Wang, Wenjie, 82, 349, 415, 648, 852, 858
Wang, Wenkai, 289, 311, 322
Wang, Wenqiang, 353, 687
Wang, Wenshuo, 749, 778
Wang, Wensong, 666
Wang, Wenxiao, 685, 909
Wang, Wenxuan, 90, 195, 261, 293, 425, 582, 687, 803, 861, 882
Wang, Wenya, 128, 186, 322, 362, 649, 694, 704, 889
Wang, William Yang, 194, 445
Wang, Wuqi, 640
Wang, Xi, 183, 394, 401, 659, 733, 848

Wang, Xiang, 145, 683
Wang, Xiangchen, 318
Wang, Xiangfeng, 560
Wang, Xiangqi, 471
Wang, Xiangsen, 377
Wang, Xiangyu, 332
Wang, Xianquan, 330, 623
Wang, Xianyang, 97
Wang, Xiao, 305, 314, 793
Wang, Xiaobao, 455
Wang, Xiaodong, 348
Wang, Xiaofei, 112
Wang, XiaoFeng, 482, 504, 593
Wang, Xiaofeng, 735
Wang, Xiaohan, 568
Wang, Xiaohua, 321, 661
Wang, Xiaohui, 326
Wang, Xiaojie, 824
Wang, Xiaolong, 462, 482, 855
Wang, Xiaonan, 834
Wang, Xiaoning, 509
Wang, Xiaopeng, 456
Wang, Xiaorui, 729
Wang, Xiaoyang, 274, 859
Wang, Xiaoyu, 166, 381
Wang, Xiaozhi, 782
Wang, Xidong, 155
Wang, Xilai, 281
Wang, Ximing, 679
Wang, Xin, 443, 519, 662, 689, 696, 717, 759
Wang, Xinbing, 361
Wang, Xinda, 303, 653, 684, 885
Wang, Xing, 538
Wang, Xingchen, 611
Wang, Xinghao, 706
Wang, Xingjun, 685
Wang, Xinglin, 405
Wang, Xingmei, 414
Wang, Xingwei, 671
Wang, Xingxing, 432
Wang, Xingyue, 777
Wang, Xinhe, 307
Wang, Xinlin, 700
Wang, Xinming, 187
Wang, Xinru, 723, 726
Wang, Xinrun, 654
Wang, Xinsheng, 748
Wang, Xintao, 671, 687
Wang, Xintong, 471
Wang, Xinyi, 107, 445, 646, 737
Wang, Xinyu, 212, 359, 587, 788
Wang, Xinyuan, 204
Wang, Xiong, 736
Wang, Xirui, 790
Wang, Xiting, 151, 498, 741, 777, 904
Wang, Xu, 333, 420, 651, 714, 759
Wang, Xuan, 211, 637
Wang, Xuanlin, 311
Wang, Xuefei, 651
Wang, Xuehai, 436
Wang, Xuehe, 305
Wang, xuelong, 108
Wang, Xuesong, 389
Wang, Xueyan, 497
Wang, Xuhong, 128, 412, 451
Wang, Xun, 486
Wang, Xunlong, 498
Wang, Xv, 318, 323
Wang, Yabiao, 724
Wang, Yakai, 138
Wang, Yalin, 204
Wang, Yan, 142, 166, 226, 276, 379, 744
Wang, Yanbo, 82, 117, 259, 316, 363, 369
Wang, Yancheng, 185
Wang, Yandi, 683
Wang, Yanfeng, 93, 322, 364, 407, 409
Wang, Yang, 313, 669, 699, 714, 767, 773
Wang, Yangfan, 397
Wang, Yanhao, 356
Wang, Yanli, 620, 729
Wang, Yanlin, 620, 729, 776
Wang, Yanmeng, 803
Wang, Yanting, 559
Wang, Yao, 230, 295
Wang, Yaoting, 877
Wang, Yaowei, 335, 393
Wang, Yaoxiang, 636
Wang, Yasha, 264, 788, 850
Wang, Yau-Shian, 175
Wang, Yawei, 442
Wang, Yawen, 475, 858
Wang, Yaxiong, 84, 330, 392, 676
Wang, Yaxuan, 215
Wang, Ye, 346, 354, 373, 516
Wang, Yejing, 695
Wang, Yeqiang, 313
Wang, Yequan, 325, 354, 355, 396, 638, 703
Wang, Yi, 388, 729, 892
Wang, Yi-Cheng, 262
Wang, Yian, 121, 591
Wang, Yibin, 844
Wang, Yibing, 743
Wang, Yibo, 262, 307
Wang, Yichao, 587
Wang, Yiding, 444
Wang, Yifan, 375, 420, 657

Wang, YiFei, 268
Wang, Yihan, 753
Wang, Yihang, 629, 724
Wang, Yihao, 720
Wang, Yijun, 221
Wang, Yike, 446, 550
Wang, Yikun, 228
Wang, Yile, 272, 802
Wang, Yili, 717
Wang, Yilin, 132, 144, 335, 673
Wang, Yilong, 854
Wang, Yimin, 235
Wang, Ying, 460
Wang, Yingchun, 451, 696
Wang, Yingjie, 388
Wang, Yingkang, 398
Wang, Yingshan Susan, 242
Wang, Yingyao, 341, 636, 733
Wang, Yingze, 134
Wang, Yining, 348, 391, 506, 684, 760
Wang, Yiqi, 214
Wang, Yitong, 752, 908
Wang, Yiwei, 146, 280, 436, 635, 636, 686, 732, 821
Wang, Yiwen, 765
Wang, Yixu, 696
Wang, Yixuan, 311
Wang, Yiyao, 687
Wang, Yiyu, 379, 666
Wang, Yizhe, 383
Wang, Yizhou, 124, 143, 495, 532, 815
Wang, Yong, 378, 623, 739, 776
Wang, Yongbo, 905
Wang, Yongcan, 569
Wang, Yonghan, 278
Wang, Yongli, 400
Wang, Yongwei, 695
Wang, Yongxin, 253
Wang, Yu, 85, 93, 153, 205, 229, 311, 322, 364, 385, 388, 399, 409, 413, 493, 867, 901
Wang, Yu-Chiang Frank, 529
Wang, Yuan, 450, 786
Wang, Yuanchun, 169
Wang, Yuandong, 737
Wang, Yuanfei, 532
Wang, Yuanlong, 403
WANG, Yuanping, 496
Wang, Yuanxiang, 462, 500
Wang, Yuanyuan, 414, 435
Wang, Yuanzhuo, 205, 371, 584, 723, 901
Wang, Yubo, 887
Wang, Yuchen, 131, 442
Wang, Yucheng, 554, 746, 901
Wang, Yuchi, 416, 686
Wang, Yudong, 758
Wang, Yue, 85, 172, 410, 425, 656, 716
Wang, Yuejiao, 456
Wang, Yueming, 330
Wang, Yuer, 682
Wang, Yufei, 186, 474, 712
Wang, Yufeng, 353
Wang, Yuguang, 649
Wang, Yuhan, 832
Wang, Yuhang, 518
Wang, Yuhao, 322, 364, 409, 653, 746
Wang, Yuheng, 617
Wang, Yuhui, 117, 820, 838
Wang, YuJie, 427
Wang, Yujie, 545
Wang, Yulong, 699, 704
Wang, Yumeng, 691
Wang, Yun-Jie, 418
Wang, Yunhe, 412
Wang, Yunke, 409
Wang, Yunzhe, 129
Wang, Yuqi, 321, 676
Wang, Yuquan, 348
Wang, Yuran, 838
Wang, Yurun, 344
Wang, Yushuo, 606
Wang, Yusong, 172
Wang, Yuting, 393
Wang, Yuxia, 95, 236, 442, 874
Wang, Yuxian, 658
Wang, Yuxiang, 194, 649
Wang, Yuxuan, 477, 760
Wang, Yuxuan Lu; Bingsheng Yao; Shao Zhang; Yisi Sang; Yun Wang; Hansu Gu; Peng Zhang; Tun Lu; Toby Jia-Jun Li; Dakuo,
Wang, Yuyao, 458
Wang, Yuzhuo, 871
wang, Yvchuan, 327
Wang, Zaixiang, 725
Wang, Ze, 348, 490
Wang, Zehan, 339
Wang, Zehao, 774
Wang, Zehong, 178
Wang, Zekun, 345, 395, 866
Wang, Zengbin, 776
Wang, Zeqing, 719
Wang, Zexin, 345, 577
WANG, Zezhong, 701
Wang, Zezhong, 184
Wang, Zezhou, 284, 494
Wang, Zhangyang, 498

Wang, Zhaohui, 578
Wang, Zhaokun, 383, 619, 671
Wang, Zhaoyang, 599
Wang, Zhefan, 209
Wang, Zhefeng, 346
Wang, Zhen, 94, 386, 585, 630, 809
Wang, Zhendong, 112
Wang, Zheng, 130, 336, 819
Wang, Zhenghua, 661
Wang, Zhengjia, 96, 170, 769
Wang, Zhengkui, 209, 569
Wang, Zhenglin, 437
Wang, Zhengren, 674
Wang, Zhenguang, 699
Wang, Zhengxiang, 533
Wang, Zhengyang, 111, 625
Wang, Zhenhailong, 802
Wang, Zhenhua, 310
Wang, Zhensheng, 137, 189
Wang, Zhenting, 141
Wang, Zhenyu, 323
Wang, Zheyu, 197
Wang, Zhi, 145, 316, 593
Wang, Zhibin, 651
Wang, Zhibo, 369
Wang, Zhicai, 645
Wang, Zhichao, 566, 678
Wang, Zhicheng, 404
Wang, Zhichun, 673
Wang, Zhigang, 358
Wang, Zhihao, 422
Wang, Zhihua, 97
Wang, Zhijun, 94
Wang, Zhilin, 756
Wang, Zhipeng, 124, 364
Wang, Zhisheng, 656
Wang, Zhitong, 449
Wang, Zhixiang, 258, 491
Wang, Zhiyong, 424, 734
Wang, Zhiyuan, 202, 870
Wang, Zhizhi, 482
Wang, Zhong, 384
Wang, Zhongpu, 107
Wang, Zhongqing, 506, 857
Wang, Zhongyuan, 479
Wang, Zhouyang, 857
Wang, Zhu, 867
Wang, Zhuo, 550
Wang, Zhuoer, 479
Wang, Zhuxuanzi, 137
Wang, Zibo, 571
Wang, Zichao, 784
Wang, Zifeng, 446
Wang, Zihan, 258, 349, 673, 692
Wang, Zihao, 482, 560, 713, 789, 845
Wang, Ziheng, 90, 651, 759
Wang, Zijian, 787
Wang, Zijie, 807
Wang, Zijing, 643, 663
Wang, Zikang, 909
Wang, Zilei, 634, 698
Wang, Zilong, 412
Wang, Ziming, 341
Wang, Zimu, 230
Wang, Ziqi, 304, 317
Wang, Ziqin, 421
Wang, Ziqing, 457, 752, 780, 793
Wang, Zirui, 545
Wang, Zitai, 535
Wang, Ziwen, 311
Wang, Zixuan, 159, 259, 347, 670, 722
Wang, Ziyan, 534
Wang, Ziyang, 124, 396, 758
Wang, Ziyi, 359, 826, 875
Wang, Ziyou, 673
Wang, Ziyu, 260, 868
Wang, Ziyuan, 836
Wang, Ziyue, 214, 462
Wang, Zizhen, 471
Wang, Zongqi, 251, 696
Wang, ZongYu, 271, 408, 863
Wang, Zun, 591
Wang;,
Wang;, Xuhong,
Wang;, Yi,
Wang;, Yixiao Huang;Lan Zhang;Chaoran,
Wang;, Zhongyu,
wangdexian, 437
wangdexian, wangdexian, 437
Wangsajaya, 667
Wangsajaya, Nadya Yuki, 667
Wanna, 820
Wanna, Selma, 820
Wanner, 555
Wanner, Miriam, 555
Wanzare, 471
Wanzare, Lilian Diana Awuor, 471
Warczyski, 138
Warczyski, Jdrzej, 138
Warren, 577
Warren, Greta, 577
Warschauer;,
Warschauer;, Mark,
Warstadt, 105, 771, 876
Warstadt, Alex, 105, 771, 876
Wasi, 135, 862

Wasi, Azmine Toushik, 135, 862
Wasnik, 731
Wasnik, Pankaj, 731
Wastl, 247, 771
Wastl, Michelle, 247, 771
WATANABE, 608
Watanabe, 99, 111, 217, 311, 455, 456, 492, 512, 518, 529, 605, 832
Watanabe, Narimasa, 311
Watanabe, Shinji, 111, 217, 455, 456, 529
Watanabe, Taro, 99, 492, 512, 605, 832
Watanabe, Yoshida Daiki; Hiroyuki Deguchi; Yusuke Sakai; Hidetaka Kamigaito; Taro,
WATANABE, YUSUKE, 608
Watanabe;,
Watanabe;, Taro,
Watrin, 173
Watrin, Patrick, 173
Weatherhead, 597
Weatherhead, Sion, 597
Weeber, 263, 548
Weeber, Franziska, 263, 548
Weeren, 631
Weeren, Niels van, 631
Weghe, 130
Weghe, Nico Van de, 130
WEI, 278, 560, 786
Wei, 86, 89, 92, 94, 104, 111, 113, 124, 131, 141, 145, 152, 162, 167, 177, 202, 214, 234, 235, 239, 241, 248, 288, 290, 291, 304, 311, 312, 315, 332, 336, 341, 344, 354, 355, 358, 381, 384, 404, 420, 421, 422, 436, 448, 455, 457, 469, 480, 491, 492, 516, 542, 602, 604, 629, 644, 653, 656, 659, 660, 673, 683, 692, 693, 714, 717, 718, 726, 739, 743, 756, 759, 771, 777, 784, 800, 826, 836, 848, 870, 871, 883, 886, 895, 900, 904, 907, 908
wei, 503, 543, 546, 562, 868
Wei, Barry, 104
Wei, Bifan, 248, 871
Wei, Bofan, 718
Wei, Bowen, 86, 131
Wei, Chen, 421
Wei, Daimeng, 756
Wei, Dennis, 771
Wei, Fanjunduo, 653
Wei, Feng, 756
Wei, Furu, 89, 145, 167, 659, 900, 907
Wei, Hanjun, 800
Wei, Hao-Ran, 94
WEI, HU, 560
Wei, Hua, 177, 288, 492, 602
Wei, Jiaheng, 152
Wei, Jianguo, 332
Wei, Jiansheng, 644, 756, 848, 908
Wei, Jinbiao, 241, 784
Wei, Jincheng, 660
Wei, Jingsong, 743
Wei, Jingxuan, 777
Wei, Jiwei, 404
Wei, Kai, 177
Wei, Kaiwen, 162, 629
Wei, Kangda, 895
Wei, Lai, 656
Wei, Lanning, 436
Wei, Lei, 344
Wei, Lingwei, 693
Wei, Liqun, 239
Wei, Peng, 604
Wei, Ruixin, 354
Wei, Shaohang, 659
Wei, Sheng-Lun, 291
Wei, Tao, 92
Wei, Tianjun, 235
Wei, Tianshuo, 717
Wei, Tianxin, 457, 469, 826
Wei, Tianyi, 358
Wei, Tzu-Chieh, 312
Wei, Wei, 604, 683
Wei, Wenqi, 141
Wei, Xiao-Yong, 542, 692
Wei, Xiaolong, 304
Wei, Xing, 448
Wei, Xinyu, 290
Wei, Xuetao, 214, 739
Wei, Yang, 455
Wei, Ye, 717
Wei, Yifan, 726
Wei, Yifei, 404, 836
Wei, Ying, 384, 422
Wei, Yiwei, 354
Wei, Yiyang, 800
Wei, Yu Zi, 759
Wei, Yuhao, 202, 870
Wei, Yukun, 714
Wei, Yunzhao, 311
WEI, YUXIANG, 278, 786
Wei, Zeming, 886
Wei, Zheng, 315, 420, 904
Wei, Zhenlin, 421
Wei, Zhepei, 111, 480
Wei, Zhewei, 234
Wei, Zhihua, 341
Wei, Zhimin, 336

Wei, Zhipeng, 883
Wei, Zhixiang, 491
Wei, Zhongyu, 113, 124, 381, 516
wei, zhongyu, 503, 543, 546, 562, 868
Wei, Zihao, 355
Wei, Zikai, 162
Wei, Zili, 673
Weicheng_Ma,
Weicheng_Ma, Weicheng_Ma,
Weidi, 329
Weidi, Sun, 329
Wein, 95
Wein, Shira, 95
Welch, 500
Welch, James, 500
Weld, 889
Weld, Daniel S, 889
Welleck, 474
Welleck, Sean, 474
Weller, 226
Weller, Nicholas, 226
Wen, 111, 122, 135, 148, 168, 205, 249, 271, 298, 303, 313, 316, 317, 326, 329, 337, 346, 348, 352, 357, 368, 369, 387, 391, 401, 418, 422, 457, 463, 470, 497, 501, 513, 534, 546, 568, 581, 587, 594, 601, 621, 656, 659, 666, 668, 699, 700, 706, 707, 717, 718, 726, 748, 749, 782, 789, 791, 793, 794, 817, 832, 847, 848, 851, 853, 884, 886, 896, 905
Wen, Bingbing, 168
Wen, Bosi, 497, 851
Wen, Chao, 659
Wen, Cheng, 316, 329
Wen, Chenyuhao, 832
Wen, Dacheng, 700
Wen, Feng, 587
Wen, Haibin, 717
Wen, Hanlin, 748
Wen, Haoyang, 817
Wen, Hongyi, 249
Wen, Ji-Rong, 337, 501, 700, 717, 847
Wen, Juan, 581, 782
Wen, Liangjian, 884
Wen, Licheng, 148, 303
Wen, Lijie, 122
Wen, Long, 369
Wen, Mi, 348
Wen, Muning, 271, 794
Wen, Pengcheng, 470, 513
Wen, Qianfeng, 568
Wen, Qianlong, 534
Wen, Qingsong, 205, 546, 621, 749
Wen, Rui, 111, 346, 817
Wen, Rujie, 463
Wen, Shuangyue, 329
Wen, Siwei, 789
Wen, Xiangyu, 298
Wen, Xiaofei, 594
Wen, Ximing, 905
Wen, Xincheng, 391
Wen, Yajie, 886
Wen, Yan, 326
Wen, Yanlong, 718
Wen, Yi, 791
Wen, Yibo, 457, 793
Wen, Yilin, 357, 422
Wen, Ying, 135, 601, 794
Wen, Yuanbo, 418
Wen, Zengxuan, 896
Wen, Zeyi, 387, 726
Wen, Zhengqi, 313, 317, 352, 656, 668, 699, 706
Wen, Zhihua, 401
Wen, Zhoufutu, 707
Wen, Zichen, 368, 666
Wen, Zijian, 853
Wen, Zujie, 848
Wendler, 821
Wendler, Chris, 821
Weng, 132, 151, 209, 335, 348, 361, 369, 434, 467, 498, 501, 687, 781, 834, 840
Weng, Chunhua, 498
Weng, Han, 434
Weng, Jiaqi, 132
Weng, Luoxuan, 687
Weng, Rongxiang, 151, 209, 335, 501
Weng, Shuchen, 348
Weng, Yixuan, 369, 840
Weng, Yu, 361
Weng, Zhenyu, 834
Weng, Zixuan, 781
Weng, Ziyan, 467
Wense, 207, 286
Wense, Katharina von der, 207, 286
Wenyue, 583
Wenyue, Li, 583
WenYujia, 687
WenYujia, WenYujia, 687
Wenzel, 822
Wenzel, Markus, 822
Werb, 505
Werb, Gabriela Alves, 505
West, 821, 838, 891
West, Peter, 891
West, Robert, 821, 838
Whalen, 533

Whalen, Ryan, 533
Whang, 202
Whang, Steven Euijong, 202
White, 85, 194
White, Matt Murtagh, 85
White, Michael, 194
Whitehouse, 869
Whitehouse, Chenxi, 869
Whitted, 609
Whitted, Jennifer, 609
Wi, 805
Wi, Hyowon, 805
Wibowo, 869
Wibowo, Haryo, 869
Wicaksono, 695
Wicaksono, Alfan Farizki, 695
Wiegreffe, 567
Wiegreffe, Sarah, 567
Wiemerslage, 799
Wiemerslage, Adam, 799
Wierzbicki, 203
Wierzbicki, Adam, 203
Wilcox, 105, 153, 242, 876
Wilcox, Ethan, 153, 242, 876
Wilcox, Ethan Gotlieb, 105
Williams, 223, 238, 548
Williams, Adina, 238
Williams, Miles, 223
Williams, Tristan, 548
Wilson, 476
Wilson, Steven, 476
Winata, 853, 862, 869
Winata, Genta, 869
Winata, Genta Indra, 853, 862
Winkler, 167
Winkler, Miriam, 167
Wintner, 660
Wintner, Shuly, 660
Wisniewski, 228
Wisniewski, Guillaume, 228
Witbrock, 125, 181
Witbrock, Michael, 125, 181
Wolf, 95, 453, 600
Wolf, Lior, 453, 600
Wolf, Nathan, 95
Wolfson, 181, 452
Wolfson, Tomer, 181, 452
WONG, 347
Wong, 129, 163, 184, 229, 243, 261, 274, 436, 440, 451, 541, 565, 577, 602, 677, 701, 724, 726, 762, 767, 773, 821, 825, 829, 862, 899
Wong, Derek F., 129, 274, 541, 565, 577, 602, 677, 762, 899
Wong, Elaine, 829
Wong, Eric, 825
Wong, Ginny Y., 243
Wong, Hon Yung, 440
Wong, Hugh Mee, 229
Wong, Kam-Fai, 184, 436, 451, 701, 726, 767
Wong, Lionel, 261
Wong, Ngai, 163, 773, 821
Wong, Tack Hwa, 862
Wong, Tien-Tsin, 724
WONG, Wang Chak, 347
Woo, 458, 596, 908
Woo, Caeley, 458
Woo, Hyeondong, 596
Woo, Tony, 908
Woodard, 636
Woodard, Damon L., 636
Woodhead, 832
Woodhead, Simon, 832
Woodland, 140
Woodland, Philip C., 140
Woody_Haosheng_Gan,
Woody_Haosheng_Gan, Woody Haosheng Gan,
Wornell, 592
Wornell, Gregory W., 592
Woydt, 298
Woydt, Tim, 298
Wright, 472, 517
Wright, Dustin, 472, 517
WU, 298, 597, 708
Wu, 91, 92, 93, 94, 95, 96, 99, 109, 111, 112, 119, 120, 126, 130, 134, 135, 136, 137, 138, 139, 140, 141, 145, 148, 149, 152, 155, 159, 161, 165, 169, 179, 180, 184, 186, 191, 197, 199, 202, 204, 208, 210, 214, 221, 224, 226, 230, 231, 232, 233, 239, 242, 244, 248, 250, 257, 258, 259, 260, 264, 266, 268, 269, 270, 274, 276, 278, 279, 280, 281, 282, 283, 285, 286, 291, 292, 295, 299, 300, 301, 303, 304, 306, 313, 315, 317, 319, 323, 326, 330, 333, 339, 343, 344, 347, 348, 351, 352, 353, 358, 359, 361, 362, 363, 369, 372, 373, 374, 375, 377, 380, 382, 386, 392, 393, 395, 396, 406, 414, 416, 417, 418, 422, 423, 424, 425, 427, 428, 429, 430, 433, 435, 444, 448, 449, 450, 451, 455, 463, 464, 467, 473, 474, 483, 487, 488, 490, 499, 501, 504, 505, 506, 507, 509, 514, 527, 533, 535, 539, 540, 544, 549, 556, 558, 559, 560, 570, 571, 573, 576, 581,

