

ACL 2025

**The 63rd Annual Meeting of the Association for
Computational Linguistics**

Proceedings of the GEM² Workshop

July 31 - August 1, 2025

©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 979-8-89176-261-9

Introduction

Introduction

Welcome to the **GEM² 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). GEM² 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;

GEM² 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 GEM².

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

<i>Towards Comprehensive Evaluation of Open-Source Language Models: A Multi-Dimensional, User-Driven Approach</i>	
Qingchen Yu	1
<i>Psycholinguistic Word Features: a New Approach for the Evaluation of LLMs Alignment with Humans</i>	
Javier Conde, Miguel González Saiz, María Grandury, Pedro Reviriego, Gonzalo Martínez and Marc Brysbaert	8
<i>Spatial Representation of Large Language Models in 2D Scene</i>	
WenyaWu WenyaWu and Weihong Deng	18
<i>The Fellowship of the LLMs: Multi-Model Workflows for Synthetic Preference Optimization Dataset Generation</i>	
Samee Arif, Sualeha Farid, Abdul Hameed Azeemi, Awais Athar and Agha Ali Raza	30
<i>Does Biomedical Training Lead to Better Medical Performance?</i>	
Amin Dada, Osman Alperen Koraş, Marie Bauer, Jean-Philippe Corbeil, Amanda Butler Contreras, Constantin Marc Seibold, Kaleb E Smith, julian.friedrich@uk-essen.de julian.friedrich@uk-essen.de and Jens Kleesiek	46
<i>HEDS 3.0: The Human Evaluation Data Sheet Version 3.0</i>	
Anya Belz and Craig Thomson	60
<i>ARGENT: Automatic Reference-free Evaluation for Open-Ended Text Generation without Source Inputs</i>	
Xinyue Zhang, Agathe Zecevic, Sebastian Zeki and Angus Roberts	82
<i>Are LLMs (Really) Ideological? An IRT-based Analysis and Alignment Tool for Perceived Socio-Economic Bias in LLMs</i>	
Jasmin Wachter, Michael Radloff, Maja Smolej and Katharina Kinder-Kurlanda	99
<i>Knockout LLM Assessment: Using Large Language Models for Evaluations through Iterative Pairwise Comparisons</i>	
Isik Baran Sandan, Tu Anh Dinh and Jan Niehues	121
<i>Free-text Rationale Generation under Readability Level Control</i>	
Yi-Sheng Hsu, Nils Feldhus and Sherzod Hakimov	129
<i>Selective Shot Learning for Code Explanation</i>	
Paheli Bhattacharya and Rishabh Gupta	151
<i>Can LLMs Detect Intrinsic Hallucinations in Paraphrasing and Machine Translation?</i>	
Evangelia Gogoulou, Shorouq Zahra, Liane Guillou, Luise Dürlich and Joakim Nivre	161
<i>Evaluating LLMs with Multiple Problems at once</i>	
Zhengxiang Wang, Jordan Kodner and Owen Rambow	178
<i>Learning and Evaluating Factual Clarification Question Generation Without Examples</i>	
Matthew Toles, Yukun Huang and Zhou Yu	200
<i>SECQUE: A Benchmark for Evaluating Real-World Financial Analysis Capabilities</i>	
Noga BenYoash, Menachem Brief, Oded Ovadia, Gil Shenderovitz, Moshik Mishaelli, Rachel Lemberg and Eitam Sheetrit	212

