NAACL 2025

# Annual Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics

## Proceedings of the Conference
## Volume 1: Long Papers

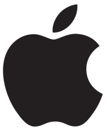
April 29 - May 4, 2025

The NAACL organizers gratefully acknowledge the support from the following sponsors.

### Platinum

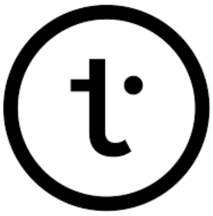
Megagon Labs

### Gold



Bloomberg
Engineering

### Bronze



## Diversity and Inclusion Ally

©2025 Association for Computational Linguistics

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
317 Sidney Baker St. S
Suite 400 - 134
Kerrville, TX 78028
USA
Tel: +1-855-225-1962
acl@aclweb.org

ISBN 979-8-89176-189-6

# Message from the General Chair

**Message from the General Chair** Welcome to the 2025 meeting of the Nations of the Americas Chapter of the Association for Computational Linguistics! I am proud to help organize the first NAACL conference to carry the new name of our organization, one that emphasizes inclusion for all of the Americas. I am also pleased to welcome you to Albuquerque, New Mexico, a state whose unique blend of cultural influences will make for an excellent backdrop for NAACL 2025, especially with this year's special theme on NLP in a Multicultural World.

This year's program benefited from a now mature ACL rolling review process. I would like to extend a big thank you to our ARR Editors in Chief, Viviane Moreira, Anna Rogers, and Michael White, who were very helpful not just with reviewing for the main conference, but who also shared their OpenReview expertise with chairs from our other tracks. We also benefited from the helpful advice of last year's NAACL general and program chairs: Katrin Erk, Kevin Duh, Helena Gomez, and Steven Bethard. Finally, Ryan Cotterell stepped up to help with the publications process and software, even while not serving as publications chair.

Of course, a conference of this magnitude cannot come together without some drama; in our case, we had some unexpected funding shortages from traditional government sources. We would like to extend a huge thank you to the boards of both the ACL and NAACL for filling those funding gaps, ensuring that our important D&I, volunteer, and student author support programs continue to thrive.

The virtual component of our conference is crucial to an inclusive, affordable experience for all NAACL authors and attendees. This year, we made small refinements to the hybrid format, mirroring NAACL 2024's choices to combine a virtual poster session via Gather with asynchronous online content via Underline. Our virtual poster session will be hosted on May 6, the Tuesday after the conference, in the hopes that promoting it at the conference's plenary sessions will help boost attendance. We opted not to have virtual oral presentations at the in-person conference, as those continue to be tricky to get right. To participants, virtual as well as in-person: Please let us know what worked for you and what did not, so we can continue to improve the hybrid experience.

The job of General Chair is a strange one, as it mostly involves cheering on many other people as they do amazing work. I have been fortunate to have been teamed with an excellent set of program chairs; to Luis Chiruzzo, Alan Ritter, and Lu Wang: thanks for everything, I'm very proud of what we've built together. I'd also like to extend my heartfelt thanks to Jenn Rachford (ACL) and Damira Mršic (Underline) who provide the knowledge, continuity and professionalism to bring all of this together.

Many thanks also to:

- Workshop chairs: Saab Mansour, Kenton Murray, and Alexis Palmer
- Tutorial chairs: Maria Lomeli, Swabha Swayamdipta, and Rui Zhang
- Demo chairs: Nouha Dziri, Shizhe Diao, and Sean (Xiang) Ren
- Industry track chairs: Weizhu Chen, Xue-Yong Fu, Mohammad Kachuee, and Yi Yang
- Student research workshop chairs: Abteen Ebrahimi, Emmy Liu, and Samar Haider, and their faculty advisors Maria Leonor Pacheco and Shira Wein
- Publication chairs: Arman Cohan, Manling Li, and Yichao Zhou
- Website chairs: Arya McCarthy and Vered Shwartz

- Publicity and social media chairs: Eleftheria Briakou, Tuhin Chakrabarty, and Ximena Gutierrez-Vasques
- Diversity and inclusion chairs: Akiko I. Eriguchi, Chi-Kiu (Jackie) Lo, and Niloofar Mireshghallah
- Sponsorship chairs: Prithviraj (Raj) Ammanabrolu and Maha Elbayad
- Volunteers chairs: Robin Jia and David Mortensen
- Ethics chairs: Manuel Mager and Yulia Tsvetkov
- Handbook chair: Winston Wu
- Best paper committee chairs: Marine Carpuat and Anna Rumshisky
- Visa chairs: Eduardo Blanco and Parisa Kordjamshidi
- Virtual infrastructure chair: Jieyu Zhao

Whenever possible, I tried to populate each committee with someone who had served in the same role in NAACL 2024, to provide continuity, so I'll extend an extra thanks to all chairs who accepted this second year of service. Thanks also to the members of the ACL and NAACL Executive Committees for their support, feedback, and advice.

Finally, I would like to thank all authors, invited speakers and panelists, area chairs and reviewers, volunteers and session chairs, and all attendees, in-person and virtual. The conference is nothing without you.

Welcome again and enjoy the conference!

**Colin Cherry**
Google
NAACL 2025 General Chair

# Message from the Program Chairs

**Message from the Program Chairs** Welcome to the 2025 Annual Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics! NAACL 2025 marks an exciting milestone as the first conference held under our new name, reflecting our commitment to greater inclusivity across the diverse communities of the Americas. NAACL 2025 is a hybrid conference, and we are excited to have attendees and presenters join us both in person in Albuquerque and online from around the globe. We are thrilled to welcome you to what promises to be a vibrant and engaging conference.

**Special Theme: NLP in a Multicultural World** Current NLP tools and models, particularly large language models (LLMs), rely on vast amounts of data for training. However, this data often over-represents a small number of dominant languages, and even within those, tends to prioritize certain geographical or cultural varieties. As a result, a long tail of under-represented languages, dialects, and cultural contexts remains largely overlooked by the NLP community. For NAACL 2025, we introduced a special theme track on "NLP in a Multicultural World." With this theme track, we sought to foster discussion and research on how NLP can better serve the linguistic and cultural diversity of the world. We encouraged contributions on topics such as cultural localization of language models, new NLP applications to support people from diverse cultures, revitalization or refunctionalization of endangered or sleeping languages, analysis of cultural biases in language models, and historical considerations and diachronic analysis. This track was dedicated to developing more inclusive, culturally aware NLP techniques that reflect and support the vibrant multicultural world we live in.

We received 71 submissions to the special theme, of which 23 were accepted for presentation at the conference. We hope these papers spark meaningful conversations and inspire future work in this important and evolving area of research.

**Review Process** 3,185 papers were submitted to the October ARR cycle, of which we estimate 3,099 were intended to be submitted to NAACL based on the "preferred venue" field in the submission form. We also received 147 papers from previous ARR cycles committed to NAACL. The program chairs recruited 98 Senior Area Chairs to view reviews and metareviews provided by ARR and make final recommendations on which papers to accept to both the main conference and Findings. 1,432 Area Chairs wrote metareviews for ARR, and 10,648 reviewers wrote reviews for the submitted papers.

**Acceptance Rate** Calculating an acceptance rate is challenging due to the multi-step ARR review process, in which papers are first submitted to ARR to get reviews, then authors commit their papers (together with reviews) to a specific *ACL conference. Of the 3,185 papers submitted to the October ARR cycle, we estimate that 3,099 intended to submit to NAACL. Based on this information, we estimate that 22% of papers submitted to the October cycle and intended for NAACL were accepted to the main conference, and another 15% were accepted to Findings, bringing the total estimated acceptance rate for papers accepted to be presented at the conference (Main + Findings) to 37%. Out of the 1,647 papers committed to NAACL with ARR reviews, 719 were accepted to the main conference, and 477 were accepted to Findings. 40 papers were desk rejected or withdrawn.

**Presentation Format** At NAACL 2025, papers were assigned one of three possible presentation modes: in-person participants could be assigned oral or poster presentations, while virtual participants could present posters. We selected 246 of the papers accepted to the main conference as oral presentations, and the rest of them were assigned as posters, together with all the Findings papers. Oral presentations were assigned a 15-minute slot, with 12 minutes for presentation and 3 minutes for questions. For choosing papers as oral presentations, we first split all papers according to track, sorted them according to overall score (considering review, metareview, SAC recommendation), and took into consideration the authors' presentation preference. Then we grouped papers in sets of 6. Some tracks had very few accepted papers, so some of them were grouped together to form areas of affinity.

**Program Format** NAACL 2025 consists both of an in-person and a virtual conference, held on different days. The virtual part of the conference is held after the in-person one and a few days later (on May 6), so participants traveling home after the in-person conference could attend the virtual conference. The conference program includes 3 keynote speakers: Rada Mihalcea (University of Michigan), Mike Lewis (Meta), and Josh Tenenbaum (Massachusetts Institute of Technology). 260 papers are scheduled to be presented as oral presentations (also including papers from TACL, CL, and the industry track), 594 papers are scheduled as in-person posters, and 256 virtual posters.

**Gratitude** NAACL 2025 would not have been possible without the hard work of all people involved. We thank everyone who contributed, including:

- The General Chair, Colin Cherry.
- The ARR Editors-in-Chief of the October 2024 cycle: Viviane Moreira, Anna Rogers, Michael White.
- The OpenReview team, especially Rachel Smart.
- The 98 Senior Area Chairs.
- The 1,432 Area Chairs and 10,648 Reviewers.
- The best paper committee chairs, Marine Carpuat and Anna Rumshisky.
- The ethics chairs, Yulia Tsvetkov and Manuel Mager, and their team of reviewers.
- The website chairs, Vered Shwartz and Arya McCarthy.
- The publication chairs, Yichao Zhou, Manling Li, and Arman Cohan.
- The publicity chairs, Ximena Gutierrez-Vasques, Eleftheria Briakou, and Tuhin Chakrabarty.
- The volunteers chairs, Robin Jia and David Mortensen.
- The visa chairs, Eduardo Blanco and Parisa Kordjamshidi.
- The ACL Anthology Director, Matt Post, and his team.
- The Program Chairs of EMNLP 2024 (Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung (Vivian) Chen) and NAACL 2024 (Kevin Duh, Helena Gomez, and Steven Bethard).
- Damira Mršic and the Underline Team.
- Jenn Rachhford and the entire conference support staff.

**Luis Chiruzzo**, Universidad de la República, Uruguay
**Alan Ritter**, Georgia Institute of Technology
**Lu Wang**, University of Michigan
NAACL 2025 Program Committee Co-Chairs
April 2025

# Table of Contents

*Understanding Figurative Meaning through Explainable Visual Entailment*
Arkadiy Saakyan, Shreyas Kulkarni, Tuhin Chakrabarty and Smaranda Muresan . . . . . . . . . . . . . . 1

*Benchmarking Distributional Alignment of Large Language Models*
Nicole Meister, Carlos Guestrin and Tatsunori Hashimoto . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 24

*World Models with Hints of Large Language Models for Goal Achieving*
Zeyuan Liu, Ziyu Huan, Xiyao Wang, Jiafei Lyu, Jian Tao, Xiu Li, Furong Huang and Huazhe Xu 50

*CogLM: Tracking Cognitive Development of Large Language Models*
Xinglin Wang, Peiwen Yuan, Shaoxiong Feng, Yiwei Li, Boyuan Pan, Heda Wang, Yao Hu and Kan Li . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 73

*Improving and Assessing the Fidelity of Large Language Models Alignment to Online Communities*
Minh Duc Chu, Zihao He, Rebecca Dorn and Kristina Lerman . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 88

*Improving Retrospective Language Agents via Joint Policy Gradient Optimization*
Xueyang Feng, bo Lan, Quanyu Dai, Lei Wang, Jiakai Tang, Xu Chen, Zhenhua Dong and Ji-Rong Wen . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 112

*CodexGraph: Bridging Large Language Models and Code Repositories via Code Graph Databases*
Xiangyan Liu, Bo Lan, Zhiyuan Hu, Yang Liu, Zhicheng Zhang, Fei Wang, Michael Qizhe Shieh and Wenmeng Zhou . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 142

*Instantly Learning Preference Alignment via In-context DPO*
Feifan Song, Yuxuan Fan, Xin Zhang, Peiyi Wang and Houfeng Wang . . . . . . . . . . . . . . . . . . . . . 161

*ALTER: Augmentation for Large-Table-Based Reasoning*
Han Zhang, Yuheng Ma and Hanfang Yang . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 179

*What the #?*!: Disentangling Hate Across Target Identities*
Yiping Jin, Leo Wanner and Aneesh Moideen Koya . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 199

*MAD Speech: Measures of Acoustic Diversity of Speech*
Matthieu Futeral, Andrea Agostinelli, Marco Tagliasacchi, Neil Zeghidour and Eugene Kharitonov 222

*The Russian-focused embedders' exploration: ruMTEB benchmark and Russian embedding model design*
Artem Snegirev, Maria Tikhonova, Maksimova Anna, Alena Fenogenova and Aleksandr Abramov 236

*PRACTIQ: A Practical Conversational Text-to-SQL dataset with Ambiguous and Unanswerable Queries*
Mingwen Dong, Nischal Ashok Kumar, Yiqun Hu, Anuj Chauhan, Chung-Wei Hang, Shuaichen Chang, Lin Pan, Wuwei Lan, Henghui Zhu, Jiarong Jiang, Patrick Ng and Zhiguo Wang . . . . . . . . . . 255

*MIRAGE-Bench: Automatic Multilingual Benchmark Arena for Retrieval-Augmented Generation Systems*
Nandan Thakur, Suleman Kazi, Ge Luo, Jimmy Lin and Amin Ahmad . . . . . . . . . . . . . . . . . . . 274

*LLMs Are Biased Towards Output Formats! Systematically Evaluating and Mitigating Output Format Bias of LLMs*
Do Xuan Long, Ngoc-Hai Nguyen, Tiviatis Sim, Hieu Dao, Shafiq Joty, Kenji Kawaguchi, Nancy F. Chen and Min-Yen Kan . . . 299

*The Impact of Visual Information in Chinese Characters: Evaluating Large Models' Ability to Recognize and Utilize Radicals*
Xiaofeng Wu, Karl Stratos and Wei Xu . . . 331

*PromptRefine: Enhancing Few-Shot Performance on Low-Resource Indic Languages with Example Selection from related Example Banks*
Soumya Suvra Ghosal, Soumyabrata Pal, Koyel Mukherjee and Dinesh Manocha . . . 351

*Unlocking Decoding-time Controllability: Gradient-Free Multi-Objective Alignment with Contrastive Prompts*
Tingchen Fu, Yupeng Hou, Julian McAuley and Rui Yan . . . 366

*Fingerspelling within Sign Language Translation*
Garrett Tanzer . . . 385

*MoDS: Moderating a Mixture of Document Speakers to Summarize Debatable Queries in Document Collections*
Nishant Balepur, Alexa Siu, Nedim Lipka, Franck Dernoncourt, Tong Sun, Jordan Lee Boyd-Graber and Puneet Mathur . . . 465

*Aligning Sentence Simplification with ESL Learner's Proficiency for Language Acquisition*
Guanlin Li, Yuki Arase and Noel Crespi . . . 492

*PeerQA: A Scientific Question Answering Dataset from Peer Reviews*
Tim Baumgärtner, Ted Briscoe and Iryna Gurevych . . . 508

*ALiiCE: Evaluating Positional Fine-grained Citation Generation*
Yilong Xu, Jinhua Gao, Xiaoming Yu, Baolong Bi, Huawei Shen and Xueqi Cheng . . . 545

*An LLM-Based Approach for Insight Generation in Data Analysis*
Alberto Sánchez Pérez, Alaa Boukhary, Paolo Papotti, Luis Castejón Lozano and Adam Elwood 562

*WebQuality: A Large-scale Multi-modal Web Page Quality Assessment Dataset with Multiple Scoring Dimensions*
Tao Zhang, Yige Wang, ZhuHangyu ZhuHangyu, Li Xin, Chen Xiang, Tian Hua Zhou and Jin Ma . . . 583

*UFO: A UI-Focused Agent for Windows OS Interaction*
Chaoyun Zhang, Liqun Li, Shilin He, Xu Zhang, Bo Qiao, Si Qin, Minghua Ma, Yu Kang, Qingwei Lin, Saravan Rajmohan, Dongmei Zhang and Qi Zhang . . . 597

*Is your benchmark truly adversarial? AdvScore: Evaluating Human-Grounded Adversarialness*
Yoo Yeon Sung, Maharshi Gor, Eve Fleisig, Ishani Mondal and Jordan Lee Boyd-Graber . . . 623

*Fact-Aware Multimodal Retrieval Augmentation for Accurate Medical Radiology Report Generation*
Liwen Sun, James Jialun Zhao, Wenjing Han and Chenyan Xiong . . . 643

*On Behalf of the Stakeholders: Trends in NLP Model Interpretability in the Era of LLMs*
Nitay Calderon and Roi Reichart . . . 656

*Direct Preference Optimization of Video Large Multimodal Models from Language Model Reward*
Ruohong Zhang, Liangke Gui, Zhiqing Sun, Yihao Feng, Keyang Xu, Yuanhan Zhang, Di Fu, Chunyuan Li, Alexander G Hauptmann, Yonatan Bisk and Yiming Yang . . . . . . . . . . . . . . . . . . . . . . . . 694

*FlexiGPT: Pruning and Extending Large Language Models with Low-Rank Weight Sharing*
James Seale Smith, Chi-Heng Lin, Shikhar Tuli, Haris Jeelani, Shangqian Gao, Yilin Shen, Hongxia Jin and Yen-Chang Hsu . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 718

*Conformalized Answer Set Prediction for Knowledge Graph Embedding*
Yuqicheng Zhu, Nico Potyka, Jiarong Pan, Bo Xiong, Yunjie He, Evgeny Kharlamov and Steffen Staab . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 731

*Parameter-free and Accessible Prompt Learning to Enhance Adversarial Robustness for Pre-trained Vision-Language Models*
Xingran Zhou, Kun Yang, Changtao Miao, Bingyu Hu, Zhuoer Xu, Shiwen Cui, Changhua Meng and Dan Hong . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 751

*Fine-grained Fallacy Detection with Human Label Variation*
Alan Ramponi, Agnese Daffara and Sara Tonelli . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 762

*Does Liking Yellow Imply Driving a School Bus? Semantic Leakage in Language Models*
Hila Gonen, Terra Blevins, Alisa Liu, Luke Zettlemoyer and Noah A. Smith . . . . . . . . . . . . . . . 785

*SELFGOAL: Your Language Agents Already Know How to Achieve High-level Goals*
Ruihan Yang, Jiangjie Chen, Yikai Zhang, Siyu Yuan, Aili Chen, Kyle Richardson, Yanghua Xiao and Deqing Yang . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 799

*Familarity: Better Evaluation of Zero-Shot Named Entity Recognition by Quantifying Label Shifts in Synthetic Training Data*
Jonas Golde, Patrick Haller, Max Ploner, Fabio Barth, Nicolaas Paul Jedema and Alan Akbik820

*Learning to Summarize from LLM-generated Feedback*
Hwanjun Song, Taewon Yun, Yuho Lee, Jihwan Oh, Gihun Lee, Jason Cai and Hang Su . . . . . 835

*Hybrid Graphs for Table-and-Text based Question Answering using LLMs*
Ankush Agarwal, Chaitanya Devaguptapu and Ganesh S . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 858

*CFinBench: A Comprehensive Chinese Financial Benchmark for Large Language Models*
Ying Nie, Binwei Yan, Tianyu Guo, Hao Liu, Haoyu Wang, Wei He, Binfan Zheng, Weihao Wang, Qiang Li, Weijian Sun, Yunhe Wang and Dacheng Tao . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 876

*LLM-Based Explicit Models of Opponents for Multi-Agent Games*
XiaoPeng Yu, Wanpeng Zhang and Zongqing Lu . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 892

*SeqAR: Jailbreak LLMs with Sequential Auto-Generated Characters*
Yan Yang, Zeguan Xiao, Xin Lu, Hongru Wang, Xuetao Wei, Hailiang Huang, Guanhua Chen and Yun Chen . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 912

*JMMMU: A Japanese Massive Multi-discipline Multimodal Understanding Benchmark for Culture-aware Evaluation*
Shota Onohara, Atsuyuki Miyai, Yuki Imajuku, Kazuki Egashira, Jeonghun Baek, Xiang Yue, Graham Neubig and Kiyoharu Aizawa . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 932

