

GEM 2026

**The Fifth Workshop on Natural Language Generation,
Evaluation, and Metrics (GEM)**

Proceedings of the Workshop

Co-located with the 64th Annual Meeting of the Association for Computational Linguistics
(ACL)

July 4, 2026

©2026 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 979-8-89176-423-1

Introduction

Welcome to the fifth iteration of the Generation, Evaluation & Metrics series workshop, GEM 2026! This year, GEM is co-located with the 64th Annual Meeting of the Association for Computational Linguistics (ACL 2026) in San Diego, California, USA on July 4th.

GEM brings together researchers and practitioners to tackle the hard problem of meaningful, efficient, and robust evaluation of large language models (LLMs). Evaluation of language models has grown to be a central theme in NLP research, while remaining far from solved. As language models (LMs) have become more powerful, errors are tougher to spot and systems harder to distinguish. Evaluation practices are evolving rapidly – from living benchmarks like Chatbot Arena to LMs being used as evaluators themselves (e.g., LM as judge, autoraters). Further research is needed to understand the interplay between metrics, benchmarks, and human-in-the-loop evaluation, and their impact in real-world settings.

This year’s edition builds on the success of earlier GEM workshops at ACL 2021, EMNLP 2022, EMNLP 2023 and ACL 2025. Similarly to last year, the workshop co-hosts the ReprONLP shared task on reproducibility of evaluations. For the first time, GEM includes a special track for opinion and statement papers.

We received a total of 124 submissions: 101 direct, 8 ReprONLP and 15 via ARR commitments. Of these 124, 6 were withdrawn, 28 were rejected, and 90 manuscripts were accepted for presentation: 69/101 direct, 8/8 ReprONLP, 13/15 ARR. From the 90 accepted submissions, we have 76 regular archival papers (including 5 opinion pieces, 8 ReprONLP papers and 13 ARR papers) and 14 non-archival extended abstracts. 5 archival papers were shortlisted for oral presentations based on the scores received in the reviews and meta-reviews. The technical programme was made possible by 141 reviewers who volunteered their time and expertise and 17 area chairs, who oversaw the meta-review process.

The final GEM programme features three keynote talks, oral and poster presentations (the latter includes the presentation of additional Findings papers), and the ReprONLP overview session. We are grateful to the ACL conference organisers for their support, and to the ReprONLP team, our reviewers and area chairs for their important work on the programme composition. Finally, we thank all the authors and participants for presenting their novel ideas and engaging in lively discussions. We are looking forward to meeting you all at an interesting and pleasant event.

GEM Workshop Organizers

Organizing Committee

Organizing Committee

Simon Mille, ADAPT, Dublin City University
Sebastian Gehrmann, Bloomberg
Patrícia Schmidtová, Charles University
Ondřej Dušek, Charles University
Marzieh Fadaee, Cohere
Kyle Lo, Allen Institute for AI
Enrico Santus, Bloomberg
Gabriel Stanovsky, Hebrew University

Program Committee

Program Chairs

Simon Mille, ADAPT, Dublin City University
Sebastian Gehrmann, Bloomberg
Patrícia Schmidtová, Charles University
Ondřej Dušek, Charles University
Marzieh Fadaee, Cohere
Kyle Lo, Allen Institute for AI
Enrico Santus, Bloomberg
Gabriel Stanovsky, Hebrew University

Area Chairs

Abigail Walsh
Alba Táboas García, Universitat Pompeu Fabra
Alexander Shvets, Barcelona Supercomputing Center
Ashley Lewis, Ohio State University, Columbus
David M Howcroft, University of Aberdeen
Eduardo Calò
Fahime Same, Trivago N.V.
Kanishk Verma
Kaustubh Dhole, Emory University
Mateusz Lango, Charles University
Mayank Jobanputra
Michela Lorandi, Dublin City University
Rudali Huidrom, Indraprastha Institute of Information Technology, Delhi
Silvia Casola
Simon Mille
Vilém Zouhar, Department of Computer Science, ETHZ - ETH Zurich
William Soto Martinez

Reviewers

Jane Adkins, Karmanya Aggarwal, Shima Asaadi

Sanket Badhe, Vanya Bannihatti Kumar, Priyam Basu, Anya Belz, Subrata Biswas, Bernd Bohnet

