TrustNLP 2025

The 5th Workshop on Trustworthy NLP

Proceedings of the Workshop (TrustNLP 2025)

May 3, 2025

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

Gold

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 979-8-89176-233-6

Introduction

We welcome all participants of TrustNLP 2025, the Fifth Workshop on Trustworthy Natural Language Processing. This year, we are excited to host our TrustNLP workshop at NAACL 2025, 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):

- Secure, Faithful & Trustworthy Generation with LLMs
- Data Privacy Preservation and Data Leakage Issues in LLMs
- Red-teaming, backdoor or adversarial attacks and defenses for LLM safety
- Fairness, LLM alignment, Human Preference Elicitation, Participatory NLP
- Toxic Language Detection and Mitigation
- Explainability and Interpretability of LLM generation
- Robustness of LLMs
- Mitigating LLM Hallucinations & Misinformation
- Fairness and Bias in multi-modal generative models: Evaluation and Treatments
- Industry applications of Trustworthy NLP
- Culturally-Aware and Inclusive LLMs

Our agenda features 3 keynote speeches, a industrial panel session, an oral presentation session, and a poster session. We received 66 submissions, out of which 45 were accepted. Among them, 37 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, panelists, and participants. We also gratefully acknowledge the generous sponsorship provided by Amazon and Capital One.

Program Committee

Program Chairs

Yang Trista Cao, University of Texas at Austin
Kai-Wei Chang, University of California, Los Angeles and Amazon
Anubrata Das, University of Texas, Austin
Jwala Dhamala, Amazon Alexa AI
Aram Galstyan, Information Sciences Institute, University of Southern California, University of Southern California, University of Southern California and Amazon Alexa
Rahul Gupta
Anoop Kumar, Amazon
Ninareh Mehrabi, Amazon
Anil Ramakrishna, Amazon
Yixin Wan, University of California, Los Angeles
Tharindu Kumarage
Satyapriya Krishna

Reviewers

Haozhe An, Berk Atıl

Connor Baumler, Gagan Bhatia

Javier Carnerero-Cano, Christina A Chance, Canyu Chen, Zizhao Chen, Pedro Cisneros-Velarde

Kaveh Eskandari Miandoab

Usman Gohar, Navita Goyal, Lavanya Gupta

Pengfei He, Zihao He

Jivitesh Jain, Siddharth D Jaiswal, Yeonsung Jung

Satyapriya Krishna, Atharva Kulkarni

Adarsh N L, Jooyoung Lee, Xiangci Li, Qin Liu, Hamed Loghmani, Yanan Long

Subhabrata Majumdar, Jennifer Mickel

Huy Nghiem, Haoran Niu

Aishwarya Padmakumar, Kartik Perisetla

Chahat Raj, Vipula Rawte, Anthony Rios, Shubhashis Roy Dipta

Erfan Shayegani, Shahriar Shayesteh, Anna Sotnikova, Tejas Srinivasan

Xin Xu

Ziping Ye

Caiqi Zhang, Lingjun Zhao, Xinlin Zhuang

Table of Contents