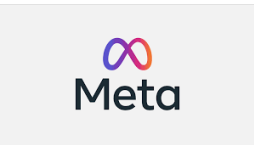
TrustNLP 2026

The 6th Workshop on Trustworthy NLP

Proceedings of the Workshop (TrustNLP 2026)

July 4, 2026

The TrustNLP 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 979-8-89176-418-7

Introduction

We welcome all participants of TrustNLP 2026, the Sixth Workshop on Trustworthy Natural Language Processing. This year, we are excited to host our TrustNLP workshop at ACL 2026, aimed at fostering discussions on these pressing challenges and driving the development of solutions that prioritize trustworthiness in NLP technologies. The workshop aspires to bring together researchers from various fields to engage in meaningful dialogue on key topics such as fairness and bias mitigation, transparency and explainability, privacy-preserving NLP methods, and the ethical deployment of AI systems. By providing a platform for sharing innovative research and practical insights, this workshop seeks to bridge the gaps between these interconnected objectives and establish a foundation for a more comprehensive and holistic approach to trustworthy NLP.

Recent advances in Natural Language Processing, and the emergence of pretrained Large Language Models (LLM) specifically, have led to significant breakthroughs in language understanding, generation, and interaction, leading to increasing usage of the models in real-life tasks. However, these advancements come with risks, including potential breaches of privacy, the propagation of bias, copyright violation, and vulnerabilities to adversarial manipulation. The demand for trustworthy NLP solutions is pressing as the public, policymakers, and organizations seek assurances that NLP systems protect data confidentiality, operate fairly, and adhere to ethical principles. In response to these challenges, we invited papers which focus on different aspects of safe and trustworthy language modeling. Topics of interest include (but are not limited to):

- Privacy-Preserving Model Training
- Unlearning and Model Editing
- Fairness and Bias: Evaluation and Treatments
- Model Explainability and Interpretability
- Culturally-Aware and Inclusive LLMs
- Accountability, Safety, and Robustness
- Red-teaming, backdoor or adversarial attacks and defenses for LLM safety
- Ethics, Social responsibility, and Dual-use
- Causal Inference and Fair ML
- Secure, Faithful, Safe, and Trustworthy Data/Language Generation
- Hallucination and Unqualified Suggestion
- Toxic Language Detection and Mitigation
- Industry applications of Trustworthy NLP

Our agenda features 2 keynote speeches, an oral presentation session, and a poster session. We received 88 submissions, out of which 53 were accepted. Among them, 41 have been included in our proceedings. These papers span a wide array of topics including fairness, robustness, jailbreaking, privacy, factuality, and uncertainty estimation in NLP. We would like to express our gratitude to all the authors, committee members, keynote speakers, and participants. We also gratefully acknowledge the generous sponsorship provided by Meta, Amazon, and Capital One.

Organizing Committee

Workshop organisers

Kai-Wei Chang, UCLA, Amazon Nova RAI
Ninareh Mehrabi, Meta
Satyapriya Krishna, Amazon Nova RAI
Anubrata Das, University of Texas at Austin
Jwala Dhamala, Amazon Nova RAI
Yang Trista Cao, Amazon Nova RAI
Tharindu Kumarage, Amazon Nova RAI
Anil Ramakrishna, Meta
Christos Christodoulopoulos, Information Commissioner's Office
Yixin Wan, UCLA
Aram Galystan, USC, Amazon AGI
Anoop Kumar, Capital One
Rahul Gupta, Amazon Nova RAI

Program Committee

Program Chairs

Yang Trista Cao, University of Texas at Austin
Kai-Wei Chang, University of California, Los Angeles and Amazon
Christos Christodoulopoulos, Information Commissioner's Office
Anubrata Das, Autodesk
Jwala Dhamala, Amazon Alexa AI
Tharindu Kumarage, Amazon and Arizona State University
Ninareh Mehrabi, Facebook
Anil Ramakrishna, Meta
Yixin Wan, University of California, Los Angeles