Michele Cafagna, Stefano Campese, Jasper Kyle Catapang, Pallabi Chakraborty, Yogen Vilas Chaudhari, Divya Chaudhary, Long Chen, Ji Yong Cho, Tahiya Chowdhury, Miruna Clinciu, Jordan Clive

Sahil Rajesh Dhayalkar, Abhishek Divekar, Srimonti Dutta, Ondřej Dušek

Cornelius Emde

Naghme Farzi

Jugal Gajjar, Cristina Garbacea, Sankalp Gilda, Leander Girrback, Aditi Gupta

Sil Hamilton, Adeep Hande, Davan Harrison, Harshavardhan, Mujtaba Hasan, Jiayi He, Xanh Ho, Soheil Hor, Kaili Huang, Tianyi Huang, Patrick Huber

Yuki Ichihara, Marina Igitkhanian, Nikolai Ilinykh, Alvi Md Ishmam, Anna A Ivanova

Rafid Ishrak Jahan, Sankalp Jajee, Anubhav Jangra, Chathuri Jayaweera, Pooja Jhunjunwala, Yangfeng Ji, Yanru Jiang, Shailza Jolly

Emil Kalbaliyev, Hsien-Te Kao, Janak Kapuriya, Charu Karakkaparambil James, Marzena Karpinska, Ali Keramati, Sergey Kovalchuk, Arjun Krishna, Pawan Kumar, Rishu Kumar, Sachin Kumar, Yonghoon Kwon

Lujun LI, Ian Lane, Alberto Lavelli, Jaeyun Lee, Jing Yang Lee, Włodzimierz Lewoniewski, Siyan Li, Yinghui Li, Weixin Liu, Zefang Liu, Giuliano Lorenzoni, Ehsan Lotfi, Fabian Lukassen, Maria Lymperaiou

Khyati Mahajan, Saad Mahamood, Abinaya Mahendiran, Siddarth Malreddy, Sushant Mehta, Guangyu Meng, Avni Mittal, Sebastien Montella

Tapas Nayak, Kosuke Nishida, Tadashi Nomoto, Erfan Nourbakhsh

Hamidah Oderinwale, Kristýna Onderková

Cheoneum Park, Yongsin Park, Tatiana Passali, Maitrik Patel, Natalie Perez, Jan-Thorsten Peter, Dina Pisarevskaya, Priya Pitre, Maja Popovic

Yinzhu Quan

Anand A. Rajasekar, Viswanathan Ranganathan, Ehud Reiter, Leonardo F. R. Ribeiro, Daniele Riboni, Fabien Ringeval, Juan Diego Rodriguez, Allen G Roush

Tasfia Seuti, Kalash Shah, Rifat Shahriyar, Raghav Sharma, Tianhao Shen, S. A. I. Shouborno, Brady Steele

Anh Ta, Craig Thomson, Rabin Tiwari, Ekaterina Trofimova

Rajendra Ugrani, Srihari Unnikrishnan

Madison Van Doren, Emiel van Miltenburg

Cong Wang, Yike Wang, Zirui Wei

Yusuke Yamauchi, Bing Yan, Guanqun Yang, Xinchun Yang, Zhuoyi Yang, Asaf Yehudai, Yan-chao Yu

Wajdi Zaghouani, Alessandra Zarcone, Yongxin Zhou

Keynote Talk

Follow the Evidence: Diagnosing the What, Where, and Why of Generative Model Failures

Vered Shwartz
University of British Columbia



Time: **09:10** – Room: **Harbor E&F**

Abstract: Generative models are improving at a remarkable pace, yet our ability to evaluate them is struggling to keep up. Proprietary, black-box, and frequently updated models limit us to evaluating outputs rather than understanding what shapes them. Meanwhile, as model outputs grow more polished, their errors become more subtle, and we resort to increasingly relying on models themselves as evaluators, with their own blind spots and biases. In this talk, I will discuss three case studies that together reveal the limitations of current evaluation practices. First, I will present Spotlight, a benchmark for fine-grained localization of errors in generated videos. We show that VLMs used as evaluators substantially lag behind humans, missing real errors while hallucinating non-existent ones. Second, I will present Value Drifts, a systematic evaluation that looks inside the LLM post-training process and finds that contrary to common belief, it is supervised fine-tuning — not preference optimization — that most shapes a model’s value profile. Third, I will discuss ongoing work on investigating what factors determine whether a multilingual LLM can answer a question about a fact acquired in one language when prompted in another. Across all three case studies, a common thread emerges: the gap between what our evaluations measure and how models may behave “in the wild” is wider than it appears.

