ACL 2025

The 63rd Annual Meeting of the Association for Computational Linguistics

Proceedings of the GEM² Workshop

July 31 - August 1, 2025

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 979-8-89176-261-9

Introduction

Welcome to the **GEM^2 Workshop at ACL 2025**! The fourth iteration of the Generation, Evaluation & Metrics series brings together researchers and practitioners to tackle the hard problem of *meaningful, efficient, and robust* evaluation of large language models (LLMs). GEM^2 is co-located with the 63rd Annual Meeting of the Association for Computational Linguistics (ACL 2025) in Vienna, Austria and online, from July 31 to August 1, 2025.

Building on the success of earlier GEM workshops at ACL 2021, EMNLP 2022, and EMNLP 2023, this edition introduces two large-scale prediction benchmarks—**DOVE** and **DataDecide**—and co-hosts the **ReproNLP** shared task on reproducibility of evaluations. These resources aim to spur research on prompt robustness, cost-effective benchmarking, and principled comparison of LLM outputs.

We received a total of **108 submissions**.Of these, **79 manuscripts were accepted for presentation** and **29 were rejected**.The exact breakdown into archival papers (**68**), and non-archival abstracts (**11**).

The technical programme was made possible by **106** reviewers who volunteered their time and expertise and **10** area chairs, who oversaw the meta-review process;

GEM^2 spans two days and features keynote talks, oral and poster presentations, an Industrial Track panel, and the ReproNLP results session. We are grateful to the conference organisers for their support in running a fully hybrid event.

Organising Team

- **Workshop Chairs**: Ofir Arviv, Miruna Clinciu, Kaustubh Dhole, Rotem Dror, Sebastian Gehrmann, Eliya Habba, Itay Itzhak, Simon Mille, Enrico Santus, João Sedoc, Michal Shmueli Scheuer, Gabriel Stanovsky,Yotam Perlitz, Oyvind Tafjord

Acknowledgements We thank the ACL 2025 organising committee, the ReproNLP team, our reviewers and area chairs, and the sponsors who provided travel grants. Finally, we are indebted to all authors for their enthusiastic participation—your work is at the heart of GEM^2.

Organizing Committee

Workshop Chairs

Ofir Arviv, IBM Research
Miruna Clinciu, Heriot-Watt University
Kaustubh Dhole, Emory University
Rotem Dror, University of Haifa
Sebastian Gehrmann, Bloomberg
Eliya Habba, Hebrew University of Jerusalem
Itay Itzhak, Hebrew University of Jerusalem
Simon Mille, Dublin City University
Yotam Perlitz, IBM Research
Enrico Santus, Bloomberg
João Sedoc, New York University
Michal Shmueli Scheuer, IBM Research
Gabriel Stanovsky, Hebrew University of Jerusalem
Oyvind Tafjord, Allen Institute for Artificial Intelligence

Program Committee

Program Chairs

Ofir Arviv, International Business Machines
Anya Belz, Dublin City University
Miruna Clinciu
Kaustubh Dhole, Emory University
Rotem Dror, University of Haifa
Sebastian Gehrmann, Bloomberg
Itay Itzhak
Simon Mille
Yotam Perlitz, International Business Machines
Enrico Santus
João Sedoc, New York University
Gabriel Stanovsky, Hebrew University of Jerusalem
Craig Thomson, Dublin City University and University of Aberdeen

Area Chairs

Ofir Arviv, International Business Machines
Anya Belz, Dublin City University
Hila Gonen, University of Washington
Javier González Corbelle, Universidad de Santiago de Compostela
John P. Lalor, University of Notre Dame
Simon Mille
Yotam Perlitz, International Business Machines
Vered Shwartz
Craig Thomson, Dublin City University and University of Aberdeen
Charles Welch, McMaster University

Reviewers

Samuel Ackerman, Noof Alfear, Anuoluwapo Aremu, Samee Arif, Shima Asaadi

Simone Balloccu, Nirajan Bekoju, Anya Belz, Noga BenYoash, Paheli Bhattacharya, Marc Brysbaert