582, 584, 590, 591, 594, 596, 599, 602, 608, 609, 617, 619, 621, 623, 626, 628, 629, 630, 636, 645, 646, 650, 656, 657, 660, 661, 665, 668, 669, 670, 671, 676, 677, 680, 681, 685, 689, 690, 693, 696, 697, 698, 700, 702, 703, 706, 707, 712, 713, 715, 723, 724, 725, 729, 730, 732, 735, 736, 738, 739, 746, 747, 749, 754, 758, 769, 773, 774, 775, 776, 777, 782, 785, 788, 789, 797, 798, 810, 817, 820, 821, 822, 827, 829, 839, 840, 843, 844, 846, 849, 851, 852, 853, 859, 860, 865, 867, 871, 873, 877, 881, 886, 887, 888, 889, 892, 895, 901, 905, 906, 907
wu, 693
Wu, Biao, 738
Wu, Bin, 119, 266, 386, 451, 629
Wu, Bingli, 126
Wu, Bingxian, 596
Wu, Bingzhe, 382
Wu, Bo, 571
Wu, Bohan, 191
Wu, Boya, 300
Wu, Bozhi, 186
Wu, Chao, 303
Wu, Chengyan, 95, 886
Wu, Chenhao, 226
WU, Chenyan, 708
Wu, Chia-Chun, 840
Wu, Chien-Sheng, 281, 609, 789
Wu, Chuan, 690
Wu, Chunwei, 501
Wu, Dapeng, 244, 774
Wu, Dayong, 395, 697
Wu, Decheng, 774
Wu, Desheng, 681
Wu, Di, 138, 483, 544, 867
Wu, Faguo, 135
Wu, Fan, 239, 286, 527
Wu, Fangyu, 725
Wu, Fangzhao, 351, 533, 700
Wu, Fei, 135, 323, 754
Wu, Gang, 671
Wu, Haixing, 660
Wu, Haiyu, 204
Wu, Han, 221
Wu, Hanqian, 724
Wu, Hanqing, 313
Wu, Hao, 257, 375
Wu, Haolun, 276, 507, 509, 584
Wu, Haoyu, 881
Wu, Haoyuan, 540
WU, Hengyu, 597
Wu, Hua, 304, 715, 724, 746
Wu, Huan, 696
Wu, Hui, 505
Wu, Huijia, 285
Wu, Ian, 474
Wu, Ji, 514, 749
Wu, Jiajun, 260, 693
Wu, Jialian, 348, 490
Wu, Jialong, 853
Wu, Jian, 155, 450, 645, 676
Wu, Jiancan, 145
Wu, Jiang, 239, 396
Wu, Jianguo, 499
Wu, Jiaqi, 169
Wu, Jiaxuan, 581, 782
Wu, Jiayi, 120, 417, 449, 713
Wu, Jiaying, 155, 248, 363, 888, 889, 901
Wu, Jie, 264
Wu, Jingjing, 377
Wu, Jingpei, 689
Wu, Jingzhuo, 698
Wu, Jinming, 775
Wu, Jinyang, 96, 313, 352, 656, 668, 706
Wu, Junchao, 602, 677
Wu, Junda, 257, 483, 514
Wu, Junfei, 698
Wu, Junjie, 425, 680, 689, 846
Wu, Junru, 94
Wu, Kaishun, 152, 646, 736
Wu, Kangyi, 775
Wu, Lei, 91, 374
Wu, Leijie, 161
Wu, Lemeng, 444
Wu, Lianwei, 330, 392, 676, 703
Wu, Libing, 619
Wu, Lijian, 888
Wu, Lijun, 239, 268, 274, 487, 829, 859
Wu, Lin, 617
Wu, Linjuan, 94, 621
Wu, Linyu, 576, 785
Wu, Meiqi, 317
Wu, Mengjia, 681, 852
Wu, Minghao, 416
Wu, Mingqi, 820, 839
Wu, Ning, 264
Wu, Ningxin, 372
Wu, Ou, 369
Wu, Pangjing, 573
Wu, Peilin, 333
Wu, Penghao, 361
Wu, Pingyu, 358
Wu, Qi, 359, 362, 738
Wu, Qianhui, 599

Wu, Qianqian, 239
Wu, Qingqiang, 665
Wu, Qingyun, 906
Wu, Qinyuan, 99, 571
Wu, Qinzhuo, 558
Wu, Qiong, 208
Wu, Qiyu, 873
Wu, Raina, 427
Wu, Rong, 148, 303
Wu, Rui, 141, 428
Wu, Ruiqi, 619
Wu, Runze, 797
Wu, Sai, 347
Wu, Sherry, 112
Wu, Shu, 473, 650
Wu, Shuang, 730
Wu, Shunlong, 535
Wu, Sihong, 269, 623
Wu, Silin, 467
Wu, Sitong, 423, 433
Wu, Siwei, 260
Wu, Siyan WuGuo Yu Bin, 374
Wu, Siyu, 608
Wu, Siyuan, 259
Wu, Sizhe, 619
Wu, Song-Li, 330, 623
Wu, Songhao, 798, 865
Wu, Steven, 539
Wu, Taiqiang, 140, 773
Wu, Tao, 754
Wu, Tianhe, 712
Wu, Tianwei, 556
Wu, Tianxing, 224
Wu, Tianyi, 233, 777
Wu, Tianyu, 685, 769
Wu, Tinglang, 180
Wu, Tingxuan, 136
Wu, Tingyu, 467
Wu, Tongshuang, 232, 306
Wu, Tongyu, 359
Wu, Tz-Ying, 152
Wu, Wei, 92, 197, 373, 865
Wu, Wei-Chi, 291
Wu, Weiheng, 560
Wu, Weiqi, 769
Wu, Weiyi, 136, 860
Wu, Wenjun, 464
Wu, Wenlong, 707
Wu, Wenqing, 871
Wu, Wenshan, 278
Wu, Wentao, 136, 429
Wu, Wenxiao, 134
Wu, Wenyi, 130
Wu, Xian, 184, 315, 867
Wu, Xiangyu, 414, 738
Wu, Xianjie, 165, 230, 295
Wu, Xiao-Kun, 747
Wu, Xiao-Ming, 250, 788
Wu, Xiaobao, 210, 851, 859
Wu, Xiaojun, 723
Wu, Xiaokai, 630
Wu, Xiaoming, 374
Wu, Xiaoqing, 282
Wu, Xiaoting, 430
Wu, Xiaoyuan, 844
Wu, Xin, 292, 359, 506
Wu, Xingyong, 278
Wu, Xingyu, 570
Wu, Xinwei, 406, 590, 729, 732
Wu, Xixi, 120
Wu, Xixin, 339, 435
Wu, Xuecheng, 657, 877
Wu, Xueyan, 463
Wu, Xun, 907
Wu, Yanan, 702
Wu, Yang, 821, 843
Wu, Yangyang, 650
Wu, Yanxi, 344
Wu, Yaozu, 488
Wu, Yaqiang, 871
Wu, Yawen, 299
Wu, Yebo, 449, 905
Wu, Yerong, 224
WU, YI, 298
Wu, Yi, 197
Wu, Yichen, 202, 887
Wu, Yifan, 258, 628
Wu, Yifu, 822
Wu, Yihao, 455
Wu, Yijian, 906
Wu, Yilong, 820
Wu, Ying Nian, 280, 424, 504, 636, 758
Wu, Yingcai, 703
Wu, Yinhao, 283
Wu, Yiquan, 749
Wu, Yixin, 111, 594, 661, 817
Wu, Yonghui, 109
Wu, Yongjiang, 582
Wu, Yongkang, 669
Wu, Yongxuan, 319
Wu, Yu, 623
Wu, Yu-Chun, 418
Wu, Yuan, 351, 623
Wu, Yuanchen, 892
Wu, Yuanhao, 559
Wu, Yuchen, 280, 676

Wu, Yuchuan, 739, 849
Wu, Yueyi, 91
Wu, Yuhan, 670
Wu, Yuhang, 414, 776
Wu, Yuhe, 670, 827
Wu, Yujiao, 330, 676
Wu, Yukang, 120
Wu, Yulong, 179
Wu, Yunfan, 414
Wu, Yunfang, 149, 270, 422, 539
Wu, Yuning, 93
Wu, Yunjia, 326
Wu, Yunpeng, 609
Wu, Yunze, 665
Wu, Yuzhang, 626
Wu, Zeyu, 677
Wu, Zezhong, 137
Wu, Zhanyu, 735
Wu, Zhaofen, 393
Wu, Zhaofeng, 139, 242, 450
Wu, Zhen, 414
Wu, Zhenyu, 591
Wu, Zhijing, 231
wu, zhiliang, 693
Wu, Zhiqiang, 380
Wu, Zhiyong, 159, 359, 435
Wu, Zhize, 353
Wu, Zhizheng, 199
Wu, Zhonghai, 448, 810
Wu, Zhuoyang, 406
Wu, Zifei, 359
Wu, Zihan, 343
Wu, Zijian, 301
Wu, Zijun, 214
Wu, Ziming, 895
Wu, Zirui, 549
Wujiajia, 540
Wujiajia, Wujiajia, 540
Wulff, 663
Wulff, Dirk U., 663
Wullach, 835
Wullach, Tomer, 835
Wullschleger, 829
Wullschleger, Pascal, 829
Wunderlich, 174
Wunderlich, Florian Valentin, 174
Wutschitz, 720
Wutschitz, Lukas, 720
Wüst, 298
Wüst, Antonia, 298

Xi, 201, 273, 387, 545, 550, 576, 717, 749, 820, 821, 838, 839, 877
Xi, Meng, 749
Xi, Ronghui, 545
Xi, Yuling, 576
Xi, Yunjia, 877
Xi, Zhiheng, 201, 387, 545, 717, 820, 821, 838, 839
Xi, Zijie, 545, 550
Xi, Ziqiao, 273
Xia, 87, 142, 180, 191, 202, 262, 276, 279, 304, 332, 349, 393, 410, 414, 415, 420, 433, 435, 437, 477, 479, 483, 528, 563, 597, 599, 606, 616, 649, 662, 665, 671, 676, 682, 725, 735, 738, 823, 832, 833, 849, 853, 857, 874, 900
Xia, Bin, 433
Xia, Congying, 262
Xia, Deguo, 849
Xia, Ding, 563, 823, 833
Xia, Feng, 415
Xia, Gus, 202
Xia, Hanchen, 599, 649
Xia, Hanyang, 738
Xia, Heming, 738
Xia, Lei, 606
Xia, Long, 142, 276, 304
Xia, Lun, 682
Xia, Peng, 528, 832
Xia, Pengfei, 437
Xia, Qiang, 332
Xia, Qifeng, 420
Xia, Qiyun, 420
Xia, Rui, 414
Xia, Runze, 725
Xia, Shan, 479
Xia, Shijie, 665
Xia, Shu-Tao, 393
Xia, Shutao, 349, 662
Xia, Sirui, 671
Xia, Tian, 735
Xia, Tianhua, 279
Xia, Tianle, 87, 616
Xia, Xiaobo, 874
Xia, Xingyou, 597
Xia, Xue, 410
Xia, Yan, 857
Xia, Yihang, 676
Xia, Ying-Jieh, 477
Xia, Yingce, 180, 191
Xia, Yingjie, 435
Xia, Yu, 483
Xia, Yuqing, 900
Xia, Ziyi, 853
Xiang, 164, 166, 287, 288, 289, 336, 338, 377, 387, 397, 577, 581, 650, 761, 790, 855, 897

Xiang, Boran, 855
Xiang, Dawei, 287, 790
Xiang, Ke, 577
Xiang, Liuyu, 377
Xiang, Lu, 387
Xiang, Mengyu, 650
Xiang, Ruoyu, 166
Xiang, Wei, 288
Xiang, Wenxin, 289
Xiang, Yang, 761
Xiang, Yixin, 338
Xiang, Zhen, 164, 581
Xiang, Zhenghao, 897
Xiang, Zhishang, 336, 397
XIAO, 353, 687
Xiao, 87, 94, 97, 126, 134, 137, 145, 178, 181, 189, 191, 197, 216, 221, 233, 245, 249, 253, 258, 268, 297, 300, 303, 304, 309, 326, 333, 335, 346, 361, 387, 391, 396, 413, 416, 418, 434, 449, 450, 456, 480, 488, 491, 493, 506, 536, 543, 551, 570, 581, 582, 591, 600, 611, 614, 626, 644, 652, 654, 661, 665, 666, 669, 671, 675, 679, 687, 688, 693, 701, 722, 726, 734, 739, 740, 741, 743, 748, 753, 756, 757, 761, 772, 790, 793, 803, 830, 847, 848, 853, 854, 868, 877, 885, 888, 896, 904, 908
Xiao, Beier, 191
Xiao, Bowen, 726
Xiao, Boyu, 761
Xiao, Changxuan, 304
Xiao, Changyi, 654
Xiao, Chaojun, 543
Xiao, Chaowei, 488, 626
Xiao, Chenghao, 268, 434, 450, 669, 830
Xiao, Chengjun, 740
Xiao, Chenjun, 734
Xiao, Chuan, 396, 480
Xiao, Congxi, 197
Xiao, Feng, 216, 303, 885
Xiao, Guang, 679
Xiao, Hanlin, 506
Xiao, Hanshen, 178
Xiao, He, 493
Xiao, Hongjiang, 346, 693
Xiao, Hongwang, 300
Xiao, Jiangfeng, 396
Xiao, Jianha, 97
Xiao, Jin, 249
Xiao, Jinfeng, 868
Xiao, Jing, 391, 803
Xiao, Jingyu, 582
Xiao, Jipeng, 741
Xiao, Juesi, 94
Xiao, Jun, 126, 189, 304, 722
Xiao, Junhao, 790
Xiao, Juxin, 793
Xiao, Kejun, 614
Xiao, Kenan, 847
Xiao, Liqiang, 181
Xiao, Qiao, 253
Xiao, Qingying, 416
Xiao, Sheng, 753
Xiao, Shijian, 600
Xiao, Shiliang, 743
Xiao, Shitao, 491, 853
Xiao, Shuyao, 570
Xiao, Sibo, 536
Xiao, Siyang, 258
Xiao, Teng, 888
Xiao, Tong, 335, 387, 413, 456, 626, 675, 739, 748
Xiao, Wang, 361
Xiao, Wenyi, 297, 661
Xiao, Xi, 137
Xiao, Xiyao, 418, 652
Xiao, Xusen, 87
XIAO, Yan, 353, 687
Xiao, Yan, 333
Xiao, Yang, 134, 145, 309, 416, 611, 665, 734, 847
Xiao, Yanghua, 216, 249, 581, 644, 666, 671, 687, 688, 756, 757, 848, 908
Xiao, Yihao, 591
Xiao, Yilin, 245, 449, 854
Xiao, Yiming, 551
Xiao, Yiyong, 326
Xiao, Yongkang, 221
Xiao, Yongzhao, 877
Xiao, Yunze, 233, 772, 904
Xiao, Yuxin, 297
Xiao, Zeguan, 739
Xiao, Zhiping, 191
Xiao, Zhou, 701
Xiao, Ziyu, 896
XiaolingBai, 830
XiaolingBai, XiaolingBai, 830
XiaotengLiu, 895
XiaotengLiu, XiaotengLiu, 895
Xiaoyong, 95
Xiaoyong, Liu, 95
Xiaoyu, 560
Xiaoyu, Sui, 560
XIE, 815
Xie, 82, 95, 103, 120, 142, 145, 159, 163, 164, 166, 172, 191, 204, 236, 255, 257, 277, 297, 300, 306, 316, 343, 363, 375, 379, 380, 384, 388, 400, 405, 412, 428, 432, 434,

442, 445, 460, 469, 473, 487, 498, 501, 515, 527, 533, 536, 546, 547, 552, 554, 562, 568, 577, 583, 591, 603, 604, 609, 619, 627, 628, 648, 650, 678, 686, 687, 690, 705, 718, 724, 735, 736, 738, 745, 770, 795, 798, 815, 821, 840, 853, 861, 864, 878, 881, 886, 901
Xie, Boyi, 82, 609
Xie, Cy, 172
Xie, Donglin, 815
Xie, Gaogang, 577
Xie, Haoyang, 297
Xie, Huiru, 552
Xie, Huiyuan, 405, 735
Xie, Jiahong, 770
Xie, Jian, 546, 690
Xie, Jianwen, 901
Xie, Jiaxuan, 255
Xie, Jiayuan, 316
Xie, JingJing, 591
Xie, Lei, 363, 736
Xie, Miao, 300
Xie, Michael, 547
Xie, Pengjun, 120, 388, 473, 853
Xie, Pengtao, 103
Xie, Qianlong, 432
Xie, Qianqian, 166, 428, 686
Xie, Qingsong, 627
Xie, Qiuhao, 343
Xie, Qiujie, 840
Xie, Qizhuo, 554
Xie, Rui, 687
Xie, Ruobing, 163, 798, 821
Xie, Shaoan, 878
Xie, Shufang, 191, 487
Xie, Tao, 306, 501
Xie, Tian, 145
Xie, Tinghao, 562
Xie, Wei, 735
Xie, Weihao, 204
Xie, Wenxuan, 277, 515, 650
Xie, Wenya, 142, 164
XIE, XIE, 815
Xie, Xike, 400, 795
Xie, Xing, 375, 460, 469, 498, 533
Xie, Xinyuan, 705
Xie, Xiurui, 343
Xie, Xurong, 724
Xie, Yaochen, 861
Xie, Yi, 648
Xie, Yifeng, 738
Xie, Yong, 583
Xie, Yu, 533
Xie, Yuan, 412, 568
Xie, Yuanbo, 384
Xie, Yueqi, 533, 562
Xie, Yuexiang, 619
Xie, Yuhang, 380
Xie, Yupeng, 628, 687
Xie, Yutong, 738
Xie, Yuxi, 445
Xie, Zequn, 604, 650, 678
Xie, Zhenda, 379
Xie, Zheng, 718
Xie, Zheyong, 745
Xie, Zhifei, 384
Xie, Zhihui, 159
Xie, Zhiwen, 577
Xie, Zhiyu, 434
Xie, Zhongle, 536
Xie, Zhouhang, 257
Xie, Zhuohan, 95, 236, 442, 527, 864, 881, 886
Xie, Zijian, 690
Xie, Ziqian, 603
Xin, 227, 400, 795
Xin, Jiayi, 227
Xin, Tao, 400
Xin, Xin, 795
Xing, 169, 216, 236, 260, 262, 269, 271, 296, 315, 341, 401, 404, 427, 431, 442, 491, 554, 600, 630, 732, 733, 809, 878
Xing, Bowen, 315, 404
Xing, Chen, 262
Xing, Eric P., 809, 878
Xing, Jie, 401
Xing, Jingxuan, 341, 733
Xing, Lekai, 600
Xing, Luxi, 216
Xing, Mingzhe, 427
Xing, Rui, 236, 442
Xing, Shuo, 630
Xing, Tiancheng, 271
Xing, Victor, 169
Xing, Xingrun, 491
Xing, Yan, 260
Xing, Yu, 554
Xing, Yue, 269, 296
Xing, Zhening, 732
Xing, Zijie, 431
Xinran, 560, 597
Xinran, Liu, 597
Xinran, Xu, 560
XIONG, 166
Xiong, 94, 108, 111, 119, 144, 163, 164, 178, 197, 205, 212, 281, 285, 303, 306, 312, 314, 320, 325, 350, 351, 369, 386, 399, 402,

412, 435, 444, 508, 521, 540, 547, 590, 593, 603, 611, 618, 626, 639, 643, 672, 674, 681, 702, 729, 730, 732, 733, 746, 758, 773, 776, 789, 821, 843, 852, 866, 883, 892
Xiong, Aiping, 547
Xiong, Beibei, 350
Xiong, Bojian, 94
Xiong, Caiming, 281, 285, 866
Xiong, Chen, 789
Xiong, Chenyan, 314, 325
Xiong, Deyi, 94, 119, 306, 369, 386, 590, 618, 681, 729, 732
Xiong, Fangzhou, 892
Xiong, Feiyu, 674, 733
Xiong, Feng, 593, 776
Xiong, Fucheng, 435, 758
XIONG, GUOJUN, 166
Xiong, Hao, 746
Xiong, Hongyu, 611
Xiong, Hui, 144, 197, 205, 351, 730
Xiong, Jia, 672
Xiong, Jiahao, 212
Xiong, Jian, 178, 312
Xiong, Jing, 163, 821
Xiong, Jingting Zheng; Deyi,
Xiong, Jingwei, 789, 843
Xiong, Kai, 399, 626, 643
Xiong, Lei, 508
Xiong, Li, 883
Xiong, Lian, 111
Xiong, Qian, 639
Xiong, Ruibin, 540
Xiong, Ruochong, 435
Xiong, Ruoxuan, 402
Xiong, Shaopan, 702
Xiong, Shasha, 320
Xiong, Xiaoyu, 681
Xiong, Xiqiao, 303
Xiong, Xuan, 773
Xiong, Yongyi, 108
Xiong, Yu-Jie, 852
Xiong, Yuanhao, 444
Xiong, Zhiwei, 412
Xiong, Zidi, 164, 603
XU, 116, 567, 661
Xu, 83, 85, 87, 97, 106, 107, 108, 109, 117, 118, 121, 124, 126, 129, 132, 134, 136, 137, 140, 144, 148, 149, 151, 155, 157, 158, 159, 163, 166, 168, 172, 186, 187, 195, 201, 205, 207, 217, 226, 231, 233, 239, 243, 245, 246, 250, 251, 253, 255, 256, 258, 261, 272, 278, 285, 286, 287, 288, 289, 293, 296, 298, 303, 304, 305, 307, 316, 319, 320, 321, 325, 328, 329, 334, 340, 343, 344, 345, 347, 348, 350, 354, 355, 356, 363, 368, 370, 377, 379, 382, 383, 387, 390, 391, 392, 393, 395, 397, 398, 399, 406, 407, 409, 413, 416, 418, 419, 422, 424, 425, 429, 430, 431, 433, 434, 436, 442, 446, 447, 448, 449, 459, 460, 462, 464, 468, 469, 470, 471, 474, 476, 479, 490, 491, 493, 500, 501, 505, 507, 537, 540, 550, 551, 554, 560, 566, 576, 578, 582, 584, 585, 588, 591, 592, 593, 599, 601, 603, 616, 626, 630, 631, 634, 637, 640, 643, 651, 654, 655, 656, 657, 663, 665, 669, 670, 673, 676, 679, 680, 684, 685, 688, 692, 708, 712, 713, 715, 718, 723, 728, 729, 732, 733, 736, 737, 738, 740, 741, 747, 751, 753, 759, 761, 762, 775, 783, 784, 791, 792, 794, 795, 798, 821, 824, 830, 835, 844, 848, 849, 850, 851, 855, 858, 863, 865, 866, 867, 870, 872, 877, 882, 883, 884, 886, 887, 896, 899, 901, 909
xu, 129, 540, 587
Xu, Austin, 285, 866
Xu, Baixuan, 253, 733
Xu, Benfeng, 729
Xu, Bin, 476, 554, 631
Xu, Bing, 387, 762
Xu, Bingbing, 319, 584, 685, 901
Xu, Binqian, 383
Xu, Binxing, 272, 347
Xu, Bo, 747
Xu, Boyan, 738, 747
Xu, Buqiang, 655
Xu, Caijun, 654
Xu, Can, 474, 712, 713
Xu, Chang, 409, 783
Xu, Changwei, 651
Xu, Chao, 239, 505
Xu, Chejian, 887
Xu, Chen, 97, 231, 233, 442
Xu, Cheng, 468
Xu, Chenghao, 634, 654
Xu, Chengjin, 723
Xu, Chenjun, 168
Xu, Chenkai, 759
Xu, Chenliang, 490
Xu, Chenning, 872
Xu, Chenye, 251
Xu, Chuanfei, 387
Xu, Chunpu, 391
Xu, Dannong, 680

