GEM 2026

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

Proceedings of the Workshop

July 4, 2026

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 Of these 124, 6 were withdrawn, 28 were rejected, and 90 manuscripts were accepted for presentation: 69/101 direct, 8/8 ReproNLP, 13/15 ARR.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

Naghmeh Farzi

Jugal Gajjar, Cristina Garbacea, Sankalp Gilda, Leander Girrbach, 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 Jhunjhunwala, 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, Xinchen Yang, Zhuoyi Yang, Asaf Yehudai, Yanchao 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

Program

Saturday, July 4, 2026

08:55 - 09:10 *Opening remarks*

09:10 - 09:50 *Invited talk #1: Vered Shwartz*

09:50 - 10:05 *Oral presentation #1*

10:05 - 10:20 *Oral presentation #2*

10:20 - 10:50 *Coffee break*

10:50 - 11:15 *ReproNLP Shared Task Overview*

11:15 - 11:30 *Oral presentation #3*

11:35 - 12:35 *Poster session #1*

12:35 - 13:55 *Lunch break*

13:55 - 14:55 *Poster session #2*

15:00 - 15:40 *Invited talk #2: Swabha Swayamdipta*

15:40 - 16:10 *Coffee break*

16:10 - 16:25 *Oral presentation #4*

16:25 - 16:40 *Oral presentation #5*

16:40 - 17:20 *Invited talk #3: Chris Callison-Burch*

17:20 - 17:30 *Closing session*