Bio: **Vered Shwartz** is an Assistant Professor of Computer Science at the University of British Columbia, a CIFAR AI Chair at the Vector Institute, and the author of “Lost in Automatic Translation: Navigating Life in English in the Age of Language Technologies”. Her current research focus is on (1) testing and improving the capabilities of large language models and vision and language models; (2) developing culturally-competent AI; and (3) responsible NLP applications in sensitive domains (e.g., legal, medical). Before joining UBC, she was a postdoctoral researcher at the Allen Institute for AI (AI2) and the University of Washington. Prior to that, she completed her PhD in Computer Science at Bar-Ilan University.

Keynote Talk
**Small Samples, Big Reveal: What can we learn from limited
observations of language model behavior?**

Swabha Swayamdipta
University of Southern California



Time: **15:00** – Room: **Harbor E&F**

Abstract: The majority of popular language models today are both large-scale and close-sourced, making studying their behavior quite challenging. This talk tries to answer how much we can learn from limited observations of language model behavior. First, we show that language models can be reliably evaluated using even randomly selected microbenchmarks of a certain size. Second, we use language model outputs, i.e. next-token probability distributions, to build prompt inversion attacks to reveal hidden prompts with high accuracy. These findings highlight the importance of scientific research into large language models without access to large computation resources, while still allowing accountability for the providers, as well as efficient and reliable evaluation.

Bio: Swabha Swayamdipta is an Assistant Professor of Computer Science and a co-Associate Director of the Center for AI and Society at the University of Southern California. Her research interests are in natural language processing and machine learning, with a primary interest in the evaluation of generative models of language, understanding the behavior of language models, and designing language technologies for societal good. At USC, Swabha leads the Data, Interpretability, Language and Learning (DILL) Lab. She received her PhD from Carnegie Mellon University, followed by a postdoc at the Allen Institute for AI and the University of Washington. Her work has received outstanding paper awards at EMNLP 2024, ICML 2022, NeurIPS 2021 and ACL 2020. Her research is supported by awards from the NIH, NSF, Apple, the Allen Institute for AI, Intel Labs, the Zumberge Foundation and a WiSE Gabilan Fellowship.

Keynote Talk

Autorubric: A Unified Framework for Rubric-Based LLM Evaluation

Chris Callison-Burch
University of Pennsylvania



Time: **16:40** – Room: **Harbor E&F**

Abstract: LLM-as-a-judge has become the default for evaluating open-ended generation, but the approach is riddled with silent failure modes, including position bias, verbosity bias, criterion conflation, sycophancy, and run-to-run inconsistency, that corrupt judgments without any visible signal. Mitigations exist, scattered across the LM-as-judge literature and decades of work in psychometrics and educational measurement, but every research group ends up paying a “Reinvention Tax,” reimplementing option shuffling, ensemble voting, calibration, and reliability metrics from scratch.

I will present Autorubric, an open-source framework that consolidates these best practices into a single library with opinionated defaults: analytic per-criterion decomposition, mixed criterion types, ensemble judging, length penalties, and a full suite of psychometric reliability metrics. Beyond measurement, Autorubric’s mandatory per-criterion explanations function as “textual gradients” for two downstream applications: rubric-guided prompt induction and RL with rubric rewards. Autorubric is available at <https://autorubric.org/>.

Bio: **Chris Callison-Burch** is the Raj and Neera Singh Professor of Artificial Intelligence at the University of Pennsylvania, where he directs the online Master’s in AI and teaches Penn Engineering’s flagship AI course to more than 500 students each fall. In 2026 he received the Lindback Award for Distinguished Teaching, Penn’s highest teaching honor. He chairs the advisory board for the Human Language Technology Center of Excellence at Johns Hopkins University. He testified before Congress in 2023 on generative AI and copyright law, and in 2026 participated in the Isaac Asimov Memorial Debate at the American Museum of Natural History, moderated by Neil deGrasse Tyson. He has authored more than 200 publications with over 36,000 citations, and is a Sloan Research Fellow with research support from DARPA, IARPA, NSF, and industry partners including Google, Microsoft, and Amazon.