Xu, Depeng, 578
Xu, Derong, 791
Xu, Dexuan, 670
Xu, Dingling, 673
Xu, En, 855
Xu, Fan, 334, 383, 469, 684
Xu, Fangzhi, 159, 591, 851
Xu, Fengli, 255, 858
Xu, Guanghui, 540
Xu, Hainiu, 157
Xu, Haiyang, 126, 158
xu, Hanlin, 587
Xu, Hanwen, 433
Xu, Hao, 137, 325, 431
Xu, Hao-Xiang, 865
Xu, Haobo, 446, 460
Xu, Haojun, 87
Xu, Haolei, 751
Xu, Haoming, 398
Xu, Haotian, 424, 431, 673
Xu, Haozhe, 354
Xu, Heng-Da, 593
Xu, Hengyuan, 392
Xu, Hongfei, 345
Xu, Hongxia, 155, 676
Xu, Hongyuan, 718
Xu, Huangming, 370, 382
Xu, Huanmin, 278
Xu, Huazhe, 656
Xu, Huixin, 631
Xu, Jasmine, 246
Xu, Jia, 148, 303
Xu, Jiahao, 576, 909
Xu, Jiajie, 132, 144, 830
Xu, Jiajun, 136, 409
Xu, Jian, 187
Xu, Jiannan, 471
Xu, Jie, 343
Xu, Jiejun, 901
XU, JIEXIANG, 116, 567
Xu, Jin, 736
Xu, Jinfeng, 159, 305
Xu, Jing, 640
Xu, Jingjun, 124
Xu, Jingtao, 107
Xu, Jingwen, 321
Xu, Jingxuan, 258, 491
Xu, Jinsha, 307
Xu, Jun, 289, 328, 334, 395, 448, 884
Xu, Jungang, 500
Xu, Junhao, 630
Xu, Junjie, 347
Xu, Junjielong, 882
Xu, Kaidi, 334
Xu, Ke, 258
Xu, Keer, 217
Xu, Kerun, 733
Xu, Kewei, 398
Xu, Kun, 347
Xu, Lijiao, 821
Xu, Liming, 679
Xu, Linli, 137
Xu, Linlong, 406, 729, 732
Xu, Liyan, 296, 425
Xu, Lu, 459
Xu, Mengjia, 108
Xu, Miao, 436
Xu, Ming, 87, 616
Xu, Mingtao, 501
Xu, Mingze, 669
Xu, Mu, 593
Xu, Naen, 121, 566, 877
Xu, Nan, 316, 792
Xu, Nuo, 118, 657
Xu, Panpan, 442
Xu, Pau Tong Lin, 462
Xu, Peilan, 430
Xu, Peng, 134, 887
Xu, Pin, 304
Xu, Qiancheng, 149
Xu, Qiang, 298
Xu, Qidi, 740
Xu, Qiming, 205
Xu, Qingyang, 592
Xu, Qiongkai, 368
Xu, Ran, 794
Xu, Renjing, 507
Xu, Rongman, 821, 851
Xu, Rongwu, 85
Xu, Ruifeng, 692, 715
Xu, Ruijie, 651
Xu, RuiLiang, 239
Xu, Ruiyao, 172
Xu, Runhui, 129
Xu, Ruonan, 651
Xu, Ruotao, 761
Xu, Ruping, 316
Xu, Shanshan, 416
Xu, Shaoqing, 775
Xu, Shaoyang, 582
Xu, Shaoyuan, 835
Xu, Shawn Chang, 320
Xu, Shicheng, 355
Xu, Shuyao, 637
Xu, Tianlong, 205
Xu, Tianyang, 243, 863

Xu, Tianyi, 166
Xu, Tianze, 665
Xu, Tingqiang, 540
Xu, Tong, 340, 436, 560, 791, 896
Xu, Wanghan, 250
Xu, Wanpeng, 288
Xu, Wei, 85, 106, 464, 753, 886
Xu, Wei-Jie, 759
Xu, Weibo, 791
Xu, Weidi, 637
Xu, Weihong, 398
Xu, Weijie, 148, 447
Xu, Weiwen, 469, 747
Xu, Wen Qi, 737
Xu, Wenxi, 163
Xu, Wenxuan, 109
Xu, Wenyan, 287
Xu, Wenzhuo, 883
Xu, Wujiang, 393
Xu, Xiao, 688
Xu, Xiao-Fan, 418
Xu, Xiaofeng, 479
Xu, Xiaogang, 566
Xu, Xiaohua, 588
Xu, Xiaoyu, 207, 419
Xu, Xin, 363
XU, XINCHI, 661
Xu, Xingle, 643, 663
Xu, Xingzhong, 688
Xu, Xinran, 393
Xu, Xinyi, 108
Xu, Xiucheng, 319, 685
Xu, Xuenan, 390
Xu, Yan, 255
Xu, Yang, 140, 356, 399, 448, 493, 550, 585, 603, 867
Xu, Yangdi, 186
Xu, Yanyan, 377
Xu, Yichen, 195, 293
Xu, Yifan, 286
Xu, Yijue, 655
Xu, Yimeng, 728
Xu, Yinfei, 855
Xu, Ying, 824
Xu, Yinuo, 844
Xu, Yitao, 226
Xu, Yixuan, 784
Xu, Yongtian, 350
Xu, Yongxin, 850
Xu, Yu, 397, 759, 821
Xu, Yuanjian, 651, 670
Xu, Yuanwu, 344
Xu, Yuchen, 83
Xu, Yuling, 261
Xu, Yunqing, 470
Xu, Yunqiu, 643
Xu, Yuzhuang, 398, 665
Xu, Zenan, 540, 848
Xu, Zenglin, 256, 550
Xu, Zhangchen, 599
Xu, Zhao, 407, 738, 798
Xu, Zhaozhuo, 245, 449, 626
Xu, Zhe, 849
Xu, Zhe-Yu, 551
Xu, Zhen, 429
Xu, Zhenbo, 285, 422
Xu, ZhenChun, 340
Xu, Zheng, 601
Xu, Zhenran, 379
Xu, Zhi, 870
Xu, Zhichao, 442
Xu, Zhihao, 151, 741
Xu, Zhijian, 298
xu, Zhijiang, 540
Xu, Zhikun, 348
Xu, Zhipeng, 434
xu, zhiqiang, 129
Xu, Zhiwei, 795
Xu, Zhiwu, 329
Xu, Zhiyu, 537
Xu, Zhongxing, 592
Xu, Zijie, 392, 740
Xu, Ziqi, 201
Xu, Ziqiang, 413
Xu, Zishan, 157, 899
Xu, Ziwen, 398, 708
Xu, Zixiang, 117
Xu, Ziyang, 233
Xu, Ziyao, 399
Xu, Zukang, 718
Xuan, 164, 186, 292, 405, 572, 772, 904
Xuan, Minh Chu, 405
Xuan, Richeng, 186, 292
Xuan, Weihao, 164, 572, 772, 904
XuanrenChen, 257
XuanrenChen, XuanrenChen, 257
XudongNiu, 713
XudongNiu, XudongNiu, 713
XUE, 436, 451, 903
Xue, 95, 132, 165, 230, 234, 278, 295, 305, 315, 363, 398, 418, 435, 443, 461, 470, 474, 540, 543, 585, 597, 605, 618, 632, 685, 708, 724, 751, 752, 775, 797, 821, 842, 846, 847, 854, 858, 886
Xue, Bing, 752
XUE, Boyang, 436, 451, 903

Xue, Chao, 165, 230, 295
Xue, Cheng, 418
Xue, Dongyun, 315, 461
Xue, Feng, 685
Xue, Guangquan, 632
Xue, Hanlin, 435
Xue, Hao, 443, 597, 854
Xue, Hongfei, 363
Xue, Hui, 132, 398, 708, 751
Xue, Jinbao, 540
Xue, Letian, 775
Xue, Liumeng, 474
Xue, Lyuxin, 846
Xue, Mengge, 605, 847
Xue, Mingfeng, 858
Xue, Nianwen, 305, 842
Xue, Qibo, 821
Xue, Wei, 470, 543, 618
Xue, Weizhi, 234
Xue, Xiangyuan, 797
Xue, Xianwei, 278
Xue, Yida, 585
Xue, Yingjie, 597
Xue, Yun, 95, 886
Xue, Zhucun, 724
Xueyun, 319
Xueyun, Tian, 319
xuye, 729
xuye, xuye, 729
XWang, 105
XWang, XWang, 105
Xypolopoulos, 865
Xypolopoulos, Christos, 865

Yabo, 830
Yabo, Li, 830
Yadav, 160, 161, 290, 441, 652
Yadav, Neemesh, 290
Yadav, Ravindra Kumar, 160, 161
Yadav, Sumit, 652
Yadav, Vikas, 441
Yadavalli, 876
Yadavalli, Aditya, 876
Yadavally, 532
Yadavally, Aashish, 532
Yaghoobzadeh, 704
Yaghoobzadeh, Yadollah, 704
Yagneswaran, 160
Yagneswaran, Sriraja, 160
Yahan_Zheng,
Yahan_Zheng, Yahan_Zheng,
Yahiaoui, 494
Yahiaoui, Rashid, 494
Yahyal, 266
Yahyal, Mohamed, 266
Yakefu, 886
Yakefu, Adina, 886
Yakun, 470
Yakun, Cui, 470
Yaldiz, 271
Yaldiz, Duygu Nur, 271
Yamagata, 854
Yamagata, Yuki, 854
Yamagishi, 768
Yamagishi, Junichi, 768
Yamagiwa, 885
Yamagiwa, Hiroaki, 885
Yamaguchi, 176, 242, 700
Yamaguchi, Atsuki, 176, 242, 700
Yamamoto, 113, 416
Yamamoto, Kento, 416
Yamamoto, Yuji, 113
Yamauchi, 454
Yamauchi, Yusuke, 454
Yamshchikov, 110, 393
Yamshchikov, Ivan P., 110, 393
YAN, 126
Yan, 87, 92, 93, 126, 129, 142, 144, 158, 178, 191, 205, 209, 214, 250, 262, 265, 268, 272, 289, 292, 295, 301, 318, 330, 331, 345, 349, 355, 363, 377, 384, 387, 406, 407, 412, 417, 440, 442, 446, 448, 468, 487, 501, 522, 538, 540, 542, 562, 624, 628, 634, 659, 664, 669, 673, 698, 702, 707, 713, 718, 719, 722, 725, 748, 749, 771, 775, 776, 787, 797, 798, 803, 817, 845, 851, 856, 865, 876, 877, 878, 884, 905
Yan, Alison, 331
Yan, An, 250
Yan, Bingyu, 272, 787
Yan, Brian, 542
Yan, Chenqian, 538
Yan, Chenyu, 330
Yan, Dong, 363
Yan, Feng, 817
Yan, Hang, 301, 851
Yan, Hanqi, 803
Yan, Hongfei, 209, 501
Yan, Jianfeng, 540
Yan, Jiexi, 634
Yan, Jun, 331
Yan, Junbing, 87, 562
Yan, Juncheng, 448
Yan, Kaiwen, 292
Yan, Kezhong, 702
Yan, Liang, 776

Yan, Lingyong, 713
Yan, Liu, 92
Yan, Meng, 628
Yan, Ming, 93, 126, 158, 214, 797
Yan, Nan, 468
Yan, Peihong Li; Yan,
Yan, Peng, 349, 659
Yan, Qiang, 407
Yan, Qiaojing, 442
Yan, Qiuyan, 262
Yan, Rong, 318
Yan, Rui, 129, 384, 487, 624, 798, 865, 876, 905
Yan, Ruiqi, 748
Yan, Ruiyi, 178, 884
Yan, Shannan, 698
Yan, Shen, 845
Yan, Shilinlu, 775
Yan, Shiqi, 856
YAN, Shuicheng, 126
Yan, Shuicheng, 384
Yan, Shuo, 142
Yan, Sikuan, 191, 289
Yan, Sixing, 268
Yan, Siyu, 877
Yan, Weicai, 355, 749
Yan, Weixiang, 878
Yan, Wenbiao, 719
Yan, Wenbo, 268
Yan, Wentao, 412
Yan, Xiangchao, 440
Yan, Xiao, 295, 664
Yan, Yibo, 144, 205, 707
Yan, Yizhou, 725
Yan, Yu, 417, 669
Yan, Yuchen, 446, 722
Yan, Yujun, 771
Yan, Yukun, 377, 387, 406, 673
Yan, Yuliang, 265
Yan, Zhenyu, 771
Yan, Zimo, 718
Yan, Zirui, 771
Yan, Ziyue, 345
Yanaka, 206, 228, 572
Yanaka, Hitomi, 206, 228, 572
Yanaka;,
Yanaka;, Jin Tanaka; Daiki Matsuoka; Ryoma Kumon; Hitomi,
Yanaka;, Kaito Watanabe; Taisei Yamamoto; Tomoki Doi; Hitomi,
yanbingren, 303, 885
yanbingren, yanbingren, 303, 885
Yanfeng, 783
Yanfeng, Ding, 783
YANG, 161, 773, 802
Yang, 85, 87, 89, 93, 94, 96, 97, 103, 105, 107, 111, 112, 118, 119, 120, 122, 126, 127, 128, 130, 131, 132, 134, 135, 137, 138, 140, 142, 143, 146, 148, 149, 152, 154, 157, 158, 159, 161, 162, 163, 165, 166, 170, 171, 172, 181, 182, 184, 185, 186, 187, 189, 191, 194, 197, 199, 205, 208, 212, 215, 216, 217, 219, 221, 228, 236, 241, 242, 248, 249, 251, 253, 254, 256, 258, 260, 264, 265, 269, 270, 271, 272, 278, 282, 283, 286, 289, 292, 296, 297, 301, 303, 304, 306, 312, 316, 321, 324, 325, 326, 333, 335, 339, 342, 343, 347, 348, 349, 350, 352, 363, 364, 369, 370, 371, 376, 377, 378, 379, 382, 386, 387, 390, 395, 396, 398, 400, 401, 402, 403, 404, 407, 409, 410, 413, 414, 415, 418, 419, 420, 421, 422, 424, 425, 431, 432, 434, 435, 441, 448, 450, 452, 453, 455, 457, 464, 466, 470, 473, 475, 479, 485, 493, 496, 497, 500, 505, 506, 507, 512, 513, 514, 515, 519, 528, 529, 531, 532, 533, 534, 535, 536, 538, 539, 540, 545, 550, 554, 556, 558, 560, 561, 572, 573, 575, 582, 590, 591, 594, 595, 596, 602, 604, 610, 617, 620, 625, 626, 627, 629, 630, 634, 635, 636, 640, 642, 643, 645, 646, 648, 650, 652, 654, 656, 657, 659, 660, 663, 668, 670, 671, 673, 674, 675, 679, 680, 681, 683, 686, 687, 688, 689, 690, 693, 695, 696, 697, 698, 703, 707, 708, 714, 715, 716, 717, 718, 719, 723, 724, 727, 730, 731, 732, 734, 737, 738, 739, 740, 741, 742, 743, 746, 748, 749, 753, 756, 759, 761, 762, 774, 775, 782, 789, 790, 794, 795, 796, 798, 801, 803, 814, 815, 817, 819, 825, 826, 827, 833, 839, 845, 846, 847, 849, 850, 853, 854, 856, 861, 864, 873, 883, 884, 885, 886, 888, 893, 895, 897, 899, 900, 907, 908
yang, 255, 742
Yang, Aijun, 695
Yang, An, 560
Yang, Andy, 165
Yang, Ao, 431
Yang, Baosong, 94
Yang, Biao, 379
Yang, Bin-Bin, 303, 885
Yang, Binhan, 435
Yang, Bo, 679
Yang, Bowen, 591, 671
Yang, Carl, 794

Yang, Cehao, 723
Yang, Chang, 642
Yang, Changyuan, 533, 798
Yang, Chao, 297, 717
Yang, Chao-Han Huck, 217, 529, 908
Yang, Chen, 316
Yang, Chenchen, 228
Yang, Cheng, 146, 157, 303
Yang, Chenghao, 707
Yang, Chengran, 515, 531
Yang, Chengyi, 514
Yang, Chenxu, 152
Yang, Chenyang, 306, 845
Yang, Chun, 403
Yang, Chunjie, 817
Yang, Chunyu, 675, 681
Yang, Dan, 604
Yang, David H., 215, 536
Yang, Dawei, 718
Yang, Deqing, 216, 249, 572
Yang, Dingwen, 545
Yang, Dingyi, 497, 861
Yang, Diyi, 464, 827
Yang, Dongchao, 339, 435
Yang, Dongdong, 883
Yang, Ellie, 199
Yang, Enhui, 740
Yang, Eric, 212, 864
Yang, Eugene, 154
Yang, Eunho, 120, 122, 217
Yang, Fan, 205, 253, 379
Yang, Fangkai, 349
Yang, Fei, 414
Yang, Fuchao, 158
Yang, Fumeng, 473
Yang, Fuwei, 400
Yang, Fuzhen, 741
Yang, Grace Hui, 182, 839
Yang, Guangyu, 594
Yang, Guanrou, 432, 455
Yang, Haihua, 333
Yang, Haiqin, 650
Yang, Haitong, 312
Yang, Han, 435
Yang, Hanqing, 707
Yang, Haofu, 135
Yang, Haojin, 636
Yang, Haoqi, 814
Yang, Hong-Ming, 187
Yang, Hongxia, 131
Yang, Howard H., 731
Yang, Hui-Ying, 265
Yang, Huiwen, 161
Yang, Hyeongjun, 690
Yang, Ivory, 466
Yang, Jiabin, 297
Yang, Jian, 194, 260, 335, 485, 660, 687, 739, 825, 853
Yang, Jianfei, 519
Yang, Jiang-Ming, 846, 856
Yang, Jianlei, 128
Yang, Jiaxi, 634
Yang, Jiaye, 849
Yang, Jicheng, 304
Yang, Jie, 448
Yang, Jin, 679
Yang, Jing, 219, 590, 900
Yang, Jing WangYaomin WuYinglin WangYitong, 660
Yang, Jingbo, 604
Yang, Jinglin, 413
Yang, Jingpu, 475
Yang, Jingwen, 421, 674
Yang, Jingyue, 670
Yang, Jinhai, 774
Yang, Jirui, 646
Yang, Jisoo, 596
Yang, Jiyan, 795
Yang, Jun, 627
Yang, Junwei, 191
Yang, Junxiao, 132, 398
Yang, Junyao, 419
Yang, Kai ChenQiang, 390
Yang, Kailai, 573, 856
Yang, Kaiming, 126
Yang, Kang, 138
Yang, Ke, 197, 364
Yang, Kejuan, 847
Yang, Kevin, 194
Yang, Kexin, 560
Yang, Kisu, 594
Yang, Kuo, 143, 714
Yang, Kyuseok, 833
YANG, LAN, 802
Yang, Lei, 94, 386, 402
Yang, Li, 534
Yang, Liang, 378, 431
Yang, Liangwei, 111
Yang, Lin, 333
Yang, Lina, 93
Yang, Ling, 656, 668
Yang, Linyi, 689
Yang, Liu, 795
Yang, Longhao, 854
Yang, Longqi, 479, 561
Yang, Lujia, 749

Yang, Luoxiao, 629
Yang, Mao, 573, 636
Yang, Meng, 671
Yang, Menglin, 424, 732
Yang, Mengxue, 403
Yang, Mengyue, 248
Yang, Min, 119, 142, 149, 348, 849
Yang, Ming-Hsuan, 470
Yang, Mingbo, 687
Yang, Morunliu, 716
Yang, Muli, 634
Yang, MuRun, 413
Yang, Murun, 626
Yang, Muyun, 762
Yang, Nakyeong, 283, 296, 900
Yang, Nan, 89, 659
Yang, Ni, 627
Yang, Ning, 719
Yang, Ningyuan, 538
Yang, Pei, 864
Yang, Peiji, 656
Yang, Peiran, 85
Yang, Peiru, 401
Yang, Pengfei, 680
Yang, Penghui, 535
Yang, Qingqing, 258
Yang, Qingsong, 886
Yang, Qingyao, 493
Yang, Qinmin, 893
Yang, Qishun, 171
Yang, Qu, 149, 425, 716
Yang, Ronghui, 654
Yang, Ruichao, 105
Yang, Ruihan, 572
Yang, Ruilin, 884
Yang, Ruiqing, 819
Yang, Ruixuan, 409
Yang, Run, 241
Yang, Runde, 617
Yang, Runhan, 734
YANG, RUNMING, 773
Yang, Runming, 140
Yang, Ruozhen, 789
Yang, Saiyong, 149, 539
Yang, Sen, 560, 703, 756
Yang, Seungbin, 595
Yang, Seunghan, 107
Yang, Shanshan, 166
Yang, Shaoxiong, 256
Yang, Shen, 554
Yang, Shidong, 378
Yang, Shiduo, 413
Yang, Shiji, 746
Yang, Shijia, 743
Yang, Shiping, 163, 264
Yang, Shu, 171, 582, 602, 803, 886
Yang, Shuai, 506
Yang, Shudong, 130
Yang, Shufan, 801
Yang, Shuhua, 854
Yang, Shujian, 432
Yang, Shuo, 159, 352, 540, 681, 815
Yang, Siling, 753
Yang, Sohee, 742
Yang, Songbo, 707
Yang, Soyoung, 301
yang, Sungkyu, 255
Yang, Tengchao, 505
Yang, Tianchang, 556
Yang, Tiankai, 172
Yang, Tianmeng, 715
Yang, Tianpei, 657
Yang, Tianyu, 269, 620, 650
Yang, Tianzhuo, 85, 748
Yang, Tong, 87, 303
Yang, Tongshen, 278
Yang, Tongtong, 690
Yang, Van, 895
Yang, Wanli, 371
Yang, Wei, 260, 306, 849
Yang, Weikai, 640
Yang, Wendy H., 826
Yang, Wenhao, 507
Yang, Wenjin, 801
Yang, Wenkai, 573
Yang, Wenmian, 137, 189, 512
Yang, Wenqi, 407
Yang, Xi, 186, 496, 782, 833
Yang, Xianglin, 333
Yang, Xianjun, 118
Yang, Xiao, 158
Yang, Xiaobo, 292
Yang, Xiaochen, 420
Yang, Xiaocheng, 441
Yang, Xiaochun, 688
Yang, Xiaocui, 326, 422, 663, 671, 673, 730
Yang, Xiaogui, 708
Yang, Xiaokang, 377
Yang, Xiaoliang, 739
Yang, Xiaoyan, 395
yang, xiaoyan, 742
Yang, Xin, 159, 475, 854
Yang, Xinchen, 103
Yang, Xingchang, 396
Yang, Xinghao, 387
Yang, Xintong, 272, 347