*EASYTOOL: Enhancing LLM-based Agents with Concise Tool Instruction*
Siyu Yuan, Kaitao Song, Jiangjie Chen, Xu Tan, Yongliang Shen, Kan Ren, Dongsheng Li and Deqing Yang . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 951

*Decoding Hate: Exploring Language Models' Reactions to Hate Speech*
Paloma Piot and Javier Parapar . . . . . . . . . . 973

*Babysit A Language Model From Scratch: Interactive Language Learning by Trials and Demonstrations*
Ziqiao Ma, Zekun Wang and Joyce Chai . . . . . . . . . . 991

*MoCE: Adaptive Mixture of Contextualization Experts for Byte-based Neural Machine Translation*
Langlin Huang, Mengyu Bu and Yang Feng . . . . . . . . . . 1011

*LLM-Human Pipeline for Cultural Grounding of Conversations*
Rajkumar Pujari and Dan Goldwasser . . . . . . . . . . 1029

*ACCESS : A Benchmark for Abstract Causal Event Discovery and Reasoning*
Vy Vo, Lizhen Qu, Tao Feng, Yuncheng Hua, Xiaoxi Kang, Songhai Fan, Tim Dwyer, Lay-Ki Soon and Gholamreza Haffari . . . . . . . . . . 1049

*Unmasking Implicit Bias: Evaluating Persona-Prompted LLM Responses in Power-Disparate Social Scenarios*
Bryan Chen Zhengyu Tan and Roy Ka-Wei Lee . . . . . . . . . . 1075

*GloCOM: A Short Text Neural Topic Model via Global Clustering Context*
Quang Duc Nguyen, Tung Nguyen, Duc Anh Nguyen, Linh Ngo Van, Sang Dinh and Thien Huu Nguyen . . . . . . . . . . 1109

*Reversed Attention: On The Gradient Descent Of Attention Layers In GPT*
Shahar Katz and Lior Wolf . . . . . . . . . . 1125

*Self-Harmonized Chain of Thought*
Ziqi Jin and Wei Lu . . . . . . . . . . 1153

*AnaScore: Understanding Semantic Parallelism in Proportional Analogies*
Liyan Wang, Haotong Wang and Yves Lepage . . . . . . . . . . 1175

*Generating Complex Question Decompositions in the Face of Distribution Shifts*
Kelvin Han and Claire Gardent . . . . . . . . . . 1189

*Diversify-verify-adapt: Efficient and Robust Retrieval-Augmented Ambiguous Question Answering*
Yeonjun In, Sungchul Kim, Ryan A. Rossi, Mehrab Tanjim, Tong Yu, Ritwik Sinha and Chanyoung Park . . . . . . . . . . 1212

*Unifying AI Tutor Evaluation: An Evaluation Taxonomy for Pedagogical Ability Assessment of LLM-Powered AI Tutors*
Kaushal Kumar Maurya, Kv Aditya Srivatsa, Kseniia Petukhova and Ekaterina Kochmar . . . 1234

*Where is the answer? An empirical study of positional bias for parametric knowledge extraction in language model*
Kuniaki Saito, Chen-Yu Lee, Kihyuk Sohn and Yoshitaka Ushiku . . . . . . . . . . 1252

*Evaluating Morphological Compositional Generalization in Large Language Models*
Mete Ismayilzada, Defne Circi, Jonne Sälevä, Hale Sirin, Abdullatif Köksal, Bhuwan Dhingra, Antoine Bosselut, Duygu Ataman and Lonneke Van Der Plas . . . . . . . . . . 1270

*Balancing Forget Quality and Model Utility: A Reverse KL-Divergence Knowledge Distillation Approach for Better Unlearning in LLMs*
Bichen Wang, Yuzhe Zi, Yixin Sun, Yanyan Zhao and Bing Qin . . . . . . . . . . 1306

*AgentMove: A Large Language Model based Agentic Framework for Zero-shot Next Location Prediction*
Jie Feng, Yuwei Du, Jie Zhao and Yong Li . . . . . . . . . . 1322

*Embedding derived animacy rankings offer insights into the sources of grammatical animacy*
Vivian G. Li . . . . . . . . . . 1339

*Generating Long-form Story Using Dynamic Hierarchical Outlining with Memory-Enhancement*
Qianyue Wang, Jinwu Hu, Zhengping Li, Yufeng Wang, Daiyuan li, Yu Hu and Mingkui Tan1352

*Little Giants: Synthesizing High-Quality Embedding Data at Scale*
Haonan Chen, Liang Wang, Nan Yang, Yutao Zhu, Ziliang Zhao, Furu Wei and Zhicheng Dou 1392

*Can LLMs Convert Graphs to Text-Attributed Graphs?*
Zehong Wang, Sidney Liu, Zheyuan Zhang, Tianyi Ma, Chuxu Zhang and Yanfang Ye . . . . . 1412

*Forest for the Trees: Overarching Prompting Evokes High-Level Reasoning in Large Language Models*
Haoran Liao, Shaohua Hu, Zhihao Zhu, Hao HE and Yaohui Jin . . . . . . . . . . 1433

*On the Role of Speech Data in Reducing Toxicity Detection Bias*
Samuel Bell, Mariano Coria Meglioli, Megan Richards, Eduardo Sánchez, Christophe Ropers, Skyler Wang, Adina Williams, Levent Sagun and Marta R. Costa-jussà . . . . . . . . . . 1454

*ITALIC: An Italian Culture-Aware Natural Language Benchmark*
Andrea Seveso, Daniele Potertì, Edoardo Federici, Mario Mezzanzanica and Fabio Mercorio1469

*RAP: A Metric for Balancing Repetition and Performance in Open-Source Large Language Models*
Donghao Huang, Thanh-Son Nguyen, Fiona Liausvia and Zhaoxia Wang . . . . . . . . . . 1479

*Improving Data Annotation for Low-Resource Relation Extraction with Logical Rule-Augmented Collaborative Language Models*
Xiyang Liu, Chunming Hu, Richong Zhang, Junfan Chen and Baowen Xu . . . . . . . . . . 1497

*CompAct: Compressed Activations for Memory-Efficient LLM Training*
Yara Shamshoum, Nitzan Hodos, Yuval Sieradzki and Assaf Schuster . . . . . . . . . . 1511

*Large Language Models Are Cross-Lingual Knowledge-Free Reasoners*
Peng Hu, Sizhe Liu, Changjiang Gao, Xin Huang, Xue Han, Junlan Feng, Chao Deng and Shujian Huang . . . . . . . . . . 1525

*What Did I Do Wrong? Quantifying LLMs' Sensitivity and Consistency to Prompt Engineering*
Federico Errica, Davide Sanvito, Giuseppe Siracusano and Roberto Bifulco . . . . . . . . . . 1543

*Detect, Disambiguate, and Translate: On-Demand Visual Reasoning for Multimodal Machine Translation with Large Vision-Language Models*
Danyang Liu, Fanjie Kong, Xiaohang Sun, Dhruva Patil, Avijit Vajpayee, Zhu Liu, Vimal Bhat and Najmeh Sadoughi . . . . . . . . . . 1559

*Mitigating Hallucinations in Multi-modal Large Language Models via Image Token Attention-Guided Decoding*
Xinhao Xu, Hui Chen, Mengyao Lyu, Sicheng Zhao, Yizhe Xiong, Zijia Lin, Jungong Han and Guiguang Ding . . . . . . . . . . 1571

*A Multi-modal Large Language Model with Graph-of-Thought for Effective Recommendation*
Zixuan Yi and Iadh Ounis . . . . . . . . . . 1591

*Investigating Human Values in Online Communities*
Nadav Borenstein, Arnav Arora, Lucie-Aimée Kaffee and Isabelle Augenstein . . . . . . . . . . 1607

*Pointwise Mutual Information as a Performance Gauge for Retrieval-Augmented Generation*
Tianyu Liu, Jirui Qi, Paul He, Arianna Bisazza, Mrinmaya Sachan and Ryan Cotterell . . . . . 1628

*MATO: A Model-Agnostic Training Optimization for Aspect Sentiment Triplet Extraction*
Shaopeng Tang, Lin Li, Xiaohui Tao, Leqi Zhong and Qing Xie . . . . . . . . . . . . . . . . . . . . . . . . . 1648

*Dynamic Data Mixing Maximizes Instruction Tuning for Mixture-of-Experts*
Tong Zhu, Daize Dong, Xiaoye Qu, Jiacheng Ruan, Wenliang Chen and Yu Cheng . . . . . . . . 1663

*EmoDynamiX: Emotional Support Dialogue Strategy Prediction by Modelling MiXed Emotions and Discourse Dynamics*
Chenwei Wan, Matthieu Labeau and Chloé Clavel . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 1678

*ReasVQA: Advancing VideoQA with Imperfect Reasoning Process*
Jianxin Liang, Xiaojun Meng, Huishuai Zhang, Yueqian Wang, Jiansheng Wei and Dongyan Zhao 1696

*Divergent Thoughts toward One Goal: LLM-based Multi-Agent Collaboration System for Electronic Design Automation*
Haoyuan Wu, Haisheng Zheng, Zhuolun He and Bei Yu . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 1710

*A Survey of QUD Models for Discourse Processing*
Yingxue Fu . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 1722

*SafetyQuizzer: Timely and Dynamic Evaluation on the Safety of LLMs*
Zhichao Shi, Shaoling Jing, Yi Cheng, Hao Zhang, Yuanzhuo Wang, Jie Zhang, Huawei Shen and Xueqi Cheng . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 1733

*Privacy Checklist: Privacy Violation Detection Grounding on Contextual Integrity Theory*
Haoran Li, Wei Fan, Yulin Chen, Cheng Jiayang, Tianshu Chu, Xuebing Zhou, Peizhao Hu and Yangqiu Song . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 1748

*Investigating the (De)Composition Capabilities of Large Language Models in Natural-to-Formal Language Conversion*
Ziyao Xu and Houfeng Wang . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 1767

*Stealthy Jailbreak Attacks on Large Language Models via Benign Data Mirroring*
Honglin Mu, Han He, Yuxin Zhou, Yunlong Feng, Yang Xu, Libo Qin, Xiaoming Shi, Zeming Liu, Xudong Han, Qi Shi, Qingfu Zhu and Wanxiang Che . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 1784

*VividMed: Vision Language Model with Versatile Visual Grounding for Medicine*
Lingxiao Luo, Bingda Tang, Xuanzhong Chen, Rong Han and Ting Chen . . . . . . . . . . . . . . . . . 1800

*Mixture of Multimodal Adapters for Sentiment Analysis*
Kezhou Chen, Shuo Wang, Huixia Ben, Shengeng Tang and Yanbin Hao . . . . . . . . . . . . . . . . . 1822

*The Impact of Inference Acceleration on Bias of LLMs*
Elisabeth Kirsten, Ivan Habernal, Vedant Nanda and Muhammad Bilal Zafar . . . . . . . . . . . . . . 1834

*AfriHate: A Multilingual Collection of Hate Speech and Abusive Language Datasets for African Languages*
Shamsuddeen Hassan Muhammad, Idris Abdulmumin, Abinew Ali Ayele, David Ifeoluwa Adelani, Ibrahim Said Ahmad, Saminu Mohammad Aliyu, Paul Röttger, Abigail Oppong, Andiswa Bukula, Chiamaka Ijeoma Chukwuneke, Ebrahim Chekol Jibril, Elyas Abdi Ismail, Esubalew Alemneh, Hagos Tesfahun Gebremichael, Lukman Jibril Aliyu, Meriem Beloucif, Oumaima Hourrane, Rooweither Mabuya, Salomey Osei, Samuel Rutunda, Tadesse Destaw Belay, Tadesse Kebede Guge, Tesfa Tegegne Asfaw, Lilian Diana Awuor Wanzare, Nelson Odhiambo Onyango, Seid Muhie Yimam and Nedjma Ousidhoum . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 1854

*Revealing the Barriers of Language Agents in Planning*
Jian Xie, Kexun Zhang, Jiangjie Chen, Siyu Yuan, Kai Zhang, Yikai Zhang, Lei Li and Yanghua Xiao . . . . . 1872

*You Only Read Once (YORO): Learning to Internalize Database Knowledge for Text-to-SQL*
Hideo Kobayashi, Wuwei Lan, Peng Shi, Shuaichen Chang, Jiang Guo, Henghui Zhu, Zhiguo Wang and Patrick Ng . . . . . 1889

*Option Symbol Matters: Investigating and Mitigating Multiple-Choice Option Symbol Bias of Large Language Models*
Zhen Yang, Ping Jian and Chengzhi Li . . . . . 1902

*DAWN-ICL: Strategic Planning of Problem-solving Trajectories for Zero-Shot In-Context Learning*
Xinyu Tang, Xiaolei Wang, Xin Zhao and Ji-Rong Wen . . . . . 1918

*LLaSA: Large Language and Structured Data Assistant*
Yao Xu, Shizhu He, Jiabei Chen, ZengXiangrong ZengXiangrong, Bingning Wang, Guang Liu, Jun Zhao and Kang Liu . . . . . 1935

*Towards Efficient and Multifaceted Computer-assisted Pronunciation Training Leveraging Hierarchical Selective State Space Model and Decoupled Cross-entropy Loss*
Fu-An Chao and Berlin Chen . . . . . 1947

*Information-Guided Identification of Training Data Imprint in (Proprietary) Large Language Models*
Abhilasha Ravichander, Jillian Fisher, Taylor Sorensen, Ximing Lu, Maria Antoniak, Bill Yuchen Lin, Niloofar Mireshghallah, Chandra Bhagavatula and Yejin Choi . . . . . 1962

*An Interpretable and Crosslingual Method for Evaluating Second-Language Dialogues*
Rena Wei Gao, Xuetong Wu, Carsten Roever, Jing Wu, Long Lv, Jingxuan Wu and Jey Han Lau 1979

*From Allies to Adversaries: Manipulating LLM Tool-Calling through Adversarial Injection*
Rupeng Zhang, Haowei Wang, Junjie Wang, Mingyang Li, Yuekai Huang, Dandan Wang and Qing Wang . . . . . 2009

*COVE: COntext and VEracity prediction for out-of-context images*
Jonathan Tonglet, Gabriel Thiem and Iryna Gurevych . . . . . 2029

*Discourse-Driven Evaluation: Unveiling Factual Inconsistency in Long Document Summarization*
Yang Zhong and Diane Litman . . . . . 2050

*Language Models are Crossword Solvers*
Soumadeep Saha, Sutanoya Chakraborty, Saptarshi Saha and Utpal Garain . . . . . 2074

*WHoW: A Cross-domain Approach for Analysing Conversation Moderation*
Ming-Bin Chen, Lea Frermann and Jey Han Lau . . . . . 2091

*Uplifting Lower-Income Data: Strategies for Socioeconomic Perspective Shifts in Large Multi-modal Models*
Joan Nwatu, Oana Ignat and Rada Mihalcea . . . . . 2127

*MSc-SQL: Multi-Sample Critiquing Small Language Models For Text-To-SQL Translation*
Satya Krishna Gorti, Ilan Gofman, Zhaoyan Liu, Jiapeng Wu, Noël Vouitsis, Guangwei Yu, Jesse C. Cresswell and Rasa Hosseinzadeh . . . . . 2145

*Mitigating Heterogeneity among Factor Tensors via Lie Group Manifolds for Tensor Decomposition Based Temporal Knowledge Graph Embedding*
Jiang Li, Xiangdong Su and Guanglai Gao . . . . . 2161

*What Goes Into a LM Acceptability Judgment? Rethinking the Impact of Frequency and Length*
Lindia Tjuatja, Graham Neubig, Tal Linzen and Sophie Hao . . . . . . . . . . . . . . . . . . . . . . . . . . . . 2173

*WaveFM: A High-Fidelity and Efficient Vocoder Based on Flow Matching*
Tianze Luo, Xingchen Miao and Wenbo Duan . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 2187

*Analyzing and Evaluating Correlation Measures in NLG Meta-Evaluation*
Mingqi Gao, Xinyu Hu, Li Lin and Xiaojun Wan . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 2199

*Cascading Large Language Models for Salient Event Graph Generation*
Xingwei Tan, Yuxiang Zhou, Gabriele Pergola and Yulan He . . . . . . . . . . . . . . . . . . . . . . . . . . . 2223

*Token-Level Density-Based Uncertainty Quantification Methods for Eliciting Truthfulness of Large Language Models*
Artem Vazhentsev, Lyudmila Rvanova, Ivan Lazichny, Alexander Panchenko, Maxim Panov, Timothy Baldwin and Artem Shelmanov . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 2246

*How Can We Diagnose and Treat Bias in Large Language Models for Clinical Decision-Making?*
Kenza Benkirane, Jackie Kay and Maria Perez-Ortiz . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 2263

*From Redundancy to Relevance: Information Flow in LVLMs Across Reasoning Tasks*
Xiaofeng Zhang, Yihao Quan, Chen Shen, Xiaosong Yuan, Shaotian Yan, Liang Xie, Wenxiao Wang, Chaochen Gu, Hao Tang and Jieping Ye . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 2289

*Patent-CR: A Dataset for Patent Claim Revision*
Lekang Jiang, Pascal A. Scherz and Stefan Goetz . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 2300

*MergeME: Model Merging Techniques for Homogeneous and Heterogeneous MoEs*
Yuhang Zhou, Giannis Karamanolakis, Victor Soto, Anna Rumshisky, Mayank Kulkarni, Furong Huang, Wei Ai and Jianhua Lu . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 2315

*Fine-Tuned LLMs are Time Capsulesfor Tracking Societal Bias Through Books*
Sangmitra Madhusudan, Robert Morabito, Skye Reid, Nikta Gohari Sadr and Ali Emami . . . 2329

*Exploring the Cost-Effectiveness of Perspective Taking in Crowdsourcing Subjective Assessment: A Case Study of Toxicity Detection*
Xiaoni Duan, Zhuoyan Li, Chien-Ju Ho and Ming Yin . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 2359

*NormAd: A Framework for Measuring the Cultural Adaptability of Large Language Models*
Abhinav Sukumar Rao, Akhila Yerukola, Vishwa Shah, Katharina Reinecke and Maarten Sap 2373

*LiPO: Listwise Preference Optimization through Learning-to-Rank*
Tianqi Liu, Zhen Qin, Junru Wu, Jiaming Shen, Misha Khalman, Rishabh Joshi, Yao Zhao, Mohammad Saleh, Simon Baumgartner, Jialu Liu, Peter J Liu and Xuanhui Wang . . . . . . . . . . . . . . . . 2404

*Adaptive Prompting: Ad-hoc Prompt Composition for Social Bias Detection*
Maximilian Spliethöver, Tim Knebler, Fabian Fumagalli, Maximilian Muschalik, Barbara Hammer, Eyke Hüllermeier and Henning Wachsmuth . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 2421

*Enhancing Discriminative Representation in Similar Relation Clusters for Few-Shot Continual Relation Extraction*
Anh Duc Le, Nam Le Hai, Thanh Xuan Nguyen, Linh Ngo Van, Nguyen Thi Ngoc Diep, Sang Dinh and Thien Huu Nguyen . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 2450

*SymBa: Symbolic Backward Chaining for Structured Natural Language Reasoning*
Jinu Lee and Wonseok Hwang . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 2468

*MEDA: Dynamic KV Cache Allocation for Efficient Multimodal Long-Context Inference*
Zhongwei Wan, Hui Shen, Xin Wang, Che Liu, Zheda Mai and Mi Zhang . . . . . . . . . . . . . . . . 2485

*Language Models Largely Exhibit Human-like Constituent Ordering Preferences*
Ada Tur, Gaurav Kamath and Siva Reddy . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 2498

*SafeQuant: LLM Safety Analysis via Quantized Gradient Inspection*
Sindhu Padakandla, Sadbhavana Babar, Rathod Darshan D and Manohar Kaul . . . . . . . . . . . . 2522

*Exploring Large Language Models for Effective Rumor Detection on Social Media*
Yirong Zeng, Xiao Ding, Bibo Cai, Ting Liu and Bing Qin . . . . . . . . . . . . . . . . . . . . . . . . . . . . 2537

*No Simple Answer to Data Complexity: An Examination of Instance-Level Complexity Metrics for Classification Tasks*
Ryan A. Cook, John P. Lalor and Ahmed Abbasi . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 2553

*NLI under the Microscope: What Atomic Hypothesis Decomposition Reveals*
Neha Srikanth and Rachel Rudinger . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 2574

*HISTOIRESMORALES: A French Dataset for Assessing Moral Alignment*
Thibaud Leteno, Irina Proskurina, Antoine Gourru, Julien Velcin, Charlotte Laclau, Guillaume Metzler and Christophe Gravier . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 2590

