

EvalEval 2026

ACL 2026 Workshop on Evaluating Evaluations (EvalEval)

Proceedings of the Workshop on Evaluating Evaluations

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 979-8-89176-429-3

Introduction

Welcome to the 2026 Workshop on Evaluating Evaluations (EvalEval), held in conjunction with ACL 2026 in San Diego, CA.

This workshop brings together researchers and practitioners to examine the practice of AI evaluation, centering the tensions and collaborations between model developers and evaluation researchers. As AI systems grow increasingly capable and widely deployed, the need for rigorous, meaningful, and community-informed evaluation has never been more pressing. EvalEval provides a space to surface practical insights from across the evaluation ecosystem, spanning methodological rigor, sociotechnical perspectives, scalability, and real-world use.

This year, we received 84 submissions. After a thorough review process involving 182 reviewers and 10 area chairs, we accepted 43 papers: 6 as oral presentations and 37 as poster presentations. All submissions received meta-reviews, and decisions were made based on reviewer scores and area chair recommendations.

The workshop also features a shared task, *Every Eval Ever*, aimed at building a unified, standardized database of LLM evaluations, as well as a panel discussion bringing together model developers and evaluation researchers.

We would like to thank all authors for their submissions, our reviewers and area chairs for their careful and thoughtful evaluations, and the ACL 2026 organizing committee for their support. We are also grateful to our invited panelists for their participation and to the broader EvalEval community for their continued engagement.

EvalEval Organizing Committee

Jennifer Mickel (Co-Chair), Ichhya Pant (Co-Chair), Usman Gohar (Publication Chair), Mubashara Akhtar, Jan Batzner, Leshem Choshen, Avijit Ghosh, Michelle Lin, Zeerak Talat

Organizing Committee

Workshop Organizer

Ichhya Pant, Independent

Usman Gohar, Iowa State University

Jan Batzner, Weizenbaum Institute, Technical University Munich

Leshem Choshen, MIT, IBM Research, MIT-IBM Watson AI Lab

Jennifer Mickel, EleutherAI

Mubashara Akhtar, ETH Zurich

Avijit Ghosh, HuggingFace

Zeeraq Talat, University of Edinburgh

Michelle Lin, Mila

Program Committee

Reviewers

Baber Abbasi, Amina A. Abdu, Tawsif Ahmed, Sanchit Ahuja, Aleksandr, Panos Alexopoulos, Malihe Alikhani, Mowafak Allaham, Ahmad Mustafa Anis, Catherine Arnett, Saeid Asgari

Agathe Balayn, Nishant Balepur, Vaibhav Balloli, Monojit Banerjee, Renata Barreto, Dan Bateyko, Glen Berman, Marta Bienkiewicz, Ahana Biswas, Julian Bitterwolf, Sam Blouir, Miranda Bogen, Shamik Bose, Olivia Beyer Bruvik, Dave Buckley, Devichand Budagam

Jane Castleman, Mahasweta Chakraborti, Khaoula Chehbouni, Mingyu Chen, Jenny Chim, Sayak Chowdhury, Shivaprasad chitta

Jocelyn D'Arcy, Aman Dalmia, Elizabeth M. Daly, Ruchira Dhar, Thaïs Distinguin

Ahmed Elhady

Benjamin Fenelon

Jatin Ganhotra, Mohit Gaur, Sushant Gautam, Marissa Gerchick, Paolo Giudici, Sujata Goswami

Leif Hancox-Li, Jose Hernandez-Orallo, Michael Hind, Aris Hofmann, Brian H Hu

Anna A Ivanova

Devina Jain, Alexander Jameson, Ayrton San Joaquin, Nari Johnson

Yatima Kagurazaka, Deniz Karabacak, Navreet Kaur, Ryan Othniel Kearns, Drew Keller, Wm. Matthew Kennedy, Dayeon Ki, Kimon Kieslich, Haein Kong, Anastassia Kornilova, Alex Korolev, Sasikanth Kotti, Satyapriya Krishna

