Eval4NLP 2025

The 5th Workshop on Evaluation and Comparison of NLP Systems

Proceedings of the Workshop

December 23, 2025

The Eval4NLP organizers gratefully acknowledge the support from the following sponsors.

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 979-8-89176-305-0

Introduction

Welcome to the Fifth Workshop on Evaluation and Comparison of NLP Systems (Eval4NLP 2025).
The current year has brought further astonishing achievements in NLP. Generative large language models (LLMs) like ChatGPT, Gemini, or LLama, continue to demonstrate wide capabilities in understanding and performing tasks from in-context descriptions without fine-tuning, bringing worldwide attention to the risks and opportunities that arise from current and ongoing research.
Given the ever growing speed of research, fair evaluations and comparisons are of fundamental importance to the NLP community in order to properly track progress. This concerns the creation of benchmark datasets that cover typical use cases and blind spots of existing systems, the designing of metrics for evaluating the performance of NLP systems on different dimensions, and the reporting of evaluation results in an unbiased manner.
We believe that new insights and methodologies, particularly in the recent years, have led to much renewed interest in the workshop topic. The first workshop in the series, Eval4NLP'20, was the first workshop to take a broad and unifying perspective on the subject matter. The second (Eval4NLP'21), third (Eval4NLP'22) and fourth (Eval4NLP'23) workshop extended this perspective. Our fifth workshop continues this tradition of being a reputed platform for presenting and discussing latest advances in NLP evaluation methods and resources.
Our workshop attracted a lot of attention from the research community. Among the 35 submissions, 14 were accepted for presentation after thorough consideration by the program committee (yielding an acceptance rate of 40
We would like to thank all of the authors for their contributions, the program committee for their thoughtful reviews, the keynote speaker for sharing their perspective, and all the attendees for their participation. We believe that all of these will contribute to a lively and successful workshop. Looking forward to meeting you all at Eval4NLP 2025!
Eval4NLP 2025 Organizing Committee, Mousumi Akter, Tahiya Chowdhury, Erion Çano, Juri Opitz, Christoph Leiter, Steffen Eger

Organizing Committee

Program Chairs

Mousumi Akter, Technische Universität Dortmund

Program Chairs

Tahiya Chowdhury, Colby College

Program Chairs

Christoph Leiter, University of Mannheim

Program Chairs

Juri Opitz, University of Zurich

Program Chairs

Erion Çano, University of Bochum

Program Chairs

Steffen Eger, University of Technology Nuremberg

Program Committee

Program Chairs

Mousumi Akter, Technische Universität Dortmund
Tahiya Chowdhury, Colby College
Steffen Eger, University of Technology Nuremberg
Christoph Leiter, Universität Mannheim
Juri Opitz, University of Zurich
Erion Çano, Ruhr-Universität Bochum

Keynote Talk
Invited 1

Iryna Gurevych
Technical University of Darmstadt
2025-12-23 10:00:00-10:45:00 – Room: **online**

Bio: Iryna Gurevych is a Professor of Computer Science at TU Darmstadt, with additional appointments at MBZUAI and INSAIT. She directs the Ubiquitous Knowledge Processing (UKP) Lab and co-directs the ELLIS Natural Language Processing program. Gurevych is a Fellow of ELLIS (2019) and the ACL (2020), a member of the German National Academy of Sciences Leopoldina and the Berlin-Brandenburg Academy of Sciences and Humanities (BBAW), and the inaugural recipient of the LOEWE Spitzenprofessur (2021).
She has received numerous prestigious honors, including the ERC Advanced Grant (2022) and the Milner Award (2025). Gurevych's research spans Natural Language Processing, Machine Learning, Multimodal Data Analysis, Digital Humanities, and Computational Social Science. She has led several major research initiatives—including CEDIFOR, the AIPHES research training group, and the CA-SG "Content Analytics for the Social Good" program—and served as President of the Association for Computational Linguistics.

Table of Contents

Program

Tuesday, December 23, 2025

09:00 - 09:10 *Opening Remarks*

09:10 - 10:00 *Session 1*

Fair Play in the Newsroom: Actor-Based Filtering Gender Discrimination in Text Corpora
Stefanie Urchs, Veronika Thurner, Matthias Aßenmacher, Christian Heumann and Stephanie Thiemichen

Reliable Inline Code Documentation with LLMs: Fine-Grained Evaluation of Comment Quality and Coverage
Rohan Patil, Gaurav Tirodkar and Shubham Gatfane

Measuring Visual Understanding in Telecom domain: Performance Metrics for Image-to-UML conversion using VLMs
H. G. Ranjani and Rutuja Prabhudesai

10:00 - 10:45 *Invited Talk - Speaker: Prof. Dr. Iryna Gurevych*

10:45 - 11:30 *Session 2*

Evaluation of Generated Poetry
David Mareček, Kateřina Motalík Hodková, Tomáš Musil and Rudolf Rosa

Simulating Training Data Leakage in Multiple-Choice Benchmarks for LLM Evaluation
Naila Shafirni Hidayat, Muhammad Dehan Al Kautsar, Alfan Farizki Wicaksono and Fajri Koto

Test Set Quality in Multilingual LLM Evaluation
Chalamalasetti Kranti, Gabriel Bernier-Colborne, Yvan Gauthier and Sowmya Vajjala

11:30 - 11:35 *Break*

11:35 - 12:30 *Session 3*

Between the Drafts: An Evaluation Framework for Identifying Quality Improvement and Stylistic Differences in Scientific Texts
Danqing Chen, Ingo Weber and Felix Dietrich

Tuesday, December 23, 2025 (continued)

SynClaimEval: A Framework for Evaluating the Utility of Synthetic Data in Long-Context Claim Verification
Mohamed Elaraby and Jyoti Prakash Maheswari

The dentist is an involved parent, the bartender is not": Revealing Implicit Biases in QA with Implicit BBQ
Aarushi Wagh and Saniya Srivastava

Beyond Tokens and Into Minds: Future Directions for Human-Centered Evaluation in Machine Translation Post-Editing
Molly Apsel, Sunil Kothari, Manish Mehta and Vasudevan Sundarababu

12:30 - 14:00 *Lunch Break*

14:00 - 15:00 *Session 4*

InFiNITE (∞): Indian Financial Narrative Inference Tasks & Evaluations
Sohom Ghosh, Arnab Maji and Sudip Kumar Naskar

Non-Determinism of "Deterministic" LLM System Settings in Hosted Environments
Berk Atıl, Sarp Aykent, Alexa Chittams, Lisheng Fu, Rebecca J. Passonneau, Evan Radcliffe, Guru Rajan Rajagopal, Adam Sloan, Tomasz Tudrej, Ferhan Ture, Zhe Wu, Lixinyu Xu and Breck Baldwin

Beyond the Rubric: Cultural Misalignment in LLM Benchmarks for Sexual and Reproductive Health
Sumon Kanti Dey, Manvi S, Zeel Mehta, Meet Shah, Unnati Agrawal, Suhani Jalota and Azra Ismail

TitleTrap: Probing Presentation Bias in LLM-Based Scientific Reviewing
Shurui Du

15:15 - 15:00 *Finding Paper Presentation - Agent-based Automated Claim Matching with Instruction-following - Dina Pisarevskaya, Arkaitz Zubiaga*

15:30 - 15:15 *Closing Session*