Reviewers

Zuhaib Akhtar, Srinivasa Rao Aravilli, Hitesh Arora, Berk Atıl, Mina Basirat, Gagan Bhatia, Raj Biswas, Abhinav Bohra, Dylan Bouchard, Keith Burghardt, Canyu Chen, Jwala Dhamala, Sahil Rajesh Dhayalkar, Ishita Doshi, Gaoyuan Du, Anupama Garani, Sankalp Gilda, Usman Gohar, Navita Goyal, Zihao He, Chi-Yang Hsu, Shengyuan Hu, Ashish Jain, Alizishaan Khatri, Alizishaan Khatri, Gyuwan Kim, Shirley Kokane, Satyapriya Krishna, Ian Lane, Jooyoung Lee, Hamed Loghmani, Parth Mehta, Zeeshan Memon, Sahil Mishra, Sahil Mishra, Chinmoy Mohapatra, Rohith Namboothiri, Huy Nghiem, Hieu Minh Nguyen, Hari Vilas Panjwani, Udita Patel, Maya Pavlova, Kartik Perisetla, Prasanth, Chahat Raj, Elakkiya Rajasekar, Shail Raval, Sathwik Reddy, Kuleen Sasse, Rahul Seetharaman, Sayambhu Sen, Pralam Shah, Shivani Shekhar, Arth Singh, Ishan Singh, Nikhil Singhal, Harguna Sood, Jingyu Tang, Jay Karan Telukunta, Poojitha Thota, Simran Tiwari, Rohith Uppala, Sohan Venkatesh, Ashwin Vinod, Qianli Wang, Michał Wiliński, Weihang Xiao, Ziping Ye, Pardis Sadat Zahraei, Xinlin Zhuang

Keynote Talk

Adversarial Arena: Driving Innovation in Responsible AI through Interactive Competition

Michael Johnston
Amazon AGI
2026-07-04 09:05:00 – Room: **Harbor G**

Abstract: As AI systems become increasingly capable, it is critical that techniques ensuring their safety and alignment to human values keep up with the exponential pace of innovation. At the same time, there is increasing concern from an evaluation science perspective that static benchmarks fail to capture the true capabilities of models with respect to both utility and safety. Also, while high quality diverse data is critical for training, it is rare and expensive to create, especially for new tasks and multi-turn conversations. In this talk, I will illustrate how these three issues can be addressed through an 'Adversarial Arena' approach to driving research and data creation, where different approaches are evaluated through interactive competition. I will draw on examples and learnings from the Amazon Nova AI Challenge: Trusted AI, an international AI competition now in its second year. In the challenge, competing teams build either secure coding agents or automated red teaming bots and their creations face off in a series of tournaments setting in motion a continuous flywheel of innovation and data generation.

Bio: Dr. Michael Johnston is Applied Science Manager in the Responsible AI team in Amazon AGI. Michael has over 30 years of experience in artificial intelligence and machine learning and research contributions spanning NLP, dialog, multimodality, fusion of human and artificial intelligence, and trustworthy AI. Before joining Amazon, he was VP of Research and Innovation at Interactions Corporation, and earlier held positions at AT&T Labs Research, Oregon Graduate Institute, Brandeis University, and Apple. Michael has over 60 U.S. patents, and has published over 80 scientific papers. He has designed and overseen multiple international challenge competitions in artificial intelligence, including the Alexa Prize, the Amazon Trusted AI Challenge, and Amazon Nova AI Challenge: Trusted Software Agents.

Keynote Talk

Toward Trustworthy Language Models through Interpretability and Control

Lilly Weng
UC San Diego
2026-07-04 11:00:00 – Room: **Harbor G**

Abstract: In this talk, I will present recent work from my lab toward trustworthy language models through representation-level interpretability, behavior steering, and principled evaluation. In particular, I will discuss: (i) recent findings showing that behaviors such as self-reflection and reasoning dynamics are encoded in the internal representations of language models and can be manipulated to steer model behavior and improve reasoning efficiency; (ii) emerging frameworks for rigorously evaluating the faithfulness of neuron- and representation-level explanations; and (iii) recent efforts toward trustworthy reasoning and interpretable-by-design language models, including training-free model editing, structured reasoning supervision, and architectures with explicit concept bottlenecks. These works illustrate a broader perspective for trustworthy language models: achieving reliable language models requires not only strong capabilities, but also principled mechanisms for understanding, evaluating, and controlling their internal behaviors.