John P. Lalor, En-Shiun Annie Lee, Hanwool Lee, Yukyung Lee, Nicole Lemke, Michelle Lin, Dongqi Liu

Temina Madon, Khyati Mahajan, Yifan Mai, David Manheim, Tasmiah Tahsin Mayeesha, Harry Mayne, Anoop Mishra, Diganta Misra, Daniela Muhaj, Tanmoy Mukherjee, Namrata Mukhija, Seph mard

Ayush Nangia, Gauri Nayak, Isar Nejadgholi, Vera Neplenbroek, Duy K. Nguyen

Justin Olive, Oluwagbemike Olowe, Robert On

Kevin Paeth, Roya Pakzad, Koyena Pal, Patricia Paskov, Aashkaben Kalpesh Patel, Valerio Pepe, David Huu Pham, Sashank Pisupati, Heila Precel

FATMA-ZOHRRA REZKELLAH, Sunny Rai, Harsh Raj, Jyoutir Raj, Mitali Raj, Deepika Raman, Aishwarya Ramasethu, Varsha Ramineni, Anka Reuel, Keith Richie, Michael Alexander Riegler, Noah Ringler, Markelle Roesti, Luis Fernando Ramirez Ruiz

PARTHA PRATIM SAHA, Pouya Sadeghi, Subramanyam Sahoo, Shreyashkar Lal Sahu, Andrew Samo, Jeba Sania, Daniel Schofield, Robert Scholz, Indira Sen, Preethi Seshadri, Colin Sheablymyer, Imama Shehzad, Imama Shehzad, Olivia Shoemaker, Amita Shukla, Maryam Sikander, Scott Simmons, Aarush Sinha, Anna Sokol, Leon Stauffer, Ryan Steed, Lily Stelling, Ilan Strauss, Nathan Suri

Vassil Tashev, Tuesday

Anahita Valakche

Andreas Waldis, Stephanie Wang, Matthew Wilde, Kyra Wilson, Cherry Wu, Zezhen Wu

Srishti Yadav, Shannon Yang, Qinyuan Ye, Evelyn Yee, Asaf Yehudai, Cheng Yu, Arda Yüksel

Table of Contents

<i>Rigorous Interpretation Is a Form of Evaluation</i>	
Isabelle Lee, Emmy Liu, Cathy Jiao, Brihi Joshi, Dani Yogatama, Fazl Barez and Michael Saxon	1
<i>Evaluating Multi-turn Human-AI Interaction</i>	
Shi Ding and Sijian Tan	12
<i>Guidelines for Whom? Rethinking AI Ethics in Resource-Constrained Migration Services</i>	
Nari Yoo, Ashley Khor, Namrata Mukhija, Aminat Adebiyi and Miri Zilka	19
<i>Evaluating Large Language Model News Sentiment in Finance under Liquidity and Market Frictions</i>	
Kemal Kirtac	26
<i>From Wordle to Fibble⁵: Evaluating LLM Reasoning Under Escalating Deception</i>	
Chang Liu	36
<i>Mind the Gap: How Elicitation Protocols Shape the Stated-Revealed Preference Gap in Language Models</i>	
Pranav Mahajan, Ihor Kendiukhov, Syed Hussain and Lydia Nottingham	46
<i>When Scanners Lie: Evaluator Instability in LLM Red-Teaming</i>	
Lidor Erez, Omer Hofman, Tamir Nizri and Roman Vainshtein	56
<i>Reasoning Model Is Superior LLM-Judge, Yet Suffers from Biases</i>	
Hui Huang, Xuanxin Wu, Muyun Yang and Yuki Arase	70
<i>From Rubrics to Recipe: Principle-Centric Benchmark for Evaluating Large Language Models</i>	
Shirley Anugrah Hayati, Ruizi Wang and Dongyeop Kang	82
<i>Too long; didn't solve</i>	
Lucía Cabrera, Jocelyn D'Arcy and Isaac Saxton-Knight	100
<i>Graduating the Benchmark Scale: Lessons from Thermometry</i>	
Sean Trott and Oisín Parkinson-Coombs	111
<i>Caged Birds and Cute Bookworms: Feminine Tropes and Implicit Gender Bias in Large Language Models</i>	
Sachita Nishal and Jack Bandy	116
<i>Scorecard of AI Benchmark Quality</i>	
Ayrton San Joaquin, Rokas Gipiškis and Ze Shen Chin	128
<i>Defining Cultural Capabilities for AI Evaluation: A Taxonomy Grounded in Intercultural Communication Theory</i>	
Isar Nejadgholi, Masoud Kianpour, Krishnapriya Vishnubhotla and Maryam Molamohammadi	161
<i>BenchNavigator: A Discovery Interface for Comparing LLM Benchmarks</i>	
Anna Sokol, Inge Vejsbjerg, Elizabeth M. Daly, David Piorkowski, Michael Hind, Nuno Moniz and Nitesh V. Chawla	174
<i>Beyond Static Benchmarks: A Validity, Reliability, and Sociotechnical Framework for Evaluating LLMs in Deployment Contexts</i>	
Ben Jenkins	201