Yang, Xinwei, 236
Yang, Xinyu, 184, 528, 591, 604
Yang, Xiufeng, 289
Yang, Xiulin, 242
Yang, Xiyuan, 533
Yang, Xu, 654
Yang, Xue, 296, 629, 652
Yang, Xuemeng, 148, 303
Yang, Xun, 171, 561
Yang, Yajie, 387
Yang, Yang, 135, 296, 404, 418, 635, 738
Yang, Yaodong, 85, 513, 532, 748
Yang, Yaoqing, 154, 208
Yang, Yazheng, 321
Yang, Yehan, 96
Yang, Yi, 401, 643, 668, 693, 734
Yang, Yibo, 648, 845
Yang, Yifan, 352, 363, 396, 431, 432, 455
Yang, Yifei, 888
Yang, Yihao, 254
Yang, Yingxuan, 271, 794
Yang, Yingzhen, 185
Yang, Yiqian, 441
Yang, Yiran, 698
Yang, Yixiang, 732
Yang, Yiying, 650
Yang, Yizhe, 97, 187
Yang, Yonghui, 171, 415, 645
YANG, YOONMO, 161
Yang, Yu, 342
Yang, Yuancheng, 333
Yang, Yue, 205, 652
Yang, Yuhang, 680
Yang, Yuhao, 907
Yang, Yujia, 500
Yang, Yujin, 610
Yang, Yujiu, 140, 251, 696, 724
Yang, Yuming, 162, 376, 545, 839
Yang, Yunfan, 723
Yang, Yuning, 343
Yang, Yunqiao, 432
Yang, Yuqin, 856
Yang, Yuqing, 452
Yang, Yutong, 270
Yang, Yuwei, 282
Yang, Yuwen, 630
Yang, Yuzhe, 181
Yang, Ze, 575
Yang, Zelai, 134
Yang, Zesheng, 248
Yang, Zeyu, 873
Yang, Zhao, 325
Yang, Zhaoyu, 727
Yang, Zheming, 761
Yang, Zhen, 286, 453, 625, 727, 737
Yang, Zhengfeng, 350
Yang, Zhengyuan, 112
Yang, Zhenhua, 324
Yang, Zhenyu, 410, 627
Yang, Zheyuan, 410
Yang, Zhi-Xin, 775
Yang, Zhibang, 850
Yang, ZhiBo, 450
Yang, Zhibo, 303, 434, 885
Yang, Zhichao, 127, 221
Yang, Zhidong, 708
Yang, Zhixiong, 897
Yang, Zhixuan, 370, 382
Yang, Zhiyuan, 759
Yang, Zhizhuo, 558
Yang, Zhongjun, 545, 550, 796
Yang, Zhongliang, 714
Yang, Zhuoran, 743
Yang, Zichen DingBowen, 591
Yang, Zijie, 369
Yang, Ziliang, 899
Yang, Ziqing, 111, 457
Yang, Zixuan, 697
Yang, Ziyi, 790
Yang, Ziyuan, 170
Yang, Zongtao, 683
Yang;,
Yang;, Cheng,
Yang;, Zonglin,
Yangjunqi, 555
Yangjunqi, Yangjunqi, 555
Yangxue, 379
Yangxue, Yangxue, 379
Yannakoudakis, 195
Yannakoudakis, Helen, 195
Yano, 452
Yano, Ken, 452
Yanqing, 95
Yanqing, He, 95
Yantosca, 655
Yantosca, Michael S., 655
Yanuka, 777
Yanuka, Moran, 777
YanyiSu, 740
YanyiSu, YanyiSu, 740
YAO, 161, 162
Yao, 87, 106, 118, 129, 135, 144, 152, 154, 157, 184, 211, 220, 221, 234, 252, 272, 274, 278, 282, 292, 295, 320, 327, 330, 347, 362, 371, 377, 384, 386, 393, 398, 404, 410, 427, 430, 454, 460, 462, 469, 472,

476, 483, 491, 492, 496, 498, 499, 514, 527, 528, 530, 560, 563, 569, 595, 596, 599, 617, 619, 624, 635, 638, 655, 676, 694, 696, 708, 716, 724, 736, 747, 750, 757, 782, 784, 792, 801, 806, 824, 832, 856, 859, 861, 871, 884, 899
Yao, Benjamin Z., 617
Yao, Bingsheng, 824, 861
Yao, Chang, 347
Yao, Chengyuan, 292, 560
Yao, Dingyu, 152
Yao, Feng, 747
Yao, Hanshu, 393
Yao, Hantao, 282
Yao, Huaiyuan, 492
Yao, Huaxiu, 528, 599, 832
Yao, Jiacheng, 410
Yao, Jiadi, 716
YAO, Jianmin, 162
Yao, Jiayi, 472
Yao, Jiayu, 635
Yao, Jing, 460, 469, 498
Yao, Jingyu, 430
Yao, Junfeng, 371
Yao, Junyi, 499
Yao, Laura, 792
Yao, Li, 724
Yao, Liang, 635
Yao, Lin, 330
Yao, Lina, 483, 514
Yao, Ling, 596
Yao, Liuyi, 619
Yao, Louie Hong, 454
Yao, Maoqing, 404
Yao, Peisen, 184
Yao, Qian, 750
Yao, Shenglong, 569
Yao, Shu, 327
Yao, Sijia, 87
Yao, Weijun, 362, 694
Yao, Wen, 135, 384
Yao, Xianrong, 491, 801, 856
Yao, Xiaozhe, 784
YAO, Xin, 161
Yao, Xin, 295
Yao, Xingbo, 144
Yao, Xingyu, 320
Yao, Xuesong, 272
Yao, Yang, 696
Yao, Yao, 184, 234, 292
Yao, Yian, 736
Yao, Yifan, 676
Yao, Yihang, 427
Yao, Yiming, 386
Yao, Yina, 806
Yao, Yiqun, 638
Yao, YiXing, 871
Yao, Yuan, 377, 527, 624
Yao, Yuanshun, 118
Yao, Yuguang, 563, 595
Yao, Yunzhi, 398, 655, 708
Yao, Yutong, 129, 274
Yao, Zhewei, 252, 496
Yao, Zhiming, 884
Yao, Zhiyuan, 157, 278, 899
Yao, Zhongjiang, 462
Yao, Zijun, 220, 476, 530, 757, 782, 859
Yao, Ziyu, 154
Yao, Zonghai, 106, 211, 221
Yap, 278
Yap, Kim-Hui, 278
Yari, 562, 589
Yari, Amir Hossein, 562, 589
Yariv, 544
Yariv, Guy, 544
Yashwante, 890
Yashwante, Pratham, 890
Yates, 163, 862
Yates, Andrew, 163, 862
Yatskar, 829
Yatskar, Mark, 829
Ye, 120, 126, 147, 149, 178, 194, 207, 243, 244, 249, 259, 260, 276, 288, 300, 309, 310, 311, 313, 327, 332, 345, 353, 360, 363, 368, 379, 381, 382, 384, 387, 389, 395, 397, 405, 425, 428, 433, 445, 448, 460, 469, 471, 483, 506, 507, 516, 527, 529, 532, 534, 535, 541, 544, 545, 561, 573, 591, 597, 607, 617, 644, 651, 659, 680, 687, 688, 716, 717, 718, 735, 746, 747, 756, 774, 790, 814, 836, 838, 848, 849, 851, 853, 897, 899, 900
Ye, Chao, 368, 680
Ye, Deheng, 756
Ye, Dengpan, 507, 597
Ye, Fangda, 384
Ye, Fanghua, 149, 276, 425, 716
Ye, Fengying, 541, 899
Ye, Fuda, 483
Ye, Guanghui, 313, 544
Ye, Guangnan, 395, 561, 716, 746
Ye, Hanrong, 529
Ye, Hao, 397
Ye, Haoran, 445
Ye, He, 389
Ye, Hongfei, 310, 345

Ye, Jiabo, 126
Ye, Jiaqi, 718
Ye, Jiayi, 259
Ye, Jing, 387
Ye, Jingheng, 460, 535
Ye, Jingyong, 178
Ye, Jinyu, 147, 814
Ye, Junhong, 363
Ye, Junjie, 382, 545, 717, 838, 897
Ye, Lingyu, 469
Ye, Maoxiao, 687
Ye, Muyang, 379, 389
Ye, Peng, 309
Ye, Qi, 327, 836
Ye, Qilang, 747
Ye, Qiming, 288
Ye, Qingqing, 207
Ye, Qiwei, 300
Ye, Rui, 120
Ye, Shenghao, 194, 774, 853
Ye, Shuo, 249, 747
Ye, Shuqi, 573
Ye, Tianzhu, 900
Ye, Tong, 532
Ye, Wei, 448, 506, 651, 688, 735, 848
Ye, Xi, 853
Ye, Xiaotian, 516
Ye, Xiaotong, 311
Ye, Xiaoyang, 607
Ye, Xinfeng, 687
Ye, Xinge, 849
Ye, Yanfang, 178, 471
Ye, Yang, 332
Ye, Yangfan, 353, 381
Ye, Yanwei, 527
Ye, Yefang, 534
Ye, Yongshi, 433
Ye, Yuxiao, 244, 405, 735
Ye, Yuxuan, 428
Ye, Ze, 617
Ye, Zekai, 360
Ye, Zhankai, 790
Ye, Zhengmao, 260, 591
Ye, Zhuangzhuang, 659
Ye, Zihe, 644
Ye, Ziping, 851
Ye, Ziyi, 243
Yeboah, 538
Yeboah, Prisca Adwoa Dufie, 538
Yeh, 151, 153, 196, 262, 467, 534, 547, 787, 807
Yeh, Chin-Chia Michael, 787
Yeh, Jia-Fong, 153, 196
Yeh, Mi-Yen, 467, 547, 807
Yeh, Samuel, 151
Yeh, Shu-ping, 534
Yeh, Yi-Ren, 262
Yehudai, 230, 526, 557, 783
Yehudai, Asaf, 230, 526, 557, 783
Yeltay, 589
Yeltay, Zhastay, 589
Yeo, 147, 177, 217, 230, 274, 456, 741, 855
Yeo, Eunjung, 177, 217, 456
Yeo, Jinyoung, 230, 741
Yeo, Juan, 147
Yeo, Woongyeong, 855
Yeo, YoonDong, 274
Yeoh, 92
Yeoh, William, 92
Yeom, 147, 553, 885
Yeom, Jewon, 147, 553, 885
Yergattikar, 616
Yergattikar, Saurabh, 616
Yeung, 425
Yeung, Dit-Yan, 425
Yi, 97, 122, 123, 149, 161, 179, 187, 204, 377, 390, 460, 462, 469, 472, 498, 500, 533, 540, 660, 723, 849, 858
Yi, Biao, 723
Yi, Bowen, 472
Yi, Hongzhu, 187, 462, 500
Yi, JI, 161
Yi, Jingwei, 533
Yi, John Seon Keun, 858
Yi, Mun Yong, 123
Yi, Munyong, 390
Yi, Qi, 540
Yi, Xiaoquan, 660
Yi, Xiaoyang, 122, 179
Yi, Xiaoyuan, 377, 460, 469, 498
Yi, Yang, 97
Yi, Zhengyuan, 204
Yi, Zihao, 149, 849
Yifan, 389
Yifan, Chen, 389
Yifei, 829
Yifei, Li S., 829
Yildiz, 199
Yildiz, Alperen, 199
Yilmaz, 119, 150, 451, 841
Yilmaz, Emine, 119, 150, 451
Yilmaz, Ozan, 841
Yim, 161, 243
Yim, Yauwai, 161, 243
Yimam, 471, 862
Yimam, Seid Muhie, 471, 862
yiming, 581, 782

yiming, Xue, 581, 782
Yin, 93, 97, 105, 107, 120, 121, 138, 159, 204, 228, 243, 255, 279, 289, 296, 310, 335, 349, 353, 355, 376, 384, 431, 433, 434, 440, 445, 451, 456, 470, 480, 486, 512, 528, 531, 546, 551, 559, 568, 579, 587, 600, 628, 635, 636, 673, 699, 713, 715, 716, 725, 739, 746, 748, 749, 758, 766, 776, 784, 795, 797, 800, 801, 809, 817, 853, 874, 881, 891, 892, 904
yin, 245, 417, 449, 597
Yin, Bing, 433
Yin, Boqin, 470
Yin, Chenlong, 93, 559
Yin, Congchi, 243
Yin, Dapeng, 874
Yin, Dawei, 289, 310, 699, 713, 725, 795
Yin, Deyi, 353
yin, di, 245, 449
Yin, Fei, 431, 636
Yin, Guojun, 384, 568, 784
Yin, Huifeng, 120, 853
Yin, Jianwei, 749
Yin, Kai, 551
Yin, Kang, 456
Yin, Linqi, 228
Yin, Ming, 107
Yin, Qingyu, 800, 817
Yin, Qixiang, 904
Yin, Quanjun, 335, 673
Yin, Rui, 121
Yin, Shangjian, 480, 579
Yin, Shunshun, 748
Yin, Sinjoy Saha; Suman Saha; Mahfuza Farooque; Wenpeng,
Yin, Sophia, 280
Yin, Tao, 628
Yin, Wei, 138
Yin, Weichong, 715
Yin, Wenbiao, 853
Yin, Wenpeng, 881
Yin, Xiaoming, 434
Yin, Xin, 349
Yin, Xinxiang, 531
Yin, Xu-Cheng, 105
Yin, Xuanwu, 355
Yin, Xuefei, 376
Yin, Xunjian, 445
Yin, Yafeng, 97, 801
Yin, Yanbin, 809
Yin, Yichun, 255, 587
Yin, Yihang, 384
Yin, Yishu, 776
Yin, Yongjing, 739
Yin, Yueqin, 486
Yin, Yuwei, 891
Yin, Zhangyue, 159, 204, 546, 600
Yin, Zhaoxia, 296, 568, 892
Yin, Zhenfei, 440, 716, 746, 797
Yin, Zhenyun, 451
yin, Zhiyi, 417, 597
Yin, Zhongrui, 758
Yin, Ziling, 635
Yin, Ziyi, 528, 766
YING, 767
Ying, 102, 205, 209, 297, 625, 768, 883, 885
Ying, Chenduo, 297
Ying, Haochao, 885
Ying, Jiahao, 625
Ying, Jie, 209
Ying, Rex, 205
YING, TING, 767
Ying, Xiaoying, 768
Ying, Zhu Irene, 102
Ying, Zonghao, 883
YingyuShan, 737
YingyuShan, YingyuShan, 737
YIWU, 652
YIWU, YIWU, 652
Yokoi, 113, 839
Yokoi, Sho, 113, 839
Yokoya, 572, 772
Yokoya, Naoto, 572, 772
Yona, 284
Yona, Itay, 284
Yoo, 104, 223, 229, 267, 280, 445, 513, 517, 590, 612
Yoo, Chuck, 612
Yoo, Dong Whi, 513
Yoo, HaeJun, 445
Yoo, Haneul, 590
Yoo, Hyunwoo, 104, 229, 267, 280
Yoo, Kang Min, 280
Yoo, Suho, 223
Yoo, YoungJoon, 517
Yoon, 113, 179, 183, 222, 230, 267, 275, 279, 297, 385, 463, 508, 549, 552, 557, 596, 601, 678, 828, 858, 903
Yoon, Chaewon, 297
Yoon, Hoonsang, 183
Yoon, Hyung-Jin, 385
Yoon, Hyungjun, 549
Yoon, Jaehong, 222
Yoon, Janghan, 552
Yoon, Jeongho, 601, 858
Yoon, Ji Won, 903

Yoon, Jinsung, 678
Yoon, Sejong, 279
Yoon, Seohee, 275
Yoon, Seunghyouk, 828
Yoon, Seunghyun, 179, 463, 557
Yoon, Sungroh, 113, 267
Yoon, Sunjae, 596
Yoon, Susik, 230
Yoon, Youngsik, 508
Yortucboylu, 609
Yortucboylu, Evren, 609
Yosha, 113
Yosha, Iddo, 113
Yoshida, 102, 639
Yoshida, Ryo, 102, 639
Yoshie, 884
Yoshie, Osamu, 884
Yoshinaga, 190
Yoshinaga, Naoki, 190
Yotam, 794
Yotam, Tom, 794
You, 124, 136, 221, 335, 348, 415, 427, 455, 533, 555, 599, 636, 765, 775, 779, 825
You, Bo, 775
You, Datao, 427
You, Fuming, 455
You, Jiaxuan, 124, 555
You, Qi, 765
You, Runyang, 415
You, Weihang, 599
You, Weiqiu, 825
You, Weitao, 533
You, Wenhao, 136
You, Wenjia, 779
You, Xiaoyu, 348
You, Yang, 636
You, Zeng, 335
You, Zhiyuan, 221
Youn, 612
Youn, Geonjae, 612
Yousef, 860
Yousef, Majdal, 860
Yu, 84, 86, 87, 111, 116, 119, 120, 121, 124, 126, 131, 138, 143, 146, 147, 151, 158, 161, 166, 187, 191, 203, 205, 208, 211, 239, 241, 248, 249, 253, 254, 258, 259, 260, 262, 271, 280, 289, 296, 297, 299, 303, 304, 305, 307, 312, 313, 316, 319, 320, 321, 324, 325, 327, 329, 330, 339, 344, 348, 349, 357, 358, 364, 370, 371, 372, 377, 379, 391, 393, 395, 406, 407, 416, 420, 423, 424, 425, 429, 431, 433, 434, 436, 439, 444, 448, 456, 458, 459, 463, 474, 476, 483, 488, 490, 508, 541, 546, 548, 551, 552, 554, 560, 564, 567, 568, 572, 576, 577, 580, 581, 591, 595, 600, 601, 605, 607, 616, 619, 621, 624, 629, 630, 632, 638, 641, 650, 651, 666, 673, 674, 676, 684, 685, 687, 693, 707, 709, 713, 714, 715, 722, 723, 724, 725, 733, 746, 747, 749, 757, 758, 759, 767, 768, 771, 773, 774, 775, 777, 788, 790, 794, 797, 801, 804, 810, 814, 817, 821, 837, 838, 843, 845, 847, 849, 852, 855, 861, 863, 865, 877, 878, 892, 893, 896, 897, 899, 908
yu, 105, 106, 221
Yu, Aldrich, 339
Yu, Baosheng, 321
Yu, Bei, 312, 433, 899
Yu, Bihui, 777
Yu, Bin, 187, 316, 329
Yu, Bingnan, 116, 567
Yu, Bohan, 147, 685, 814
Yu, Bowen, 560, 897
Yu, Chang, 577
Yu, Changlong, 878
Yu, Chao, 296
Yu, Cheng, 707
Yu, Chenglin, 131
Yu, Chenxiao, 488
Yu, Chenyun, 638
Yu, Dazhou, 146, 299
Yu, Dian, 709, 722
Yu, Dianhai, 297
Yu, Dianyu, 391
Yu, Ding, 320, 630
Yu, Dong, 86, 143, 289, 357, 372, 436, 576, 687, 709, 722, 838
Yu, Fan, 757
Yu, Fei, 581, 666
Yu, Fenggang, 391
Yu, Ge, 358, 377, 406, 434
Yu, Guanghua, 774
Yu, Guangjun, 416
Yu, Guangya, 254, 327, 568, 591, 632, 768
Yu, Guo, 253, 861
Yu, Guoxin, 621
Yu, Haeun, 572
Yu, Haiyang, 126, 289
Yu, Hanchao, 845
Yu, Hang, 161, 395
Yu, Hao, 166
Yu, Haonan, 771
Yu, Haotian, 131
Yu, Hong, 211

yu, hong, 105, 106, 221
Yu, HongChen, 120
Yu, Hongyao, 349
Yu, Huan, 87, 605, 607, 616, 847
Yu, Huiqun, 715
Yu, Huiyu, 371
Yu, Hwanjo, 121, 151
Yu, Jianjun, 319
Yu, Jianxing, 463, 852
Yu, Jiarui, 746
Yu, Jiayang, 474
Yu, Jie, 733
Yu, Jifan, 476, 554, 804
Yu, Jiguo, 324
Yu, Jing, 364
Yu, Jinghan, 790
Yu, Jinghe, 801
Yu, Jinhong, 843
Yu, Joe, 650
Yu, Jun, 407
Yu, Junqing, 260, 849
Yu, Kaicheng, 313
Yu, Kaichi, 458
Yu, Kaidong, 370
Yu, Kaixian, 241
Yu, Kangrui, 249
Yu, Le, 897
Yu, Lei, 439, 629, 724
Yu, Li, 685
Yu, Liang-Chih, 551
Yu, Licheng, 444
Yu, Linhao, 119, 715
Yu, Linlin, 837
Yu, Lixing, 203
Yu, Lu, 817
Yu, Miao, 621, 714
Yu, Min, 429
Yu, Minshen, 757
Yu, Mo, 296, 425
Yu, Nenghai, 707
Yu, Ning, 564
Yu, Ningzhe, 344
Yu, Peijie, 431
Yu, Peilin, 439
Yu, Peiying, 304
Yu, Peng, 855
Yu, Philip S., 111, 191, 205, 262, 307, 344, 393, 488, 788, 893
Yu, Ping, 624
Yu, Puxuan, 572
Yu, Qiming, 693
Yu, Ruijie, 377
Yu, Seunguk, 865
Yu, Shanhong, 370
Yu, Shaozuo, 423
Yu, Shengquan, 320
Yu, Shi, 673
Yu, Shuhua, 208
Yu, Sijin, 349
Yu, Suyang, 908
Yu, Ta, 84
Yu, Tao, 749
Yu, Tianxiu, 725
Yu, Tinghao, 431
Yu, Tingting, 790
Yu, Tong, 483
Yu, Tony, 794
Yu, Wanhao, 790
Yu, Wanlong, 723
Yu, Wenbo, 541
Yu, Wenhan, 303
Yu, Wenhao, 289, 548
Yu, Wenqian, 364
Yu, Xianzhi, 312, 372, 759
Yu, Xiaodong, 348, 490
Yu, Xiaohang, 797
Yu, Xiaoyan, 630
Yu, Xin, 896
Yu, Xingkai, 379
Yu, Xinlei, 714
Yu, Xinyu, 420
Yu, Xuanqing, 424
Yu, Yahan, 896
Yu, Yangyang, 166
Yu, Yanhao, 330
Yu, Yanhua, 580
Yu, Yi, 619
Yu, Yichen, 258
Yu, YingJie, 124
Yu, Yipeng, 239
Yu, Yong, 248, 271, 863, 877, 892
Yu, Yongan, 674
Yu, Youngjae, 259, 552, 595, 601, 775
Yu, Yu, 138, 299
Yu, Yuanhao, 773
Yu, Yue, 444, 552, 621
Yu, Yunwen, 158
Yu, Zeping, 821
Yu, Zewei, 758
Yu, Zhenghan, 684
Yu, Zhengtao, 456, 713
Yu, Zhiwei, 280
Yu, Zhiwen, 651
Yu, Zhiyin, 448, 810
Yu, Zhiyuan, 305, 546, 600, 767
Yu, Zhou, 459, 641