Pengshan Cai, Silvia Casola, Miruna Clinciu, Jordan Clive, Jane Arleth Dela Cruz

Amin Dada, Daniel Deutsch, Jing Ding, Susana Sotelo Docio, Ondrej Dusek

Micha Elsner

Nils Feldhus, Lucie Flek, Martin Forell

Ioana Giurgiu, John Glover, Evangelia Gogoulou, Javier González Corbelle

Behnam Hedayatnia, David M Howcroft, Kaili Huang, Shulin Huang, Rudali Huidrom

Nikolai Ilinykh

Yuu Jinnai, Mayank Jobanputra, Minsuh Joo, Brihi Joshi

Emil Kalbaliyev, Jihyun Kim, Juae Kim, Yekyung Kim, Frederic Kirstein, Sergey Kovalchuk, Saurabh Kulshreshtha

Alberto Lavelli, Hwanhee Lee, Jing Yang Lee, Yinghui Li, Xiaoyu Lin, Yixin Liu, Michela Lorandi, Ehsan Lotfi, Nishant Luitel

Vittesh Maganti, Khyati Mahajan, Saad Mahamood, Potsawee Manakul, Andreas Marfurt, Gonzalo Martínez, Sebastien Montella, Seyed Mahed Mousavi

Tapas Nayak, Joakim Nivre, Naveen Jafer Nizar, Tadashi Nomoto

Soham Kamlesh Parikh, Cheoneum Park, Tatiana Passali, Diogo Pernes, Dina Pisarevskaya, Jiashu Pu

Mostafa Rahgouy, Nishant Raj, Vikas Raunak, Ehud Reiter, Fabien Ringeval, Sean Rooney

Isik Baran Sandan, Sashank Santhanam, Somdeb Sarkhel, Asad B. Sayeed, Patrícia Schmidtová, Monika Shah, Samira Shaikh, Samira Shaikh, Tatiana Shavrina, Tianhao Shen, Barkavi Sundararajan

Sotaro Takeshita, Katherine Thai, Craig Thomson, Cagri Toraman, Yuma Tsuta

Emiel Van Miltenburg, Anastasia Voznyuk

Zhengxiang Wang, Genta Indra Winata

Bing Yan, Guanqun Yang, Yao Yao

Alessandra Zarcone, Xinyue Zhang, Justin Zhao, Yongxin Zhou

Table of Contents

Program

Thursday, July 31, 2025

09:00 - 10:25 *Opening Remarks & Keynotes by Barbara Plank (Ambiguity, Consistency and Reasoning in LLMs) and Leshem Choshen (Evaluation at the Heart of the AI Wave)*

10:25 - 10:55 *Coffee Break*

10:55 - 11:30 *Talk Session 1: Anya Belz – (ReproNLP Shared Task Overview)*

10:55 - 11:30 *Talk Session 1: Minsuh Joo – Cleanse: Uncertainty Estimation Approach Using Clustering-based Semantic Consistency in LLMs*

11:30 - 12:30 *Poster Session Part 1*

12:30 - 14:00 *Lunch Break*

14:00 - 15:00 *Poster Session Part 2*

15:00 - 15:30 *Talk Session 2: Joshi Brihi, Sriram Venkatapathy, Mohit Bansal, Nanyun Peng, Haw-Shiuan Chang – CoKe: Customizable Fine-Grained Story Evaluation via Chain-of-Keyword Rationalization*

15:00 - 15:30 *Talk Session 2: Junho Myung, Yeon Su, Sunwoo Kim, Shin Yoo, Alice Oh – PapersPlease: A Benchmark for Evaluating Motivational Values of Large Language Models Based on ERG Theory*

15:30 - 16:00 *Coffee Break*

16:00 - 16:15 *Talk Session 3: Javier Conde, Miguel Gonzalez, Maria Grandury, Pedro Reviriego, Gonzalo Martinez, Marc Brysbaert – Psycholinguistic Word Features: a New Approach for the Evaluation of LLMs Alignment with Humans*

16:15 - 16:55 *Keynote by Ehud Reiter (We Should Evaluate Real-World Impact)*

16:55 - 17:40 *Panel Discussion*

17:40 - 17:50 *Closing Remarks*