

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**



©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](mailto: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

<i>Beyond Text-to-SQL for IoT Defense: A Comprehensive Framework for Querying and Classifying IoT Threats</i>	
Ryan Pavlich, Nima Ebadi, Richard Tarbell, Billy Linares, Adrian Tan, Rachael Humphreys, Jayanta Das, Rambod Ghandiparsi, Hannah Haley, Jerris George, Rocky Slavin, Kim-Kwang Raymond Choo, Glenn Dietrich and Anthony Rios .....	1
<i>Gibberish is All You Need for Membership Inference Detection in Contrastive Language-Audio Pretraining</i>	
Ruoxi Cheng, Yizhong Ding, Shuirong Cao, Zhiqiang Wang and Shitong Shao .....	13
<i>PBI-Attack: Prior-Guided Bimodal Interactive Black-Box Jailbreak Attack for Toxicity Maximization</i>	
Ruoxi Cheng, Yizhong Ding, Shuirong Cao, Ranjie Duan, Xiaoshuang Jia, Shaowei Yuan, Zhiqiang Wang and Xiaojun Jia .....	23
<i>Ambiguity Detection and Uncertainty Calibration for Question Answering with Large Language Models</i>	
Zhengyan Shi, Giuseppe Castellucci, Simone Filice, Saar Kuzi, Elad Kravi, Eugene Agichtein, Oleg Rokhlenko and Shervin Malmasi .....	41
<i>Smaller Large Language Models Can Do Moral Self-Correction</i>	
Guangliang Liu, Zhiyu Xue, Xitong Zhang, Rongrong Wang and Kristen Johnson .....	56
<i>Error Detection for Multimodal Classification</i>	
Thomas Bonnier .....	66
<i>Break the Breakout: Reinventing LM Defense Against Jailbreak Attacks with Self-Refine</i>	
Heegy Kim and Hyunsouk Cho .....	82
<i>Minimal Evidence Group Identification for Claim Verification</i>	
Xiangci Li, Sihao Chen, Rajvi Kapadia, Jessica Ouyang and Fan Zhang .....	103
<i>Cracking the Code: Enhancing Implicit Hate Speech Detection through Coding Classification</i>	
Lu Wei, Liangzhi Li, Tong Xiang, Liu Xiao and Noa Garcia .....	112
<i>Line of Duty: Evaluating LLM Self-Knowledge via Consistency in Feasibility Boundaries</i>	
Sahil Kale and vrn@stride.ai vrn@stride.ai .....	127
<i>Multi-lingual Multi-turn Automated Red Teaming for LLMs</i>	
Abhishek Singhania, Christophe Dupuy, Shivam Sadashiv Mangale and Amani Namboori .....	141
<i>Rainbow-Teaming for the Polish Language: A Reproducibility Study</i>	
Aleksandra Krasnodębska, Maciej Chrabaszcz and Wojciech Kusa .....	155
<i>BiasEdit: Debiasing Stereotyped Language Models via Model Editing</i>	
Xin Xu, Wei Xu, Ningyu Zhang and Julian McAuley .....	166
<i>Do Voters Get the Information They Want? Understanding Authentic Voter FAQs in the US and How to Improve for Informed Electoral Participation</i>	
Vipula Rawte, Deja N Scott, Gaurav Kumar, Aishneet Juneja, Bharat Sowrya Yaddanapalli and Biplav Srivastava .....	185
<i>ViBe: A Text-to-Video Benchmark for Evaluating Hallucination in Large Multimodal Models</i>	
Vipula Rawte, Sarthak Jain, Aarush Sinha, Garv Kaushik, Aman Bansal, Prathiksha Rumale Vishwanath, Samyak Rajesh Jain, Aishwarya Naresh Reganti, Vinija Jain, Aman Chadha, Amit Sheth and Amitava Das .....	232