*Leveraging Allophony in Self-Supervised Speech Models for Atypical Pronunciation Assessment*
Kwanghee Choi, Eunjung Yeo, Kalvin Chang, Shinji Watanabe and David R Mortensen . . . 2613

*SAPIENT: Mastering Multi-turn Conversational Recommendation with Strategic Planning and Monte Carlo Tree Search*
Hanwen Du, Bo Peng and Xia Ning . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 2629

*Reliability of Topic Modeling*
Kayla Schroeder and Zach Wood-Doughty . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 2649

*Style Transfer with Multi-iteration Preference Optimization*
Shuai Liu and Jonathan May . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 2663

*DTELS: Towards Dynamic Granularity of Timeline Summarization*
Chenlong Zhang, Tong Zhou, Pengfei Cao, Zhuoran Jin, Yubo Chen, Kang Liu and Jun Zhao2682

*ALERT: An LLM-powered Benchmark for Automatic Evaluation of Recommendation Explanations*
Yichuan Li, Xinyang Zhang, Chenwei Zhang, Mao Li, Tianyi Liu, Pei Chen, Yifan Gao, Kyumin Lee, Kaize Ding, Zhengyang Wang, Zhihan Zhang, Jingbo Shang, Xian Li and Trishul Chilimbi . 2704

*DETQUS: Decomposition-Enhanced Transformers for QUery-focused Summarization*
Yasir Khan, Xinlei Wu, Sangpil Youm, Justin Ho, Aryaan Mehboob Shaikh, Jairo Garciga, Rohan Sharma and Bonnie J Dorr . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 2720

*IrokoBench: A New Benchmark for African Languages in the Age of Large Language Models*
David Ifeoluwa Adelani, Jessica Ojo, Israel Abebe Azime, Jian Yun Zhuang, Jesujoba Oluwadara Alabi, Xuanli He, Millicent Ochieng, Sara Hooker, Andiswa Bukula, En-Shiun Annie Lee, Chiamaka Ijeoma Chukwuneke, Happy Buzaaba, Blessing Kudzaishe Sibanda, Godson Koffi Kalipe, Jonathan Mukiibi, Salomon Kabongo Kabenamualu, Foutse Yuehgoh, Mmasibidi Setaka, Lolwethu Ndolela, Nkiruka Odu, Rooweither Mabuya, Salomey Osei, Shamsuddeen Hassan Muhammad, Sokhar Samb, Tadesse Kebede Guge, Tombekai Vangoni Sherman and Pontus Stenetorp . . . . . . . . . . . . . . . . . . . . 2732

*The Impact of Domain-Specific Terminology on Machine Translation for Finance in European Languages*
Arturo Oncevay, Charese Smiley and Xiaomo Liu . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 2758

*Benchmarking Language Model Creativity: A Case Study on Code Generation*
Yining Lu, Dixuan Wang, Tianjian Li, Dongwei Jiang, Sanjeev Khudanpur, Meng Jiang and Daniel Khashabi . . . . . 2776

*Have LLMs Reopened the Pandora's Box of AI-Generated Fake News?*
Xinyu Wang, Wenbo Zhang, Sai Koneru, Hangzhi Guo, Bonam Mingole, S. Shyam Sundar, Sarah Rajtmajer and Amulya Yadav . . . . . 2795

*Probe-Free Low-Rank Activation Intervention*
Chonghe Jiang, Bao Nguyen, Anthony Man-Cho So and Viet Anh Nguyen . . . . . 2812

*FactTrack: Time-Aware World State Tracking in Story Outlines*
Zhiheng Lyu, Kevin Yang, Lingpeng Kong and Dan Klein . . . . . 2825

*A Bayesian Optimization Approach to Machine Translation Reranking*
Julius Cheng, Maike Züfle, Vilém Zouhar and Andreas Vlachos . . . . . 2849

*Multi-Conditional Ranking with Large Language Models*
Pouya Pezeshkpour and Estevam Hruschka . . . . . 2863

*ReGLA: Refining Gated Linear Attention*
Peng Lu, Ivan Kobyzev, Mehdi Rezagholizadeh, Boxing Chen and Philippe Langlais . . . . . 2884

*Intrinsic Bias is Predicted by Pretraining Data and Correlates with Downstream Performance in Vision-Language Encoders*
Kshitish Ghate, Isaac Slaughter, Kyra Wilson, Mona T. Diab and Aylin Caliskan . . . . . 2899

*Benchmarking Failures in Tool-Augmented Language Models*
Eduardo Treviño, Hugo Contant, James Ngai, Graham Neubig and Zora Zhiruo Wang . . . . . 2916

*Entity Decomposition with Filtering: A Zero-Shot Clinical Named Entity Recognition Framework*
Reza Averly and Xia Ning . . . . . 2935

*Towards Knowledge Checking in Retrieval-augmented Generation: A Representation Perspective*
Shenglai Zeng, Jiankun Zhang, Bingheng Li, Yuping Lin, Tianqi Zheng, Dante Everaert, Hanqing Lu, Hui Liu, Hui Liu, Yue Xing, Monica Xiao Cheng and Jiliang Tang . . . . . 2952

*The Power of Many: Multi-Agent Multimodal Models for Cultural Image Captioning*
Longju Bai, Angana Borah, Oana Ignat and Rada Mihalcea . . . . . 2970

*Prepending or Cross-Attention for Speech-to-Text? An Empirical Comparison*
Tsz Kin Lam, Marco Gaido, Sara Papi, Luisa Bentivogli and Barry Haddow . . . . . 2994

*CORRECT: Context- and Reference-Augmented Reasoning and Prompting for Fact-Checking*
Delvin Ce Zhang and Dongwon Lee . . . . . 3007

*Racing Thoughts: Explaining Contextualization Errors in Large Language Models*
Michael A. Lepori, Michael Curtis Mozer and Asma Ghandeharioun . . . . . 3020

*DREAM: Improving Video-Text Retrieval Through Relevance-Based Augmentation Using Large Foundation Models*
Yimu Wang, Shuai Yuan, Bo Xue, Xiangru Jian, Wei Pang, Mushi Wang and Ning Yu . . . . . 3037

*ToW: Thoughts of Words Improve Reasoning in Large Language Models*
Zhikun Xu, Ming Shen, Jacob Dineen, Zhaonan Li, Xiao Ye, Shijie Lu, Aswin Rrv, Chitta Baral and Ben Zhou . . . . . 3057

*A Probabilistic Framework for LLM Hallucination Detection via Belief Tree Propagation*
Bairu Hou, Yang Zhang, Jacob Andreas and Shiyu Chang . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 3076

*ERAS: Evaluating the Robustness of Chinese NLP Models to Morphological Garden Path Errors*
Qinchan Li and Sophie Hao . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 3100

*Superlatives in Context: Modeling the Implicit Semantics of Superlatives*
Valentina Pyatkin, Bonnie Webber, Ido Dagan and Reut Tsarfaty . . . . . . . . . . . . . . . . . . . . . . . . . 3112

*LLMs Are Not Intelligent Thinkers: Introducing Mathematical Topic Tree Benchmark for Comprehensive Evaluation of LLMs*
Arash Gholami Davoodi, Seyed Pouyan Mousavi Davoudi and Pouya Pezeshkpour . . . . . . . . 3127

*Specializing Large Language Models to Simulate Survey Response Distributions for Global Populations*
Yong Cao, Haijiang Liu, Arnav Arora, Isabelle Augenstein, Paul Röttger and Daniel Hershcovich 3141

*Representing Rule-based Chatbots with Transformers*
Dan Friedman, Abhishek Panigrahi and Danqi Chen . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 3155

*Incremental Sentence Processing Mechanisms in Autoregressive Transformer Language Models*
Michael Hanna and Aaron Mueller . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 3181

*Entangled Relations: Leveraging NLI and Meta-analysis to Enhance Biomedical Relation Extraction*
William P Hogan and Jingbo Shang . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 3204

*Multimodal Needle in a Haystack: Benchmarking Long-Context Capability of Multimodal Large Language Models*
Hengyi Wang, Haizhou Shi, Shiwei Tan, Weiyi Qin, Wenyuan Wang, Tunyu Zhang, Akshay Nambi, Tanuja Ganu and Hao Wang . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 3221

*WorldCuisines: A Massive-Scale Benchmark for Multilingual and Multicultural Visual Question Answering on Global Cuisines*
Genta Indra Winata, Frederikus Hudi, Patrick Amadeus Irawan, David Anugraha, Rifki Afina Putri, Wang Yutong, Adam Nohejl, Ubaidillah Ariq Prathama, Nedjma Ousidhoum, Afifa Amriani, Anar Rzayev, Anirban Das, Ashmari Pramodya, Aulia Adila, Bryan Wilie, Candy Olivia Mawalim, Cheng Ching Lam, Daud Abolade, Emmanuele Chersoni, Enrico Santus, Fariz Ikhwantri, Garry Kuwanto, Hanyang Zhao, Haryo Akbarianto Wibowo, Holy Lovenia, Jan Christian Blaise Cruz, Jan Wira Gotama Putra, Junho Myung, Lucky Susanto, Maria Angelica Riera Machin, Marina Zhukova, Michael Anugraha, Muhammad Farid Adilazuarda, Natasha Christabelle Santosa, Peerat Limkonchotiwat, Raj Dabre, Rio Alexander Audino, Samuel Cahyawijaya, Shi-Xiong Zhang, Stephanie Yulia Salim, Yi Zhou, Yinxuan Gui, David Ifeoluwa Adelani, En-Shiun Annie Lee, Shogo Okada, Ayu Purwarianti, Alham Fikri Aji, Taro Watanabe, Derry Tanti Wijaya, Alice Oh and Chong-Wah Ngo . . . . . . . . . . . . . . . . . . . . . . 3242

*Extracting and Understanding the Superficial Knowledge in Alignment*
Runjin Chen, Gabriel Jacob Perin, Xuxi Chen, Xilun Chen, Yan Han, Nina S. T. Hirata, Junyuan Hong and Bhavya Kailkhura . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 3265

*Smurfs: Multi-Agent System using Context-Efficient DFSDT for Tool Planning*
Junzhi Chen, Juhao Liang and Benyou Wang . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 3281

*From Introspection to Best Practices: Principled Analysis of Demonstrations in Multimodal In-Context Learning*
Nan Xu, Fei Wang, Sheng Zhang, Hoifung Poon and Muhao Chen . . . . . . . . . . . . . . . . . . . . . . . 3299

*Upsample or Upweight? Balanced Training on Heavily Imbalanced Datasets*
Tianjian Li, Haoran Xu, Weiting Tan, Kenton Murray and Daniel Khashabi . . . . . . . . . . . . . . . 3325

*LLM The Genius Paradox: A Linguistic and Math Expert's Struggle with Simple Word-based Counting Problems*
Nan Xu and Xuezhe Ma . . . . . 3344

*PAPILLON: Privacy Preservation from Internet-based and Local Language Model Ensembles*
Siyan Li, Vethavikashini Chithrra Raghuram, Omar Khattab, Julia Hirschberg and Zhou Yu 3371

*When2Call: When (not) to Call Tools*
Hayley Ross, Ameya Sunil Mahabaleshwarkar and Yoshi Suhara . . . . . 3391

*Mitigating Hallucinated Translations in Large Language Models with Hallucination-focused Preference Optimization*
Zilu Tang, Rajen Chatterjee and Sarthak Garg . . . . . 3410

*Large Language Models Can Solve Real-World Planning Rigorously with Formal Verification Tools*
Yilun Hao, Yongchao Chen, Yang Zhang and Chuchu Fan . . . . . 3434

*Who Relies More on World Knowledge and Bias for Syntactic Ambiguity Resolution: Humans or LLMs?*
So Young Lee, Russell Scheinberg, Amber Shore and Ameeta Agrawal . . . . . 3484

*Beyond Benchmarks: Building a Richer Cross-Document Event Coreference Dataset with Decontextualization*
Jin Zhao, Jingxuan Tu, Bingyang Ye, Xinrui Hu, Nianwen Xue and James Pustejovsky . . . . . 3499

*Can Unconfident LLM Annotations Be Used for Confident Conclusions?*
Kristina Gligoric, Tijana Zrnic, Cinoo Lee, Emmanuel Candes and Dan Jurafsky . . . . . 3514

*Beyond End-to-End VLMs: Leveraging Intermediate Text Representations for Superior Flowchart Understanding*
Junyi Ye, Ankan Dash, Wenpeng Yin and Guiling Wang . . . . . 3534

*Ihquin tlahtouah in Tetelahtzincocah: An annotated, multi-purpose audio and text corpus of Western Sierra Puebla Nahuatl*
Robert Pugh, Cheyenne Wing, María Ximena Juárez Huerta, Angeles Márquez Hernandez and Francis M. Tyers . . . . . 3549

*Benchmarking Large Language Models on Answering and Explaining Challenging Medical Questions*
Hanjie Chen, Zhouxiang Fang, Yash Singla and Mark Dredze . . . . . 3563

*Unfamiliar Finetuning Examples Control How Language Models Hallucinate*
Katie Kang, Eric Wallace, Claire Tomlin, Aviral Kumar and Sergey Levine . . . . . 3600

*Reasoning Aware Self-Consistency: Leveraging Reasoning Paths for Efficient LLM Sampling*
Guangya Wan, Yuqi Wu, Jie Chen and Sheng Li . . . . . 3613

MATVIX*: Multimodal Information Extraction from Visually Rich Articles*
Ghazal Khalighinejad, Sharon Scott, Ollie Liu, Kelly L. Anderson, Rickard Stureborg, Aman Tyagi and Bhuwan Dhingra . . . . . 3636

*Towards Rationality in Language and Multimodal Agents: A Survey*
Bowen Jiang, Yangxinyu Xie, Xiaomeng Wang, Yuan Yuan, Zhuoqun Hao, Xinyi Bai, Weijie J Su, Camillo Jose Taylor and Tanwi Mallick . . . . . 3656

*CluSanT: Differentially Private and Semantically Coherent Text Sanitization*
Ahmed Musa Awon, Yun Lu, Shera Potka and Alex Thomo . . . . . 3676

*TurkingBench: A Challenge Benchmark for Web Agents*
Kevin Xu, Yeganeh Kordi, Tanay Nayak, Adi Asija, Yizhong Wang, Kate Sanders, Adam Byerly, Jingyu Zhang, Benjamin Van Durme and Daniel Khashabi . . . . . . . . . . 3694

*CodeTree: Agent-guided Tree Search for Code Generation with Large Language Models*
Jierui Li, Hung Le, Yingbo Zhou, Caiming Xiong, Silvio Savarese and Doyen Sahoo . . . . . . 3711

*DPL: Diverse Preference Learning Without A Reference Model*
Abhijnan Nath, Andrey Volozin, Saumajit Saha, Albert Aristotle Nanda, Galina Grunin, Rahul Bhotika and Nikhil Krishnaswamy . . . . . . . . . . 3727

*Verifiable by Design: Aligning Language Models to Quote from Pre-Training Data*
Jingyu Zhang, Marc Marone, Tianjian Li, Benjamin Van Durme and Daniel Khashabi . . . . . 3748

*VoCoT: Unleashing Visually Grounded Multi-Step Reasoning in Large Multi-Modal Models*
Zejun Li, Ruipu Luo, Jiwen Zhang, Minghui Qiu, Xuanjing Huang and Zhongyu Wei . . . . . . 3769

*ACCORD: Closing the Commonsense Measurability Gap*
François Roewer-Després, Jinyue Feng, Zining Zhu and Frank Rudzicz . . . . . . . . . . 3799

*CRMArena: Understanding the Capacity of LLM Agents to Perform Professional CRM Tasks in Realistic Environments*
Kung-Hsiang Huang, Akshara Prabhakar, Sidharth Dhawan, Yixin Mao, Huan Wang, Silvio Savarese, Caiming Xiong, Philippe Laban and Chien-Sheng Wu . . . . . . . . . . 3830

*Mamba-Shedder: Post-Transformer Compression for Efficient Selective Structured State Space Models*
Juan Pablo Munoz, Jinjie Yuan and Nilesh Jain . . . . . . . . . . 3851

*CBT-Bench: Evaluating Large Language Models on Assisting Cognitive Behavior Therapy*
Mian Zhang, Xianjun Yang, Xinlu Zhang, Travis Labrum, Jamie C. Chiu, Shaun M. Eack, Fei Fang, William Yang Wang and Zhiyu Chen . . . . . . . . . . 3864

*An Efficient Gloss-Free Sign Language Translation Using Spatial Configurations and Motion Dynamics with LLMs*
Eui Jun Hwang, Sukmin Cho, Junmyeong Lee and Jong C. Park . . . . . . . . . . 3901

*Sketch2Code: Evaluating Vision-Language Models for Interactive Web Design Prototyping*
Ryan Li, Yanzhe Zhang and Diyi Yang . . . . . . . . . . 3921

*Design2Code: Benchmarking Multimodal Code Generation for Automated Front-End Engineering*
Chenglei Si, Yanzhe Zhang, Ryan Li, Zhengyuan Yang, Ruibo Liu and Diyi Yang . . . . . . . . . 3956

*Temporal-Aware Soft Prompt Tuning for Automatic Text Dating*
Hai Wang, Yuzhi Liang and Han Ren . . . . . . . . . . 3975

*Sparser Mixture-of-Adapters with Cross-Layer Generalization*
Ziyue Li and Tianyi Zhou . . . . . . . . . . 3988

*How to Align Multiple Signed Language Corpora for Better Sign-to-Sign Translations?*
Mert Inan, Yang Zhong, Vidya Ganesh and Malihe Alikhani . . . . . . . . . . 4003

*Communication Makes Perfect: Persuasion Dataset Construction via Multi-LLM Communication*
Weicheng Ma, Hefan Zhang, Ivory Yang, Shiyu Ji, Joice Chen, Farnoosh Hashemi, Shubham Mohole, Ethan Gearey, Michael Macy, Saeed Hassanpour and Soroush Vosoughi . . . . . . . . . . 4017

*Soft Prompting for Unlearning in Large Language Models*
Karuna Bhaila, Minh-Hao Van and Xintao Wu . . . . . . . . . . 4046

*Mutual-pairing Data Augmentation for Fewshot Continual Relation Extraction*
Nguyen Hoang Anh, Quyen Tran, Thanh Xuan Nguyen, Nguyen Thi Ngoc Diep, Linh Ngo Van, Thien Huu Nguyen and Trung Le . . . . . . . . . . 4057

*KMMLU: Measuring Massive Multitask Language Understanding in Korean*
Guijin Son, Hanwool Lee, Sungdong Kim, Seungone Kim, Niklas Muennighoff, Taekyoon Choi, Cheonbok Park, Kang Min Yoo and Stella Biderman . . . . . . . . . . 4076

*Protecting Privacy in Multimodal Large Language Models with MLLMU-Bench*
Zheyuan Liu, Guangyao Dou, Mengzhao Jia, Zhaoxuan Tan, Qingkai Zeng, Yongle Yuan and Meng Jiang . . . . . . . . . . 4105

*LLM4DistReconfig: A Fine-tuned Large Language Model for Power Distribution Network Reconfiguration*
Panayiotis Christou, Md. Zahidul Islam, Yuzhang Lin and Jingwei Xiong . . . . . . . . . . 4136

*WaterPool: A Language Model Watermark Mitigating Trade-Offs among Imperceptibility, Efficacy and Robustness*
Baizhou Huang and Xiaojun Wan . . . . . . . . . . 4156

*Tricking Retrievers with Influential Tokens: An Efficient Black-Box Corpus Poisoning Attack*
Cheng Wang, Yiwei Wang, Yujun Cai and Bryan Hooi . . . . . . . . . . 4183

*The Good, The Bad, and The Greedy: Evaluation of LLMs Should Not Ignore Non-Determinism*
Yifan Song, Guoyin Wang, Sujian Li and Bill Yuchen Lin . . . . . . . . . . 4195

*CVE-Bench: Benchmarking LLM-based Software Engineering Agent's Ability to Repair Real-World CVE Vulnerabilities*
Peiran Wang, Xiaogeng Liu and Chaowei Xiao . . . . . . . . . . 4207

*PROMPTEVALS: A Dataset of Assertions and Guardrails for Custom Production Large Language Model Pipelines*
Reya Vir, Shreya Shankar, Harrison Chase, William Hinthorn and Aditya Parameswaran . . . 4225