Table of Contents

<i>CoSy: Conversational Synthesis for Grounded Question Answering</i> Patrick Huber, Arash Einolghozati, Rylan Conway, Kanika Narang, Matt Smith, Waqar Nayyar, Adithya Sagar, Ahmed A Aly and Akshat Shrivastava	1
<i>VAIDYA: Validated Agents for Intelligent Diagnosis and Yielded Analysis</i> Kalash Shah, Gautam Bhutani, Rohitaswa Sarbhangia and J Snehan	11
<i>Self-Anchoring Calibration Drift in Large Language Models: How Multi-Turn Conversations Reshape Model Confidence</i> Harshavardhan	34
<i>Temporal Tokenization Strategies for Event Sequence Modeling with Large Language Models</i> Zefang Liu, Nam H Nguyen, Yinzhu Quan and Shi-Xiong Zhang	41
<i>“Be My Cheese?”: Cultural Nuance Benchmarking for Machine Translation in Multilingual LLMs</i> Madison Van Doren, Casey Ford, Jennifer Barajas, Riley VanMeter and Cory Holland	52
<i>Component Transfer Can Exceed Full Model Performance: Investigating Post-Trained Mixture-of-Experts</i> Rabin Tiwari	77
<i>Reassessing Extractive QA Datasets at Scale: LLM-as-a-Judge and In-Depth Analyses</i> Xanh Ho, Jiahao Huang, Florian Boudin and Akiko Aizawa	84
<i>IndicMMLU-Pro: Benchmarking Indic Large Language Models on Multi-Task Language Understand- ing</i> Sankalp Jajee, Ashutosh Kumar, Nikunj Kotecha, Vinija Jain, Aman Chadha and Sreyoshi Bhaduri 102	
<i>Identifying Where Large Language Models Struggle in Answering Complex Questions</i> Xanh Ho, Florian Boudin, Saku Sugawara, Khoa Duong and Akiko Aizawa	112
<i>More Yap Less Meaning: Uncovering Self-Improvement Behavior in SLMs</i> Marina Igithkhanian and Erik Arakelyan	124
<i>Reinforced Agent: Inference-Time Feedback for Tool-Calling Agents</i> Anh Ta, Junjie Zhu and Shahin Shayandeh	136
<i>RE-AD: Real-Time Requirement Adherence for Data Labeling</i> Siddarth Malreddy, Ishan Nigam, Akshay Arora, Nikhil Mittal and Subrat Sahu	148
<i>Lost in Space: Finding the Right Tokens for Structured Output</i> Sil Hamilton and David Mimno	155
<i>An Empirical Study of LLM-as-a-Judge: How Design Choices Impact Evaluation Reliability</i> Yusuke Yamauchi, Taro Yano and Masafumi Oyamada	167
<i>Capturing Epistemic Uncertainty in LLM-Based Soft Labeling</i> Yanru Jiang and Siyu Liang	177
<i>Mind the Gap... or Not? How Translation Errors and Evaluation Details Skew Multilingual Results</i> Jan-Thorsten Peter, David Vilar, Tobias Domhan, Dan Malkin and Markus Freitag	191
<i>MCJudgeBench: A Benchmark for Constraint-Level Judge Evaluation in Multi-Constraint Instruction Following</i> Jaeyun Lee, Junyoung Koh, Zeynel Tok, Hunar Batra and Ronald Clark	205