Bio: Lily Weng is an Assistant Professor in the Halıcıoğlu Data Science Institute at UC San Diego and she leads the Trustworthy Machine Learning Lab at UC San Diego. She received her PhD in Electrical Engineering and Computer Science (EECS) from MIT in August 2020, and her Bachelor and Master degree both in Electrical Engineering at National Taiwan University. Prior to UCSD, she spent 1 year in MIT-IBM Watson AI Lab and several research internships in Google DeepMind, IBM Research and Mitsubishi Electric Research Lab. Her research interest is in machine learning and deep learning, with primary focus on Trustworthy AI. Her vision is to make the next generation AI systems and deep learning algorithms more robust, reliable, explainable, trustworthy and safer. Her work has been recognized and supported by multiple NSF awards, ARL award, Intel Rising Star Faculty Award, Hellman Fellowship, and Nvidia Academic award. For more details, please see https://lilywenglab.github.io/

Table of Contents

Program

Saturday, July 4, 2026

09:00 - 09:05 *Opening Remarks*

09:05 - 09:50 *Keynote 1 - Michael Johnston*

09:50 - 10:30 *Poster Session (continues during the break)*

Evaluating Cross-Lingual Behavior and Consistency of Multimodal Large Language Models
Hao Wang, Pinzhi Huang and Daisuke Kawahara

Responsible Federated LLMs via Safety Filtering and Constitutional AI
Eunchung Noh and Jeonghun Baek

Uncertainty-Aware Proxy Attribute Reasoning for Reliable Media Bias Detection
Chin-Po Chen, Jeng-Lin Li and Ming-Ching Chang

Quantifying LLM Safety Degradation Under Repeated Attacks Using Survival Analysis
Zvi Topol

ChatbotManip: a Dataset to Facilitate Evaluation and Oversight of Manipulative Chatbot Behaviour
Jack Luigi Henry Contro, Simrat Deol, Martim Brandao and Yulan He

Controllable Pareto Trade-off between Fairness and Accuracy
Yongkang Du, Jieyu Zhao, Yijun Yang and Tianyi Zhou

What are They Thinking? Delineation, Probing, and Tracking of Concepts in LLMs
Mohamed Abdelwahab, Michelle Yu Collins, Sihan Chen, Yi Cheng Zhao, Zafarullah Mahmood, Jiading Zhu, Soliman Ali and Jonathan Rose

Hair-Trigger Alignment: Black-Box Evaluation Cannot Guarantee Post-Update Alignment
Yavuz Faruk Bakman, Duygu Nur Yaldiz, Salman Avestimehr and Sai Praneeth Karimireddy

Teaching People LLM's Errors and Getting it Right
Nathan Stringham, Fateme Hashemi Chaleshtori, Xinyuan Yan, Zhichao Xu, Bei Wang and Ana Marasovic

Re-Mask and Redirect: Exploiting Denoising Irreversibility in Diffusion Language Models
Arth Singh

Multilingual Steering by Design: Multilingual Sparse Autoencoders and Principled Layer Selection
Yusser Al Ghussin, Daniil Gurgurov, Tanja Baeumel, Josef Van Genabith, Patrick Schramowski and Simon Ostermann

CARE: A Conformal Safety Layer for Medical Summarization
Suhana Bedi, Bridget Lin, Anson Zhou, Jenelle A Jindal, Chloe O'Connell Stanwyck, Sanmi Koyejo and Nigam Shah

Deactivating Refusal Triggers: Understanding and Mitigating Overrefusal in Safety Alignment
Zhiyu Xue, Zimo Qi, Guangliang Liu, Bocheng Chen and Ramtin Pedarsani

A Systematic Taxonomy of Failure Modes in Retrieval-Augmented Generation Systems
Anupama Garani

Improving the Faithfulness of LLM-based Abstractive Summarization with Span-level Unlikelihood Training
Sicong Huang, Qianqi Yan, Shengze Wang and Ian Lane

Context Misleads LLMs: The Role of Context Filtering in Maintaining Safe Alignment of LLMs
Jinhwa Kim and Ian Harris

MASCOT: Towards Trustworthy Multi-Agent Socio-Collaborative Companion Systems
Yiyang Wang, Yiqiao Jin, Alex Cabral and Josiah Hester

Multi-Turn Jailbreaking of Aligned LLMs via Lexical Anchor Tree Search
Devang Kulshreshtha, Hang Su, Chinmay Hegde and Haohan Wang

The ACUTE Protocol: Operationalizing Language Model Activations for Better Calibration, Utility, and Trust
Nishant Subramani, Palash Goyal, Yiwen Song, Mani Malek, Yuan Xue, Tomas Pfister and Hamid Palangi