Yu, Zhouliang, 676
Yu, Zhuoyun, 436
Yu, Zhuyang, 508
Yu, Zichun, 325
Yu, Zitong, 747
Yu, Ziyang, 205
Yuan, 115, 134, 136, 144, 145, 153, 178, 188, 189, 191, 209, 219, 226, 239, 248, 249, 255, 262, 264, 266, 303, 304, 305, 312, 313, 321, 323, 325, 332, 347, 354, 359, 364, 379, 381, 405, 410, 419, 422, 434, 440, 442, 449, 467, 470, 471, 500, 507, 508, 509, 525, 543, 560, 569, 584, 587, 612, 623, 632, 642, 643, 644, 645, 652, 670, 675, 686, 695, 698, 713, 718, 725, 726, 759, 761, 762, 804, 821, 863, 879, 882, 892, 904
yuan, 340
Yuan, Ann, 500
Yuan, Bo, 189
Yuan, Changsen, 136
Yuan, Chaohao, 434
Yuan, Chenchen, 762
Yuan, Cheng, 632
Yuan, Chun, 560
Yuan, Chunfeng, 134
Yuan, Danlong, 145
Yuan, Fangfang, 304
Yuan, Fei, 248
Yuan, Gao, 347
Yuan, Guan, 262
Yuan, Hangjie, 643
Yuan, Hanzhang, 219
Yuan, Haobo, 470
Yuan, Haohan, 525
Yuan, Haowei, 863
Yuan, Haoyang, 442
Yuan, Hongbang, 642
Yuan, Huaying, 508
Yuan, Jia, 410
Yuan, Jiahao, 313
Yuan, Jiakang, 321, 440
Yuan, Jialuo, 726
Yuan, Jiawei, 144
Yuan, Jiaxin, 226, 804
Yuan, Jiayin, 670
Yuan, Jiyao, 264
Yuan, Junyi, 725
Yuan, Li, 359, 422
yuan, li, 340
Yuan, Lin, 623
Yuan, Long, 332
Yuan, Michelle, 136
Yuan, Mingxuan, 312, 759
Yuan, Peiwen, 405
Yuan, Ruibin, 686
Yuan, Ruifeng, 381, 434
Yuan, Shaozu, 354
Yuan, Shuai, 670
Yuan, Shuguang, 364
Yuan, Shuzhou, 821
Yuan, Siyu, 644
Yuan, Tangming, 266
Yuan, Tao, 612
Yuan, Wei, 379
Yuan, Weikang, 467
Yuan, Wendi, 713
Yuan, Wenhao, 305
Yuan, Xiangchi, 115, 188, 569
Yuan, Xiaojie, 718
Yuan, Xiaokun, 303
Yuan, Xiaosong, 419, 879
Yuan, Xiaowei, 325, 652
Yuan, Xihao, 587
Yuan, Ye, 191, 255, 323, 507, 509, 584, 725
Yuan, Yi, 904
Yuan, Yichen, 239
Yuan, Yizhen, 804
Yuan, Youliang, 249, 882
Yuan, Yubai, 892
Yuan, Yuchen, 761
Yuan, Yujin, 675, 695, 698
Yuan, Yuqian, 189
Yuan, Zekun, 381
Yuan, Zheng, 449
Yuan, Zhengqing, 178, 471
Yuan, Zhihang, 153
Yuan, Zhonghang, 209
Yuan, Ziqi, 543
Yuan, Zizhen, 645
Yuanyang, 560
Yuanyang, Qiu, 560
Yue, 87, 90, 98, 203, 244, 249, 250, 280, 293, 332, 361, 368, 428, 474, 539, 562, 599, 675, 740, 741, 743
Yue, Guanglei, 249
Yue, Juwei, 332
Yue, Kun, 203, 428
Yue, Shengbin, 562
Yue, Xiang, 361, 474, 599
Yue, Xianghu, 332, 740
Yue, Xiangyu, 244, 280, 743
Yue, Xiaofei, 539
Yue, Xiaoyu, 250
Yue, Xing, 368
Yue, Yanwei, 675

Yue, Yuanhao, 87, 741
Yue, Zeyang, 98
Yue, Zihao, 90, 293
YuechenLi, 713
YuechenLi, YuechenLi, 713
Yugay, 685
Yugay, Aleksandr, 685
Yukhymenko, 593
Yukhymenko, Hanna, 593
YukiWangsajaya, 469
YukiWangsajaya, Nadya, 469
Yun, 116, 124, 156, 181, 209, 214, 223, 236, 290, 301, 509, 516, 803, 888, 891, 908
Yun, Dogyeong, 223
Yun, Heeseung, 908
Yun, Hyokun, 181
Yun, Jeongin, 516
Yun, Jooyeol, 301
Yun, Jun-Hak, 209
Yun, JungMin, 116, 803, 891
Yun, Longfei, 236
Yun, Sangdoo, 156, 509
Yun, Se-Young, 214, 290, 888
Yun, Taewon, 124
Yunfan, 319
Yunfan, Li, 319
Yunyi, 161
Yunyi, Hong, 161
Yusuf_Simsek,
Yusuf_Simsek, Yusuf_Simsek,
Yuxuan, 606
Yuxuan, Han, 606
Yvon, 477
Yvon, François, 477
Yüksel, 505
Yüksel, Arda, 505

zablocki, 883
zablocki, eloi, 883
Zafar, 571
Zafar, Muhammad Bilal, 571
Zafor, 211
Zafor, Adrika, 211
Zagni, 881
Zagni, Giovanni, 881
Zahraei, 509
Zahraei, Pardis Sadat, 509
Zaiane, 699
Zaiane, Osmar, 699
Zaidan, 864
Zaidan, Omar, 864
Zaki, 731
Zaki, Mohammadi, 731
Zalmanson, 551
Zalmanson, Danit Berger, 551
Zaman, 453
Zaman, Kerem, 453
Zamaraeva, 187
Zamaraeva, Olga, 187
Zamir, 166
Zamir, Syed Waqas, 166
Zampieri, 484
Zampieri, Marcos, 484
Zan, 345, 647
Zan, Hongying, 345
Zan, Wen, 647
Zang, 135, 239, 437
Zang, Chengxi, 135
Zang, Hui, 437
Zang, Yuhang, 239
Zangari, 200, 488
Zangari, Angelo, 200, 488
Zanuttigh, 537
Zanuttigh, Pietro, 537
Zanzotto, 193, 426
Zanzotto, Fabio Massimo, 193, 426
Zaradoukas, 878
Zaradoukas, Efstratios, 878
Zaraket, 590, 860
zaraket, 494
Zaraket, Fadi, 590
zaraket, Fadi, 494
Zaraket, Fadi A, 860
Zarch, 488
Zarch, Hossein Entezari, 488
Zareian, 562
Zareian, Alireza, 562
Zaremoodi, 320
Zaremoodi, Poorya, 320
Zarharan, 829
Zarharan, Majid, 829
ZarrieSS, 218, 241, 580, 766
ZarrieSS, Sina, 218, 241, 580, 766
Zarrinkalam, 277
Zarrinkalam, Fattane, 277
Zaytsev, 685
Zaytsev, Alexey, 685
Zaïane, 203
Zaïane, Osmar R., 203
Zdongxing, 762
Zdongxing, Zdongxing, 762
ZEESHAN, 471
ZEESHAN, MD ARFEEN, 471
Zeinalipour, 634
Zeinalipour, Kamyar, 634
Zeldes, 765, 875

Zeldes, Amir, 765, 875
Zelinsky, 533
Zelinsky, Gregory, 533
Zeng, 98, 110, 124, 125, 143, 157, 164, 170, 175, 180, 190, 192, 206, 219, 222, 243, 268, 269, 272, 283, 292, 297, 329, 330, 335, 345, 378, 379, 391, 406, 427, 431, 435, 457, 475, 491, 497, 533, 540, 546, 548, 552, 572, 581, 588, 607, 613, 614, 618, 624, 626, 629, 636, 646, 650, 654, 657, 658, 664, 669, 670, 685, 691, 705, 713, 726, 740, 758, 759, 772, 773, 786, 790, 793, 817, 824, 832, 838, 845, 849, 851, 888, 889, 904
Zeng, Anxiang, 624
Zeng, Aohan, 646, 786, 851
Zeng, Bo, 406
Zeng, Bowen, 125, 497
Zeng, Chuxuan, 726
Zeng, Daniel Dajun, 650
Zeng, Daojian, 685
Zeng, Delong, 849
Zeng, Guangtao, 427
Zeng, Guoyang, 758
Zeng, Hanqing, 192
Zeng, Hengqi, 170
Zeng, Hongchuan, 243
Zeng, Huimin, 124, 143, 618, 824
Zeng, Jiajie, 654
Zeng, Jiali, 626
Zeng, Jichuan, 391
Zeng, Jie, 190, 581
Zeng, Jingjie, 378, 431
Zeng, Jingying, 269
Zeng, Jinle, 297
Zeng, Kaide, 832
Zeng, Ke, 329, 345, 435, 475, 636, 691, 758
Zeng, Min, 888
Zeng, Peter, 533
Zeng, Qi, 206, 607
Zeng, Qiming, 664
Zeng, Qingcheng, 110, 164, 572, 772, 838
Zeng, Qingkai, 817
Zeng, Qinglin, 219
Zeng, Qiuhai, 268
Zeng, Qunhong, 740
Zeng, Ruilin, 272
Zeng, Shaoning, 657
Zeng, Shenglai, 175
Zeng, Shiyu (Vivienne), 292
Zeng, Shuang, 824
Zeng, Wangding, 379
Zeng, Wei, 670
Zeng, Wenhao, 222, 283
Zeng, Xia, 705
Zeng, Xiangxiang, 406, 548
Zeng, Xianlin, 491
Zeng, Xiaoyi, 614
Zeng, Xingchen, 793
Zeng, Xingshan, 157
Zeng, Xinyi, 629
Zeng, Yi, 98, 552
Zeng, Yiming, 790
Zeng, Yirong, 658
Zeng, Yizhe, 793
Zeng, Yongcheng, 335
Zeng, Yu, 759
Zeng, Yuan, 845
Zeng, Yunsheng, 330
Zeng, Yuting, 588
Zeng, Yuyuan, 540
Zeng, Zhen, 180, 613, 904
Zeng, Zhi, 888, 889
Zeng, Zhichen, 457
Zeng, Zhiyuan, 546
Zeng, Ziqian, 713, 773
Zeng, Ziyang, 669
Zha, 609, 673, 776
Zha, Daren, 673
Zha, Rui, 609
Zha, Xinyu, 776
Zhai, 122, 171, 197, 201, 252, 263, 312, 313, 339, 341, 370, 389, 559, 783, 844
Zhai, Bohan, 252
Zhai, ChengXiang, 197
Zhai, Guangtao, 339, 559
Zhai, Guocheng, 313
Zhai, Jie, 341
Zhai, Juan, 263, 844
Zhai, Naixin, 171
Zhai, Skylar, 783
Zhai, Wei, 312
Zhai, Yanbo, 370
Zhai, Yitao, 201
Zhai, Yuanzhao, 389
Zhai, Yunpeng, 122
Zhan, 140, 147, 232, 260, 262, 268, 338, 403, 432, 498, 541, 565, 587, 641, 683, 754, 790, 837, 848, 852, 883
Zhan, Chen, 852
Zhan, Chengjun, 587
Zhan, Clovis Barbour; Huixin,
Zhan, Hao, 403
Zhan, Jian, 754
Zhan, Jiaqing, 837
Zhan, Liang, 268

Zhan, Libin, 683
Zhan, Mingjie, 432
Zhan, Qiusi, 262
Zhan, Runzhe, 541, 565
Zhan, Wangzhi, 232
Zhan, Weixiao, 883
Zhan, Xiao, 140
Zhan, Xinglang, 260
Zhan, Yi, 338
Zhan, Ying, 641
Zhan, Yuliang, 848
Zhan, Zhihao, 147
Zhan, Zhonghao, 790
ZHANG, 144, 148, 202, 359, 378, 778, 865
Zhang, 82, 83, 85, 86, 87, 88, 90, 92, 93, 94, 95, 96, 98, 100, 103, 104, 105, 107, 111, 114, 115, 117, 118, 120, 122, 124, 125, 126, 127, 128, 129, 130, 132, 133, 134, 136, 137, 141, 142, 143, 145, 146, 148, 151, 152, 153, 155, 156, 157, 158, 159, 161, 162, 163, 164, 165, 166, 167, 168, 169, 173, 175, 178, 179, 181, 182, 184, 185, 186, 187, 189, 191, 192, 195, 196, 198, 199, 201, 202, 203, 204, 205, 206, 207, 209, 210, 211, 212, 213, 214, 215, 219, 220, 221, 224, 228, 230, 231, 234, 235, 238, 239, 241, 248, 249, 250, 251, 252, 255, 258, 259, 260, 261, 263, 264, 266, 268, 269, 271, 272, 273, 274, 276, 278, 279, 280, 281, 282, 283, 284, 285, 286, 287, 289, 290, 292, 293, 294, 297, 298, 300, 301, 302, 303, 304, 305, 306, 307, 309, 310, 313, 315, 317, 318, 319, 321, 322, 323, 324, 325, 326, 327, 328, 329, 331, 332, 333, 334, 335, 336, 338, 339, 340, 341, 344, 345, 346, 347, 348, 349, 350, 352, 353, 354, 356, 357, 359, 360, 361, 362, 363, 368, 369, 370, 371, 372, 373, 374, 377, 378, 379, 380, 381, 382, 383, 384, 385, 387, 388, 389, 390, 391, 392, 393, 394, 395, 396, 397, 398, 399, 401, 403, 404, 405, 407, 408, 409, 410, 411, 412, 413, 414, 417, 418, 419, 421, 423, 424, 427, 428, 429, 430, 431, 433, 434, 436, 437, 438, 439, 440, 442, 444, 448, 450, 453, 454, 455, 456, 457, 459, 462, 463, 464, 465, 467, 470, 471, 472, 473, 474, 475, 476, 478, 479, 480, 481, 482, 483, 488, 490, 491, 494, 495, 497, 498, 499, 500, 501, 503, 505, 506, 507, 508, 509, 512, 514, 515, 516, 518, 519, 520, 524, 525, 527, 528, 529, 530, 531, 534, 535, 538, 540, 541, 542, 545, 546, 548, 550, 552, 554, 555, 559, 560, 561, 564, 565, 566, 568, 569, 571, 572, 573, 575, 576, 578, 582, 583, 584, 585, 586, 587, 588, 590, 591, 594, 595, 596, 597, 598, 599, 600, 601, 602, 604, 606, 607, 609, 612, 617, 618, 619, 620, 621, 622, 623, 624, 625, 626, 627, 628, 629, 630, 631, 632, 634, 635, 636, 641, 642, 643, 644, 646, 647, 648, 649, 650, 651, 652, 653, 655, 656, 657, 658, 659, 660, 661, 663, 665, 666, 667, 668, 669, 670, 671, 672, 673, 674, 675, 676, 677, 678, 680, 681, 682, 683, 684, 685, 686, 687, 688, 689, 690, 692, 693, 694, 695, 698, 699, 700, 701, 702, 703, 704, 706, 707, 708, 712, 714, 715, 716, 717, 718, 719, 720, 721, 722, 723, 724, 725, 726, 727, 728, 729, 730, 731, 732, 733, 735, 736, 738, 740, 741, 742, 745, 747, 748, 749, 750, 751, 752, 753, 754, 755, 756, 757, 758, 759, 760, 761, 762, 765, 767, 768, 769, 771, 773, 774, 775, 776, 777, 780, 781, 782, 785, 786, 787, 788, 789, 790, 791, 792, 793, 794, 795, 796, 797, 800, 802, 803, 804, 807, 809, 810, 815, 817, 818, 820, 821, 822, 824, 827, 828, 831, 832, 833, 834, 836, 837, 838, 839, 840, 844, 845, 846, 847, 848, 849, 850, 851, 852, 853, 854, 856, 859, 860, 862, 863, 864, 867, 870, 871, 873, 874, 875, 876, 877, 878, 879, 880, 883, 884, 885, 886, 887, 888, 892, 893, 894, 895, 896, 897, 899, 900, 901, 902, 905, 906, 907, 908, 909
zhang, 193, 224, 379, 403, 427, 621, 784, 786
Zhang, Aijia, 259
Zhang, An, 883
Zhang, Angela, 209
Zhang, Anyi, 134
Zhang, Baolei, 723
Zhang, Beining Wu;Fuyou Mao; Jiong Lin;Cheng Yang; Jiaxuan Lu; Yifu Guo;Siyu,
Zhang, Binbin, 363
Zhang, Binchi, 585, 824
Zhang, Bingquan, 686
Zhang, Bo, 239, 309, 384, 440, 448, 545, 810
Zhang, Bo-Wen, 105
Zhang, Bohan, 146, 607, 646
Zhang, Bolin, 630
Zhang, Bolun, 863
Zhang, Bowei, 836
Zhang, Bowen, 434, 620, 729
Zhang, Boxuan, 248, 871
Zhang, Boyuan, 687

Zhang, Boyun, 678
Zhang, Briana, 85
Zhang, Buyun, 795
Zhang, Caiqi, 104, 503, 572, 845, 864
Zhang, Can, 636
Zhang, Ce, 436
Zhang, Changqing, 326
Zhang, Chao, 92, 181, 627, 669, 731, 748
Zhang, Chaoning, 404
Zhang, Chaoqi, 821
Zhang, Chaowei, 673
Zhang, Chaoyi, 757
Zhang, Chen, 359, 384, 456, 714, 797, 810
Zhang, Chen Bo Calvin, 560
Zhang, Chenchen, 540, 676, 900
Zhang, Cheng, 369, 723, 732
Zhang, Chenghao, 103, 319
Zhang, Chengzhi, 871
Zhang, Chenyuan, 404
Zhang, Chi, 146, 280, 372, 507, 516, 732
Zhang, Chong, 680
ZHANG, Chongsheng, 778
Zhang, Chuanxin, 727
Zhang, Chuanyi, 145
Zhang, Chunhong, 669
Zhang, Chunhui, 115, 136, 323, 534, 569, 860
Zhang, Chuxu, 178
Zhang, Crystina, 161, 519
Zhang, Dailing, 436
Zhang, Daipeng, 332
Zhang, Dandan, 625
Zhang, Daoxin, 404
Zhang, David, 285
Zhang, Dazhi, 893
Zhang, Delvin Ce, 456
Zhang, Denghui, 754
Zhang, Deyue, 883
Zhang, Di, 888, 907
Zhang, Dong, 204, 228, 708
Zhang, Dongdong, 167, 440, 821
Zhang, Dongmei, 211, 264, 349, 684, 720, 773
Zhang, Dongxu, 127, 719
ZHANG, Doudou, 144
Zhang, Duzhen, 716
Zhang, Fa, 391
Zhang, Fan, 430, 442, 527, 886
Zhang, Fanjin, 169
Zhang, Fanrui, 134, 844
Zhang, Fei, 717
Zhang, Feiran, 661
Zhang, Feng, 352, 431, 769, 884
Zhang, Fengchun, 377
Zhang, Fengzhuo, 535
Zhang, Fu, 370, 382, 438, 671
Zhang, Fuzheng, 672
Zhang, Gaifan, 173
Zhang, Gaowei, 688
Zhang, Ge, 324, 490, 561, 676
Zhang, Gehao, 263
Zhang, Geyuan, 831, 906
Zhang, Gong, 161
Zhang, Gongbo, 498
Zhang, Guang, 651, 670, 827
Zhang, Guangjun, 403
Zhang, Guangquan, 524
Zhang, Guangwei, 626
Zhang, Guanhua, 403
Zhang, Guannan, 344
Zhang, Guibin, 126, 714, 797
Zhang, Guilin, 239
Zhang, Guofeng, 676
Zhang, Guoqiang, 382, 838
Zhang, Guoxi, 85
Zhang, Haifeng, 385
Zhang, Haijun, 719
Zhang, Hailong, 607
Zhang, Haiming, 635
Zhang, Hainan, 753
Zhang, Haipeng, 886
Zhang, Haiyang, 230
Zhang, Han, 202, 870
zhang, han, 379
Zhang, Hanchen, 286
Zhang, Hang, 541, 631
Zhang, Hanling, 153
Zhang, Hanning, 559
Zhang, Hanrong, 393
Zhang, Hao, 143, 252, 677, 742, 836
Zhang, Haodi, 646
Zhang, Haofeng, 305
Zhang, Haonan, 405, 849
Zhang, HaoPeng, 892
Zhang, Haopeng, 525
Zhang, Haotian, 472
Zhang, Haowei, 130
Zhang, Haoxiang, 901
Zhang, Haoxin, 652
Zhang, Haoyu, 601
Zhang, Haoyue, 133
Zhang, Haozhen, 186
Zhang, He, 319
Zhang, Hejia, 444
Zhang, Hengrun, 715
Zhang, Hengyuan, 163, 264, 758, 821
Zhang, Hongbo, 152
Zhang, Hongfei, 479

Zhang, Honggang, 802
Zhang, Hongming, 436
Zhang, Hongyu, 628, 729
zhang, Hu, 403, 427
Zhang, Hua, 202
Zhang, Huadong, 322
Zhang, Hualei, 133
Zhang, Huan, 670
Zhang, Hui, 680
Zhang, Huishuai, 145, 379, 382
Zhang, Huixuan, 334, 383
Zhang, Huopu, 442, 902
Zhang, Jennifer Yuntong, 514
Zhang, Ji, 214, 467
Zhang, Jiacheng, 480
Zhang, Jiahao, 531, 854
Zhang, Jiaheng, 576, 777, 785
Zhang, Jiahuan, 313, 344
Zhang, Jiajie, 571, 880
Zhang, Jiajun, 491, 634, 698
Zhang, Jiaming, 206, 396
Zhang, Jian, 122, 179, 203, 205, 261, 725, 851
Zhang, Jianan, 410
Zhang, Jianbing, 867
Zhang, Jiangning, 724
Zhang, Jianhao, 369, 718
Zhang, Jianing, 134, 833
Zhang, Jianjun, 319
Zhang, Jianke, 698
Zhang, Jiankun, 529
Zhang, Jianpeng, 381
Zhang, Jianqing, 745
Zhang, Jianshu, 436, 726, 775
Zhang, Jiapeng, 439
Zhang, Jiaqi, 647
zhang, jiaqi, 193
Zhang, Jiaqing, 273
Zhang, Jiaquan, 404
Zhang, Jiaran, 210
Zhang, JiaRui, 765
Zhang, Jiatong, 827
Zhang, Jiawei, 384
Zhang, Jiaxin, 281, 609
Zhang, Jiayi, 602
Zhang, Jiaying, 395
Zhang, Jiayu, 747
Zhang, Jiazheng, 382, 545, 838
ZHANG, Jie, 148, 359
Zhang, Jie, 133, 162, 195, 212, 235, 555, 617
Zhang, Jie M., 478
Zhang, Jielun, 260
Zhang, Jihai, 185
Zhang, Jin, 410
Zhang, Jing, 146, 642, 646, 647
Zhang, Jinghan, 95
Zhang, Jinghao, 414
Zhang, Jinghuai, 566, 781
Zhang, Jinghui, 475, 876
Zhang, Jingren, 631
Zhang, Jingyu, 118, 499
Zhang, Jinhong, 322
Zhang, Jinjian, 352, 564
Zhang, Jinshan, 749
Zhang, Jintian, 317
Zhang, Jinwei, 377
Zhang, Jinxia, 656
Zhang, Jinyang, 264
Zhang, Jipeng, 736
Zhang, Jiwen, 503, 546
Zhang, Jue, 720
Zhang, Jun, 125, 128, 221, 259, 384, 497, 597, 699
Zhang, Junge, 317
Zhang, Junpeng, 319
Zhang, Junru, 159
Zhang, Junshuo, 324
Zhang, Juntian, 202, 384, 887
Zhang, Junwei, 371
Zhang, Junxiang, 456
Zhang, Junyi, 483
Zhang, Junyuan, 239, 582
Zhang, Jusheng, 219, 590, 899, 900
Zhang, Kai, 212, 353, 546, 689, 893
Zhang, Kaifu, 379, 660, 701, 738
Zhang, Kaike, 421
Zhang, Kaipeng, 114, 134, 844
Zhang, Kaitai, 846
Zhang, Kaiwei, 559
Zhang, Kaiyan, 623, 893
Zhang, Kaiyi, 905
Zhang, Kaiyuan, 762
Zhang, Kangning, 248, 892
Zhang, Kanjian, 656
Zhang, Kechi, 260, 796
Zhang, Kexin, 457, 894
Zhang, Kun, 205, 454, 878
Zhang, Lan, 682
Zhang, Le, 609
Zhang, Lechen, 133, 450, 472
Zhang, Lefei, 184, 294, 362, 888
Zhang, Lei, 634
Zhang, Leixin, 213
Zhang, Li, 151, 542, 560, 582, 790
Zhang, Liang, 293
Zhang, Liang-Jie, 302
Zhang, Liangjie, 700
Zhang, Libo, 702