*ToolFlow: Boosting LLM Tool-Calling Through Natural and Coherent Dialogue Synthesis*
Zezhong Wang, Xingshan Zeng, Weiwen Liu, Liangyou Li, Yasheng Wang, Lifeng Shang, Xin Jiang, Qun Liu and Kam-Fai Wong . . . . . . . . . . 4246

*Fighting Spurious Correlations in Text Classification via a Causal Learning Perspective*
Yuqing Zhou and Ziwei Zhu . . . . . . . . . . 4264

*Knowledge-Aware Query Expansion with Large Language Models for Textual and Relational Retrieval*
Yu Xia, Junda Wu, Sungchul Kim, Tong Yu, Ryan A. Rossi, Haoliang Wang and Julian McAuley 4275

*SVD-LLM V2: Optimizing Singular Value Truncation for Large Language Model Compression*
Xin Wang, Samiul Alam, Zhongwei Wan, Hui Shen and Mi Zhang . . . . . . . . . . 4287

*AudioBench: A Universal Benchmark for Audio Large Language Models*
Bin Wang, Xunlong Zou, Geyu Lin, Shuo Sun, Zhuohan Liu, Wenyu Zhang, Zhengyuan Liu, AiTi Aw and Nancy F. Chen . . . . . . . . . . 4297

*Efficient Prompting for Continual Adaptation to Missing Modalities*
Zirun Guo, Shulei Wang, Wang Lin, Weicai Yan, Yangyang Wu and Tao Jin . . . . . . . . . . 4317

*Benchmarking and Building Zero-Shot Hindi Retrieval Model with Hindi-BEIR and NLLB-E5*
Arkadeep Acharya, Rudra Murthy, Vishwajeet Kumar and Jaydeep Sen . . . . . . . . . . 4328

*Retrieval, Reasoning, Re-ranking: A Context-Enriched Framework for Knowledge Graph Completion*
Muzhi Li, Cehao Yang, Chengjin Xu, Xuhui Jiang, Yiyan Qi, Jian Guo, Ho-fung Leung and Irwin King . . . . . 4349

*See-Saw Modality Balance: See Gradient, and Sew Impaired Vision-Language Balance to Mitigate Dominant Modality Bias*
Junehyoung Kwon, MiHyeon Kim, Eunju Lee, Juhwan Choi and YoungBin Kim . . . . . 4364

*Harnessing and Evaluating the Intrinsic Extrapolation Ability of Large Language Models for Vehicle Trajectory Prediction*
Jiawei Liu, Yanjiao Liu, Xun Gong, Tingting Wang, Hong Chen and Yunfeng hu . . . . . 4379

*Stronger Models are Not Always Stronger Teachers for Instruction Tuning*
Zhangchen Xu, Fengqing Jiang, Luyao Niu, Bill Yuchen Lin and Radha Poovendran . . . . . 4392

*Efficient and Effective Prompt Tuning via Prompt Decomposition and Compressed Outer Product*
Pengxiang Lan, Haoyu Xu, Enneng Yang, Yuliang Liang, Guibing Guo, Jianzhe Zhao and Xingwei Wang . . . . . 4406

*Threshold Filtering Packing for Supervised Fine-Tuning: Training Related Samples within Packs*
Jiancheng Dong, Lei Jiang, Wei Jin and Lu Cheng . . . . . 4422

*Transferable Post-training via Inverse Value Learning*
Xinyu Lu, Xueru Wen, Yaojie Lu, Bowen Yu, Hongyu Lin, Haiyang Yu, Le Sun, Xianpei Han and Yongbin Li . . . . . 4436

*FLEX: Expert-level False-Less EXecution Metric for Text-to-SQL Benchmark*
Heegyu Kim, Jeon Taeyang, SeungHwan Choi, Seungtaek Choi and Hyunsouk Cho . . . . . 4448

*AID: Adaptive Integration of Detectors for Safe AI with Language Models*
Xinran Wang, Enmao Diao, Qi Le, Jie Ding and Ali Anwar . . . . . 4476

*SSMLoRA: Enhancing Low-Rank Adaptation with State Space Model*
Jiayang Yu, Yihang Zhang, Bin Wang, Peiqin Lin, YongKang Liu and Shi Feng . . . . . 4493

*Sharpness-Aware Minimization for Topic Models with High-Quality Document Representations*
Tung Nguyen, Tue Le, Hoang Tran Vuong, Quang Duc Nguyen, Duc Anh Nguyen, Linh Ngo Van, Sang Dinh and Thien Huu Nguyen . . . . . 4507

*$C^2$: Scalable Auto-Feedback for LLM-based Chart Generation*
Woosung Koh, Janghan Yoon, MinHyung Lee, Youngjin Song, Jaegwan Cho, Jaehyun Kang, Taehyeon Kim, Se-Young Yun, Youngjae Yu and Bongshin Lee . . . . . 4525

*A Top-down Graph-based Tool for Modeling Classical Semantic Maps: A Case Study of Supplementary Adverbs*
Zhu Liu, Cunliang Kong, Ying Liu and Maosong Sun . . . . . 4567

*UniHGKR: Unified Instruction-aware Heterogeneous Knowledge Retrievers*
Dehai Min, Zhiyang Xu, Guilin Qi, Lifu Huang and Chenyu You . . . . . 4577

*Improving Model Evaluation using SMART Filtering of Benchmark Datasets*
Vipul Gupta, Candace Ross, David Pantoja, Rebecca J. Passonneau, Megan Ung and Adina Williams . . . . . 4595

*Entropy-Based Decoding for Retrieval-Augmented Large Language Models*
Zexuan Qiu, Zijing Ou, Bin Wu, Jingjing Li, Aiwei Liu and Irwin King . . . . . 4616

*What We Talk About When We Talk About LMs: Implicit Paradigm Shifts and the Ship of Language Models*
Shengqi Zhu and Jeffrey Rzeszotarski . . . . . 4628

*Diversity Helps Jailbreak Large Language Models*
Weiliang Zhao, Daniel Ben-Levi, Wei Hao, Junfeng Yang and Chengzhi Mao . . . . . 4647

*Constrained Decoding with Speculative Lookaheads*
Nishanth Sridhar Nakshatri, Shamik Roy, Rajarshi Das, Suthee Chaidaroon, Leonid Boytsov and Rashmi Gangadharaiah . . . . . 4681

*DyPCL: Dynamic Phoneme-level Contrastive Learning for Dysarthric Speech Recognition*
Wonjun Lee, Solee Im, Heejin Do, Yunsu Kim, Jungseul Ok and Gary Lee . . . . . 4701

*Revisiting Early Detection of Sexual Predators via Turn-level Optimization*
JinMyeong AN, Sangwon Ryu, Heejin Do, Yunsu Kim, Jungseul Ok and Gary Lee . . . . . 4713

*StyleTTS-ZS: Efficient High-Quality Zero-Shot Text-to-Speech Synthesis with Distilled Time-Varying Style Diffusion*
Yinghao Aaron Li, Xilin Jiang, Cong Han and Nima Mesgarani . . . . . 4725

*Fact, Fetch, and Reason: A Unified Evaluation of Retrieval-Augmented Generation*
Satyapriya Krishna, Kalpesh Krishna, Anhad Mohananey, Steven Schwarcz, Adam Stambler, Shyam Upadhyay and Manaal Faruqui . . . . . 4745

*ReachAgent: Enhancing Mobile Agent via Page Reaching and Operation*
Qinzhuo Wu, Wei Liu, Jian Luan and Bin Wang . . . . . 4760

*Learning to Solve Domain-Specific Calculation Problems with Knowledge-Intensive Programs Generator*
Chengyuan Liu, Shihang Wang, Lizhi Qing, Jun Lin, Ji Zhang, Fei Wu and Kun Kuang . . . . . 4776

*SLIM: Let LLM Learn More and Forget Less with Soft LoRA and Identity Mixture*
Jiayi Han, Liang Du, Hongwei Du, Xiangguo Zhou, Yiwen Wu, Yuanfang Zhang, Weibo Zheng and Donghong Han . . . . . 4792

*MMEvalPro: Calibrating Multimodal Benchmarks Towards Trustworthy and Efficient Evaluation*
Jinsheng Huang, Liang Chen, Taian Guo, Fu Zeng, Yusheng Zhao, Bohan Wu, Ye Yuan, Haozhe Zhao, Zhihui Guo, Yichi Zhang, Jingyang Yuan, Wei Ju, Luchen Liu, Tianyu Liu, Baobao Chang and Ming Zhang . . . . . 4805

*MiLoRA: Harnessing Minor Singular Components for Parameter-Efficient LLM Finetuning*
Hanqing Wang, Yixia Li, Shuo Wang, Guanhua Chen and Yun Chen . . . . . 4823

*Analyzing (In)Abilities of SAEs via Formal Languages*
Abhinav Menon, Manish Shrivastava, David Krueger and Ekdeep Singh Lubana . . . . . 4837

*Multimodal Cognitive Reframing Therapy via Multi-hop Psychotherapeutic Reasoning*
Subin Kim, Hoonrae Kim, Heejin Do and Gary Lee . . . . . 4863

*Explanation based In-Context Demonstrations Retrieval for Multilingual Grammatical Error Correction*
Wei Li, Wen Luo, Guangyue Peng and Houfeng Wang . . . . . 4881

*A Unified Supervised and Unsupervised Dialogue Topic Segmentation Framework Based on Utterance Pair Modeling*
Shihao Yang, Ziyi Zhang, Yue Jiang, Chunsheng Qin and Shuhua Liu . . . . . 4898

*Evaluating Small Language Models for News Summarization: Implications and Factors Influencing Performance*
Borui Xu, Yao Chen, Zeyi Wen, Weiguo Liu and Bingsheng He . . . . . . . . . . . . . . . . . . . . . . . . 4909

*Dynamic Fisher-weighted Model Merging via Bayesian Optimization*
Sanwoo Lee, Jiahao Liu, Qifan Wang, Jingang Wang, Xunliang Cai and Yunfang Wu . . . . . . 4923

*AI-Assisted Human Evaluation of Machine Translation*
Vilém Zouhar, Tom Kocmi and Mrinmaya Sachan . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 4936

*MLLM-Bench: Evaluating Multimodal LLMs with Per-sample Criteria*
Wentao Ge, Shunian Chen, Hardy Chen, Nuo Chen, Junying Chen, Zhihong Chen, Wenya Xie, Shuo Yan, ChenghaoZhu ChenghaoZhu, Ziyue Lin, Dingjie Song, Xidong Wang, Anningzhe Gao, Zhang Zhiyi, Jianquan Li, Xiang Wan and Benyou Wang . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 4951

*AgentSense: Benchmarking Social Intelligence of Language Agents through Interactive Scenarios*
Xinyi Mou, Jingcong Liang, Jiayu Lin, Xinnong Zhang, Xiawei Liu, Shiyue Yang, Rong Ye, Lei Chen, Haoyu Kuang, Xuanjing Huang and Zhongyu Wei . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 4975

*FactCG: Enhancing Fact Checkers with Graph-Based Multi-Hop Data*
Deren Lei, Yaxi Li, Siyao Li, Mengya Hu, Rui Xu, Ken Archer, Mingyu Wang, Emily Ching and Alex Deng . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 5002

*Label Drop for Multi-Aspect Relation Modeling in Universal Information Extraction*
lu Yang, Jiajia Li, En Ci, Lefei Zhang, Zuchao Li and Ping Wang . . . . . . . . . . . . . . . . . . . . . . . . 5021

*Test-Time Code-Switching for Cross-lingual Aspect Sentiment Triplet Extraction*
Dongming Sheng, Kexin Han, Hao Li, Yan Zhang, Yucheng Huang, Jun Lang and Wenqiang Liu 5041

*VisCGEC: Benchmarking the Visual Chinese Grammatical Error Correction*
Xiaoman Wang, Dan Yuan, Xin Liu, Yike Zhao, Xiaoxiao Zhang, Xizhi Chen and Yunshi Lan 5054

*Are We Done with MMLU?*
Aryo Pradipta Gema, Joshua Ong Jun Leang, Giwon Hong, Alessio Devoto, Alberto Carlo Maria Mancino, Rohit Saxena, Xuanli He, Yu Zhao, Xiaotang Du, Mohammad Reza Ghasemi Madani, Claire Barale, Robert McHardy, Joshua Harris, Jean Kaddour, Emile Van Krieken and Pasquale Minervini 5069

*MeNTi: Bridging Medical Calculator and LLM Agent with Nested Tool Calling*
Yakun Zhu, Shaohang Wei, Xu Wang, Kui Xue, Shaoting Zhang and Xiaofan Zhang . . . . . . . 5097

*Steering Knowledge Selection Behaviours in LLMs via SAE-Based Representation Engineering*
Yu Zhao, Alessio Devoto, Giwon Hong, Xiaotang Du, Aryo Pradipta Gema, Hongru Wang, Xuanli He, Kam-Fai Wong and Pasquale Minervini . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 5117

*MoDification: Mixture of Depths Made Easy*
Chen Zhang, Meizhi Zhong, Qimeng Wang, Xuantao Lu, Zheyu Ye, Chengqiang Lu, Yan Gao, Yao Hu, Kehai Chen, Min Zhang and Dawei Song . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 5137

*On the Vulnerability of Text Sanitization*
Meng Tong, Kejiang Chen, Xiaojian Yuan, Jiayang Liu, Weiming Zhang, Nenghai Yu and Jie Zhang . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 5150

*Multilingual Needle in a Haystack: Investigating Long-Context Behavior of Multilingual Large Language Models*
Amey Hengle, Prasoon Bajpai, Soham Dan and Tanmoy Chakraborty . . . . . . . . . . . . . . . . . . . . 5165

*Verify-in-the-Graph: Entity Disambiguation Enhancement for Complex Claim Verification with Interactive Graph Representation*
Hoang Pham, Thanh-Do Nguyen and Khac-Hoai Nam Bui . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 5181

*Exploring the Potential of Large Language Models for Heterophilic Graphs*
Yuxia Wu, Shujie Li, Yuan Fang and Chuan Shi . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 5198

*Exploiting Edited Large Language Models as General Scientific Optimizers*
Qitan Lv, Tianyu Liu and Hong Wang . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 5212

*DIRAS: Efficient LLM Annotation of Document Relevance for Retrieval Augmented Generation*
Jingwei Ni, Tobias Schimanski, Meihong Lin, Mrinmaya Sachan, Elliott Ash and Markus Leippold . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 5238

*Hello Again! LLM-powered Personalized Agent for Long-term Dialogue*
Hao Li, Chenghao Yang, An Zhang, Yang Deng, Xiang Wang and Tat-Seng Chua . . . . . . . . . 5259

*My LLM might Mimic AAE - But When Should It?*
Sandra Camille Sandoval, Christabel Acquaye, Kwesi Adu Cobbina, Mohammad Nayeem Teli and Hal Daumé Iii . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 5277

*High-Dimension Human Value Representation in Large Language Models*
Samuel Cahyawijaya, Delong Chen, Yejin Bang, Leila Khalatbari, Bryan Wilie, Ziwei Ji, Etsuko Ishii and Pascale Fung . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 5303

*Not all Hallucinations are Good to Throw Away When it Comes to Legal Abstractive Summarization*
Nihed Bendahman, Karen Pinel-Sauvagnat, Gilles Hubert and Mokhtar Boumedyen Billami5331

*Query-focused Referentiability Learning for Zero-shot Retrieval*
Jaeyoung Kim, Dohyeon Lee and Seung-won Hwang . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 5345

*A Novel Computational Modeling Foundation for Automatic Coherence Assessment*
Aviya Maimon . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 5359

*Token-based Decision Criteria Are Suboptimal in In-context Learning*
Hakaze Cho, Yoshihiro Sakai, Mariko Kato, Kenshiro Tanaka, Akira Ishii and Naoya Inoue 5378

*CSEval: Towards Automated, Multi-Dimensional, and Reference-Free Counterspeech Evaluation using Auto-Calibrated LLMs*
Amey Hengle, Aswini Kumar Padhi, Anil Bandhakavi and Tanmoy Chakraborty . . . . . . . . . . 5402

*Multilingual Machine Translation with Open Large Language Models at Practical Scale: An Empirical Study*
Menglong Cui, Pengzhi Gao, Wei Liu, Jian Luan and Bin Wang . . . . . . . . . . . . . . . . . . . . . . . . . 5420

*RAG LLMs are Not Safer: A Safety Analysis of Retrieval-Augmented Generation for Large Language Models*
Bang An, Shiyue Zhang and Mark Dredze . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 5444

*Evaluating Evidence Attribution in Generated Fact Checking Explanations*
Rui Xing, Timothy Baldwin and Jey Han Lau . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 5475

*ETHIC: Evaluating Large Language Models on Long-Context Tasks with High Information Coverage*
Taewhoo Lee, Chanwoong Yoon, Kyochul Jang, Donghyeon Lee, Minju Song, Hyunjae Kim and Jaewoo Kang . . . . . 5497

*Aggregation Artifacts in Subjective Tasks Collapse Large Language Models' Posteriors*
Georgios Chochlakis, Alexandros Potamianos, Kristina Lerman and Shrikanth Narayanan . . 5513

*Arabic Dataset for LLM Safeguard Evaluation*
Yasser Ashraf, Yuxia Wang, Bin Gu, Preslav Nakov and Timothy Baldwin . . . . . 5529

*Anticipating Future with Large Language Model for Simultaneous Machine Translation*
Siqi Ouyang, Oleksii Hrinchuk, Zhehuai Chen, Vitaly Lavrukhin, Jagadeesh Balam, Lei Li and Boris Ginsburg . . . . . 5547

*GuideLLM: Exploring LLM-Guided Conversation with Applications in Autobiography Interviewing*
Jinhao Duan, Xinyu Zhao, Zhuoxuan Zhang, Eunhye Grace Ko, Lily Boddy, Chenan Wang, Tianhao Li, Alexander Rasgon, Junyuan Hong, Min Kyung Lee, Chenxi Yuan, Qi Long, Ying Ding, Tianlong Chen and Kaidi Xu . . . . . 5558

*Fine-Tuning Large Language Models with Sequential Instructions*
Hanxu Hu, Simon Yu, Pinzhen Chen and Edoardo Ponti . . . . . 5589

*Diverse In-Context Example Selection After Decomposing Programs and Aligned Utterances Improves Semantic Parsing*
Mayank Kothyari, Sunita Sarawagi, Soumen Chakrabarti, Gaurav Arora and Srujana Merugu5611

*Elevating Legal LLM Responses: Harnessing Trainable Logical Structures and Semantic Knowledge with Legal Reasoning*
Rujing Yao, Yang Wu, Chenghao Wang, Jingwei Xiong, Fang Wang and Xiaozhong Liu . . . 5630

*Efficient One-shot Compression via Low-Rank Local Feature Distillation*
Yaya SY, Christophe Cerisara and Irina Illina . . . . . 5643

*Waste Not, Want Not; Recycled Gumbel Noise Improves Consistency in Natural Language Generation*
Damien De Mijolla, Hannan Saddiq and Kim Moore . . . . . 5662

*ConQRet: A New Benchmark for Fine-Grained Automatic Evaluation of Retrieval Augmented Computational Argumentation*
Kaustubh Dhole, Kai Shu and Eugene Agichtein . . . . . 5687

*SynthDetoxM: Modern LLMs are Few-Shot Parallel Detoxification Data Annotators*
Daniil Moskovskiy, Nikita Sushko, Sergey Pletenev, Elena Tutubalina and Alexander Panchenko 5714

*BEMEAE: Moving Beyond Exact Span Match for Event Argument Extraction*
Enfa Fane, Md Nayem Uddin, Oghenevovwe Ikumariegbe, Daniyal Kashif, Eduardo Blanco and Steven Corman . . . . . 5734

*uDistil-Whisper: Label-Free Data Filtering for Knowledge Distillation in Low-Data Regimes*
Abdul Waheed, Karima Kadaoui, Bhiksha Raj and Muhammad Abdul-Mageed . . . . . 5750

*Iterative Self-Tuning LLMs for Enhanced Jailbreaking Capabilities*
Chung-En Sun, Xiaodong Liu, Weiwei Yang, Tsui-Wei Weng, Hao Cheng, Aidan San, Michel Galley and Jianfeng Gao . . . . . 5768

*VoiceTextBlender: Augmenting Large Language Models with Speech Capabilities via Single-Stage Joint Speech-Text Supervised Fine-Tuning*
Yifan Peng, Krishna C Puvvada, Zhehuai Chen, Piotr Zelasko, He Huang, Kunal Dhawan, Ke Hu, Shinji Watanabe, Jagadeesh Balam and Boris Ginsburg . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 5787