<i>MedAct: Removing the Human Bottleneck in Benchmarking Clinical LLM Safety</i>	
Arjun Krishna, Brian Pridgen and Max Silverstein	222
<i>Response Content Units: Evaluating Completeness and Proactiveness in Medical Open-Response Question Answering</i>	
Yongsin Park, Wen-wai Yim, Emma McKibbin, Asma Ben Abacha and Fei Xia	231
<i>NanoFlux: Adversarial Dual-LLM Evaluation and Distillation for Multi-Domain Reasoning</i>	
Raviteja Anantha, Soheil Hor, Teodor Nicola Antoniu and Layne C Price	253
<i>Evaluating the Reliability of LLMs in Faithfully Updating Text: An Empirical Study</i>	
Ayan Datta, Paheli Bhattacharya and Rishabh Gupta	271
<i>Not All Tokens Are Equal: Per-Dimension Top-K Pooling for Adversarially Robust BERT Classification</i>	
Manoranjan Dash, Shivam Anand Aralikatti, Shanay Sheth and Pranav Shinde	285
<i>Near-Miss: Latent Policy Failure Detection in Agentic Workflows</i>	
Ella Rabinovich, David Boaz, Naama Zwerdling and Ateret Anaby Tavor	296
<i>Evaluating Counterfactual Strategic Reasoning in Large Language Models</i>	
Dimitrios Georgousis, Maria Lymperaiou, Angeliki Dimitriou, Giorgos Filandrianos and Giorgos Stamou	309
<i>Speculative Refinement: A Hybrid Autoregressive Diffusion Decoding Strategy and Its Behavior Across Benchmarks</i>	
Aditi Gupta, Neel Mishra, Kushagra Trivedi and Pawan Kumar	355
<i>SAUCE: Summary Analysis Using Conversation Entailment</i>	
Man-Ling Sung, Hemanth Kandula, Jeff Ma, William Hartmann and Matthew Snover	364
<i>Evaluating ASR Quality at Scale on TV Entertainment Platforms</i>	
Adeep Hande, Kishorekumar Sundararajan, Yidnekachew Endale, Akshatha Babu KrishnaSwamy, Sachin Dabral, Dawn Reed and Michael Pereira	378
<i>Fine-Tuning vs. RAG for Multi-Hop Question Answering with Novel Knowledge</i>	
Zhuoyi Yang, Yurun Song, Kyler G. Harris, Iftekhar Ahmed and Ian Harris	384
<i>MHGraphBench: Knowledge Graph-Grounded Benchmarking of Mental Health Knowledge in Large Language Models</i>	
Weixin Liu, Congning Ni, Shelagh A. Mulvaney, Susannah L. Rose, Murat Kantarcioglu, Bradley A. Malin and Zhijun Yin	393
<i>A Progressive Evaluation Framework for Multicultural Analysis of Story Visualization</i>	
Janak Kapuriya, Ali Hatami and Paul Buitelaar	410
<i>Is GraphRAG Needed? From Basic RAG to Graph-/Agentic Solutions with Context Optimization</i>	
Long Chen, Ryan Razkenari, Yuxuan Zhou, Yuan Tian, Rahul Ghosh, Venkatesh Pappakrishnan, Disha Ahuja and Vidya Sagar Ravipati	428
<i>Cross-Domain Semantic Fidelity Evaluation for Meaning-to-Text Generation</i>	
Davan Harrison and Marilyn Walker	443
<i>E-star 12B: Reliable Rubric-Following and Domain-Adaptive SLM Evaluator for Korean Industrial Settings</i>	
Yonghoon Kwon, Heondeuk Lee and Barom Kang	456