Zhang, Licheng, 689
Zhang, Lin, 354, 707, 708
Zhang, Linfeng, 239, 368, 423, 649, 655, 666, 750
Zhang, Ling, 381
Zhang, Lingfeng, 479
Zhang, Linglin, 437
Zhang, Lingling, 248, 464, 730, 821, 871
Zhang, Lingyin, 301
Zhang, Lingzhe, 122
Zhang, Lingzhi, 495
Zhang, Linhai, 374, 668
Zhang, Linhao, 701
Zhang, Liqing, 211
Zhang, Lirui, 382
Zhang, Litian, 787, 880
Zhang, Liwen, 120, 853
Zhang, Liyu, 284
Zhang, Lizhe, 378
Zhang, Lizhu, 192, 588
zhang, long, 224
Zhang, Lu, 370
Zhang, Luyang, 762
Zhang, Lyuxin, 181
Zhang, Man, 776
Zhang, Manman, 315
Zhang, Manyi, 372
Zhang, Manyuan, 280
Zhang, Maosen, 583
Zhang, Meicong, 411
Zhang, Meiru, 390
Zhang, Meishan, 403, 404, 428, 630, 668, 680, 689
Zhang, Meng, 429
Zhang, Mengdi, 353
Zhang, Menghao, 251
Zhang, Mengjie, 752
Zhang, Mengqi, 516
Zhang, Mengxiang, 336
Zhang, Mi, 348, 529
Zhang, Mian, 333
Zhang, Miao, 86, 326, 514, 902
Zhang, Mike, 238, 788, 862
Zhang, Min, 338, 354, 356, 379, 393, 397, 403, 404, 428, 578, 630, 632, 653, 668, 678, 692, 716, 722, 733, 754, 759, 761, 774
Zhang, Ming, 191, 255, 448, 717, 820, 838, 897
Zhang, Ming-Liang, 341, 636
Zhang, Mingfei, 373
Zhang, Minghan, 727
Zhang, Mingjie, 735
Zhang, Mingrui, 588
Zhang, Mingyang, 286
Zhang, Mingyuan, 124
Zhang, Minjia, 130
Zhang, Minxing, 439
Zhang, Minying, 290
Zhang, Mu, 463, 568
Zhang, Muhan, 444, 476, 818, 859
Zhang, Mulei, 797
Zhang, Ning, 92, 767
Zhang, Ningyu, 317, 334, 398, 436, 585, 708
Zhang, Peinan, 810
Zhang, Peirong, 324
Zhang, Peixuan, 348
Zhang, Peng, 642, 665
Zhang, Penghao, 471
Zhang, Pengkun, 886
Zhang, Poppy, 892
zhang, puzhen, 786
Zhang, Qi, 201, 382, 387, 545, 586, 595, 596, 700, 717, 728, 740, 758, 792, 820, 821, 838, 839, 896, 897
Zhang, Qian, 98, 434, 630, 833
Zhang, Qianchi, 753
Zhang, Qianen, 873
Zhang, Qiang, 192
Zhang, Qichao, 249, 568, 621
Zhang, Qifei, 624, 749
Zhang, Qiming, 507
Zhang, Qin, 837
Zhang, Qinggang, 336, 854
Zhang, QingHao, 800
Zhang, QingLin, 413
Zhang, Qintong, 239
Zhang, Qixin, 141, 424
Zhang, Qixun, 658
Zhang, Qiyuan, 474, 712
Zhang, Quanshi, 319
Zhang, Ran, 622
Zhang, Richong, 619, 735
Zhang, Rongyang, 298
Zhang, Rongyi, 350
Zhang, Rongzhi, 817
Zhang, Rongzhou, 414
Zhang, Rui, 239, 318, 328, 418, 660
Zhang, Ruidong, 599
Zhang, Ruihan, 220
Zhang, Ruijia, 892
Zhang, Ruiqing, 684, 722
Zhang, Ruiyi, 103, 483, 584, 832
Zhang, Ruochen, 482
Zhang, Ruqing, 721, 754
Zhang, Ruwen, 88
Zhang, Saiqian, 279
Zhang, Senbo, 677
Zhang, Shanghang, 479
zhang, shaojie, 621

Zhang, Shaolei, 899
Zhang, Shaoting, 255, 528
Zhang, Shaowei, 94
ZHANG, SHAOWU, 378
Zhang, Sheng, 587, 792
Zhang, Shenglin, 134
Zhang, Shengming, 609
Zhang, Shengxiang, 88
Zhang, Shengyu, 289, 323
Zhang, Shengzhi, 499
Zhang, Shi-Xiong, 853
Zhang, Shikun, 651, 688, 735
Zhang, Shilei, 678
Zhang, Shiqi, 470
Zhang, Shiyu, 207
Zhang, Shiyue, 821
Zhang, Shouzheng HuangMeishan ZhangBaotian HuMin, 658
Zhang, Shuai, 313, 352, 656, 668, 699, 706, 791
Zhang, Shuaiyu, 440
Zhang, Shuang, 368
Zhang, Shubo, 184, 726
Zhang, Shufei, 440, 774
Zhang, Shuhai, 335
Zhang, Shuhao, 683
Zhang, Shujian, 297
Zhang, Shumin, 520
Zhang, Shuo, 157, 405, 467
Zhang, Shuyu, 723, 726
Zhang, Siyu DingWenyuan, 715
Zhang, Siyue, 241, 473, 800
Zhang, Songming, 96
Zhang, Songyang, 209, 221
Zhang, Tao, 128, 773, 774, 853
Zhang, Taolin, 714, 755
Zhang, Terry Jingchen, 156, 475, 600
Zhang, Tianfu, 629
Zhang, Tianhui, 100, 273
Zhang, Tianle, 756, 761
Zhang, Tianwei, 583
Zhang, TianYi, 202
Zhang, Tianyi, 129, 264, 294, 421, 586
Zhang, Tianyun, 85
Zhang, TianZe, 380
Zhang, Ting, 515, 531, 874
Zhang, Tong, 555, 559, 750, 807
Zhang, Vincent Jim, 166
Zhang, Wei, 260, 287, 412, 793, 877
Zhang, Weidong, 675, 695
Zhang, Weijia, 767
Zhang, Weijie, 703
Zhang, Weijuan, 499
Zhang, Weili, 252
Zhang, Weiming, 721, 782
Zhang, Weinan, 157, 248, 271, 328, 389, 601, 721, 794, 863, 877, 892
Zhang, Weipeng, 345, 636
Zhang, Weiran, 527
Zhang, Weitian, 292
Zhang, Weiwei, 329
Zhang, Weixing, 414
Zhang, Weixu, 276, 509, 599
Zhang, Weiyan, 327, 568, 591, 768
Zhang, Weiyu, 374
Zhang, Weizhi, 111, 191, 393, 488, 893
Zhang, Wen, 326, 334, 354, 624, 671, 704, 715, 720
Zhang, Wenbin, 584
Zhang, Wenhao, 219, 681, 760
Zhang, Wenhu, 423, 433
Zhang, Wenhui, 287
Zhang, Wenjie, 310, 680
Zhang, Wenlin, 214, 587, 791
Zhang, Wenlong, 250, 440
Zhang, Wenqi, 304, 631, 722
Zhang, Wenqiao, 189
Zhang, Wensheng, 385
Zhang, Wentao, 210, 239, 274, 303, 340, 373, 388, 434, 642, 674, 750, 789, 853
Zhang, Wenxuan, 168, 182, 582
Zhang, Wenyu, 765
Zhang, Wenyuan, 332
Zhang, Wenzheng, 239
Zhang, Wuyang, 155
Zhang, Xi, 198, 880
Zhang, Xiang, 130
Zhang, Xiangbo, 824, 834
Zhang, Xiangdong, 809
Zhang, Xiangliang, 82, 259, 269, 357, 471, 505
Zhang, Xiangwen, 459, 793
Zhang, Xiangxu, 867
Zhang, Xiangyu, 292, 560, 612, 708, 828
Zhang, Xiangzheng, 87, 303, 437, 883
Zhang, Xianren, 268
Zhang, Xianyin, 372
Zhang, Xiao, 300
Zhang, Xiaocheng, 659
Zhang, Xiaodan, 462
Zhang, Xiaode, 107
Zhang, Xiaofan, 255, 528
Zhang, Xiaofeng, 359, 361, 419, 629, 879
Zhang, Xiaojin, 124, 290
Zhang, Xiaojun, 363
Zhang, Xiaokang, 146
Zhang, Xiaolu, 349, 501
Zhang, Xiaoming, 272, 787
Zhang, Xiaoping, 583

Zhang, Xiaowang, 391
Zhang, Xiaoyi, 284, 494
Zhang, Xiaoyu, 844
Zhang, Xichen, 423, 433
Zhang, Xin, 278, 380, 481, 708, 771, 846, 856
Zhang, Xingchen, 437
Zhang, Xinghua, 289
Zhang, Xingjian, 307
Zhang, Xinglang, 849
Zhang, Xinglin, 331, 569
Zhang, Xingong, 620
ZHANG, Xingsheng, 202
Zhang, Xingxing, 659
Zhang, Xingyu, 304
Zhang, Xinhao, 604
Zhang, Xinliang Frederick, 133, 575
Zhang, Xinrui, 350
Zhang, Xinyan, 202, 870
Zhang, Xinyu, 164, 248, 383, 871
Zhang, Xinyuan, 438
Zhang, Xinyue, 453
Zhang, Xinyun, 518
Zhang, Xiyuan, 690
Zhang, Xu, 201, 389, 410, 652, 684
Zhang, Xu-Yao, 187, 401
Zhang, Xuan, 636
Zhang, Xuanming, 459
Zhang, Xuanyu, 759
Zhang, Xuchao, 599
Zhang, Xue, 96, 583, 699
Zhang, Xuelin, 676
Zhang, Xuemiao, 209, 501
Zhang, Xueqiao, 669
Zhang, Xueru, 215, 907
Zhang, Xuesong, 327
Zhang, Xuewen, 845
Zhang, Xueyao, 199
Zhang, Xueyi, 181
Zhang, Xujie, 408
Zhang, Xulong, 392
Zhang, Xun, 384
Zhang, Xuteng, 283
Zhang, Ya, 409
Zhang, Ya-Qin, 93
Zhang, Yadong, 635
Zhang, Yage, 628
Zhang, Yan, 450, 641, 675, 681
Zhang, Yanfei, 719
ZHANG, Yang, 865
Zhang, Yang, 159, 224, 280, 417, 427, 457, 594, 628, 647, 648, 684, 751, 817, 822
Zhang, Yangshijie, 303, 353, 687, 885
Zhang, Yangtao, 818
Zhang, Yanlin, 815
Zhang, Yansong, 751
Zhang, Yanxin, 338, 534
Zhang, Yanyong, 433
Zhang, Yanzhao, 473
Zhang, Yanzhe, 352, 464
Zhang, Yanzhi, 656
Zhang, Yaocheng, 568
Zhang, Yaping, 387
Zhang, Yaqi, 383
Zhang, Yazhou, 565
Zhang, YeXuan, 506
Zhang, Yi, 158, 302, 363, 598, 681, 852, 884
Zhang, Yibo, 775
Zhang, Yice, 715
Zhang, Yichang, 852, 897
Zhang, Yichi, 334, 676, 704, 720, 902
Zhang, Yichun, 212
Zhang, Yidan, 596
Zhang, YiFan, 205, 749
Zhang, Yifan, 266, 531, 676
Zhang, Yifei, 326, 671, 673, 730
Zhang, Yihang, 850
Zhang, Yihua, 360, 795
Zhang, Yijing, 561
Zhang, Yijun, 632
Zhang, Yikai, 644, 671
Zhang, Yilin, 279
Zhang, Yimeng, 491, 875
Zhang, Yiming, 405, 719, 728, 756
Zhang, Yin, 228, 404, 621, 760
Zhang, Yindong, 512
Zhang, Ying, 306, 310, 374, 497, 552, 680, 851
Zhang, Yinger, 158
Zhang, Yingfang, 895
Zhang, Yingjie, 384
Zhang, Yingyi, 214, 464, 587, 791
Zhang, Yingying, 867
Zhang, Yinmin, 292, 560
Zhang, Yinqi, 877
Zhang, Yiqian, 142
Zhang, Yiqing, 529
Zhang, Yiquan, 137, 189, 512
Zhang, Yiqun, 474, 663
Zhang, Yiran, 471
Zhang, Yitian, 124, 143, 524, 618
Zhang, Yiwei, 729
Zhang, Yiwen, 353
Zhang, Yizhou, 397
Zhang, Yizhuo, 554, 847
Zhang, Yong, 803
Zhang, Yongdong, 282, 729
Zhang, Yongfeng, 644

Zhang, Yonggang, 303, 395
Zhang, Yongqi, 162, 483
Zhang, Yongyan, 90
Zhang, Yu, 310, 345, 359, 377, 475, 476, 538, 596, 650, 745, 804, 899, 901
Zhang, Yuan, 346, 347, 693
Zhang, Yuanhan, 361
Zhang, Yuanhe, 775
Zhang, Yuanqi MuBingfeng PiDefei XiaLei ZuoYongqi, 325
Zhang, Yuanxing, 606, 900
Zhang, Yuanzhe, 685
Zhang, Yuchen, 676
Zhang, Yuchong, 148
Zhang, Yudong, 714
Zhang, Yue, 21, 348, 369, 373, 774, 820, 840, 847, 853
Zhang, Yueke, 531
Zhang, Yueqi, 405
Zhang, Yueyang, 142, 276
zhang, Yufei, 784
Zhang, Yuhang, 530
Zhang, Yuhao, 776
Zhang, Yuheng, 780
Zhang, Yuji, 573
Zhang, Yujie, 467
Zhang, Yujing, 854
Zhang, Yun, 412
Zhang, Yunbei, 137
Zhang, Yuning, 707
Zhang, Yunkai, 192
Zhang, Yunke, 838
Zhang, Yunruo, 508
Zhang, Yunxiang, 133, 450, 738
Zhang, Yunyao, 260, 849
Zhang, Yunyi, 789
Zhang, Yupeng, 643
Zhang, Yuqing, 323
Zhang, Yushun, 464
Zhang, Yuting, 310
Zhang, Yutong, 92, 134, 464, 827
Zhang, Yuwei, 548
Zhang, Yuxi, 184
Zhang, Yuxin, 221
Zhang, Yuxing, 448
Zhang, Yuxing Lu; Yushuhong Lin; Jason,
Zhang, Yuxuan, 370
Zhang, Yuyao, 657, 728
Zhang, Yuyi, 124, 324
Zhang, Yuyu, 475, 599, 853
Zhang, Yuyue, 545
Zhang, Yuzhe, 278, 475
Zhang, Zaibin, 797
Zhang, Zaixi, 717
Zhang, Ze Yu, 294
Zhang, Zecheng, 667
Zhang, Zefan, 541
Zhang, Zehua, 234
Zhang, Zeliang, 490
Zhang, Zeyu, 117, 675, 738, 851
Zhang, Zhang, 749
Zhang, Zhao, 141
Zhang, Zhaodan, 410
Zhang, Zhaohan, 465, 719, 807
Zhang, Zhaoning, 665, 702, 725
Zhang, Zhaowei, 424
Zhang, Zhaoxiang, 676
Zhang, Zhaoyi, 538
Zhang, Zhejun, 354
Zhang, Zhen, 454
Zhang, Zhenduo, 848
Zhang, Zheng, 321, 338, 756
Zhang, Zhengchen, 209
Zhang, Zhenghao, 500
Zhang, Zhengxin, 83
Zhang, Zhenliang, 334
Zhang, Zhenyu, 231, 304
Zhang, Zheqi, 541
Zhang, Zhexin, 132, 398, 413
Zhang, Zheyu, 540, 762
Zhang, Zheyuan, 178
Zhang, Zhi, 165
Zhang, Zhibo, 394, 407, 585
Zhang, Zhicheng, 304, 535, 698
Zhang, Zhihan, 196, 505, 879
Zhang, Zhihao, 545, 820, 821
Zhang, Zhiheng, 685
Zhang, Zhihong, 336
Zhang, Zhijing, 145
Zhang, Zhipeng, 732, 741
Zhang, Zhiqiang, 694, 848
Zhang, Zhirui, 516
Zhang, Zhisong, 436
Zhang, Zhiwei, 701, 738
Zhang, Zhixin, 680
Zhang, Zhiyu, 786
Zhang, Zhizhong, 412
Zhang, Zhongwang, 120, 853
Zhang, Zhongxing, 234
Zhang, Zhuo, 550
Zhang, Zhuosheng, 159, 470, 721
Zhang, Zicheng, 339
Zhang, Zihan, 399
Zhang, Zihong, 294
Zhang, Zijian, 332, 888
Zhang, Zikai, 909

Zhang, Zining, 155
Zhang, Zipei, 714
Zhang, Ziqi, 134
Zhang, Ziqiang, 412
Zhang, Zishuai Zhang;Sihao Yu;Wenyi Xie;Ying Nie;Junfeng Wang;Zhiming Zheng; Dawei Yin;Hainan,
Zhang, Zixi, 488
Zhang, Zixing, 414
Zhang, Zixuan, 817
Zhang, Ziyi, 831
Zhang, Ziyin, 161
Zhang, Ziyu, 456
Zhang, Ziyun, 284, 494
Zhang, Zizhao, 786
Zhang, Zory, 908
Zhang, Zuming, 453
Zhang-Li, 476, 554, 804
Zhang-Li, Daniel, 476, 554, 804
Zhang., 507
Zhang., Jing, 507
Zhang1, 111
Zhang1, Chenwei, 111
Zhang1, Xinyang, 111
Zhang;,
Zhang;, Denghui,
Zhang;, Gewen Liang; Mufan Xu; Kehai Chen; Wei Wang; Yuwei Wang; Muyun Yang; Tiejun Zhao; Min,
Zhang;, Yang,
zhangchenxi, 337
zhangchenxi, Zhu, 337
ZhangLiujie, 638
ZhangLiujie, ZhangLiujie, 638
ZhangYi, 560
zhangyi, 116, 567
zhangyi, Wang, 116, 567
ZhangYi, ZhangYi, 560
Zhao, 87, 88, 97, 98, 103, 114, 119, 125, 137, 141, 145, 146, 147, 150, 155, 157, 162, 164, 166, 168, 169, 172, 181, 184, 185, 191, 192, 200, 205, 212, 214, 221, 224, 239, 241, 247, 249, 250, 260, 262, 264, 268, 269, 271, 278, 292, 294, 299, 303, 304, 308, 311, 313, 315, 323, 325, 331, 332, 334, 335, 338, 339, 340, 341, 349, 352, 353, 354, 358, 359, 362, 368, 370, 372, 374, 378, 379, 380, 382, 384, 385, 394, 396, 397, 398, 403, 405, 406, 407, 408, 409, 410, 413, 421, 424, 427, 435, 443, 446, 455, 457, 461, 464, 469, 470, 472, 473, 477, 479, 486, 488, 499, 501, 507, 511, 514, 527, 531, 535, 540, 542, 543, 545, 550, 560, 561, 568, 569, 572, 575, 581, 583, 587, 588, 592, 594, 599, 611, 614, 621, 622, 623, 625, 626, 631, 635, 636, 641, 642, 643, 644, 645, 649, 650, 651, 652, 654, 662, 665, 669, 673, 675, 678, 684, 685, 692, 694, 697, 698, 701, 702, 703, 704, 705, 707, 708, 716, 717, 718, 719, 720, 725, 726, 727, 728, 729, 730, 732, 733, 736, 737, 740, 742, 746, 751, 755, 756, 757, 758, 759, 760, 761, 762, 771, 772, 775, 777, 780, 783, 784, 787, 791, 792, 794, 795, 797, 800, 812, 814, 819, 824, 826, 827, 829, 832, 836, 837, 839, 841, 847, 848, 850, 851, 853, 859, 868, 871, 885, 888, 894, 901, 907, 908
zhao, 234, 357, 701, 745
Zhao, Bing, 560
Zhao, Bo, 464, 760
Zhao, Boran, 409
Zhao, Chen, 241, 269, 837, 853
Zhao, Chengshuai, 185
Zhao, Chenxu, 827
Zhao, Chenye, 703
Zhao, ChenZhuo, 303, 684, 885
Zhao, Chunguang, 756
Zhao, Chuyang, 185
zhao, danyang, 357
Zhao, Di, 407
Zhao, Ding, 427
Zhao, Dongbin, 249, 568, 621
Zhao, Dongyan, 145, 379
zhao, Fei, 701, 745
Zhao, Feifei, 98
Zhao, Feiyu, 332
Zhao, Feng, 372, 623, 759, 859
Zhao, Fuheng, 572
Zhao, Guangxiang, 87
Zhao, Guangzhen, 359
Zhao, Guojiang, 599, 649
Zhao, Guoliang, 368
Zhao, H. Vicky, 829
Zhao, Hai, 184, 292, 294, 470, 814
zhao, hai, 234
Zhao, Haishu, 456
Zhao, Haiteng, 851
Zhao, Haiyu, 669
Zhao, Hantao, 88
Zhao, Hao, 784
Zhao, Haodong, 435, 758
Zhao, Haoran, 271
Zhao, Haoyu, 675
Zhao, Haozhe, 772