*Rethinking Word Similarity: Semantic Similarity through Classification Confusion*
Kaitlyn Zhou, Haishan Gao, Sarah Li Chen, Dan Edelstein, Dan Jurafsky and Chen Shani . . 5803

*SUNAR: Semantic Uncertainty based Neighborhood Aware Retrieval for Complex QA*
Venktesh V, Mandeep Rathee and Avishek Anand . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 5818

*Do RAG Systems Cover What Matters? Evaluating and Optimizing Responses with Sub-Question Coverage*
Kaige Xie, Philippe Laban, Prafulla Kumar Choubey, Caiming Xiong and Chien-Sheng Wu 5836

*Stronger Universal and Transferable Attacks by Suppressing Refusals*
David Huang, Avidan Shah, Alexandre Araujo, David Wagner and Chawin Sitawarin . . . . . . 5850

*The BiGGen Bench: A Principled Benchmark for Fine-grained Evaluation of Language Models with Language Models*
Seungone Kim, Juyoung Suk, Ji Yong Cho, Shayne Longpre, Chaeeun Kim, Dongkeun Yoon, Guijin Son, Yejin Cho, Sheikh Shafayat, Jinheon Baek, Sue Hyun Park, Hyeonbin Hwang, Jinkyung Jo, Hyowon Cho, Haebin Shin, Seongyun Lee, Hanseok Oh, Noah Lee, Namgyu Ho, Se June Joo, Miyoung Ko, Yoonjoo Lee, Hyungjoo Chae, Jamin Shin, Joel Jang, Seonghyeon Ye, Bill Yuchen Lin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee and Minjoon Seo . . . . . . . . . . . . . . . . . 5877

*DreamSync: Aligning Text-to-Image Generation with Image Understanding Feedback*
Jiao Sun, Deqing Fu, Yushi Hu, Su Wang, Royi Rassin, Da-Cheng Juan, Dana Alon, Charles Herrmann, Sjoerd Van Steenkiste, Ranjay Krishna and Cyrus Rashtchian . . . . . . . . . . . . . . . . . . . . . . 5920

*Uncovering Bias in Large Vision-Language Models at Scale with Counterfactuals*
Phillip Howard, Kathleen C. Fraser, Anahita Bhiwandiwalla and Svetlana Kiritchenko . . . . . 5946

*AEGIS2.0: A Diverse AI Safety Dataset and Risks Taxonomy for Alignment of LLM Guardrails*
Shaona Ghosh, Prasoon Varshney, Makesh Narsimhan Sreedhar, Aishwarya Padmakumar, Traian Rebedea, Jibin Rajan Varghese and Christopher Parisien . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 5992

*UOREX: Towards Uncertainty-Aware Open Relation Extraction*
Rebii Jamal, Mounir Ourekouch and Mohammed Erradi . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 6027

*Hephaestus: Improving Fundamental Agent Capabilities of Large Language Models through Continual Pre-Training*
Yuchen Zhuang, Jingfeng Yang, Haoming Jiang, Xin Liu, Kewei Cheng, Sanket Lokegaonkar, Yifan Gao, Qing Ping, Tianyi Liu, Binxuan Huang, Zheng Li, Zhengyang Wang, Pei Chen, Ruijie Wang, Rongzhi Zhang, Nasser Zalmout, Priyanka Nigam, Bing Yin and Chao Zhang . . . . . . . . . . . 6041

*TinyThinker: Distilling Reasoning through Coarse-to-Fine Knowledge Internalization with Self-Reflection*
Shengmin Piao and Sanghyun Park . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 6069

*VisDoM: Multi-Document QA with Visually Rich Elements Using Multimodal Retrieval-Augmented Generation*
Manan Suri, Puneet Mathur, Franck Dernoncourt, Kanika Goswami, Ryan A. Rossi and Dinesh Manocha . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 6088

*VTechAGP: An Academic-to-General-Audience Text Paraphrase Dataset and Benchmark Models*
Ming Cheng, Jiaying Gong, Chenhan Yuan, William A Ingram, Edward Fox and Hoda Eldardiry 6110

*Large Language Models Share Representations of Latent Grammatical Concepts Across Typologically Diverse Languages*
Jannik Brinkmann, Chris Wendler, Christian Bartelt and Aaron Mueller . . . . . . . . . . . . . . . . . . 6131

*Examining and Adapting Time for Multilingual Classification via Mixture of Temporal Experts*
Weisi Liu, Guangzeng Han and Xiaolei Huang . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 6151

*FLEURS-ASL: Including American Sign Language in Massively Multilingual Multitask Evaluation*
Garrett Tanzer . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 6167

*EvoAgent: Towards Automatic Multi-Agent Generation via Evolutionary Algorithms*
Siyu Yuan, Kaitao Song, Jiangjie Chen, Xu Tan, Dongsheng Li and Deqing Yang . . . . . . . . . 6192

*EmoCharacter: Evaluating the Emotional Fidelity of Role-Playing Agents in Dialogues*
Qiming Feng, Qiujie Xie, Xiaolong Wang, Qingqiu Li, Yuejie Zhang, Rui Feng, Tao Zhang and Shang Gao . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 6218

*Language Models can Categorize System Inputs for Performance Analysis*
Dominic Sobhani, Ruiqi Zhong, Edison Marrese-Taylor, Keisuke Sakaguchi and Yutaka Matsuo 6241

*FinEval: A Chinese Financial Domain Knowledge Evaluation Benchmark for Large Language Models*
Xin Guo, Haotian Xia, Zhaowei Liu, Hanyang Cao, Zhi Yang, Zhiqiang Liu, Sizhe Wang, Jinyi Niu, Chuqi Wang, Yanhui Wang, Xiaolong Liang, Xiaoming Huang, Bing Zhu, Zhongyu Wei, Yun Chen, Weining Shen and Liwen Zhang . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 6258

*Rethinking the Role of LLMs for Document-level Relation Extraction: a Refiner with Task Distribution and Probability Fusion*
Fu Zhang, Xinlong Jin, Jingwei Cheng, Hongsen Yu and Huangming Xu . . . . . . . . . . . . . . . . 6293

*Decomposition Dilemmas: Does Claim Decomposition Boost or Burden Fact-Checking Performance?*
Qisheng Hu, Quanyu Long and Wenya Wang . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 6313

*Model Surgery: Modulating LLM's Behavior Via Simple Parameter Editing*
Huanqian Wang, Yang Yue, Rui Lu, Jingxin Shi, Andrew Zhao, Shenzhi Wang, Shiji Song and Gao Huang . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 6337

*Effective Skill Unlearning through Intervention and Abstention*
Yongce Li, Chung-En Sun and Tsui-Wei Weng . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 6358

*CharacterBox: Evaluating the Role-Playing Capabilities of LLMs in Text-Based Virtual Worlds*
Lei Wang, Jianxun Lian, Yi Huang, Yanqi Dai, Haoxuan Li, Xu Chen, Xing Xie and Ji-Rong Wen 6372

*A Cognitive Evaluation Benchmark of Image Reasoning and Description for Large Vision-Language Models*
Xiujie Song, Mengyue Wu, Kenny Q. Zhu, Chunhao Zhang and Yanyi Chen . . . . . . . . . . . . . . 6392

*CoME: An Unlearning-based Approach to Conflict-free Model Editing*
Dahyun Jung, Jaehyung Seo, Jaewook Lee, Chanjun Park and Heuiseok Lim . . . . . . . . . . . . . . 6410

*On The Origin of Cultural Biases in Language Models: From Pre-training Data to Linguistic Phenomena*
Tarek Naous and Wei Xu . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 6423

*Adapting Sentence-level Automatic Metrics for Document-level Simplification Evaluation*
Mounica Maddela and Fernando Alva-Manchego . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 6444

*Decoding Speculative Decoding*
Minghao Yan, Saurabh Agarwal and Shivaram Venkataraman . . . . . . . . . . . . . . . . . . . . . . . . . . . 6460

*Leveraging LLM For Synchronizing Information Across Multilingual Tables*
Siddharth Khincha, Tushar Kataria, Ankita Anand, Dan Roth and Vivek Gupta . . . . . . . . . . . . 6474

*ConMeC: A Dataset for Metonymy Resolution with Common Nouns*
Saptarshi Ghosh and Tianyu Jiang . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 6493

*Self-DC: When to Reason and When to Act? Self Divide-and-Conquer for Compositional Unknown Questions*
Hongru Wang, Boyang Xue, Baohang Zhou, Tianhua Zhang, Cunxiang Wang, Huimin Wang, Guanhua Chen and Kam-Fai Wong . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 6510

*TRANSIENTTABLES: Evaluating LLMs' Reasoning on Temporally Evolving Semi-structured Tables*
Abhilash Shankarampeta, Harsh Mahajan, Tushar Kataria, Dan Roth and Vivek Gupta . . . . . 6526

*AdvisorQA: Towards Helpful and Harmless Advice-seeking Question Answering with Collective Intelligence*
Minbeom Kim, Hwanhee Lee, Joonsuk Park, Hwaran Lee and Kyomin Jung . . . . . . . . . . . . . . 6545

*tRAG: Term-level Retrieval-Augmented Generation for Domain-Adaptive Retrieval*
Dohyeon Lee, Jongyoon Kim, Jihyuk Kim, Seung-won Hwang and Joonsuk Park . . . . . . . . . . 6566

*JRE-L: Journalist, Reader, and Editor LLMs in the Loop for Science Journalism for the General Audience*
Gongyao Jiang, Xinran Shi and Qiong Luo . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 6579

*Take the essence and discard the dross: A Rethinking on Data Selection for Fine-Tuning Large Language Models*
Ziche Liu, Rui Ke, Yajiao Liu, Feng Jiang and Haizhou Li . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 6595

*Graph Neural Network Enhanced Retrieval for Question Answering of Large Language Models*
Zijian Li, Qingyan Guo, Jiawei Shao, Lei Song, Jiang Bian, Jun Zhang and Rui Wang . . . . . . 6612

*Pula: Training Large Language Models for Setswana*
Nathan Brown and Vukosi Marivate . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 6634

*LegalViz: Legal Text Visualization by Text To Diagram Generation*
Eri Onami, Taiki Miyanishi, Koki Maeda and Shuhei Kurita . . . . . . . . . . . . . . . . . . . . . . . . . . . . 6657

*Active Few-Shot Learning for Text Classification*
Saeed Ahmadnia, Arash Yousefi Jordehi, Mahsa Hosseini Khasheh Heyran, Seyed Abolghasem Mirroshandel, Owen Rambow and Cornelia Caragea . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 6677

*Enhancing Multimodal Entity Linking with Jaccard Distance-based Conditional Contrastive Learning and Contextual Visual Augmentation*
Cong-Duy T Nguyen, Xiaobao Wu, Thong Thanh Nguyen, Shuai Zhao, Khoi M. Le, Nguyen Viet Anh, Feng Yichao and Anh Tuan Luu . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 6695

*ResearchAgent: Iterative Research Idea Generation over Scientific Literature with Large Language Models*
Jinheon Baek, Sujay Kumar Jauhar, Silviu Cucerzan and Sung Ju Hwang . . . . . . . . . . . . . . . . . 6709

*Logit Separability-Driven Samples and Multiple Class-Related Words Selection for Advancing In-Context Learning*
Zixiao Zhu, Zijian Feng, Hanzhang Zhou, Junlang Qian and Kezhi Mao . . . . . . . . . . . . . . . . . . 6739

*Identifying Emerging Concepts in Large Corpora*
Sibo Ma and Julian Nyarko . . . . . . . . . . 6760

*CodeSCM: Causal Analysis for Multi-Modal Code Generation*
Mukur Gupta, Noopur Bhatt and Suman Jana . . . . . . . . . . 6779

*From Distributional to Overton Pluralism: Investigating Large Language Model Alignment*
Thom Lake, Eunsol Choi and Greg Durrett . . . . . . . . . . 6794

*Advancing MoE Efficiency: A Collaboration-Constrained Routing (`C2R`) Strategy for Better Expert Parallelism Design*
Mohan Zhang, Pingzhi Li, Jie Peng, Mufan Qiu and Tianlong Chen . . . . . . . . . . 6815

*LibEvolutionEval: A Benchmark and Study for Version-Specific Code Generation*
Sachit Kuhar, Wasi Uddin Ahmad, Zijian Wang, Nihal Jain, Haifeng Qian, Baishakhi Ray, Murali Krishna Ramanathan, Xiaofei Ma and Anoop Deoras . . . . . . . . . . 6826

*Evaluating and Mitigating Object Hallucination in Large Vision-Language Models: Can They Still See Removed Objects?*
Yixiao He, Haifeng Sun, Pengfei Ren, Jingyu Wang, Huazheng Wang, Qi Qi, Zirui Zhuang and Jing Wang . . . . . . . . . . 6841

*Self-Pluralising Culture Alignment for Large Language Models*
Shaoyang Xu, Yongqi Leng, Linhao Yu and Deyi Xiong . . . . . . . . . . 6859

*K-COMP: Retrieval-Augmented Medical Domain Question Answering With Knowledge-Injected Compressor*
Jeonghun Cho and Gary Lee . . . . . . . . . . 6878

*DrawEduMath: Evaluating Vision Language Models with Expert-Annotated Students' Hand-Drawn Math Images*
Sami Baral, Li Lucy, Ryan Knight, Alice Ng, Luca Soldaini, Neil Heffernan and Kyle Lo . . 6902

*Knowledge Graph Guided Evaluation of Abstention Techniques*
Kinshuk Vasisht, Navreet Kaur and Danish Pruthi . . . . . . . . . . 6921

*Wav2Prompt: End-to-End Speech Prompt Learning and Task-based Fine-tuning for Text-based LLMs*
Keqi Deng, Guangzhi Sun and Phil Woodland . . . . . . . . . . 6940

*Legal Judgment Prediction based on Knowledge-enhanced Multi-Task and Multi-Label Text Classification*
Ang Li, Yiquan Wu, Ming Cai, Adam Jatowt, Xiang Zhou, Weiming Lu, Changlong Sun, Fei Wu and Kun Kuang . . . . . . . . . . 6957

*SPeCtrum: A Grounded Framework for Multidimensional Identity Representation in LLM-Based Agent*
Keyeun Lee, Seo Hyeong Kim, Seolhee Lee, Jinsu Eun, Yena Ko, Hayeon Jeon, Esther Hehsun Kim, Seonghye Cho, Soeun Yang, Eun-mee Kim and Hajin Lim . . . . . . . . . . 6971

*Beemo: Benchmark of Expert-edited Machine-generated Outputs*
Ekaterina Artemova, Jason S Lucas, Saranya Venkatraman, Jooyoung Lee, Sergei Tilga, Adaku Uchendu and Vladislav Mikhailov . . . . . . . . . . 6992

*SANDWiCH: Semantical Analysis of Neighbours for Disambiguating Words in Context ad Hoc*
Daniel Guzman Olivares, Lara Quijano and Federico Liberatore . . . . . . . . . . 7019

*Towards Automatic Evaluation for Image Transcreation*
Simran Khanuja, Vivek Iyer, Xiaoyu He and Graham Neubig . . . . . . . . . . 7034

*ImgTrojan: Jailbreaking Vision-Language Models with ONE Image*
Xijia Tao, Shuai Zhong, Lei Li, Qi Liu and Lingpeng Kong . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 7048

*RAG-Star: Enhancing Deliberative Reasoning with Retrieval Augmented Verification and Refinement*
Jinhao Jiang, Jiayi Chen, Junyi Li, Ruiyang Ren, Shijie Wang, Xin Zhao, Yang Song and Tao Zhang . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 7064

*Mitigating Biases of Large Language Models in Stance Detection with Counterfactual Augmented Calibration*
Ang Li, Jingqian Zhao, Bin Liang, Lin Gui, Hui Wang, Xi Zeng, Xingwei Liang, Kam-Fai Wong and Ruifeng Xu . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 7075

*Beyond the Next Token: Towards Prompt-Robust Zero-Shot Classification via Efficient Multi-Token Prediction*
Junlang Qian, Zixiao Zhu, Hanzhang Zhou, Zijian Feng, Zepeng Zhai and Kezhi Mao . . . . . 7093

*Investigating Hallucinations in Simultaneous Machine Translation: Knowledge Distillation Solution and Components Analysis*
Donglei Yu, Xiaomian Kang, Yuchen Liu, Feifei Zhai, Nanchang Cheng, Yu Zhou and Chengqing Zong . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 7116

*Markov Chain of Thought for Efficient Mathematical Reasoning*
Wen Yang, Minpeng Liao and Kai Fan . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 7132

*Towards Inducing Long-Context Abilities in Multilingual Neural Machine Translation Models*
Varun Gumma, Pranjal A Chitale and Kalika Bali . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 7158

*Yeah, Un, Oh: Continuous and Real-time Backchannel Prediction with Fine-tuning of Voice Activity Projection*
Koji Inoue, Divesh Lala, Gabriel Skantze and Tatsuya Kawahara . . . . . . . . . . . . . . . . . . . . . . . . . 7171

*Prompt Compression for Large Language Models: A Survey*
Zongqian Li, Yinhong Liu, Yixuan Su and Nigel Collier . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 7182

*Goal-Conditioned DPO: Prioritizing Safety in Misaligned Instructions*
Joo Bon Maeng, Seongmin Lee, Seokin Seo and Kee-Eung Kim . . . . . . . . . . . . . . . . . . . . . . . . . 7196

*K-Level Reasoning: Establishing Higher Order Beliefs in Large Language Models for Strategic Reasoning*
Yadong Zhang, Shaoguang Mao, Tao Ge, Xun Wang, Yan Xia, Man Lan and Furu Wei . . . . 7212

*SylloBio-NLI: Evaluating Large Language Models on Biomedical Syllogistic Reasoning*
Magdalena Wysocka, Danilo Carvalho, Oskar Wysocki, Marco Valentino and Andre Freitas 7235

*The State and Fate of Summarization Datasets: A Survey*
Noam Dahan and Gabriel Stanovsky . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 7259

*MGM: Global Understanding of Audience Overlap Graphs for Predicting the Factuality and the Bias of News Media*
Muhammad Arslan Manzoor, Ruihong Zeng, Dilshod Azizov, Preslav Nakov and Shangsong Liang . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 7279

*A Logical Fallacy-Informed Framework for Argument Generation*
Luca Mouchel, Debjit Paul, Shaobo Cui, Robert West, Antoine Bosselut and Boi Faltings . . 7296

*LLaMA-Berry: Pairwise Optimization for Olympiad-level Mathematical Reasoning via O1-like Monte Carlo Tree Search*
Di Zhang, Jianbo Wu, Jingdi Lei, Tong Che, Jiatong LI, Tong Xie, Xiaoshui Huang, Shufei Zhang, Marco Pavone, Yuqiang Li, Wanli Ouyang and Dongzhan Zhou . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 7315

*Generative Prompt Internalization*
Haebin Shin, Lei Ji, Yeyun Gong, Sungdong Kim, Eunbi Choi and Minjoon Seo . . . . . . . . . . 7338

*Script-Agnosticism and its Impact on Language Identification for Dravidian Languages*
Milind Agarwal, Joshua Otten and Antonios Anastasopoulos . . . . . . . . . . . . . . . . . . . . . . . . . . . . 7364

*NAT: Enhancing Agent Tuning with Negative Samples*
Renxi Wang, Xudong Han, Yixuan Zhang, Timothy Baldwin and Haonan Li . . . . . . . . . . . . . . 7385

*Hazards in Daily Life? Enabling Robots to Proactively Detect and Resolve Anomalies*
Zirui Song, Guangxian Ouyang, Meng Fang, Hongbin Na, Zijing Shi, Zhenhao Chen, fu Yujie, Zeyu Zhang, Shiyu Jiang, Miao Fang, Ling Chen and Xiuying Chen . . . . . . . . . . . . . . . . . . . . . . . . . . 7399

*How to Make the Most of LLMs' Grammatical Knowledge for Acceptability Judgments*
Yusuke Ide, Yuto Nishida, Justin Vasselli, Miyu Oba, Yusuke Sakai, Hidetaka Kamigaito and Taro Watanabe . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 7416

*Is Your LLM Outdated? A Deep Look at Temporal Generalization*
ChenghaoZhu ChenghaoZhu, Nuo Chen, Yufei Gao, Yunyi Zhang, Prayag Tiwari and Benyou Wang . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 7433