<i>Pressure-Testing Deception Probes in LLMs: Scaling, Robustness, and the Geometry of Deceptive Representations</i>	
Sachin Kumar	472
<i>Sycophancy Negatively Affects LLM-as-a-Judge in Conflict Evaluation</i>	
Naghmeh Farzi, Laura Dietz and Samuel Carton	490
<i>Concord: An Agreement-Aware Multi-Adjudication Pipeline for LLM Evaluation</i>	
Tyler Bliss, Mahit Verma, Aila Iyer-Singh, Subrata Biswas, Sheikh Asif Imran and Bashima Islam	502
<i>The Silent Vote: Improving Zero-Shot LLM Reliability by Aggregating Semantic Neighborhoods</i>	
Sanket Badhe, Priyanka Tiwari and Deep Shah	511
<i>Are LLM Benchmarks Already Contaminated? A Systematic Review of Contamination Detection Methods</i>	
Erfan Nourbakhsh, Mohammad Sadegh Sirjani, Amir Mousavi, Khoa Nguyen, John Quarles, Mimi Xie and Rocky Slavin	518
<i>RBCorr: Response Bias Correction in Language Models</i>	
Om Bhatt and Anna A Ivanova	540
<i>Exploring Coherence of LLMs in Multilingual Question Answering</i>	
Stefano Campese and Ivano Lauriola	554
<i>Token Cost Inequality: Measuring Tokenization Disparities Across Scripts in Roman Urdu and Urdu</i>	
Waleed Jamil, Saima Rafi and Yanchao Yu	563
<i>Semantic vs. Structural Signals: Log-Probability and LLM-as-a-Judge for Reference-Free Code Evaluation</i>	
Dmitriy Fedrushkov, Yulong He, Ivan Smirnov, Artem Aliev and Sergey Kovalchuk	574
<i>Stability vs. Manipulability: Evaluating Robustness Under Post-Decision Interaction in LLM Judges</i>	
Srimonti Dutta and Akshata Kishore Moharir	582
<i>Permutation-Consensus Listwise Judging for Robust Factuality Evaluation</i>	
Tianyi Huang, Nathan Huang, Justin Tang, Wenqian Chen and Elsa Fan	595
<i>MedFact: Benchmarking the Fact-Checking Capabilities of Large Language Models on Chinese Medical Texts</i>	
Jiayi He, Yangmin Huang, Qianyun Du, Xiangying Zhou, Zhiyang He, Jiaxue Hu, Xiaodong Tao and Lixian Lai	604
<i>Early-Token Confidence Predicts Reasoning Quality in Multi-Agent LLM Debate</i>	
Ali Keramati, Justin Cheok, Jacob Horne and Mark Warschauer	653
<i>Complex-IF and Beyond: Expert Rubrics for RLVR</i>	
Sushant Mehta, Liudas Panavas, Eleanor Fleming, Paul Mains and Edwin Chen	668
<i>C2-Faith: Benchmarking LLM Judges for Causal and Coverage Faithfulness in Chain-of-Thought Reasoning</i>	
Avni Mittal and Rauno Arike	678
<i>Evaluating Multilingual Sentiment Classifiers Using an LLM-Annotated Wikipedia Benchmark</i>	
Milena Stróżyńska, Włodzimierz Lewoniewski and Izabela Czumalowska	692
<i>Process Standardisation for Human Evaluation of NLP System Outputs</i>	
Craig Thomson, Javier González Corbelle and Anya Belz	704

<i>Language Modeling for the Future of Finance: A Survey into Metrics, Tasks, and Data Opportunities</i> Nikita Tatarinov, Siddhant Sukhani, Agam Shah and Sudheer Chava	718
<i>WildIFEval: Instruction Following in the Wild</i> Gili Lior, Asaf Yehudai, Ariel Gera and Liat Ein-Dor	745
<i>EconWebArena: Benchmarking Autonomous Agents on Economic Tasks in Realistic Web Environments</i> Zefang Liu and Yinzhu Quan	779
<i>ISO-Bench: Benchmarking Multimodal Causal Reasoning in Visual–Language Models through Procedural Plans</i> Ananya Sadana, Yash Kumar Lal and Jiawei Zhou	797
<i>Text Analytics Evaluation Framework: A Case Study on LLMs and Social Media</i> Yuefeng Shi, Nedjma Ousidhoum and Jose Camacho-Collados	808
<i>Teaching Values to Machines: Simulating Human-Like Behavior in LLMs</i> Asaf Yehudai, Naama Rozen and Ariel Gera	825
<i>MetaGraph: A Large-Scale Meta-Analysis of GenAI in Financial NLP (2022–2025)</i> Paolo Pedinotti, Peter Baumann, Nathan Jessurun, Leslie Barrett and Enrico Santus	848
<i>When Users Are Happy but Agents Are Wrong: Multi-Dimensional Evaluation of Tool-Augmented Dialogue</i> Tanya Shourya, Yingfan Wang, Zhaoyi Joey Hou, Shamik Roy, Vinayshekhar Bannihatti Kumar and Rashmi Gangadharaiah	862
<i>Tool-Aware Planning for Contact-Center Analytics: Evaluating LLMs through Lineage-Guided Query Decomposition</i> Varun Nathan, Shreyas Guha and Ayush Kumar	893
<i>TSAQA: Time Series Analysis Question And Answering Benchmark</i> Baoyu Jing, Sanhorn Chen, Lecheng Zheng, Boyu Liu, Zihao Li, Jiaru Zou, Tianxin Wei, Zhining Liu, Zhichen Zeng, Ruizhong Qiu, Xiao Lin, Yuchen Yan, Dongqi Fu, Jingchao Ni, Jingrui He and Hanghang Tong	944
<i>Who Endorsed It? Measuring Authority Bias Across Expertise Levels in Language Models</i> Priyanka Mary Mammen, Emil Joswin and Shankar Venkitachalam	980
<i>Reference Games as a Testbed for the Alignment of Model Uncertainty and Clarification Requests</i> Manar Ali, Judith Sieker, Sina Zarrieß and Hendrik Buschmeier	990
<i>Mapping Out the NLP Evaluation Landscape with a Standard Taxonomy of Quality Criteria</i> Anya Belz, Simon Mille and Craig Thomson	999
<i>Position: Toward a Metric Typology for Language Model Evaluation</i> Jasper Kyle Catapang	1015
<i>Position: What Are We Measuring? Rethinking Evaluation in Natural Language Generation</i> Wajdi Zaghouani	1021
<i>Position: Evaluation Scores Are Perishable Knowledge Claims</i> Sankalp Gilda and Shlok Gilda	1029
<i>Position: A Semiotic-Hermeneutic Approach to Evaluating Meaning in LLM Summaries via the Inductive Conceptual Rating Metric</i> Natalie Perez, Sreyoshi Bhaduri and Aman Chadha	1036