Zhao, Henry Hengyuan, 888
Zhao, Hongbo, 212
Zhao, Huaipeng, 614
Zhao, Huan, 313
Zhao, Huijun, 575, 611
Zhao, Huiqiang, 155
Zhao, Jeff, 166
Zhao, Jiahui, 141
Zhao, Jiale, 421
Zhao, Jian, 385, 756, 761, 819
Zhao, Jiaqi, 407
Zhao, Jieyu, 479, 561
Zhao, Jihao, 733
Zhao, Jin, 443, 472, 812
Zhao, Jinghua, 352
Zhao, Jinman, 531
Zhao, Jujia, 349
Zhao, Jun, 97, 147, 325, 358, 396, 424, 631, 642, 652, 685, 703, 814, 819, 839, 853
Zhao, Junbo, 719, 728, 758
Zhao, Junchen, 486
Zhao, Junchuan, 354
Zhao, Junfeng, 264, 850
Zhao, Junning, 527
Zhao, Kai, 200, 239, 334, 405
Zhao, Kaili, 847
Zhao, Kaiqi, 125, 181, 641
Zhao, Kaiyan, 542
Zhao, Kangwen, 315, 461
Zhao, Kun, 268
Zhao, Li, 184
Zhao, Liang, 146, 299, 560, 697
Zhao, Lin, 137, 775
Zhao, Liyan, 705
Zhao, Liye, 694
Zhao, Mark, 568
Zhao, Meng, 703
Zhao, Mengchen, 340, 654
Zhao, Minda, 868
Zhao, Mingkuan, 164, 370
Zhao, Minjun, 397, 759
Zhao, Ningyuan, 398
Zhao, Peilin, 756
Zhao, Peizhi, 550
Zhao, Pengpeng, 394
Zhao, Pengxiang, 583, 759, 894
Zhao, Pengyu, 740
Zhao, Pinlong, 662
Zhao, Pu, 349, 684
Zhao, Pukun, 755
Zhao, Qihao, 278
Zhao, Qingfei, 673
Zhao, Qinghua, 308
Zhao, Raoyuan, 583, 812
Zhao, Rui, 384, 720, 751
Zhao, Runcong, 157
Zhao, Runsong, 675, 698
Zhao, Ruochen, 168
Zhao, Sendong, 378, 761
zhao, sendong, 357
Zhao, Shangqing, 635
Zhao, Shanshan, 701
Zhao, Shihan, 757
Zhao, Shiju, 501
Zhao, Shiwan, 352
Zhao, Shu, 169, 625, 727, 746
Zhao, Shuaibing, 702
Zhao, Shuyuan, 221, 703
Zhao, Sihang, 249
Zhao, Siyu, 704
Zhao, Tianwei, 908
Zhao, Tianzhe, 730
Zhao, Tiejun, 323
Zhao, Tong, 103, 162, 446, 511, 644, 708
Zhao, Tuo, 794
Zhao, Wangbo, 636
Zhao, Wanru, 665
Zhao, Wanxu, 839
Zhao, Wayne Xin, 746, 847
Zhao, Wei, 622
Zhao, Weijia, 380
Zhao, Weilin, 543
Zhao, Weilun, 716
Zhao, Weixiang, 755, 762
Zhao, Wenshuo, 224
Zhao, Wenting, 150, 262
Zhao, Xiangyu, 214, 250, 464, 587, 592, 791, 797
Zhao, Xiaohan, 338
Zhao, Xiaohu, 406, 729, 732
Zhao, Xiaomeng, 239
Zhao, Xiaoqi, 514
Zhao, Xin, 119, 717, 848
Zhao, Xinghao, 185
Zhao, Xingmeng, 477
Zhao, XinLong, 651
Zhao, Xinping, 403
Zhao, Xinyang, 645
Zhao, Xinyi, 868
Zhao, Xinyu, 787, 792
Zhao, Xiutian, 780
Zhao, Xu, 569
Zhao, Xuandong, 592
Zhao, Xuanle, 716
Zhao, Xujiang, 771, 824, 837
Zhao, Xukai, 824
Zhao, Xunyi, 362

Zhao, Yafei, 665
Zhao, Yang, 382, 626, 643
Zhao, Yanjun, 469, 826
Zhao, Yanpeng, 760
Zhao, Yanxiao, 403, 540
Zhao, Yanyan, 755, 762
Zhao, Yi, 777, 871
Zhao, Yibo, 697
Zhao, Yida, 841
Zhao, Yiheng, 331
Zhao, Yijia, 166
Zhao, Yilun, 166, 205, 241, 269, 410, 473, 623, 783, 784, 800, 901, 907
Zhao, Yiming, 535
Zhao, Yingxiu, 341
Zhao, Yiping, 581
Zhao, Yiren, 488
Zhao, Yiwen, 716, 746
Zhao, Yizheng, 141
Zhao, Yizhuo, 545
Zhao, Yong, 335
Zhao, Yongxin, 697
Zhao, Yu, 374
Zhao, Yuan, 499
Zhao, Yuanhe, 413
Zhao, Yue, 172, 191, 488, 836
Zhao, Yuhan, 644
Zhao, Yuhang, 650
Zhao, Yukun, 725
Zhao, Yunke, 408
Zhao, Yunxiao, 403, 427
Zhao, Yusheng, 756
Zhao, Yutian, 184
Zhao, Yuxin, 737
Zhao, Yuyang, 311
Zhao, Yuze, 353
Zhao, Zhe, 594, 678, 717
Zhao, Zhejun, 304, 692
Zhao, Zheng, 742
Zhao, Zhengyi, 184, 726
Zhao, Zheyu, 260
Zhao, Zhixiong, 718
Zhao, Zhixue, 247
Zhao, Zhiyuan, 239
Zhao, Zhou, 114, 339, 650, 707, 780, 832
Zhao, Zhuokai, 192, 588
Zhao, Zichen, 507
Zhao, Zifeng, 649
Zhao, Zihan, 457
Zhao, Zijia, 736
Zhao, Zijing, 669
Zhao, Ziqi, 795
Zhao, Ziran, 372
Zhao, Ziwei, 692
Zhao;,
Zhao;, Jieyu,
ZHE, 565
ZHE, YU, 565
Zhefang, 569
Zhefang, Zhefang, 569
Zhelnin, 330, 342
Zhelnin, Maxim, 330, 342
Zhen, 150, 312, 372, 391, 580, 759
Zhen, Dongyang, 150
Zhen, Genpeng, 391
Zhen, Hui-Ling, 312, 372, 759
Zhen, Shuai, 580
Zheng, 87, 91, 100, 110, 119, 124, 125, 126, 129, 161, 171, 175, 181, 184, 189, 214, 234, 239, 240, 241, 248, 253, 260, 295, 315, 317, 318, 321, 323, 327, 329, 330, 332, 334, 336, 339, 340, 341, 346, 350, 369, 373, 382, 398, 401, 402, 404, 406, 410, 411, 412, 413, 414, 418, 436, 443, 448, 454, 456, 460, 466, 469, 474, 479, 496, 508, 527, 535, 540, 544, 551, 554, 561, 565, 585, 599, 604, 620, 624, 631, 635, 636, 646, 648, 653, 660, 661, 666, 675, 676, 680, 684, 690, 691, 695, 698, 702, 703, 706, 707, 712, 720, 729, 733, 751, 753, 758, 760, 769, 775, 787, 805, 823, 828, 843, 847, 854, 856, 872, 882, 885, 886, 897, 901
Zheng, Andy, 599
Zheng, Bang, 110
Zheng, Baolin, 660
Zheng, Binbin, 171, 329, 691
Zheng, Bo, 234, 341, 369, 402, 636, 660, 675, 695, 698, 702, 707, 758, 885
Zheng, Bolong, 332
Zheng, Chujie, 897
Zheng, Congmin, 248
Zheng, Da, 334, 436
Zheng, Dajun, 340
Zheng, Dian, 100
Zheng, Dixin, 847
Zheng, Feng, 318
Zheng, Guangmin, 551
Zheng, Guanjie, 448
Zheng, Guanyu, 318, 323
Zheng, Gui, 635
Zheng, Guozhou, 398
Zheng, Hai-Tao, 460, 535, 698
Zheng, Haizhong, 604
Zheng, Hanwen, 91
Zheng, Hao, 805

Zheng, Haonan, 87
Zheng, Haoran, 401
Zheng, Haorui, 456
Zheng, Haoyu, 189, 327
Zheng, Hongbo, 823
Zheng, Hongwei, 753, 769
Zheng, Huan, 646
Zheng, Huawei, 703
Zheng, Huiwen, 680
Zheng, Huiyuan, 382
Zheng, Jian Liao;JianXing,
Zheng, Jianjie, 684
Zheng, Jiasheng, 653
Zheng, Jie, 336
Zheng, Jingsheng, 317
Zheng, Jingting, 119
Zheng, Jinguang, 680
Zheng, Jingyi, 854
Zheng, Jinquan, 410, 411
Zheng, Jiyang, 340
Zheng, Kai, 474, 712
Zheng, Kexin, 129
Zheng, Lecheng, 240
Zheng, Leqi, 698
Zheng, Li, 856
Zheng, Lingchao, 406
Zheng, Mao, 872
Zheng, Mengxin, 882
Zheng, Mingqian, 466
Zheng, Mingwei, 828
Zheng, Mingyu, 872
Zheng, Nanning, 775
Zheng, Peizhen, 406
Zheng, Pengcheng, 404
Zheng, Pengfei, 414
Zheng, Pujun, 410, 411
Zheng, Qian, 565, 843
Zheng, Qiaoyu, 561
Zheng, Shiyi, 253
Zheng, Shuntian, 544
Zheng, Shuxin, 886
Zheng, Tao, 554
Zheng, Tianqi, 175
Zheng, Tianshi, 161, 253, 733
Zheng, Tianyu, 260
Zheng, Tong, 241, 413
Zheng, Weihua, 469
Zheng, Weijian, 872
Zheng, Wenxiang, 690
Zheng, Wenzhen, 350
Zheng, Wuqiang, 648
Zheng, Xianda, 125, 181
Zheng, Xiangxi, 624
Zheng, Xiao, 805
Zheng, Xiaofan, 706
Zheng, Xiaoqing, 321, 661
Zheng, Xiaorou, 336
Zheng, Xiaoyun, 720
Zheng, Xin, 653
Zheng, Xin TongWeidong ZhangJiaang LiHaibin ChenShilei LiuLangming LiuKangtao LvYujin YuanWenbo SuBo, 751
Zheng, Xinyu, 479
Zheng, Xu, 666
Zheng, Xunwen, 527
Zheng, Yan, 787
Zheng, Yefeng, 184, 315
Zheng, Yi, 346, 508
Zheng, Ying, 443
Zheng, Yongsen, 702
Zheng, Yu, 327
Zheng, Yuanhong, 239
Zheng, Yuanjin, 214
Zheng, Yuanlei, 124
Zheng, Yueyuan, 496
Zheng, Yujia, 454
Zheng, Yushuo, 339
Zheng, Zaiyi, 585
Zheng, Zeyu, 248
Zheng, Zhangkai, 418
Zheng, Zhe, 631
Zheng, Zhedong, 330, 676
Zheng, Zhen, 901
Zheng, Zhi, 340, 436
Zheng, Zhicheng, 733
Zheng, Zhiming, 753
Zheng, ZhuHang, 373
Zheng, Zibin, 620, 729
Zheng, Zihan, 412
Zheng, Zihao, 540
Zheng, Zilong, 760
Zheng, Zishuo, 295
Zheng, Ziwei, 126
Zheng;,
Zheng;, Ziqian,
Zhenyu, 318
Zhenyu, Wang, 318
Zhi, 107, 298, 630, 762
Zhi, Gong, 107
Zhi, Xuda, 762
Zhi, Xuyang, 298
Zhi, Zhuo, 630
Zhixiao, 186
Zhixiao, Jiang, 186
ZhixueSong, 732
ZhixueSong, ZhixueSong, 732

Zhong, 92, 136, 162, 205, 206, 274, 283, 298, 340, 368, 403, 404, 424, 429, 513, 532, 534, 546, 576, 602, 611, 629, 645, 655, 656, 659, 660, 674, 690, 697, 720, 736, 739, 751, 783, 785, 815, 832, 859, 874
Zhong, Aoxiao, 205, 546
Zhong, Cheng, 783
Zhong, Ding, 832
Zhong, Fangwei, 424, 532, 815
Zhong, Hanmeng, 136, 429
Zhong, Heming, 206
Zhong, Hongbin, 534
Zhong, Hongqiong, 660
Zhong, Jiang, 162, 629
Zhong, Jianyuan, 298
Zhong, Jin, 690
Zhong, Lin, 645, 674
Zhong, Lingfeng, 368
Zhong, Linhao, 576, 785
Zhong, Linqing, 404
Zhong, Qimin, 659
Zhong, Qiuyue, 513
Zhong, Ruobin, 655
Zhong, Shuxin, 736
Zhong, Weihong, 697, 739
Zhong, Wenliang, 283
Zhong, Xu, 340
Zhong, Xuewen, 720, 751
Zhong, Yan, 403
Zhong, Yaoyao, 874
Zhong, Yicheng, 656
Zhong, Yishan, 602
Zhong, Yuan, 611
Zhong, Zhanping, 274, 859
Zhong, Zhenyu, 92
Zhongteng_Cai1, Zhongteng_Cai1, Zhongteng_Cai1, 815
Zhongwei, Zhongwei, 815
ZHOU, 501
Zhou, 81, 82, 87, 88, 92, 94, 96, 97, 102, 108, 118, 119, 120, 121, 128, 130, 134, 135, 136, 137, 138, 143, 145, 146, 147, 149, 150, 153, 161, 162, 163, 178, 179, 189, 208, 217, 219, 225, 230, 233, 236, 239, 245, 250, 268, 272, 273, 279, 281, 282, 284, 285, 292, 295, 296, 297, 309, 311, 312, 313, 315, 318, 319, 322, 323, 328, 332, 334, 339, 340, 341, 346, 348, 349, 350, 352, 354, 357, 358, 362, 363, 369, 373, 374, 379, 382, 387, 388, 389, 395, 406, 407, 409, 410, 413, 414, 418, 419, 421, 422, 423, 425, 429, 432, 436, 437, 440, 449, 455, 457, 458, 461, 463, 466, 470, 471, 474, 479, 486, 487, 496, 506, 507, 514, 515, 516, 519, 527, 531, 532, 537, 538, 540, 543, 551, 555, 561, 563, 564, 566, 568, 570, 572, 581, 588, 593, 594, 595, 596, 599, 601, 609, 610, 617, 620, 621, 625, 626, 627, 629, 636, 640, 643, 645, 646, 657, 659, 666, 668, 672, 674, 680, 684, 686, 689, 692, 693, 707, 709, 714, 719, 721, 723, 725, 731, 732, 734, 735, 739, 740, 744, 745, 747, 748, 750, 751, 753, 758, 760, 773, 775, 777, 787, 789, 790, 792, 793, 797, 799, 801, 805, 809, 814, 817, 831, 832, 838, 839, 841, 848, 850, 853, 854, 856, 857, 859, 862, 867, 875, 877, 881, 886, 893, 896, 900, 906, 909
zhou, 298
Zhou, Aimin, 311, 514
Zhou, Ao, 97, 128, 801
Zhou, Aojun, 339
Zhou, Baichuan, 599
Zhou, Baohang, 374
Zhou, Ben, 348
Zhou, Bin, 373, 496
Zhou, Bing, 334
Zhou, Bingsen, 456
Zhou, Biyu, 143, 389, 793
Zhou, Bo, 540, 753, 760, 848, 900
Zhou, Bowen, 239, 413, 423, 750, 790, 893
Zhou, Cangqi, 414
Zhou, Chenghu, 596
Zhou, Chengming, 407
Zhou, Chenxi, 147, 814
Zhou, Chenyue, 692
Zhou, Chuan, 319
Zhou, Chuang, 245, 449, 854
Zhou, Chunyi, 121, 566, 877
Zhou, Chuyan, 841
Zhou, Deyu, 374, 437, 668
Zhou, Dongruo, 309
Zhou, Dongzhuoran, 422
Zhou, Dunzhi, 118
Zhou, Duo, 457
Zhou, Dylan, 179
Zhou, Ej, 862
Zhou, Enyu, 838
Zhou, Gengze, 362
Zhou, Giulio, 875
Zhou, Grace Ziqi, 230
Zhou, Guodong, 506, 731, 748
Zhou, Guorui, 379, 414, 672, 777
Zhou, Haiqi, 463

Zhou, Hanyu, 217
Zhou, Hanzhang, 410
Zhou, Hao, 94, 268, 537, 543, 723
Zhou, Haojie, 350
Zhou, Haowu, 856
Zhou, Hebin, 292
Zhou, Heng, 797
Zhou, Hong-Yu, 284
Zhou, Hongkuan, 422
Zhou, Hongyu, 292
Zhou, Huan, 150
Zhou, Huichi, 790
Zhou, Irene, 909
Zhou, James, 233
Zhou, Jia, 719
Zhou, Jiaming, 352, 744
Zhou, Jiamu, 721
Zhou, Jian-Yu, 551
Zhou, Jiang, 94, 406, 732
Zhou, Jianqun, 760
Zhou, Jiawei, 88, 507
Zhou, Jiaxi, 896
Zhou, Jiaying, 147
Zhou, Jie, 96, 245, 281, 296, 425, 537, 626, 629, 758, 877
Zhou, Jiepeng, 409
Zhou, Jing, 719, 740
Zhou, Jingbo, 178, 609
Zhou, Jingren, 120, 178, 296, 388, 853
Zhou, Jingtao, 601
Zhou, Jinxing, 136
Zhou, Jiuan, 568
Zhou, Joey Tianyi, 146
ZHOU, JUN, 501
Zhou, Jun, 145, 349, 817, 848
Zhou, Junyi, 474
Zhou, Kaitlyn, 295, 458
Zhou, Karen, 620
Zhou, Kun, 119, 130, 137, 189, 236, 273, 809
Zhou, Le, 747
Zhou, Lei, 686
Zhou, Linna, 714
Zhou, Long, 432
Zhou, Mengyu, 643, 773
Zhou, Mingyang, 659
Zhou, Mingyuan, 486
Zhou, Nianjun, 610
Zhou, Pan, 350, 640
Zhou, Pei, 479, 561
Zhou, Peilin, 230
zhou, Peilun, 298
Zhou, Peng, 136, 429
Zhou, Peng Yuan, 857
Zhou, Pengfei, 636
Zhou, Qiang, 707
Zhou, Qiji, 369
Zhou, Qingshan, 797
Zhou, Rongkun, 555
Zhou, Rui, 751
Zhou, Ruihua, 279, 593
Zhou, Sashuai, 707
Zhou, Shengli, 318
Zhou, Shilin, 149
Zhou, Shiqi, 627
Zhou, Shize, 532
Zhou, Shu, 519
Zhou, Shuang, 311
Zhou, Shuhan, 328
Zhou, Shuheng, 626
Zhou, Sicheng, 87
Zhou, Silin, 108
Zhou, Sizhuo, 134
Zhou, Taotao, 350
Zhou, Tianle, 617
Zhou, Tianyang, 831
Zhou, Tianyi, 102, 138
Zhou, Tong, 358
Zhou, Wanyun, 162
Zhou, Wei, 97, 322, 352, 363, 564, 684, 792
Zhou, Weidong, 657, 689
Zhou, Weien, 135
Zhou, Wenbo, 358
Zhou, Wengang, 315, 461
Zhou, Wenqing, 354
Zhou, Wenrui, 886
Zhou, Wenxuan, 225, 297, 594, 792, 886
Zhou, Wenze, 625
Zhou, Wuyang, 538
Zhou, Xiao, 161, 867
Zhou, Xiaofan, 595
Zhou, Xiaofei, 831, 906
Zhou, Xiaohuan, 340, 657
Zhou, Xiaoling, 735
Zhou, Xiaoyu, 395, 407, 734
Zhou, Xin, 515, 531
Zhou, Xingcheng, 881
Zhou, Xingzhi, 723
Zhou, Xiuze, 674
Zhou, Xuan, 282
Zhou, Xuanhe, 239
Zhou, Xuanru, 143
Zhou, Xueyang, 350
Zhou, Xuhui, 466
Zhou, Xuwen, 805
Zhou, Yan, 341
Zhou, Yang, 421

Zhou, Yanghao, 136
Zhou, Yangjie, 487
Zhou, Yanlin, 725
Zhou, Yanyun, 867
Zhou, Yayu, 153
Zhou, Yifan, 599, 797
Zhou, Yigeng, 354
Zhou, Yilun, 285, 789
Zhou, Yiming, 382
Zhou, Ying, 387, 516
Zhou, Yingbo, 312
Zhou, Yinghan, 581
Zhou, Yixi, 527, 886
Zhou, Yixiao, 693
Zhou, Yixuan, 758
Zhou, Yiyang, 832
Zhou, Yongcheng, 334
Zhou, Youchao, 437
Zhou, Yu, 570
Zhou, Yuan, 859
Zhou, Yuanchun, 163
Zhou, Yubo, 666
Zhou, Yucheng, 625, 646, 689
Zhou, Yuhang, 395, 561, 588
Zhou, Yuhao, 250, 440, 850
Zhou, Yujia, 373
Zhou, Yujin, 470
Zhou, Yujun, 357, 471, 709
Zhou, Yunhua, 839
Zhou, Yuyan, 745
Zhou, Yuyin, 346
Zhou, Zekun, 739
Zhou, Zeqi, 572
Zhou, Zetong, 799
Zhou, Zhanchao, 92
Zhou, Zhanhui, 81
Zhou, Zhaomeng, 418, 419, 436
Zhou, Zhehao, 323
Zhou, Zheli, 877
Zhou, Zhengyang, 313
Zhou, Zhengyi, 208
Zhou, Zhenhao, 906
Zhou, Zhenhong, 82, 621, 775
Zhou, zhijie, 645
Zhou, Zhongliang, 219
Zhou, Zhongyi, 563, 601
Zhou, Zineng, 423
Zhou, Ziqin, 136
Zhou, Zisheng, 516
Zhou, Zixuan, 422
Zhou, Ziyang, 744
Zhou, Ziyi, 272, 787
Zhou, Zuyi, 332, 680
Zhou;,
Zhou;, Dawei,
Zhu, 86, 88, 95, 103, 105, 110, 117, 123, 130, 131, 132, 144, 151, 153, 157, 158, 162, 163, 164, 180, 185, 187, 189, 197, 203, 204, 213, 215, 216, 217, 224, 225, 226, 237, 240, 241, 248, 252, 254, 255, 260, 261, 264, 271, 272, 274, 275, 284, 290, 295, 303, 304, 307, 309, 310, 313, 314, 315, 322, 327, 330, 332, 335, 336, 337, 339, 341, 342, 347, 350, 352, 361, 363, 368, 369, 370, 372, 373, 375, 376, 380, 381, 383, 386, 391, 395, 398, 402, 403, 405, 413, 414, 418, 421, 422, 423, 436, 441, 445, 452, 456, 459, 461, 464, 469, 470, 475, 480, 501, 503, 504, 511, 513, 527, 528, 536, 540, 542, 543, 545, 548, 559, 564, 569, 572, 576, 582, 599, 607, 618, 619, 626, 627, 632, 642, 645, 646, 648, 653, 655, 660, 662, 669, 671, 673, 674, 675, 676, 677, 678, 679, 687, 690, 697, 699, 700, 715, 718, 719, 723, 726, 728, 729, 730, 732, 733, 734, 737, 738, 739, 742, 748, 758, 760, 761, 767, 769, 774, 777, 779, 786, 788, 789, 790, 797, 798, 803, 818, 830, 833, 835, 838, 840, 845, 849, 857, 859, 863, 864, 872, 874, 877, 880, 885, 886, 892
zhu, 724
Zhu, Andrew, 151
Zhu, Beier, 645
Zhu, Bin, 352, 414, 798
Zhu, Chenyu, 677, 790
Zhu, Chiwei, 729
Zhu, Dantong, 264
Zhu, Dawei, 542, 678
Zhu, Derui, 88
Zhu, Dingwei, 545, 838
Zhu, Dong, 373
Zhu, Erle, 786
Zhu, Fa, 380
Zhu, Fengbin, 303, 648
Zhu, Fengqi, 501
zhu, Gaoyu, 724
Zhu, Han, 513, 543, 569
Zhu, Hang, 441
Zhu, Hangxiao, 475
Zhu, Hao, 260, 742
Zhu, He, 284
Zhu, Hengshu, 163, 718, 730
Zhu, Hongtu, 241, 528
Zhu, Huijia, 626
Zhu, Huining, 405