<i>From Guidelines to Guarantees: A Graph-Based Evaluation Harness for Domain-Specific Evaluation of LLMs</i>	
Jessica M. Lundin, Usman Nasir Nakakana and Guillaume Chabot-Couture	211
<i>Document Overlap Is Not Evidence Continuity: Measuring Retrieval Jitter in Citation-Based RAG Evaluation</i>	
Punitha Ponnuraj	221
<i>Measuring AI-Induced Disempowerment: A Framework and Proposed Metrics</i>	
Je Qin Chooi, Jaeho Lee and Jasmine Xinze Li	227
<i>Position: Evaluations of AI Moral Reasoning Still Miss Half of the Picture</i>	
Aidan Kierans, Ritam Dutt, Kaley Rittichier, Shiri Dori-Hacohen and Avijit Ghosh	237
<i>Evaluation Cards for XAI Metrics</i>	
Rokas Gipiškis and Olga Kurasova	245

Program

Saturday, July 4, 2026

14:00 - 14:05 *Welcome and Introduction*

14:05 - 14:45 *Panel Presentation*

14:45 - 14:50 *Break*

14:50 - 15:20 *Oral Presentations*

Rigorous Interpretation Is a Form of Evaluation

Isabelle Lee, Emmy Liu, Cathy Jiao, Brihi Joshi, Dani Yogatama, Fazl Barez and Michael Saxon

Graduating the Benchmark Scale: Lessons from Thermometry

Sean Trott and Oisín Parkinson-Coombs

One Persona, Many Cues, Different Results: How Sociodemographic Cues Impact LLM Personalization

Franziska Weeber, Vera Neplenbroek, Jan Batzner and Sebastian Padó

Becoming Experienced Judges: Selective Test-Time Learning for Evaluators

Seungyeon Jwa, Daechul Ahn, Reokyoung Kim, Dongyeop Kang and Jonghyun Choi

LLMs Gaming Verifiers: RLVR can Lead to Reward Hacking

Lukas Helff, Quentin Delfosse, David Steinmann, Ruben Härle, Hikaru Shindo, Patrick Schramowski, Wolfgang Stammer, Kristian Kersting and Felix Friedrich

Evaluating AI-Generated Images of Cultural Artifacts with Community-Informed Rubrics

Nari Johnson, Deepthi Sudharsan, Hamna , Samantha Dalal, Theo Holroyd, Anja Thieme, Hoda Heidari, Daniela Massiceti, Jennifer Wortman Vaughan and Cecily Morrison

15:20 - 16:20 *Poster Presentations*

16:20 - 16:25 *Break*

16:25 - 17:15 *Shared Task: Every Eval Ever*

Saturday, July 4, 2026 (continued)

17:15 - 17:25 *EvalEval Community Awards*

17:25 - 17:30 *Closing Remarks*