<i>Position: Scores Without Context? Rethinking the Role of Evaluation in the Era of LLMs</i>	
Jiawei Zhou	1048
<i>The Shared Task on Reproducibility of Evaluations in NLP (ReproNLP) 2026: Overview and Results</i>	
Anya Belz, Craig Thomson and Javier González Corbelle	1055
<i>Do Nugget-Based Evaluation Patterns Generalize to List-QA?</i>	
MohammadJavad Ardestani, Ehsan Kamaloo and Davood Rafiei	1071
<i>ReproNLP 2026: A Third Replication of the Human Evaluation of a QAG System for Children’s Story-books</i>	
Marcel Mroczek, Chiara Albarello, Paul-Emmanuel Floch and Maciej Gawinecki	1082
<i>ReproHum #0124-03: Reproducing Human Scores on Neural REG Models</i>	
Maurice Langner	1094
<i>ReproHum #0866-04: Variability in Human Judgments of Sociopolitical Acceptability Across Studies</i>	
Rui Fan and Guanyi Chen	1104
<i>ReproHum #0031–01: Reproducing a Human Readability Evaluation for Question–Answer Generation Systems</i>	
Manuela Hürlimann and Mark Cieliebak	1111
<i>ReproHum #0033-05: Human Evaluation Report on Generating Scientific Definitions with Controllable Complexity"</i>	
Ines Arous and Jackie Chi Kit Cheung	1117
<i>ReproHum #0669-08: Reproducing a Recipe for Arbitrary Text Style Transfer with LLMs</i>	
Saad Mahamood	1127

Program

Saturday, July 4, 2026

08:55 - 09:10	<i>Opening remarks</i>
09:10 - 09:50	<i>Invited talk #1: Vered Shwartz</i>
09:50 - 10:05	<i>Oral presentation #1</i>
10:05 - 10:20	<i>Oral presentation #2</i>
10:20 - 10:50	<i>Coffee break</i>
10:50 - 11:15	<i>ReproNLP Shared Task Overview</i>
11:15 - 11:30	<i>Oral presentation #3</i>
11:35 - 12:35	<i>Poster session #1</i>
12:35 - 13:55	<i>Lunch break</i>
13:55 - 14:55	<i>Poster session #2</i>
15:00 - 15:40	<i>Invited talk #2: Swabha Swayamdipta</i>
15:40 - 16:10	<i>Coffee break</i>
16:10 - 16:25	<i>Oral presentation #4</i>
16:25 - 16:40	<i>Oral presentation #5</i>
16:40 - 17:20	<i>Invited talk #3: Chris Callison-Burch</i>
17:20 - 17:30	<i>Closing session</i>

Saturday, July 4, 2026 (continued)