<i>Measure only what is measurable: towards conversation requirements for evaluating task-oriented dialogue systems</i>	
Emiel Van Miltenburg, Anouck Braggaar, Emmelyn Croes, Florian Kunneman, Christine Liebrecht and Gabriella Martijn	231
<i>Can Perplexity Predict Finetuning Performance? An Investigation of Tokenization Effects on Sequential Language Models for Nepali</i>	
Nishant Luitel, Nirajan Bekoju, Anand Kumar Sah and Subarna Shakya	239
<i>Are Bias Evaluation Methods Biased ?</i>	
Lina Berrayana, Sean Rooney, Luis Garcés-Erice and Ioana Giurgiu	249
<i>IRSum: One Model to Rule Summarization and Retrieval</i>	
Sotaro Takeshita, Simone Paolo Ponzetto and Kai Eckert	262
<i>Modeling the One-to-Many Property in Open-Domain Dialogue with LLMs</i>	
Jing Yang Lee, Kong Aik Lee and Woon-Seng Gan	276
<i>Cleanse: Uncertainty Estimation Approach Using Clustering-based Semantic Consistency in LLMs</i>	
Minsuh Joo and Hyunsoo Cho	291
<i>Metric assessment protocol in the context of answer fluctuation on MCQ tasks</i>	
Ekaterina Goliakova, Xavier Renard, Marie-Jeanne Lesot, Thibault Laugel, Christophe Marsala and Marcin Detyniecki	302
<i>(Towards) Scalable Reliable Automated Evaluation with Large Language Models</i>	
Bertil Braun and Martin Forell	320
<i>Clustering Zero-Shot Uncertainty Estimations to Assess LLM Response Accuracy for Yes/No Q&A</i>	
Christopher T. Franck, Amy Vennos, W. Graham Mueller and Daniel Dakota	337
<i>Using LLM Judgements for Sanity Checking Results and Reproducibility of Human Evaluations in NLP</i>	
Rudali Huidrom and Anya Belz	354
<i>CoKe: Customizable Fine-Grained Story Evaluation via Chain-of-Keyword Rationalization</i>	
Brihi Joshi, Sriram Venkatapathy, Mohit Bansal, Nanyun Peng and Haw-Shiuan Chang	366
<i>HuGME: A benchmark system for evaluating Hungarian generative LLMs</i>	
Noémi Ligeti-Nagy, Gabor Madarasz, Flora Foldesi, Mariann Lengyel, Matyas Osvath, Bence Sarossy, Kristof Varga, Győző Zijian Yang, Enikő Héja, Tamás Váradi and Gábor Prószéky	385
<i>Judging the Judges: Evaluating Alignment and Vulnerabilities in LLMs-as-Judges</i>	
Aman Singh Thakur, Kartik Choudhary, Venkat Srinik Ramayapally, Sankaran Vaidyanathan and Dieuwke Hupkes	404
<i>Analyzing the Sensitivity of Vision Language Models in Visual Question Answering</i>	
Monika Shah, Sudarshan Balaji, Somdeb Sarkhel, Sanorita Dey and Deepak Venugopal	431
<i>Investigating the Robustness of Retrieval-Augmented Generation at the Query Level</i>	
Sezen Perçin, Xin Su, Qutub Sha Syed, Phillip Howard, Aleksei Kuvshinov, Leo Schwinn and Kay-Ulrich Scholl	439
<i>ELAB: Extensive LLM Alignment Benchmark in Persian Language</i>	
Zahra Pourbahman, Fatemeh Rajabi, Mohammadhossein Sadeghi, Omid Ghahroodi, Somayeh Bakhshaei, Arash Amini, Reza Kazemi and Mahdih Soleymani Baghshah	458