*Towards a Perspectivist Turn in Argument Quality Assessment*
Julia Romberg, Maximilian Maurer, Henning Wachsmuth and Gabriella Lapesa . . . . . . . . . . . 7458

*A Picture is Worth A Thousand Numbers: Enabling LLMs Reason about Time Series via Visualization*
Haoxin Liu, Chenghao Liu and B. Aditya Prakash . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 7486

*PlagBench: Exploring the Duality of Large Language Models in Plagiarism Generation and Detection*
Jooyoung Lee, Toshini Agrawal, Adaku Uchendu, Thai Le, Jinghui Chen and Dongwon Lee 7519

*Commonality and Individuality! Integrating Humor Commonality with Speaker Individuality for Humor Recognition*
Haohao Zhu, Xiaokun Zhang, Zeyuan Zeng, Junyu Lu, Zewen Bai, Liang Yang and Hongfei Lin 7535

*CAST: Corpus-Aware Self-similarity Enhanced Topic modelling*
Yanan Ma, Chenghao Xiao, Chenhan Yuan, Sabine N Van Der Veer, Lamiece Hassan, Chenghua Lin and Goran Nenadic . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 7548

*A Zero-Shot Open-Vocabulary Pipeline for Dialogue Understanding*
Abdulfattah Safa and Gözde Gül Şahin . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 7562

*Navigating the Cultural Kaleidoscope: A Hitchhiker's Guide to Sensitivity in Large Language Models*
Somnath Banerjee, Sayan Layek, Hari Shrawgi, Rajarshi Mandal, Avik Halder, Shanu Kumar, Sagnik Basu, Parag Agrawal, Rima Hazra and Animesh Mukherjee . . . . . . . . . . . . . . . . . . . . . . . . . . 7580

*Padding Tone: A Mechanistic Analysis of Padding Tokens in T2I Models*
Michael Toker, Ido Galil, Hadas Orgad, Rinon Gal, Yoad Tewel, Gal Chechik and Yonatan Belinkov . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 7618

*In-Context Learning (and Unlearning) of Length Biases*
Stephanie Schoch and Yangfeng Ji . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 7633

*AdTEC: A Unified Benchmark for Evaluating Text Quality in Search Engine Advertising*
Peinan Zhang, Yusuke Sakai, Masato Mita, Hiroki Ouchi and Taro Watanabe . . . . . . . . . . . . . 7672

*Empowering Retrieval-based Conversational Recommendation with Contrasting User Preferences*
Heejin Kook, Junyoung Kim, Seongmin Park and Jongwuk Lee . . . . . . . . . . . . . . . . . . . . . . . . . . 7692

*LRQ: Optimizing Post-Training Quantization for Large Language Models by Learning Low-Rank Weight-Scaling Matrices*
Jung Hyun Lee, Jeonghoon Kim, June Yong Yang, Se Jung Kwon, Eunho Yang, Kang Min Yoo and Dongsoo Lee . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 7708

*Towards Robust Knowledge Representations in Multilingual LLMs for Equivalence and Inheritance based Consistent Reasoning*
Gaurav Arora, Srujana Merugu, Shreya Jain and Vaibhav Saxena . . . . . . . . . . . . . . . . . . . . . . . . 7744

*LLMs as Meta-Reviewers' Assistants: A Case Study*
Eftekhar Hossain, Sanjeev Kumar Sinha, Naman Bansal, R. Alexander Knipper, Souvika Sarkar, John Salvador, Yash Mahajan, Sri Ram Pavan Kumar Guttikonda, Mousumi Akter, Md. Mahadi Hassan, Matthew Freestone, Matthew C. Williams Jr., Dongji Feng and Santu Karmaker . . . . . . . . . . . . . . . . 7763

*A Survey of NLP Progress in Sino-Tibetan Low-Resource Languages*
Shuheng Liu and Michael Best . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 7804

*Enhancing Language Model Hypernetworks with Restart: A Study on Optimization*
Yihan Zhang, Jie Fu, Rongrong Ji and Jie Chen . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 7826

*Functional Lexicon in Subword Tokenization*
Zachary William Hopton, Yves Scherrer and Tanja Samardzic . . . . . . . . . . . . . . . . . . . . . . . . . . . 7839

*Getting More Juice Out of Your Data: Hard Pair Refinement Enhances Visual-Language Models Without Extra Data*
Haonan Wang, Minbin Huang, Runhui Huang, Lanqing Hong, Hang Xu, Tianyang Hu, Xiaodan Liang, Zhenguo Li, Hong Cheng and Kenji Kawaguchi . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 7854

*Evaluating the Prompt Steerability of Large Language Models*
Erik Miehling, Michael Desmond, Karthikeyan Natesan Ramamurthy, Elizabeth M. Daly, Kush R. Varshney, Eitan Farchi, Pierre Dognin, Jesus Rios, Djallel Bouneffouf, Miao Liu and Prasanna Sattigeri 7874

*A Data-Driven Method for Analyzing and Quantifying Lyrics-Dance Motion Relationships*
Kento Watanabe and Masataka Goto . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 7901

*CROPE: Evaluating In-Context Adaptation of Vision and Language Models to Culture-Specific Concepts*
Malvina Nikandrou, Georgios Pantazopoulos, Nikolas Vitsakis, Ioannis Konstas and Alessandro Suglia . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 7917

*PicPersona-TOD : A Dataset for Personalizing Utterance Style in Task-Oriented Dialogue with Image Persona*
Jihyun Lee, Yejin Jeon, Seungyeon Seo and Gary Lee . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 7937

*Scaling LLM Inference Efficiently with Optimized Sample Compute Allocation*
Kexun Zhang, Shang Zhou, Danqing Wang, William Yang Wang and Lei Li . . . . . . . . . . . . . . 7959

*Large Language Models for Persian ↔ English Idiom Translation*
Sara Rezaeimanesh, Faezeh Hosseini and Yadollah Yaghoobzadeh . . . . . . . . . . . . . . . . . . . . . . . 7974

*Follow the Beaten Path: The Role of Route Patterns on Vision-Language Navigation Agents Generalization Abilities*
Kourosh T Baghaei, Dieter Pfoser and Antonios Anastasopoulos . . . . . . . . . . . . . . . . . . . . . . . . 7986

*Sneaking Syntax into Transformer Language Models with Tree Regularization*
Ananjan Nandi, Christopher D Manning and Shikhar Murty . . . . . . . . . . . . . . . . . . . . . . . . . . . . 8006

*Meta-Cultural Competence: Climbing the Right Hill of Cultural Awareness*
Sougata Saha, Saurabh Kumar Pandey and Monojit Choudhury . . . . . . . . . . . . . . . . . . . . . . . . . 8025

*Reading between the Lines: Can LLMs Identify Cross-Cultural Communication Gaps?*
Sougata Saha, Saurabh Kumar Pandey, Harshit Gupta and Monojit Choudhury . . . . . . . . . . . . 8043

*HMT: Hierarchical Memory Transformer for Efficient Long Context Language Processing*
Zifan He, Yingqi Cao, Zongyue Qin, Neha Prakriya, Yizhou Sun and Jason Cong . . . . . . . . . 8068

*Faux Polyglot: A Study on Information Disparity in Multilingual Large Language Models*
Nikhil Sharma, Kenton Murray and Ziang Xiao . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 8090

*Teaching Models to Balance Resisting and Accepting Persuasion*
Elias Stengel-Eskin, Peter Hase and Mohit Bansal . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 8108

*Making Language Models Robust Against Negation*
MohammadHossein Rezaei and Eduardo Blanco . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 8123

*Through the Lens of History: Methods for Analyzing Temporal Variation in Content and Framing of State-run Chinese Newspapers*
Shijia Liu and David A. Smith . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 8143

*PoisonedParrot: Subtle Data Poisoning Attacks to Elicit Copyright-Infringing Content from Large Language Models*
Michael-Andrei Panaitescu-Liess, Pankayaraj Pathmanathan, Yigitcan Kaya, Zora Che, Bang An, Sicheng Zhu, Aakriti Agrawal and Furong Huang . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 8173

*Towards Operationalizing Right to Data Protection*
Abhinav Java, Simra Shahid and Chirag Agarwal . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 8191

*Learning vs Retrieval: The Role of In-Context Examples in Regression with Large Language Models*
Aliakbar Nafar, K. Brent Venable and Parisa Kordjamshidi . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 8206

*GLiREL - Generalist Model for Zero-Shot Relation Extraction*
Jack Boylan, Chris Hokamp and Demian Gholipour Ghalandari . . . . . . . . . . . . . . . . . . . . . . . . 8230

*ComPO: Community Preferences for Language Model Personalization*
Sachin Kumar, Chan Young Park, Yulia Tsvetkov, Noah A. Smith and Hannaneh Hajishirzi . 8246

*GroundCocoa: A Benchmark for Evaluating Compositional & Conditional Reasoning in Language Models*
Harsh Kohli, Sachin Kumar and Huan Sun . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 8280

*ALPACA AGAINST VICUNA: Using LLMs to Uncover Memorization of LLMs*
Aly M. Kassem, Omar Mahmoud, Niloofar Mireshghallah, Hyunwoo Kim, Yulia Tsvetkov, Yejin Choi, Sherif Saad and Santu Rana . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 8296

*Evaluating Contextualized Representations of (Spanish) Ambiguous Words: A New Lexical Resource and Empirical Analysis*
Pamela D Riviere, Anne L. Beatty-Martínez and Sean Trott . . . . . . . . . . . . . . . . . . . . . . . . . . . . 8322

*Understanding LLMs' Fluid Intelligence Deficiency: An Analysis of the ARC Task*
Junjie Wu, Mo Yu, Lemao Liu, Dit-Yan Yeung and Jie Zhou . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 8339

*FedSpaLLM: Federated Pruning of Large Language Models*
Guangji Bai, Yijiang Li, Zilinghan Li, Liang Zhao and Kibaek Kim . . . . . . . . . . . . . . . . . . . . . . 8361

*IHEval: Evaluating Language Models on Following the Instruction Hierarchy*
Zhihan Zhang, Shiyang Li, Zixuan Zhang, Xin Liu, Haoming Jiang, Xianfeng Tang, Yifan Gao, Zheng Li, Haodong Wang, Zhaoxuan Tan, Yichuan Li, Qingyu Yin, Bing Yin and Meng Jiang . . . 8374

*Afrispeech-Dialog: A Benchmark Dataset for Spontaneous English Conversations in Healthcare and Beyond*
Mardhiyah Sanni, Tassallah Abdullahi, Devendra Deepak Kayande, Emmanuel Ayodele, Naome A Etori, Michael Samwel Mollel, Moshood O. Yekini, Chibuzor Okocha, Lukman Enegi Ismaila, Folafunmi Omofoye, Boluwatife A. Adewale and Tobi Olatunji . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 8399

*THREAD: Thinking Deeper with Recursive Spawning*
Philip Schroeder, Nathaniel W. Morgan, Hongyin Luo and James R. Glass . . . . . . . . . . . . . . . . 8418

*CORG: Generating Answers from Complex, Interrelated Contexts*
Hyunji Lee, Franck Dernoncourt, Trung Bui and Seunghyun Yoon . . . . . . . . . . . . . . . . . . . . . . . 8443

*Generating Diverse Hypotheses for Inductive Reasoning*
Kang-il Lee, Hyukhun Koh, Dongryeol Lee, Seunghyun Yoon, Minsung Kim and Kyomin Jung 8461

*On the Analysis and Distillation of Emergent Outlier Properties in Pre-trained Language Models*
Tianyang Zhao, Kunwar Yashraj Singh, Srikar Appalaraju, Peng Tang, Ying Nian Wu and Li Erran Li . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 8475

*Open-World Evaluation for Retrieving Diverse Perspectives*
Hung-Ting Chen and Eunsol Choi . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 8508

*Analyzing the Inner Workings of Transformers in Compositional Generalization*
Ryoma Kumon and Hitomi Yanaka . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 8529

*Substance Beats Style: Why Beginning Students Fail to Code with LLMs*
Francesca Lucchetti, Zixuan Wu, Arjun Guha, Molly Q Feldman and Carolyn Jane Anderson8541

*Reverse Thinking Makes LLMs Stronger Reasoners*
Justin Chen, Zifeng Wang, Hamid Palangi, Rujun Han, Sayna Ebrahimi, Long Le, Vincent Perot, Swaroop Mishra, Mohit Bansal, Chen-Yu Lee and Tomas Pfister . . . . . . . . . . . . . . . . . . . . . . . . . . . . 8611

*Towards Lifelong Dialogue Agents via Timeline-based Memory Management*
Kai Tzu-iunn Ong, Namyoung Kim, Minju Gwak, Hyungjoo Chae, Taeyoon Kwon, Yohan Jo, Seung-won Hwang, Dongha Lee and Jinyoung Yeo . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 8631

*StyleDistance: Stronger Content-Independent Style Embeddings with Synthetic Parallel Examples*
Ajay Patel, Jiacheng Zhu, Justin Qiu, Zachary Horvitz, Marianna Apidianaki, Kathleen McKeown and Chris Callison-Burch . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 8662

*FiNE: Filtering and Improving Noisy Data Elaborately with Large Language Models*
Junliang He, Ziyue Fan, Shaohui Kuang, Li Xiaoqing, Kai Song, Yaqian Zhou and Xipeng Qiu 8686

*CAMIEval: Enhancing NLG Evaluation through Multidimensional Comparative Instruction-Following Analysis*
Ziyue Fan, Junliang He, Li Xiaoqing, Shaohui Kuang, Kai Song, Yaqian Zhou and Xipeng Qiu 8708

*LongLeader: A Comprehensive Leaderboard for Large Language Models in Long-context Scenarios*
Pei Chen, Hongye Jin, Cheng-Che Lee, Rulin Shao, Jingfeng Yang, Mingyu Zhao, Zhaoyu Zhang, Qin Lu, Kaiwen Men, Ning Xie, Huasheng Li, Bing Yin, Han Li and Lingyun Wang . . . . . . . . . . . 8734

*Language Models Can Infer Action Semantics for Symbolic Planners from Environment Feedback*
Wang Bill Zhu, Ishika Singh, Robin Jia and Jesse Thomason . . . . . . . . . . . . . . . . . . . . . . . . . . . 8751

*SLM-Mod: Small Language Models Surpass LLMs at Content Moderation*
Xianyang Zhan, Agam Goyal, Yilun Chen, Eshwar Chandrasekharan and Koustuv Saha . . . . 8774

*On Positional Bias of Faithfulness for Long-form Summarization*
David Wan, Jesse Vig, Mohit Bansal and Shafiq Joty . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 8791

*BPO: Towards Balanced Preference Optimization between Knowledge Breadth and Depth in Alignment*
Sizhe Wang, Yongqi Tong, Hengyuan Zhang, Dawei Li, Xin Zhang and Tianlong Chen . . . . 8811

*UNDIAL: Self-Distillation with Adjusted Logits for Robust Unlearning in Large Language Models*
Yijiang River Dong, Hongzhou Lin, Mikhail Belkin, Ramon Huerta and Ivan Vulić . . . . . . . . 8827

*H-STAR: LLM-driven Hybrid SQL-Text Adaptive Reasoning on Tables*
Nikhil Abhyankar, Vivek Gupta, Dan Roth and Chandan K. Reddy . . . . . . . . . . . . . . . . . . . . . . . 8841

*Kill two birds with one stone: generalized and robust AI-generated text detection via dynamic perturbations*
Yinghan Zhou, Juan Wen, Wanli Peng, Xue Yiming, ZiWei Zhang and Wu Zhengxian . . . . . 8864

*Vision-Language Models Can Self-Improve Reasoning via Reflection*
Kanzhi Cheng, Li YanTao, Fangzhi Xu, Jianbing Zhang, Hao Zhou and Yang Liu . . . . . . . . . . 8876

*Emergence of Episodic Memory in Transformers: Characterizing Changes in Temporal Structure of Attention Scores During Training*
Deven Mahesh Mistry, Anooshka Bajaj, Yash Aggarwal, Sahaj Singh Maini and Zoran Tiganj 8893

*Knowledge Graph-Guided Retrieval Augmented Generation*
Xiangrong Zhu, Yuexiang Xie, Yi Liu, Yaliang Li and Wei Hu . . . . . . . . . . . . . . . . . . . . . . . . . . 8912

*Amphista: Bi-directional Multi-head Decoding for Accelerating LLM Inference*
Zeping Li, Xinlong Yang, Ziheng Gao, Ji Liu, Guanchen Li, Zhuang Liu, Dong Li, Jinzhang Peng, Lu Tian and Emad Barsoum . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 8925

*CAVE: Controllable Authorship Verification Explanations*
Sahana Ramnath, Kartik Pandey, Elizabeth Boschee and Xiang Ren . . . . . . . . . . . . . . . . . . . . . . 8939

*Are LLM-Judges Robust to Expressions of Uncertainty? Investigating the effect of Epistemic Markers on LLM-based Evaluation*
Dongryeol Lee, Yerin Hwang, Yongil Kim, Joonsuk Park and Kyomin Jung . . . . . . . . . . . . . . . 8962

*Dynamic Uncertainty Ranking: Enhancing Retrieval-Augmented In-Context Learning for Long-Tail Knowledge in LLMs*
Shuyang Yu, Runxue Bao, Parminder Bhatia, Taha Kass-Hout, Jiayu Zhou and Cao Xiao . . . 8985

*Seq1F1B: Efficient Sequence-Level Pipeline Parallelism for Large Language Model Training*
Sun Ao, Weilin Zhao, Xu Han, Cheng Yang, Xinrong Zhang, Zhiyuan Liu, Chuan Shi and Maosong Sun . . . . . 8998

*Differentially Private Learning Needs Better Model Initialization and Self-Distillation*
Ivoline C. Ngong, Joseph Near and Niloofar Mireshghallah . . . . . 9009

*Is a Peeled Apple Still Red? Evaluating LLMs' Ability for Conceptual Combination with Property Type*
Seokwon Song, Taehyun Lee, Jaewoo Ahn, Jae Hyuk Sung and Gunhee Kim . . . . . 9028

*CRScore: Grounding Automated Evaluation of Code Review Comments in Code Claims and Smells*
Atharva Naik, Marcus Alenius, Daniel Fried and Carolyn Rose . . . . . 9049

*KS-Lottery: Finding Certified Lottery Tickets for Multilingual Transfer in Large Language Models*
Fei Yuan, Chang Ma, Shuai Yuan, Qiushi Sun and Lei Li . . . . . 9077

*PA-RAG: RAG Alignment via Multi-Perspective Preference Optimization*
Jiayi Wu, Hengyi Cai, Lingyong Yan, Hao Sun, Xiang Li, Shuaiqiang Wang, Dawei Yin and Ming Gao . . . . . 9091

*$B^4$: A Black-Box Scrubbing Attack on LLM Watermarks*
Baizhou Huang, Xiao Pu and Xiaojun Wan . . . . . 9113

*IMRRF: Integrating Multi-Source Retrieval and Redundancy Filtering for LLM-based Fake News Detection*
Dayang Li, Fanxiao Li, Bingbing Song, Li Tang and Wei Zhou . . . . . 9127

*Matina: A Large-Scale 73B Token Persian Text Corpus*
Sara Bourbour Hosseinbeigi, Fatemeh Taherinezhad, Heshaam Faili, Hamed Baghbani, Fatemeh Nadi and Mostafa Amiri . . . . . 9143

*SMAB: MAB based word Sensitivity Estimation Framework and its Applications in Adversarial Text Generation*
Saurabh Kumar Pandey, Sachin Vashistha, Debrup Das, Somak Aditya and Monojit Choudhury 9158

*ManaTTS Persian: a recipe for creating TTS datasets for lower resource languages*
Mahta Fetrat Qharabagh, Zahra Dehghanian and Hamid R. Rabiee . . . . . 9177

*CultureInstruct: Curating Multi-Cultural Instructions at Scale*
Viet Thanh Pham, Zhuang Li, Lizhen Qu and Gholamreza Haffari . . . . . 9207

*Lost in Inference: Rediscovering the Role of Natural Language Inference for Large Language Models*
Lovish Madaan, David Esiobu, Pontus Stenetorp, Barbara Plank and Dieuwke Hupkes . . . . . 9229

*DenseSSM: State Space Models with Dense Hidden Connection for Efficient Large Language Models*
Wei He, Kai Han, Yehui Tang, Chengcheng Wang, Yujie Yang, Tianyu Guo and Yunhe Wang9243

*A Mixed-Language Multi-Document News Summarization Dataset and a Graphs-Based Extract-Generate Model*
Shengxiang Gao, Fang Nan, Yongbing Zhang, Yuxin Huang, Kaiwen Tan and Zhengtao Yu 9255