Lexical Familiarity Predicts Processing Depth for Nonliteral Language in Large Language Models
Lang-Ching Yeh, Yu-Chieh Wang and Shu-Kai Hsieh

Did You Forget What I Asked? Prospective Memory Failures in Large Language Models
Avni Mittal

Cultural Counterfactuals: Evaluating Cultural Biases in Large Vision-Language Models with Counterfactual Examples
Phillip Howard, Xin Su and Kathleen C. Fraser

Don't Want Your LLM to Recommend Nuclear Strike? Try Asking It in Japanese
Rian Touchent

Toward Dialect-Aware Safety Evaluation for Arabic Large Language Models
Wajdi Zaghouani

Single-Layer Activation Edits Easily Corrupt Factual Recall but Rarely Repair It
Zacharie Bugaud

Truth or Dare: Analyzing LLM Susceptibility to External Evidence of Varying Factuality
Han-Yu Su, Kuan-Yu Chu, Yung-Hui Li and Lun-Wei Ku

The Halo Effect and Language Takeover: Spatiotemporal Attention Decay Explains Vision-Language Model Failures in Simple Visual Counting
Haochen Zhao and Sujian Li

Why is ChicagoPredictive of Deceptive Reviews? Using LLMs to Discover Language Phenomena from Lexical Cues
Jiaming Qu, Mengtian Guo and Yue Wang

Domain-Dependent Safety Behavior in Open-Weight LLMs: An Empirical Study Across Seven Ethical Domains
Zacharie Bugaud

Reward-Robust Reinforcement Learning in LLMs via Uncertainty Set Modeling
Yuzi Yan, Xingzhou Lou, Jialian Li, Yipin Zhang, Jian Xie, Dong Yan and Yuan Shen

A Systematic Comparison between Extractive Self-Explanations and Human Rationales in Text Classification
Stephanie Brandl and Oliver Eberle

Saturday, July 4, 2026 (continued)

CRISP: Persistent Concept Unlearning via Sparse Autoencoders
Tomer Ashuach, Dana Arad, Aaron Mueller, Martin Tutek and Yonatan Belinkov

Guiding Giants: Lightweight Controllers for Weighted Activation Steering in LLMs
Amr Hegazy, Mostafa Elhoushi and Amr Alanwar

SURGELLM: Rethinking Multi-Task Evaluation through Task-Aware Feature Gating with Class-Balanced Normalization
Noor Islam S. Mohammad and Ulug Bayazit

With a Grain of SALT: Are LLMs Fair Across Social Dimensions?
Samee Arif, Zohaib Khan, Maaidah Kaleem Butt, Muhammad Suhaib Rashid, Agha Ali Raza and Awais Athar

GateKD: Confidence-Gated Closed-Loop Distillation for Robust Reasoning
Kasidit Sermsri and Teerapong Panboonyuen

The Geometry of Refusal: Linear Instability in Safety-Aligned LLMs
Shivam Ratnakar and Kartikeya Vats

The Conservative AI: Diagnosing Hold Bias and Reliability Limits in Persona-Based Monetary Policy Simulation
Giyong Kim and Sojung Kim

Confidence as Control: A Survey of Confidence Utilization in Large Language Models
Yubo Li, Tianyang Zhou, Xiaobin Shen, Yidi Miao, Rema Padman and Ramayya Krishnan

10:30 - 11:00 *Break*

11:00 - 11:45 *Keynote 2 - Lilly Weng*

11:45 - 12:25 *Oral Session*

Through a Compressed Lens: Investigating The Impact of Quantization on Factual Knowledge Recall
Qianli Wang, Mingyang Wang, Nils Feldhus, Simon Ostermann, Yuan Cao, Hinrich Schuetze, Sebastian Möller and Vera Schmitt

Saturday, July 4, 2026 (continued)

ClaimCLAIRE: A Trust-Aware Multi-Component Fact-Checking Agent for Open-World Claims
Xinman Liu and Mayank Sharma

Fairness Failure Modes of Multimodal LLMs
Canyu Chen, Anglin Cai, Joan Nwatu, Yale Li, Jessica Hullman, Rada Mihalcea, Kathleen McKeown and Manling Li

12:25 - 12:30 *Closing Remarks*