<i>Evaluating the Quality of Benchmark Datasets for Low-Resource Languages: A Case Study on Turkish</i> Elif Ecem Umutlu, Ayse Aysu Cengiz, Ahmet Kaan Sever, Seyma Erdem, Burak Aytan, Busra Tufan, Abdullah Topraksoy, Esra Darıcı and Cagri Toraman	471
<i>Big Escape Benchmark: Evaluating Human-Like Reasoning in Language Models via Real-World Escape Room Challenges</i> Zinan Tang and QiYao Sun	488
<i>Event-based evaluation of abstractive news summarization</i> Huiling You, Samia Touileb, Lilja Øvrelid and Erik Velldal	504
<i>Fine-Tune on the Format: First Improving Multiple-Choice Evaluation for Intermediate LLM Checkpoints</i> Alec Bunn, Sarah Wiegrefe and Ben Bogin	511
<i>PapersPlease: A Benchmark for Evaluating Motivational Values of Large Language Models Based on ERG Theory</i> Junho Myung, Yeon Su Park, Sunwoo Kim, Shin Yoo and Alice Oh	522
<i>Shallow Preference Signals: Large Language Model Aligns Even Better with Truncated Data?</i> Xuan Qi, Jiahao Qiu, Xinzhe Juan, Yue Wu and Mengdi Wang	532
<i>Improving Large Language Model Confidence Estimates using Extractive Rationales for Classification</i> Jane Arleth Dela Cruz, Iris Hendrickx and Martha Larson	549
<i>ReproHum #0729-04: Human Evaluation Reproduction Report for MemSum: Extractive Summarization of Long Documents Using Multi-Step Episodic Markov Decision Processes"</i> Simeon Junker	561
<i>ReproHum #0744-02: A Reproduction of the Human Evaluation of Meaning Preservation in "Factoring Meaning and Form for Intent-Preserving Paraphrasing"</i> Julius Steen and Katja Markert	568
<i>ReproHum #0031-01: Reproducing the Human Evaluation of Readability from It is AI's Turn to Ask Humans a Question"</i> Daniel Braun	576
<i>ReproHum #0033-05: Human Evaluation of Factuality from A Multidisciplinary Perspective</i> Andra-Maria Florescu, Marius Micluta-Câmpeanu, Stefana Arina Tabusca and Liviu P Dinu	583
<i>ReproHum: #0744-02: Investigating the Reproducibility of Semantic Preservation Human Evaluations</i> Mohammad Arvan and Natalie Parde	590
<i>ReproHum #0669-08: Reproducing Sentiment Transfer Evaluation</i> Kristýna Onderková, Mateusz Lango, Patrícia Schmidtová and Ondrej Dusek	601
<i>ReproHum #0067-01: A Reproduction of the Evaluation of Cross-Lingual Summarization</i> Supryadi , Chuang Liu and Deyi Xiong	609
<i>ReproHum #0729-04: Partial reproduction of the human evaluation of the MemSum and NeuSum summarisation systems</i> Simon Mille and Michela Lorandi	615
<i>Curse of bilinguality: Evaluating monolingual and bilingual language models on Chinese linguistic benchmarks</i> Yuwen Zhou and Yevgen Matusevych	622