*Measuring memorization in language models via probabilistic extraction*
Jamie Hayes, Marika Swanberg, Harsh Chaudhari, Itay Yona, Ilia Shumailov, Milad Nasr, Christopher A. Choquette-Choo, Katherine Lee and A. Feder Cooper . . . . . 9266

*Audio Is the Achilles' Heel: Red Teaming Audio Large Multimodal Models*
Hao Yang, Lizhen Qu, Ehsan Shareghi and Gholamreza Haffari . . . . . 9292

*EMS-SD: Efficient Multi-sample Speculative Decoding for Accelerating Large Language Models*
Yunsheng Ni, Chuanjian Liu, Yehui Tang, Kai Han and Yunhe Wang . . . . . . . . . . . . . . . . . . . . . 9307

*Regularized Best-of-N Sampling with Minimum Bayes Risk Objective for Language Model Alignment*
Yuu Jinnai, Tetsuro Morimura, Kaito Ariu and Kenshi Abe . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 9321

*MAPWise: Evaluating Vision-Language Models for Advanced Map Queries*
Srija Mukhopadhyay, Abhishek Rajgaria, Prerana Khatiwada, Manish Shrivastava, Dan Roth and Vivek Gupta . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 9348

*Pay More Attention to Images: Numerous Images-Oriented Multimodal Summarization*
Min Xiao, Junnan Zhu, Feifei Zhai, Chengqing Zong and Yu Zhou . . . . . . . . . . . . . . . . . . . . . . . 9379

*$S^2$-MAD: Breaking the Token Barrier to Enhance Multi-Agent Debate Efficiency*
Yuting Zeng, Weizhe Huang, Lei Jiang, Tongxuan Liu, XiTai Jin, Chen Tianying Tiana, Jing Li and Xiaohua Xu . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 9393

*MASTER: A Multi-Agent System with LLM Specialized MCTS*
Bingzheng Gan, Yufan Zhao, Tianyi Zhang, Jing Huang, LI Yusu, Shu Xian Teo, Changwang Zhang and Wei Shi . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 9409

*ScreenQA: Large-Scale Question-Answer Pairs Over Mobile App Screenshots*
Yu-Chung Hsiao, Fedir Zubach, Gilles Baechler, Srinivas Sunkara, Victor Carbune, Jason Lin, Maria Wang, Yun Zhu and Jindong Chen . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 9427

*Cross-Lingual and Cross-Cultural Variation in Image Descriptions*
Uri Berger and Edoardo Ponti . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 9453

*Soft Syntactic Reinforcement for Neural Event Extraction*
Anran Hao, Jian Su, Shuo Sun and Teo Yong Sen . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 9466

*Not All Adapters Matter: Selective Adapter Freezing for Memory-Efficient Fine-Tuning of Language Models*
Hyegang Son, Yonglak Son, Changhoon Kim and Young Geun Kim . . . . . . . . . . . . . . . . . . . . . 9479

*Bridging the Gap between Expert and Language Models: Concept-guided Chess Commentary Generation and Evaluation*
Jaechang Kim, Jinmin Goh, Inseok Hwang, Jaewoong Cho and Jungseul Ok . . . . . . . . . . . . . . 9497

*TCProF:Time-Complexity Prediction SSL Framework*
Joonghyuk Hahn, Hyeseon Ahn, Jungin Kim, Soohan Lim and Yo-Sub Han . . . . . . . . . . . . . . . 9517

*Culture-TRIP: Culturally-Aware Text-to-Image Generation with Iterative Prompt Refinement*
Suchae Jeong, Inseong Choi, Youngsik Yun and Jihie Kim . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 9543

*Behavior-SD: Behaviorally Aware Spoken Dialogue Generation with Large Language Models*
Sehun Lee, Kang-wook Kim and Gunhee Kim . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 9574

*Is Translation All You Need? A Study on Solving Multilingual Tasks with Large Language Models*
Chaoqun Liu, Wenxuan Zhang, Yiran Zhao, Anh Tuan Luu and Lidong Bing . . . . . . . . . . . . . 9594

*AlgoPuzzleVQA: Diagnosing Multimodal Reasoning Challenges of Language Models with Algorithmic Multimodal Puzzles*
Deepanway Ghosal, Vernon Toh, Yew Ken Chia and Soujanya Poria . . . . . . . . . . . . . . . . . . . . . 9615

*Towards Quantifying Commonsense Reasoning with Mechanistic Insights*
Abhinav Joshi, Areeb Ahmad, Divyaksh Shukla and Ashutosh Modi . . . . . . . . . . . . . . . . . . . . . 9633

*Beyond Logit Lens: Contextual Embeddings for Robust Hallucination Detection & Grounding in VLMs*
Anirudh Phukan, Divyansh Divyansh, Harshit Kumar Morj, Vaishnavi Vaishnavi, Apoorv Saxena and Koustava Goswami . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 9661

*M2Lingual: Enhancing Multilingual, Multi-Turn Instruction Alignment in Large Language Models*
Rishabh Maheshwary, Vikas Yadav, Hoang H Nguyen, Khyati Mahajan and Sathwik Tejaswi Madhusudhan . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 9676

*Multi$^3$Hate: Multimodal, Multilingual, and Multicultural Hate Speech Detection with Vision–Language Models*
Minh Duc Bui, Katharina Von Der Wense and Anne Lauscher . . . . . . . . . . . . . . . . . . . . . . . . . . 9714

*Grounding Fallacies Misrepresenting Scientific Publications in Evidence*
Max Glockner, Yufang Hou, Preslav Nakov and Iryna Gurevych . . . . . . . . . . . . . . . . . . . . . . . . . 9732

*Has this Fact been Edited? Detecting Knowledge Edits in Language Models*
Paul Youssef, Zhixue Zhao, Christin Seifert and Jörg Schlötterer . . . . . . . . . . . . . . . . . . . . . . . . . 9768

*AdaMergeX: Cross-Lingual Transfer with Large Language Models via Adaptive Adapter Merging*
Yiran Zhao, Wenxuan Zhang, Huiming Wang, Kenji Kawaguchi and Lidong Bing . . . . . . . . . 9785

*Coverage-based Fairness in Multi-document Summarization*
Haoyuan Li, Yusen Zhang, Rui Zhang and Snigdha Chaturvedi . . . . . . . . . . . . . . . . . . . . . . . . . . 9801

*Grammar Control in Dialogue Response Generation for Language Learning Chatbots*
Dominik Glandorf, Peng Cui, Detmar Meurers and Mrinmaya Sachan . . . . . . . . . . . . . . . . . . . . 9820

*Does Mapo Tofu Contain Coffee? Probing LLMs for Food-related Cultural Knowledge*
Li Zhou, Taelin Karidi, Wanlong Liu, Nicolas Garneau, Yong Cao, Wenyu Chen, Haizhou Li and Daniel Hershcovich . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 9840

*Palette of Language Models: A Solver for Controlled Text Generation*
Zhe Yang, Yi Huang, Yaqin Chen, XiaotingWu XiaotingWu, Junlan Feng and Chao Deng . . 9868

*MAMM-Refine: A Recipe for Improving Faithfulness in Generation with Multi-Agent Collaboration*
David Wan, Justin Chen, Elias Stengel-Eskin and Mohit Bansal . . . . . . . . . . . . . . . . . . . . . . . . . 9882

*MADial-Bench: Towards Real-world Evaluation of Memory-Augmented Dialogue Generation*
Junqing He, Liang Zhu, Rui Wang, Xi Wang, Gholamreza Haffari and Jiaxing Zhang . . . . . . 9902

*Assessing the State of the Art in Scene Segmentation*
Albin Zehe, Elisabeth Fischer and Andreas Hotho . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 9922

*DCE-LLM: Dead Code Elimination with Large Language Models*
Minyu Chen, Guoqiang Li, Ling-I Wu and Ruibang Liu . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 9942

*Instruct-of-Reflection: Enhancing Large Language Models Iterative Reflection Capabilities via Dynamic-Meta Instruction*
Liping Liu, Chunhong Zhang, Likang Wu, Chuang Zhao, Zheng Hu, Ming He and Jianping Fan 9956

*Correcting Negative Bias in Large Language Models through Negative Attention Score Alignment*
Sangwon Yu, Jongyoon Song, Bongkyu Hwang, Hoyoung Kang, Sooah Cho, Junhwa Choi, Seong-gho Joe, Taehee Lee, Youngjune Gwon and Sungroh Yoon . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 9979

*MiCEval: Unveiling Multimodal Chain of Thought's Quality via Image Description and Reasoning Steps*
Xiongtao Zhou, Jie He, Lanyu Chen, Jingyu li, Haojing Chen, Victor Gutierrez Basulto, Jeff Z. Pan and Hanjie Chen . . . . . 10002

*CartesianMoE: Boosting Knowledge Sharing among Experts via Cartesian Product Routing in Mixture-of-Experts*
Zhenpeng Su, Xing W, Zijia Lin, Yizhe Xiong, Minxuan Lv, Guangyuan Ma, Hui Chen, Songlin Hu and Guiguang Ding . . . . . 10040

*Measuring and Benchmarking Large Language Models' Capabilities to Generate Persuasive Language*
Amalie Brogaard Pauli, Isabelle Augenstein and Ira Assent . . . . . 10056

*MILU: A Multi-task Indic Language Understanding Benchmark*
Sshubam Verma, Mohammed Safi Ur Rahman Khan, Vishwajeet Kumar, Rudra Murthy and Jaydeep Sen . . . . . 10076

*AutoEval-ToD: Automated Evaluation of Task-oriented Dialog Systems*
Arihant Jain, Purav Aggarwal, Rishav Sahay, Chaosheng Dong and Anoop Saladi . . . . . 10133

*Self-calibration for Language Model Quantization and Pruning*
Miles Williams, George Chrysostomou and Nikolaos Aletras . . . . . 10149

*Logic-of-Thought: Injecting Logic into Contexts for Full Reasoning in Large Language Models*
Tongxuan Liu, Wenjiang Xu, Weizhe Huang, Yuting Zeng, Jiaxing Wang, Xingyu Wang, Hailong Yang and Jing Li . . . . . 10168

*IFIR: A Comprehensive Benchmark for Evaluating Instruction-Following in Expert-Domain Information Retrieval*
Tingyu Song, Guo Gan, Mingsheng Shang and Yilun Zhao . . . . . 10186

*QAVA: Query-Agnostic Visual Attack to Large Vision-Language Models*
Yudong Zhang, Ruobing Xie, Jiansheng Chen, Xingwu Sun, Zhanhui Kang and Yu Wang . 10205

*Evaluating and Improving Graph to Text Generation with Large Language Models*
Jie He, Yijun Yang, Wanqiu Long, Deyi Xiong, Victor Gutierrez Basulto and Jeff Z. Pan . . 10219

*The Plagiarism Singularity Conjecture*
Sriram Ranga, Rui Mao, Erik Cambria and Anupam Chattopadhyay . . . . . 10245

*Ensembling Large Language Models with Process Reward-Guided Tree Search for Better Complex Reasoning*
Sungjin Park, Xiao Liu, Yeyun Gong and Edward Choi . . . . . 10256

*One Unified Model for Diverse Tasks: Emotion Cause Analysis via Self-Promote Cognitive Structure Modeling*
Zhaoxin Yu, Xinglin Xiao and Wenji Mao . . . . . 10278

*Soft Language Prompts for Language Transfer*
Ivan Vykopal, Simon Ostermann and Marian Simko . . . . . 10294

*PICLe: Pseudo-annotations for In-Context Learning in Low-Resource Named Entity Detection*
Sepideh Mamooler, Syrielle Montariol, Alexander Mathis and Antoine Bosselut . . . . . 10314

*Can Large Language Models Invent Algorithms to Improve Themselves?*
Yoichi Ishibashi, Taro Yano and Masafumi Oyamada . . . . . 10332

*Simulating Classroom Education with LLM-Empowered Agents*
Zheyuan Zhang, Daniel Zhang-Li, Jifan Yu, Linlu Gong, Jinchang Zhou, Zhanxin Hao, Jianxiao Jiang, Jie Cao, Huiqin Liu, Zhiyuan Liu, Lei Hou and Juanzi Li . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 10364

*A Grounded Typology of Word Classes*
Coleman Haley, Sharon Goldwater and Edoardo Ponti . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 10380

*SSH: Sparse Spectrum Adaptation via Discrete Hartley Transformation*
Yixian Shen, Qi Bi, Jia-hong Huang, Hongyi Zhu, Andy D. Pimentel and Anuj Pathania . . 10400

*LLM-guided Plan and Retrieval: A Strategic Alignment for Interpretable User Satisfaction Estimation in Dialogue*
Sangyeop Kim, Sohhyung Park, Jaewon Jung, Jinseok Kim and Sungzoon Cho . . . . . . . . . . 10416

*LCIRC: A Recurrent Compression Approach for Efficient Long-form Context and Query Dependent Modeling in LLMs*
Sumin An, Junyoung Sung, Wonpyo Park, Chanjun Park and Paul Hongsuck Seo . . . . . . . . . 10431

*A Template Is All You Meme*
Luke Bates, Peter Ebert Christensen, Preslav Nakov and Iryna Gurevych . . . . . . . . . . . . . . . . 10443

*LLMs vs Established Text Augmentation Techniques for Classification: When do the Benefits Outweight the Costs?*
Jan Cegin, Jakub Simko and Peter Brusilovsky . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 10476

*Bridging the Visual Gap: Fine-Tuning Multimodal Models with Knowledge-Adapted Captions*
Moran Yanuka, Assaf Ben-Kish, Yonatan Bitton, Idan Szpektor and Raja Giryes . . . . . . . . . 10497

*Self-Training Meets Consistency: Improving LLMs' Reasoning with Consistency-Driven Rationale Evaluation*
Jaehyeok Lee, Keisuke Sakaguchi and JinYeong Bak . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 10519

*Evaluating Defeasible Reasoning in LLMs with DEFREASING*
Emily Allaway and Kathleen McKeown . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 10540

*Evaluating Input Feature Explanations through a Unified Diagnostic Evaluation Framework*
Jingyi Sun, Pepa Atanasova and Isabelle Augenstein . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 10559

*From Evidence to Belief: A Bayesian Epistemology Approach to Language Models*
Minsu Kim, Sangryul Kim and James Thorne . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 10578

*Private Synthetic Text Generation with Diffusion Models*
Sebastian Ochs and Ivan Habernal . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 10612

*Mitigating Tail Narrowing in LLM Self-Improvement via Socratic-Guided Sampling*
Yiwen Ding, Zhiheng Xi, Wei He, Lizhuoyuan Lizhuoyuan, Yitao Zhai, Shi Xiaowei, Xunliang Cai, Tao Gui, Qi Zhang and Xuanjing Huang . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 10627

*FactEval: Evaluating the Robustness of Fact Verification Systems in the Era of Large Language Models*
Mamta Mamta and Oana Cocarascu . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 10647

*Analyzing Memorization in Large Language Models through the Lens of Model Attribution*
Tarun Ram Menta, Susmit Agrawal and Chirag Agarwal . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 10661

*Track-SQL: Enhancing Generative Language Models with Dual-Extractive Modules for Schema and Context Tracking in Multi-turn Text-to-SQL*
Bingfeng Chen, Shaobin Shi, Yongqi Luo, Boyan Xu, Ruichu Cai and Zhifeng Hao . . . . . . 10690

*Prototypical Extreme Multi-label Classification with a Dynamic Margin Loss*
Kunal Dahiya, Diego Ortego and David Jimenez-Cabello . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 10709

*MCQG-SRefine: Multiple Choice Question Generation and Evaluation with Iterative Self-Critique, Correction, and Comparison Feedback*
Zonghai Yao, Aditya Parashar, Huixue Zhou, Won Seok Jang, Feiyun Ouyang , Zhichao Yang and Hong yu . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 10728

*Main Predicate and Their Arguments as Explanation Signals For Intent Classification*
Sameer Pimparkhede and Pushpak Bhattacharyya . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 10778

*Handling Missing Entities in Zero-Shot Named Entity Recognition: Integrated Recall and Retrieval Augmentation*
Ruichu Cai, Junhao Lu, Zhongjie Chen, Boyan Xu and Zhifeng Hao . . . . . . . . . . . . . . . . . . . . 10790

*KMI: A Dataset of Korean Motivational Interviewing Dialogues for Psychotherapy*
Hyunjong Kim, Suyeon Lee, Yeongjae Cho, Eunseo Ryu, Yohan Jo, Suran Seong and Sungzoon Cho . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 10803

*Automatic Input Rewriting Improves Translation with Large Language Models*
Dayeon Ki and Marine Carpuat . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 10829

*HIGGS: Pushing the Limits of Large Language Model Quantization via the Linearity Theorem*
Vladimir Malinovskii, Andrei Panferov, Ivan Ilin, Han Guo, Peter Richtárik and Dan Alistarh 10857

*The LLM Language Network: A Neuroscientific Approach for Identifying Causally Task-Relevant Units*
Badr AlKhamissi, Greta Tuckute, Antoine Bosselut and Martin Schrimpf . . . . . . . . . . . . . . . . 10887

*MixLLM: Dynamic Routing in Mixed Large Language Models*
Xinyuan Wang, Yanchi Liu, Wei Cheng, Xujiang Zhao, Zhengzhang Chen, Wenchao Yu, Yanjie Fu and Haifeng Chen . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 10912

*Continual Learning in Multilingual Sign Language Translation*
Shakib Yazdani, Josef Van Genabith and Cristina España-Bonet . . . . . . . . . . . . . . . . . . . . . . . . 10923

*Few-Shot Natural Language to First-Order Logic Translation via Code Generation*
Junnan Liu . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 10939

*How Good Are LLMs for Literary Translation, Really? Literary Translation Evaluation with Humans and LLMs*
Ran Zhang, Wei Zhao and Steffen Eger . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 10961

*PORT: Preference Optimization on Reasoning Traces*
Salem Lahlou, Abdalgader Abubaker and Hakim Hacid . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 10989

*Guiding Through Complexity: What Makes Good Supervision for Hard Reasoning Tasks?*
Xuan He, Da Yin and Nanyun Peng . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 11006

*Fine-Grained Transfer Learning for Harmful Content Detection through Label-Specific Soft Prompt Tuning*
Faeze Ghorbanpour, Viktor Hangya and Alexander Fraser . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 11047

*A Systematic Examination of Preference Learning through the Lens of Instruction-Following*
Joongwon Kim, Anirudh Goyal, Aston Zhang, Bo Xiong, Rui Hou, Melanie Kambadur, Dhruv Mahajan, Hannaneh Hajishirzi and Liang Tan . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 11062

*Lived Experience Not Found: LLMs Struggle to Align with Experts on Addressing Adverse Drug Reactions from Psychiatric Medication Use*
Mohit Chandra, Siddharth Sriraman, Gaurav Verma, Harneet Singh Khanuja, Jose Suarez Campayo, Zihang Li, Michael L. Birnbaum and Munmun De Choudhury . . . . . . . . . . . . . . . . . . . . . . . . . 11083

*Latent Factor Models Meets Instructions: Goal-conditioned Latent Factor Discovery without Task Supervision*
Zhouhang Xie, Tushar Khot, Bhavana Dalvi Mishra, Harshit Surana, Julian McAuley, Peter Clark and Bodhisattwa Prasad Majumder . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 11114

*LLM-Supported Natural Language to Bash Translation*
Finnian Westenfelder, Erik Hemberg, Stephen Moskal, Una-May O'Reilly and Silviu Chiricescu 11135

*REL-A.I.: An Interaction-Centered Approach To Measuring Human-LM Reliance*
Kaitlyn Zhou, Jena D. Hwang, Xiang Ren, Nouha Dziri, Dan Jurafsky and Maarten Sap . . 11148

*Eliciting Critical Reasoning in Retrieval-Augmented Generation via Contrastive Explanations*
Leonardo Ranaldi, Marco Valentino and Andre Freitas . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 11168

*A Distributional Perspective on Word Learning in Neural Language Models*
Filippo Ficarra, Ryan Cotterell and Alex Warstadt . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 11184

*Disentangling language change: sparse autoencoders quantify the semantic evolution of indigeneity in French*
Jacob A. Matthews, Laurent Dubreuil, Imane Terhmina, Yunci Sun, Matthew Wilkens and Marten Van Schijndel . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 11208

*Planetarium: A Rigorous Benchmark for Translating Text to Structured Planning Languages*
Max Zuo, Francisco Piedrahita Velez, Xiaochen Li, Michael Littman and Stephen Bach . . . 11223