Zhu, Jerry, 237
Zhu, Jia, 132, 144, 216, 830
Zhu, Jiachen, 248, 381, 877
Zhu, Jiahao, 673
Zhu, Jian, 217, 304, 456
Zhu, Jianchen, 540, 774
Zhu, Jiangcai, 679
Zhu, Jianing, 626
Zhu, Jiaqi, 403
Zhu, Jiayin, 464
Zhu, Jihua, 719
Zhu, Jinchang, 732
Zhu, JingBo, 335, 413, 739, 748
Zhu, Jingbo, 456, 626, 675
Zhu, Jingzhe, 421
Zhu, Jinhua, 315, 461
Zhu, Jun, 779
Zhu, Junlin, 697
Zhu, Junnan, 162, 391, 777
Zhu, Juno, 559
Zhu, Junyi, 677
Zhu, Kaijie, 767
Zhu, Kejian, 642
Zhu, Kenny, 527
Zhu, Kevin, 576
Zhu, Kun, 459
Zhu, kun, 697
Zhu, Kunhong, 368
Zhu, Lanyi, 892
Zhu, Lanyun, 607
Zhu, Li, 330, 441, 676
Zhu, Liang, 849
Zhu, Linchao, 164, 370
Zhu, Linjie, 375
Zhu, Liuwan, 153
Zhu, Mengxiao, 314, 361
Zhu, Minfeng, 370, 687
Zhu, Ming, 761
Zhu, Minghao, 224
Zhu, Minjun, 840
Zhu, Nengjun, 395
Zhu, Pengcheng, 363
Zhu, Pengyu, 383
Zhu, Pengyun, 369, 386
Zhu, Qi, 264, 398, 452, 797, 880
Zhu, Qianchao, 332
Zhu, Qiang, 352, 564
Zhu, Qiannan, 237
Zhu, Qiaoming, 123, 240
Zhu, Qiming, 252, 653
Zhu, Qin, 313
Zhu, Qing, 734
Zhu, Qingfu, 350, 627, 662
Zhu, Qinglin, 157
Zhu, Qiyuan, 197, 272, 347, 619
Zhu, Renjin, 674
Zhu, Rongyi, 117
Zhu, Rui, 789
Zhu, Ruihan, 675
Zhu, Ruike, 261
Zhu, Shang, 295
Zhu, Shaolin, 618
Zhu, shaolin, 690
Zhu, Shenzhe, 599
Zhu, Shucheng, 86
Zhu, Shuyue, 874
Zhu, Sibo, 130
Zhu, Siyi, 254
Zhu, Siyu, 599, 645
Zhu, Song-Chun, 818
Zhu, Songchun, 760
Zhu, Suxia, 342
Zhu, Tianqing, 423
Zhu, Tiantian, 158
Zhu, Tinghui, 671
Zhu, Tingyuan, 767
Zhu, Tong, 777
Zhu, Tongyao, 95, 241
Zhu, Wang Bill, 582
Zhu, Wei, 203
Zhu, Wenhao, 445
Zhu, Wenhui, 204
Zhu, Wenpeng, 632
Zhu, Xiangrong, 737
Zhu, Xiangyang, 559
Zhu, Xiaoao, 336
Zhu, Xiaobin, 105
Zhu, Xiaobing, 290
Zhu, Xiaochen, 503, 572, 864
Zhu, Xiaodan, 180
Zhu, Xiaowei, 327
Zhu, Xiaoyong, 660, 885
Zhu, Xingyu, 645
Zhu, Xinrui, 646
Zhu, Xinyu, 480
Zhu, Xiyue, 261
Zhu, Xuyuan, 418
Zhu, Yada, 469
Zhu, Yakun, 255, 528
Zhu, Yangfu, 737
Zhu, Yanmin, 723, 726
Zhu, Yanming, 376
Zhu, Yaoming, 863
Zhu, Yi, 398, 673
Zhu, Yichi, 715
Zhu, Yifan, 548, 669

Zhu, Yifei, 872
Zhu, Yihua, 213
Zhu, Yilun, 835
Zhu, Yinan, 845
Zhu, Yinghao, 264, 423
Zhu, Yingjian, 187
Zhu, Yinglun, 309
Zhu, Yitong, 857
Zhu, Yongchun, 769
Zhu, Yongpeng, 110
Zhu, Yuanjie, 307
Zhu, Yuchen, 662
Zhu, Yucheng, 339
Zhu, Yue, 185
Zhu, Yueqi, 475
Zhu, Yuke, 402, 758
Zhu, Yun, 189, 274, 675, 803, 859
Zhu, Yuqi, 436, 655
Zhu, Yuqicheng, 422
Zhu, Yutao, 103, 310, 337, 511, 699, 700, 728
Zhu, Yuxuan, 215, 271, 536
Zhu, Zekun, 341, 733
Zhu, Zerun, 304
Zhu, Zexuan, 158
Zhu, Zeyu, 363
Zhu, Zhenghao, 470
Zhu, Zhengqiu, 335, 673
Zhu, Zhengzhou, 788
Zhu, Zhihong, 315, 738
Zhu, Zhiyuan, 322
Zhu, Zhongrui, 261
Zhu, Zhou, 833
Zhu, Zikun, 886
Zhu, Zining, 226, 275
Zhu, Ziqi, 372
Zhu, Ziwei, 86, 131, 225, 504
Zhuang, 126, 132, 161, 259, 304, 327, 373, 454, 507, 517, 584, 631, 650, 668, 675, 751, 773, 778, 835, 897
Zhuang, Bohan, 675
Zhuang, Haomin, 259
Zhuang, Huiping, 773
Zhuang, Jiabao, 897
Zhuang, Jiawei, 517
Zhuang, Jun, 132, 584
Zhuang, Nan, 778
Zhuang, Qiao, 507
Zhuang, Shengyao, 161
Zhuang, Xialie, 650
Zhuang, Yan, 373
Zhuang, Yuan, 454, 835
Zhuang, Yueting, 126, 304, 631, 751
Zhuang, Ziqing, 668
Zhuang, Zirui, 327
ZHUANGZHUORAN, 705
ZHUANGZHUORAN, ZHUANGZHUORAN, 705
Zhuo, 114, 222, 620, 638, 696, 832
Zhuo, Chengxiang, 638
Zhuo, Fan, 114, 832
Zhuo, Ruiyu, 696
Zhuo, Terry Yue, 222, 620
zhuoning, 290
zhuoning, Zhu, 290
Zi, 186, 465
Zi, Evelyn H., 465
Zi, Kangli, 186
Zia, 355
Zia, Zeeshan, 355
Ziabari, 226, 539
Ziabari, Alireza Salkhordeh, 226, 539
Ziakas, 512
Ziakas, Christos, 512
Ziegenbein, 857
Ziegenbein, Timon, 857
Ziheng, 424, 768, 815
Ziheng, Zhou, 424, 768, 815
ZijianWen, 774
ZijianWen, ZijianWen, 774
Zimerman, 453, 600
Zimerman, Itamar, 453, 600
Zimmer, 822, 881
Zimmer, Jonas, 822
Zimmer, Walter, 881
Zimmermann, 414
Zimmermann, Roger, 414
Zimny, 817
Zimny, Jan, 817
Ziser, 270
Ziser, Yftah, 270
Zisserman, 903
Zisserman, Andrew, 903
Zitnik, 667
Zitnik, Marinka, 667
Zong, 161, 380, 382, 387, 462, 500, 741, 808, 892
Zong, Chengqing, 382, 387
Zong, Hao, 741
Zong, Peiyuan, 380
Zong, Qing, 161, 808
Zong, Tianyu, 462, 500, 892
Zongze, 116, 567
Zongze, Li, 116, 567
Zotter, 246
Zotter, Christof, 246
Zou, 100, 111, 144, 161, 212, 243, 278, 284, 295, 312, 393, 400, 404, 441, 469, 487, 488,

514, 623, 625, 631, 666, 704, 707, 725, 727, 746, 747, 748, 776, 795, 826, 839, 883, 893, 898, 907
Zou, An, 747
Zou, Bin, 704
Zou, Bowei, 212, 469
Zou, Chenye, 487
Zou, Guanyun, 725
Zou, Haochen, 400
Zou, Haodong, 625, 727, 746
Zou, Henry Peng, 111, 393, 488, 893
Zou, James, 284, 295, 441
Zou, Jiahong, 795
Zou, Jiaru, 826
Zou, Jie, 404
Zou, Junjie, 748
Zou, Kai, 100, 161
Zou, Lancheng, 312
Zou, Lei, 278
Zou, Quanchen, 883
Zou, Shihao, 514
Zou, Shun, 623
Zou, Sirui, 898
Zou, Xiaohan, 776
Zou, Xin, 144, 666, 707
Zou, Xingchen, 907
Zou, Xinyu, 631
Zou, Yicheng, 839
Zou, Zhiying, 243
Zouhar, 156
Zouhar, Vilém, 156
Zu, 783
Zu, Lizheng, 783
Zugarini, 614
Zugarini, Andrea, 614
Zuo, 309, 629, 811
Zuo, Bowen, 309
Zuo, Haomin, 629
Zuo, Longfei, 811
Zuromski, 813
Zuromski, Kelly L., 813
Zweigenbaum, 413
Zweigenbaum, Pierre, 413
Zwingenberger, 517
Zwingenberger, Niklas, 517
Zymet, 566
Zymet, Jesse, 566
Züfle, 126, 779
Züfle, Maike, 126, 779

Özsu, 399
Özsu, M. Tamer, 399
Özyürek, 302
Özyürek, Asl, 302

Üstün, 826
Üstün, Ahmet, 826

urech, 784
urech, Eduard Frank, 784

redniawa;,
redniawa;, Daniel Cielak; Wadysaw,

ahinuç, 143
ahinuç, Furkan, 143

tefánik, 891
tefánik, Michal, 891
torek, 91, 532
torek, Adam, 91, 532

AI at Bloomberg

Bloomberg is building the world's most trusted information network for financial professionals. Our 9,000+ engineers, developers, and data scientists are dedicated to advancing and building new solutions and systems for the Bloomberg Terminal and other products in order to solve complex, real-world problems.

One of the first on the street to use AI, we have 15+ years of experience building and using AI solutions to help process and organize the ever-increasing volume of structured and unstructured financial information. Our team of 400+ AI researchers and engineers deploy different AI technologies, including ML, NLP, computer vision, information retrieval, generative models, and agentic AI, as well as neural networks and time-series analysis. This is how we provide our customers around the globe with the reliable, high-quality data, news, and information they need to make well-informed decisions and be more productive.

Learn more here:

#MAKEITHAPPENHERE

Make it happen here.

Bloomberg

Engineering

Capital One

OTHERS ARE USING AI.
WE'RE ADVANCING IT.

Alibaba Cloud

Qwen Cloud

Qwen Cloud

Ship the Next_

Qwen Cloud is an **AI-native Cloud Platform** built for AI agents and the developers behind them, delivering Qwen-family and third-party models through **three zero-friction entry points**: Skills for agents, CLI tools for terminal-native workflows, and a web experience for humans

Feature Deep Dive

Model Marketplace - Side by Side, Data Speaks, HappyHorse ready out of the box

OpenAI-compatible API - swap the URL, keep your code, go live.

Skills - Your Agent's AI Toolkit

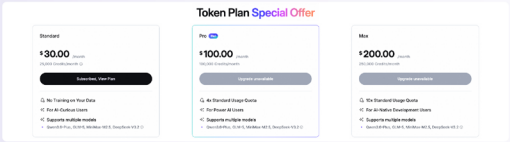
Install Qwen Cloud Skills with one command - your AI agent becomes a MaaS expert that routes models, runs API calls, and returns outputs in one flow.

CLI - Check Usage from the Terminal

Query usage, billing, and cost breakdown with a single command.

Qwen Cloud offers free model usage to new users, allowing new users to try out a wide variety of models before deployment.

Token Plan is Available



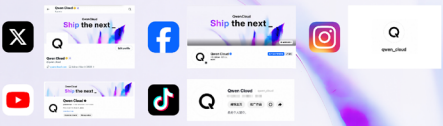
Token Plan supports a variety of models and offers controllable monthly costs, making it ideal for advanced AI programming and agent-based tool use cases.

The Frontier in Your Pocket

is coming soon

Discover. Try. Build. Manage AI capabilities — agents, images, videos, and more — in one consumer-friendly Qwen Cloud app.

Follow Qwen Cloud to stay updated with the latest news



Scan to launch Qwen Cloud

Join our Discord

Megagon Labs

Connecting AI, Data, and People
for Better Opportunities

blue

Turn messy data into agent-ready workflows.

ORACLE

Real careers in artificial intelligence

Oracle is proud to sponsor the **64th Annual Meeting of the Association for Computational Linguistics**. And we're just as proud to help shape what's next in AI. From advancing reasoning and planning to leading benchmarks in Spider 2.0 for enterprise NL2SQL, our teams are pushing the boundaries of applied AI.

Now, it's your turn. Grow your career and thrive with the AI, cloud, and database leader.

Visit oracle.com/acl or scan to discover our latest opportunities

Create the **future** with us

Appen

Human data for frontier AI

The world's leading AI models are built on more than algorithms; they're built on human expertise. We deliver the expert-validated data that trains frontier models, ensuring AI systems understand nuance, context, and complexity at scale.

Data products to build foundational AI

- Frontier Alignment
- Agentic AI
- Speech & Audio
- Multimodal AI
- Physical AI
- Model Integrity

30 Years of Pioneering Data

appen.com

文心大模型

BAIDU WENXIN

A series of knowledge-enhanced large models, from general-purpose to industry-specific, independently developed by Baidu, speed up AI innovations and empower the industry to upgrade.

ERNIE

2026.5
ERNIE 5.1

2026.1
ERNIE 5.0

2025.11
ERNIE 5.0 PREVIEW
- Natively Multimodal Model
- Elastic Training

2025.9
ERNIE X1.1

2025.6
ERNIE 4.5 Series Open-Sourced

2025.4
ERNIE 4.5 Turbo / X1 Turbo

2025.3
ERNIE 4.5 / X1
- Deep Thinking
- Multimodality

2024.6
ERNIE 4.0 Turbo
- Knowledge Distillation

2023.10
ERNIE 4.0
- Incremental Learning
- Multi-dimensional Reward Model
- Agent

2023.5
ERNIE 3.5
- Factual Knowledge Enhancement
- Reasoning Enhancement

2023.3
ERNIE Bot
- Knowledge Enhancement
- Retrieval Augmentation
- Dialogue Enhancement

2021.7
ERNIE 3.0
- Knowledge-infused Learning

2019.7
ERNIE 2.0
- Knowledge-guided Multi-task Learning

2019.3
ERNIE 1.0
- Multi-granularity Knowledge Fusion Learning

Knowledge Enhancement	Retrieval Augmentation	Dialogue Enhancement	Agents
Supervised Fine-Tuning	Reinforcement Learning	Prompt	
Pre-Training Models			

腾讯青云计划
TENCENT PROJECT UP

TENCENT PROJECT UP

腾 讯 青 云 计 划

About PROJECT UP

Project Up is a Tencent talent program with the aim of recruiting top technology students from around the world.

FOR WHO

Bachelor's, Master's, and PhD students graduating after September 2026

AT WHERE

Shenzhen, Beijing, Shanghai, Guangzhou, Chengdu, Hangzhou, the United States, Singapore, the United Kingdom, the Netherlands, etc

Tencent Project UP Official Website

Core Application Scenarios

Tencent HY
Building AGI for the Greater Good

Games
Exploring hyper-digital scenarios

Yuanbao
Unlocking the infinite possibilities of AI-driven social experiences

Weixin/WeChat
New AI-empowered product ecosystem

FinTech
AI + finance, creating the next-generation financial experiences

Tencent Cloud
Strengthening the digital foundation of industries

Content
Redefining future content creation

Tencent Advertising
Making business smarter with AI

From “Advancing regulatory variant effect prediction with a unified DNA sequence model”

Published | BioRxiv

to revealing the hidden genetic triggers of diseases like leukemia.

Work at Google

g.co/AIcareers

ANT GROUP | 2025 INCLUSION AI

ANT STAR
Plan A Talent Program

Join Ant Group and become a member of "Plan A"
Light up the future with technology together!

"Plan A" is a new elite AI talent initiative under "Ant Star." **It aims to recruit world-class artificial intelligence researchers from leading global universities to join Ant Group's exploration of AGI.**

Scan the code to learn about Ant Group

The Minds We Seek

Top talents in LLM, Multimodal, Infra, RL, Diffusion, Sparse Attention,Robotics, and Synthetic data.

Recruitment Position:

Research Scientist -

- Reinforcement Learning
- Deep Learning
- Synthetic Data
- Inference
- LLM Pre-training
- LLM Post-training
- Diffusion
- Multimodal

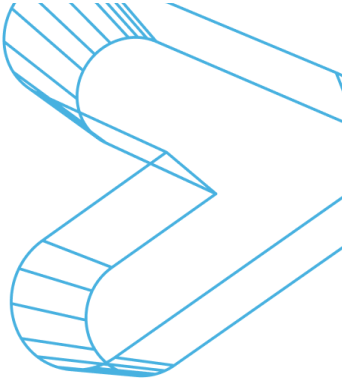
What We Offer

Frontier Technology Arena
- Industry pioneer mentorship
- Mission-critical R&D challenges
- Strategic core tech domains

Unconstrained Innovation
- Personalized development roadmaps
- Full research freedom
- Unlimited GPU resources

Elite Growth Ecosystem
- Levels and salaries based on ability, not seniority
- Get recognized faster for what you achieve

Global R&D network
- Hubs across Hangzhou/ Beijing/Shanghai/Singapore
- Integrated research labs
- Cross-disciplinary exchanges



snowflake®

LET'S BUILD THE NEXT BIG THING TOGETHER.

careers.snowflake.com

Joining Snowflake means joining our mission: to empower every enterprise to achieve its full potential through data and AI. And that mission starts with you.

Snowflake's AI Research team explores the frontier of artificial intelligence, conducting open, foundational, and applied research to turn big questions into bigger breakthroughs. This is where research across agents, post-training, and AI training and inference systems moves beyond theory and into real-world impact. Where ideas become easy, connected, and trusted AI features that power our platform and inspire innovation.

Join our team and expand what's possible in enterprise AI—at Snowflake and beyond.

CHASE YOUR NEXT PEAK

Scan the QR code to unlock opportunities at Snowflake.

GLOBAL ENGINEERING HUBS

- Menlo Park, California
- Bellevue, Washington
- Toronto, Canada
- Berlin, Germany
- Warsaw, Poland
- San José, Costa Rica

WHERE DATA DOES MORE

北京中关村学院 Zhongguancun Academy x 中关村人工智能研究院 Zhongguancun Institute of Artificial Intelligence

ZGCA • ZGCI

PER ASPERA AD ASTRA

A New Model for Innovation, Learning, and Leadership
One Vision Exponential Impact

ZGCA & ZGCI:
A Unified Approach to AI Advancement

Zhongguancun Academy and the Zhongguancun Institute of Artificial Intelligence operate as one unified development community—combining strengths and compounding impact and resources. ZGCA reimagines how AI talent is developed. ZGCI turns research into real-world solutions.

Guided by the shared principle—Extreme Fundamentals, Extreme Applications, Extreme Interdisciplinarity—the two interweave faculty, research, and technology incubation. This tight integration forges a seamless path from academic exploration to practical impact— all while cultivating top AI talent and delivering breakthrough innovation.

Our Pedagogy

EXTREMELY CLASSIC **EXTREMELY CUTTING-EDGE** **EXTREMELY PRACTICAL**

Our Research Approach

EXTREMELY FUNDAMENTAL **EXTREMELY APPLIED** **EXTREMELY INTERDISCIPLINARITY**

Academic Divisions

AI Core

01 AI Theory
02 Foundation Models
03 Embodied AI
04 AI Infrastructure & System
05 AI Safety

AI+Science and Technology

01 Life Science
02 Materials Science
03 Energy Science
04 Neuroscience
05 Science Computing
06 Agentic Science

AI+Humanities and Social Sciences

01 Education
02 Economy and Finance
03 Governance and Law
04 AI Ethics

Progress & Momentum

600+
PhD Students

19
Joint Laboratories & Research Centers

120+
Research Projects

Address: Area C, Haidian Dayue Information Technology Park, No. 17, Second Ring Road, Danufang, Haidian District, Beijing, China
Website: www.bza.edu.cn
www.zgci.ac.cn

Zhongguancun Academy

Zhongguancun Institute of Artificial Intelligence

JPMorganChase

CHIEF ANALYTICS OFFICE

Meet leaders from the Chief Analytics Office! Visit us to learn how our diverse teams are delivering AI/ML solutions to transform the finance industry.

We're looking for curious problem-solvers with a passion for developing innovative AI & ML solutions.

Stop by our booth!

Speak

Build the world's best AI language tutor at Speak

We're reinventing how people learn to speak a new language! We teach real conversation with correct pronunciation, adaptive lessons, and fluency tracking for millions of learners across 6+ languages.

Come say hi at our booth!

Open roles at speak.com/careers

ARC

AI Training Data Scanning Solutions

Turn Paper Into AI Training Data in Days

High Velocity Data Capture for AI

- Daily structured data delivery to cloud environment
- Fast ramp up, new scan centers live in 60 days
- Historical data digitization for LLM training
- Multiple formats & complex media capabilities

Built for AI Data Pipelines

- 40+ secure scan centers across North America
- 1.5 million books per month capacity
- 10M+ images per day processing capability

Enterprise Grade Security

- ISO/IEC 27001 certified processes
- SOC 2 compliant workflows
- HIPAA secure facilities

e-arc.com | (833) 272-8880

Babelscape®

Language generation, semantically grounded

Chat Minerva

The First Italian-centric Large Language Model

https://chatminerva.org

Find out more about our technologies

DATAmundi | Proud Sponsor of ACL 2026

Build AI with Multilingual Human Data Solutions

DATAmundi enables global AI teams to train, fine-tune, and evaluate models with high quality data across the full data lifecycle.

200+ Languages

270k Expert AI Contributors

- Ethical & human-led
- Domain expertise
- All data types

DATAmundi

translated.

$100,000

to fund language technology innovators who share the goal of making it easier for everyone to understand and be understood by all others.

Find out more.

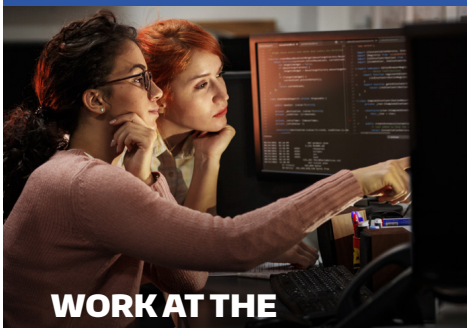
NEC

\Orchestrating a brighter world

Solving real-world challenges with core and applied AI research

Reliable AI | Human-Centric AI | AI4Science
Biomedical AI | Cyber Security

#NECLabs
neclab.eu



WORK AT THE INTERSECTION OF LANGUAGE AND TECHNOLOGY

Master of Science in Computational Linguistics

LEARN MORE AT
COMPLING.UW.EDU

UNIVERSITY of WASHINGTON

Adobe

Join us in empowering everyone to create.

An ideal balance of academic discovery and industry impact: that's what drives our groundbreaking work at Adobe Research. Our collaborative projects advance the state of the art across a range of 12 research areas. Many of Adobe Research's early-stage technologies become important elements of Adobe's products, used by millions across the globe.

Interested in future career opportunities?

Join the Adobe ACL network!

Adobe Research Areas

Artificial Intelligence & Machine Learning
AR, VR & 360 Photography
Audio
Computer Vision, Imaging & Video
Content Intelligence
Data Intelligence
Document Intelligence
Graphics (2D & 3D)
Human Computer Interaction
Intelligent Agents & Assistants
Natural Language Processing
Systems & Languages

Visit **research.adobe.com** *for more information.*

Copyright © 2025 Adobe. All rights reserved.

Thank You to Our Sponsors

DIAMOND

Bloomberg

PLATINUM

ORACLE

Appen

Tencent

Google

Meta

GOLD

JPMorganChase

SILVER

Adobe

BRONZE