<i>Towards Better Open-Ended Text Generation: A Multicriteria Evaluation Framework</i>	
Esteban Garces Arias, Hannah Blocher, Julian Rodemann, Meimingwei Li, Christian Heumann and Matthias Aßenmacher	631
<i>Bridging the LLM Accessibility Divide? Performance, Fairness, and Cost of Closed versus Open LLMs for Automated Essay Scoring</i>	
Kezia Oketch, John P. Lalor, Yi Yang and Ahmed Abbasi	655
<i>Prompt, Translate, Fine-Tune, Re-Initialize, or Instruction-Tune? Adapting LLMs for In-Context Learning in Low-Resource Languages</i>	
Christopher Toukmaji and Jeffrey Flanigan	670
<i>Winning Big with Small Models: Knowledge Distillation vs. Self-Training for Reducing Hallucination in QA Agents</i>	
Ashley Lewis	705
<i>Ad-hoc Concept Forming in the Game Codenames as a Means for Evaluating Large Language Models</i>	
Sherzod Hakimov, Lara Pfennigschmidt and David Schlangen	728
<i>Evaluating Intermediate Reasoning of Code-Assisted Large Language Models for Mathematics</i>	
Zena Al Khalili, Nick Howell and Dietrich Klakow	741
<i>From Calculation to Adjudication: Examining LLM Judges on Mathematical Reasoning Tasks</i>	
Andreas Stephan, Dawei Zhu, Matthias Aßenmacher, Xiaoyu Shen and Benjamin Roth	759
<i>PersonaTwin: A Multi-Tier Prompt Conditioning Framework for Generating and Evaluating Personalized Digital Twins</i>	
Sihan Chen, John P. Lalor, Yi Yang and Ahmed Abbasi	774
<i>Coreference as an indicator of context scope in multimodal narrative</i>	
Nikolai Ilinykh, Shalom Lappin, Asad B. Sayeed and Sharid Loáiciga	789
<i>PATCH! Psychometrics-AssisTed BenCHmarking of Large Language Models against Human Populations: A Case Study of Proficiency in 8th Grade Mathematics</i>	
Qixiang Fang, Daniel Oberski and Dong Nguyen	808
<i>MCQFormatBench: Robustness Tests for Multiple-Choice Questions</i>	
Hiroo Takizawa, Saku Sugawara and Akiko Aizawa	824
<i>(Dis)improved?! How Simplified Language Affects Large Language Model Performance across Languages</i>	
Miriam Anschütz, Anastasiya Damaratskaya, Chaeun Joy Lee, Arthur Schmalz, Edoardo Mosca and Georg Groh	847
<i>Fine-Grained Constraint Generation-Verification for Improved Instruction-Following</i>	
Zhixiang Liang, Zhenyu Hou and Xiao Wang	862
<i>Finance Language Model Evaluation (FLaME)</i>	
Glenn Matlin, Mika Okamoto, Huzafa Pardawala, Yang Yang and Sudheer Chava	880
<i>sPhinX: Sample Efficient Multilingual Instruction Fine-Tuning Through N-shot Guided Prompting</i>	
Sanchit Ahuja, Kumar Tanmay, Hardik Hansrajibhai Chauhan, Barun Patra, Kriti Aggarwal, Luciano Del Corro, Arindam Mitra, Tejas Indulal Dhamecha, Ahmed Hassan Awadallah, Monojit Choudhury, Vishrav Chaudhary and Sunayana Sitaram	927
<i>Single- vs. Dual-Prompt Dialogue Generation with LLMs for Job Interviews in Human Resources</i>	
Joachim De Baer, A. Seza Doğruöz, Thomas Demeester and Chris Develder	947

<i>Natural Language Counterfactual Explanations in Financial Text Classification: A Comparison of Generators and Evaluation Metrics</i>	
Karol Dobiczek, Patrick Altmeyer and Cynthia C. S. Liem	958
<i>An Analysis of Datasets, Metrics and Models in Keyphrase Generation</i>	
Florian Boudin and Akiko Aizawa	973
<i>U-MATH: A University-Level Benchmark for Evaluating Mathematical Skills in Large Language Models</i>	
Konstantin Chernyshev, Vitaliy Polshkov, Vlad Stepanov, Alex Myasnikov, Ekaterina Artemova, Alexei Miasnikov and Sergei Tilga	974
<i>The 2025 ReprONLP Shared Task on Reproducibility of Evaluations in NLP: Overview and Results</i>	
Anya Belz, Craig Thomson, Javier González Corbelle and Malo Ruelle	1002

Program

Thursday, July 31, 2025

09:00 - 10:25	<i>Opening Remarks & Keynotes by Barbara Plank (Ambiguity, Consistency and Reasoning in LLMs) and Leshem Choshen (Evaluation at the Heart of the AI Wave)</i>
10:25 - 10:55	<i>Coffee Break</i>
10:55 - 11:30	<i>Talk Session 1: Anya Belz – (ReproNLP Shared Task Overview)</i>
10:55 - 11:30	<i>Talk Session 1: Minsuh Joo – Cleanse: Uncertainty Estimation Approach Using Clustering-based Semantic Consistency in LLMs</i>
11:30 - 12:30	<i>Poster Session Part 1</i>
12:30 - 14:00	<i>Lunch Break</i>
14:00 - 15:00	<i>Poster Session Part 2</i>
15:00 - 15:30	<i>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</i>
15:00 - 15:30	<i>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</i>
15:30 - 16:00	<i>Coffee Break</i>
16:00 - 16:15	<i>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</i>
16:15 - 16:55	<i>Keynote by Ehud Reiter (We Should Evaluate Real-World Impact)</i>
16:55 - 17:40	<i>Panel Discussion</i>
17:40 - 17:50	<i>Closing Remarks</i>