*One fish, two fish, but not the whole sea: Alignment reduces language models' conceptual diversity*
Sonia Krishna Murthy, Tomer Ullman and Jennifer Hu . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 11241

*Using Text-Based Causal Inference to Disentangle Factors Influencing Online Review Ratings*
Linsen Li, Aron Culotta and Nicholas Mattei . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 11259

*Unlearning as multi-task optimization: A normalized gradient difference approach with an adaptive learning rate*
Xiaomeng Jin, Zhiqi Bu, Bhanukiran Vinzamuri, Anil Ramakrishna, Kai-Wei Chang, Volkan Cevher and Mingyi Hong . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 11278

*REFFLY: Melody-Constrained Lyrics Editing Model*
Songyan Zhao, Bingxuan Li, Yufei Tian and Nanyun Peng . . . . . . . . . . . . . . . . . . . . . . . . . . . . 11295

*Exploring Safety-Utility Trade-Offs in Personalized Language Models*
Anvesh Rao Vijjini, Somnath Basu Roy Chowdhury and Snigdha Chaturvedi . . . . . . . . . . . . 11316

*MultiChartQA: Benchmarking Vision-Language Models on Multi-Chart Problems*
Zifeng Zhu, Mengzhao Jia, Zhihan Zhang, Lang Li and Meng Jiang . . . . . . . . . . . . . . . . . . . . . 11341

*It Is Not Only the Negative that Deserves Attention! Understanding, Generation & Evaluation of (Positive) Moderation*
Iman Jundi, Eva Maria Vecchi, Carlotta Quensel, Neele Falk and Gabriella Lapesa . . . . . . . 11360

*Social Norms in Cinema: A Cross-Cultural Analysis of Shame, Pride and Prejudice*
Sunny Rai, Khushang Zaveri, Shreya Havaldar, Soumna Nema, Lyle Ungar and Sharath Chandra Guntuku . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 11396

*The Stochastic Parrot on LLM's Shoulder: A Summative Assessment of Physical Concept Understanding*
Mo Yu, Lemao Liu, Junjie Wu, Tsz Ting Chung, Shunchi Zhang, Jiangnan Li, Dit-Yan Yeung and Jie Zhou . . . . . 11416

*mHumanEval - A Multilingual Benchmark to Evaluate Large Language Models for Code Generation*
Md Nishat Raihan, Antonios Anastasopoulos and Marcos Zampieri . . . . . 11432

*What Do VLMs NOTICE? A Mechanistic Interpretability Pipeline for Gaussian-Noise-free Text-Image Corruption and Evaluation*
Michal Golovanevsky, William Rudman, Vedant Palit, Carsten Eickhoff and Ritambhara Singh 11462

*Are explicit belief representations necessary? A comparison between Large Language Models and Bayesian probabilistic models*
Dingyi Pan and Ben Bergen . . . . . 11483

*Self-Generated Critiques Boost Reward Modeling for Language Models*
Yue Yu, Zhengxing Chen, Aston Zhang, Liang Tan, Chenguang Zhu, Richard Yuanzhe Pang, Yundi Qian, Xuewei Wang, Suchin Gururangan, Chao Zhang, Melanie Kambadur, Dhruv Mahajan and Rui Hou . . . . . 11499

*Characterizing the Role of Similarity in the Property Inferences of Language Models*
Juan Diego Rodriguez, Aaron Mueller and Kanishka Misra . . . . . 11515

*SimRAG: Self-Improving Retrieval-Augmented Generation for Adapting Large Language Models to Specialized Domains*
Ran Xu, Hui Liu, Sreyashi Nag, Zhenwei Dai, Yaochen Xie, Xianfeng Tang, Chen Luo, Yang Li, Joyce C. Ho, Carl Yang and Qi He . . . . . 11534

*Learning to Substitute Words with Model-based Score Ranking*
Hongye Liu and Ricardo Henao . . . . . 11551

*Multilingual Reasoning via Self-training*
Leonardo Ranaldi and Giulia Pucci . . . . . 11566

*xLAM: A Family of Large Action Models to Empower AI Agent Systems*
Jianguo Zhang, Tian Lan, Ming Zhu, Zuxin Liu, Thai Quoc Hoang, Shirley Kokane, Weiran Yao, Juntao Tan, Akshara Prabhakar, Haolin Chen, Zhiwei Liu, Yihao Feng, Tulika Manoj Awalgaonkar, Rithesh R N, Zeyuan Chen, Ran Xu, Juan Carlos Niebles, Shelby Heinecke, Huan Wang, Silvio Savarese and Caiming Xiong . . . . . 11583

*ProMQA: Question Answering Dataset for Multimodal Procedural Activity Understanding*
Kimihiro Hasegawa, Wiradee Imrattanatrai, Zhi-Qi Cheng, Masaki Asada, Susan Holm, Yuran Wang, Ken Fukuda and Teruko Mitamura . . . . . 11598

*Ethical Concern Identification in NLP: A Corpus of ACL Anthology Ethics Statements*
Antonia Karamolegkou, Sandrine Schiller Hansen, Ariadni Christopoulou, Filippos Stamatiou, Anne Lauscher and Anders Søgaard . . . . . 11618

*AdaCAD: Adaptively Decoding to Balance Conflicts between Contextual and Parametric Knowledge*
Han Wang, Archiki Prasad, Elias Stengel-Eskin and Mohit Bansal . . . . . 11636

*Are Multimodal LLMs Robust Against Adversarial Perturbations? RoMMath: A Systematic Evaluation on Multimodal Math Reasoning*
Yilun Zhao, Guo Gan, Chen Zhao and Arman Cohan . . . . . 11653

*LBC: Language-Based-Classifier for Out-Of-Variable Generalization*
Kangjun Noh, Baekryun Seong, Hoyoon Byun, Youngjun Choi, Sungjin Song and Kyungwoo Song . . . . . 11666

*On the Impact of Fine-Tuning on Chain-of-Thought Reasoning*
Elita Lobo, Chirag Agarwal and Himabindu Lakkaraju . . . . . 11679

*InfoPO: On Mutual Information Maximization for Large Language Model Alignment*
Teng Xiao, Zhen Ge, Sujay Sanghavi, Tian Wang, Julian Katz-Samuels, Marc Versage, Qingjun Cui and Trishul Chilimbi . . . . . 11699

*Is In-Context Learning a Type of Error-Driven Learning? Evidence from the Inverse Frequency Effect in Structural Priming*
Zhenghao Zhou, Robert Frank and R. Thomas McCoy . . . . . 11712

*Guiding Medical Vision-Language Models with Diverse Visual Prompts: Framework Design and Comprehensive Exploration of Prompt Variations*
Kangyu Zhu, Ziyuan Qin, Huahui Yi, Zekun Jiang, Qicheng Lao, Shaoting Zhang and Kang Li 11726

*Analyzing and Improving Coherence of Large Language Models in Question Answering*
Ivano Lauriola, Stefano Campese and Alessandro Moschitti . . . . . 11740

*ALinFiK: Learning to Approximate Linearized Future Influence Kernel for Scalable Third-Parity LLM Data Valuation*
Yanzhou Pan, Huawei Lin, Yide Ran, Jiamin Chen, Xiaodong Yu, Weijie Zhao, Denghui Zhang and Zhaozhuo Xu . . . . . 11756

*E-Gen: Leveraging E-Graphs to Improve Continuous Representations of Symbolic Expressions*
Hongbo Zheng, Suyuan Wang, Neeraj Gangwar and Nickvash Kani . . . . . 11772

*Robust and Unbounded Length Generalization in Autoregressive Transformer-Based Text-to-Speech*
Eric Battenberg, RJ Skerry-Ryan, Daisy Stanton, Soroosh Mariooryad, Matt Shannon, Julian Salazar and David Teh-Hwa Kao . . . . . 11789

*PromptOptMe: Error-Aware Prompt Compression for LLM-based MT Evaluation Metrics*
Daniil Larionov and Steffen Eger . . . . . 11807

*AutoParLLM: GNN-guided Context Generation for Zero-Shot Code Parallelization using LLMs*
Quazi Ishtiaque Mahmud, Ali TehraniJamsaz, Hung D Phan, Le Chen, Mihai Capotă, Theodore L. Willke, Nesreen K. Ahmed and Ali Jannesari . . . . . 11821

*Causally Modeling the Linguistic and Social Factors that Predict Email Response*
Yinuo Xu, Hong Chen, Sushrita Rakshit, Aparna Ananthasubramaniam, Omkar Yadav, Mingqian Zheng, Michael Jiang, Lechen Zhang, Bowen Yi, Kenan Alkiek, Abraham Israeli, Bangzhao Shu, Hua Shen, Jiaxin Pei, Haotian Zhang, Miriam Schirmer and David Jurgens . . . . . 11842

*AI-LieDar : Examine the Trade-off Between Utility and Truthfulness in LLM Agents*
Zhe Su, Xuhui Zhou, Sanketh Rangreji, Anubha Kabra, Julia Mendelsohn, Faeze Brahman and Maarten Sap . . . . . 11867

*Beyond the Safety Bundle: Auditing the Helpful and Harmless Dataset*
Khaoula Chehbouni, Jonathan Colaço Carr, Yash More, Jackie CK Cheung and Golnoosh Farnadi 11895

*FollowIR: Evaluating and Teaching Information Retrieval Models to Follow Instructions*
Orion Weller, Benjamin Chang, Sean MacAvaney, Kyle Lo, Arman Cohan, Benjamin Van Durme, Dawn Lawrie and Luca Soldaini . . . . . . . . . . 11926

*Few-shot Personalization of LLMs with Mis-aligned Responses*
Jaehyung Kim and Yiming Yang . . . . . . . . . . 11943

*Prompting with Phonemes: Enhancing LLMs' Multilinguality for Non-Latin Script Languages*
Hoang H Nguyen, Khyati Mahajan, Vikas Yadav, Julian Salazar, Philip S. Yu, Masoud Hashemi and Rishabh Maheshwary . . . . . . . . . . 11975

*SHADES: Towards a Multilingual Assessment of Stereotypes in Large Language Models*
Margaret Mitchell, Giuseppe Attanasio, Ioana Baldini, Miruna Clinciu, Jordan Clive, Pieter Delobelle, Manan Dey, Sil Hamilton, Timm Dill, Jad Doughman, Ritam Dutt, Avijit Ghosh, Jessica Zosa Forde, Carolin Holtermann, Lucie-Aimée Kaffee, Tanmay Laud, Anne Lauscher, Roberto L Lopez-Davila, Maraim Masoud, Nikita Nangia, Anaelia Ovalle, Giada Pistilli, Dragomir Radev, Beatrice Savoldi, Vipul Raheja, Jeremy Qin, Esther Ploeger, Arjun Subramonian, Kaustubh Dhole, Kaiser Sun, Amirbek Djanibekov, Jonibek Mansurov, Kayo Yin, Emilio Villa Cueva, Sagnik Mukherjee, Jerry Huang, Xudong Shen, Jay Gala, Hamdan Al-Ali, Tair Djanibekov , Nurdaulet Mukhituly, Shangrui Nie, Shanya Sharma, Karolina Stanczak, Eliza Szczechla, Tiago Timponi Torrent, Deepak Tunuguntla, Marcelo Viridiano, Oskar Van Der Wal, Adina Yakefu, Aurélie Névéol, Mike Zhang, Sydney Zink and Zeerak Talat . . . . . . . . . . 11995

*Speculative Diffusion Decoding: Accelerating Language Generation through Diffusion*
Jacob K Christopher, Brian R. Bartoldson, Tal Ben-Nun, Michael Cardei, Bhavya Kailkhura and Ferdinando Fioretto . . . . . . . . . . 12042

*Bayelemabaga: Creating Resources for Bambara NLP*
Allahsera Auguste Tapo, Kevin Assogba, Christopher M Homan, M. Mustafa Rafique and Marcos Zampieri . . . . . . . . . . 12060

*Single Ground Truth Is Not Enough: Adding Flexibility to Aspect-Based Sentiment Analysis Evaluation*
Soyoung Yang, Hojun Cho, Jiyoung Lee, Sohee Yoon, Edward Choi, Jaegul Choo and Won Ik Cho . . . . . . . . . . 12071

*DREAM: Disentangling Risks to Enhance Safety Alignment in Multimodal Large Language Models*
Jianyu Liu, Hangyu Guo, Ranjie Duan, Xingyuan Bu, Yancheng He, Shilong Li, Hui Huang, Jiaheng Liu, Yucheng Wang, Chenchen Jing, Xingwei Qu, Xiao Zhang, Pei Wang, Yanan Wu, Jihao Gu, Yangguang Li and Jianke Zhu . . . . . . . . . . 12097

*In-Context Learning with Long-Context Models: An In-Depth Exploration*
Amanda Bertsch, Maor Ivgi, Emily Xiao, Uri Alon, Jonathan Berant, Matthew R. Gormley and Graham Neubig . . . . . . . . . . 12119

*Preference Consistency Matters: Enhancing Preference Learning in Language Models with Automated Self-Curation of Training Corpora*
JoonHo Lee, JuYoun Son, Juree Seok, Wooseok Jang and Yeong-Dae Kwon . . . . . . . . . . 12150

*TurtleBench: A Visual Programming Benchmark in Turtle Geometry*
Sina Rismanchian, Yasaman Razeghi, Sameer Singh and Shayan Doroudi . . . . . . . . . . 12170

*Automatically Discovering How Misogyny is Framed on Social Media*
Rakshitha Rao Ailneni and Sanda M. Harabagiu . . . . . . . . . . 12189

*Faithful, Unfaithful or Ambiguous? Multi-Agent Debate with Initial Stance for Summary Evaluation*
Mahnaz Koupaee, Jake W. Vincent, Saab Mansour, Igor Shalyminov, Han He, Hwanjun Song, Raphael Shu, Jianfeng He, Yi Nian, Amy Wing-mei Wong, Kyu J. Han and Hang Su . . . . . . . . . . 12209

*ReIFE: Re-evaluating Instruction-Following Evaluation*
Yixin Liu, Kejian Shi, Alexander Fabbri, Yilun Zhao, PeiFeng Wang, Chien-Sheng Wu, Shafiq Joty and Arman Cohan . . . . . 12247

*Language Models Predict Empathy Gaps Between Social In-groups and Out-groups*
Yu Hou, Hal Daumé Iii and Rachel Rudinger . . . . . 12288

*HARP: Hesitation-Aware Reframing in Transformer Inference Pass*
Romain Storaï and Seung-won Hwang . . . . . 12305

*JAWAHER: A Multidialectal Dataset of Arabic Proverbs for LLM Benchmarking*
Samar Mohamed Magdy, Sang Yun Kwon, Fakhraddin Alwajih, Safaa Taher Abdelfadil, Shady Shehata and Muhammad Abdul-Mageed . . . . . 12320

*EmojiPrompt: Generative Prompt Obfuscation for Privacy-Preserving Communication with Cloud-based LLMs*
Sam Lin, Wenyue Hua, Zhenting Wang, Mingyu Jin, Lizhou Fan and Yongfeng Zhang . . . . 12342

*MICE for CATs: Model-Internal Confidence Estimation for Calibrating Agents with Tools*
Nishant Subramani, Jason Eisner, Justin Svegliato, Benjamin Van Durme, Yu Su and Sam Thomson . . . . . 12362

*PAT: Parameter-Free Audio-Text Aligner to Boost Zero-Shot Audio Classification*
Ashish Seth, Ramaneswaran Selvakumar, Sonal Kumar, Sreyan Ghosh and Dinesh Manocha12376

*Language Model Council: Democratically Benchmarking Foundation Models on Highly Subjective Tasks*
Justin Zhao, Flor Miriam Plaza-del-Arco and Amanda Cercas Curry . . . . . 12395

*SCIURus: Shared Circuits for Interpretable Uncertainty Representations in Language Models*
Carter Teplica, Yixin Liu, Arman Cohan and Tim G. J. Rudner . . . . . 12451

*ProSE: Diffusion Priors for Speech Enhancement*
Sonal Kumar, Sreyan Ghosh, Utkarsh Tyagi, Anton Jeran Ratnarajah, Chandra Kiran Reddy Evuru, Ramani Duraiswami and Dinesh Manocha . . . . . 12470

*Mastering the Craft of Data Synthesis for CodeLLMs*
Meng Chen, Philip Arthur, Qianyu Feng, Cong Duy Vu Hoang, Yu-Heng Hong, Mahdi Kazemi Moghaddam, Omid Nezami, Duc Thien Nguyen, Gioacchino Tangari, Duy Vu, Thanh Vu, Mark Johnson, Krishnaram Kenthapadi, Don Dharmasiri, Long Duong and Yuan-Fang Li . . . . . 12484

*ParaICL: Towards Parallel In-Context Learning*
Xingxuan Li, Xuan-Phi Nguyen, Shafiq Joty and Lidong Bing . . . . . 12501

*CausalEval: Towards Better Causal Reasoning in Language Models*
Longxuan Yu, Delin Chen, Siheng Xiong, Qingyang Wu, Dawei Li, Zhikai Chen, Xiaoze Liu and Liangming Pan . . . . . 12512

*Layer-Level Self-Exposure and Patch: Affirmative Token Mitigation for Jailbreak Attack Defense*
Yang Ouyang, Hengrui Gu, Shuhang Lin, Wenyue Hua, Jie Peng, Bhavya Kailkhura, Meijun Gao, Tianlong Chen and Kaixiong Zhou . . . . . 12541

*DeCAP: Context-Adaptive Prompt Generation for Debiasing Zero-shot Question Answering in Large Language Models*
Suyoung Bae, YunSeok Choi and Jee-Hyong Lee . . . . . 12555

*Reward-Guided Tree Search for Inference Time Alignment of Large Language Models*
Chia-Yu Hung, Navonil Majumder, Ambuj Mehrish and Soujanya Poria . . . . . . . . . . . . . . . . . 12575

*Typographic Attacks in a Multi-Image Setting*
Xiaomeng Wang, Zhengyu Zhao and Martha Larson . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 12594

*Tonguescape: Exploring Language Models Understanding of Vowel Articulation*
Haruki Sakajo, Yusuke Sakai, Hidetaka Kamigaito and Taro Watanabe . . . . . . . . . . . . . . . . . . 12605

*CoRAC: Integrating Selective API Document Retrieval with Question Semantic Intent for Code Question Answering*
YunSeok Choi, CheolWon Na and Jee-Hyong Lee . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 12620

*Pipeline Analysis for Developing Instruct LLMs in Low-Resource Languages: A Case Study on Basque*
Ander Corral, Ixak Sarasua Antero and Xabier Saralegi . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 12636

*How to Make LLMs Forget: On Reversing In-Context Knowledge Edits*
Paul Youssef, Zhixue Zhao, Jörg Schlötterer and Christin Seifert . . . . . . . . . . . . . . . . . . . . . . . 12656

*PerCul: A Story-Driven Cultural Evaluation of LLMs in Persian*
Erfan Moosavi Monazzah, Vahid Rahimzadeh, Yadollah Yaghoobzadeh, Azadeh Shakery and Mohammad Taher Pilehvar . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 12670

*Towards Sustainable NLP: Insights from Benchmarking Inference Energy in Large Language Models*
Soham Poddar, Paramita Koley, Janardan Misra, Niloy Ganguly and Saptarshi Ghosh . . . . . 12688

*CSR-Bench: Benchmarking LLM Agents in Deployment of Computer Science Research Repositories*
Yijia Xiao, Runhui Wang, Luyang Kong, Davor Golac and Wei Wang . . . . . . . . . . . . . . . . . . . 12705

*SALAD: Improving Robustness and Generalization through Contrastive Learning with Structure-Aware and LLM-Driven Augmented Data*
Suyoung Bae, YunSeok Choi, Hyojun Kim and Jee-Hyong Lee . . . . . . . . . . . . . . . . . . . . . . . . . 12724

*Rationale-Guided Retrieval Augmented Generation for Medical Question Answering*
Jiwoong Sohn, Yein Park, Chanwoong Yoon, Sihyeon Park, Hyeon Hwang, Mujeen Sung, Hyunjae Kim and Jaewoo Kang . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 12739

*Prototype Conditioned Generative Replay for Continual Learning in NLP*
Xi Chen and Min Zeng . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 12754

*KODIS: A Multicultural Dispute Resolution Dialogue Corpus*
James Anthony Hale, Sushrita Rakshit, Kushal Chawla, Jeanne M Brett and Jonathan Gratch 12771