<i>Know What You do Not Know: Verbalized Uncertainty Estimation Robustness on Corrupted Images in Vision-Language Models</i>	
Mirko Borszukovszki, Ivo Pascal De Jong and Matias Valdenegro-Toro . . . . .	247
<i>Summary the Savior: Harmful Keyword and Query-based Summarization for LLM Jailbreak Defense</i>	
Shagoto Rahman and Ian Harris . . . . .	266
<i>Bias A-head? Analyzing Bias in Transformer-Based Language Model Attention Heads</i>	
Yi Yang, Hanyu Duan, Ahmed Abbasi, John P. Lalor and Kar Yan Tam . . . . .	276
<i>Mimicking How Humans Interpret Out-of-Context Sentences Through Controlled Toxicity Decoding</i>	
Maria Mihaela Trusca and Liesbeth Allein . . . . .	291
<i>On the Robustness of Agentic Function Calling</i>	
Ella Rabinovich and Ateret Anaby Tavor . . . . .	298
<i>Monte Carlo Temperature: a robust sampling strategy for LLM’s uncertainty quantification methods</i>	
Nicola Cecere, Andrea Bacciu, Ignacio Fernández-Tobías and Amin Mantrach . . . . .	305
<i>Know Thyself: Validating Knowledge Awareness of LLM-based Persona Agents</i>	
Savita Bhat, Ishaan Shukla and Shirish Karande . . . . .	321
<i>Building Safe GenAI Applications: An End-to-End Overview of Red Teaming for Large Language Models</i>	
Alberto Purpura, Sahil Wadhwa, Jesse Zymet, Akshay Gupta, Andy Luo, Melissa Kazemi Rad, Swapnil Shinde and Mohammad Shahed Sorower . . . . .	335
<i>Difficulty Estimation in Natural Language Tasks with Action Scores</i>	
Aleksandar Angelov, Tsegaye Misikir Tashu and Matias Valdenegro-Toro . . . . .	351
<i>Are Small Language Models Ready to Compete with Large Language Models for Practical Applications?</i>	
Neelabh Sinha, Vinija Jain and Aman Chadha . . . . .	365
<i>A Calibrated Reflection Approach for Enhancing Confidence Estimation in LLMs</i>	
Umesh Bodhwani, Yuan Ling, Shujing Dong, Yarong Feng and Hongfei Li . . . . .	399
<i>Evaluating Design Choices in Verifiable Generation with Open-source Models</i>	
Shuyang Cao and Lu Wang . . . . .	412
<i>Battling Misinformation: An Empirical Study on Adversarial Factuality in Open-Source Large Language Models</i>	
Shahnewaz Karim Sakib, Anindya Bijoy Das and Shibbir Ahmed . . . . .	432
<i>Will the Prince Get True Love’s Kiss? On the Model Sensitivity to Gender Perturbation over Fairytale Texts</i>	
Christina A Chance, Da Yin, Dakuo Wang and Kai-Wei Chang . . . . .	444
<i>Disentangling Linguistic Features with Dimension-Wise Analysis of Vector Embeddings</i>	
Saniya Karwa and Navpreet Singh . . . . .	461
<i>Gender Encoding Patterns in Pretrained Language Model Representations</i>	
Mahdi Zakizadeh and Mohammad Taher Pilehvar . . . . .	489
<i>Defining and Quantifying Visual Hallucinations in Vision-Language Models</i>	
Vipula Rawte, Aryan Mishra, Amit Sheth and Amitava Das . . . . .	501



<i>Revitalizing Saturated Benchmarks: A Weighted Metric Approach for Differentiating Large Language Model Performance</i>	
Bryan Etzine, Masoud Hashemi, Nishanth Madhusudhan, Sagar Davasam, Roshnee Sharma, Sathwik Tejaswi Madhusudhan and Vikas Yadav .....	511
<i>Synthetic Lyrics Detection Across Languages and Genres</i>	
Yanis Labrak, Markus Frohmann, Gabriel Meseguer-Brocal and Elena V. Epure .....	524
<i>A Lightweight Multi Aspect Controlled Text Generation Solution For Large Language Models</i>	
Chenyang Zhang, Jiayi Lin, Haibo Tong, Bingxuan Hou, Dongyu Zhang, Jialin Li and Junli Wang	542
<i>Gender Bias in Large Language Models across Multiple Languages: A Case Study of ChatGPT</i>	
YiTian Ding, Jinman Zhao, Chen Jia, Yining Wang, Zifan Qian, Weizhe Chen and Xingyu Yue	552
<i>Investigating and Addressing Hallucinations of LLMs in Tasks Involving Negation</i>	
Neeraj Varshney, Satyam Raj, Venkatesh Mishra, Agneet Chatterjee, Amir Saeidi, Ritika Sarkar and Chitta Baral.....	580
<i>FACTOID: FACTual enTailment fOr hallucInation Detection</i>	
Vipula Rawte, S.m Towhidul Islam Tonmoy, Shravani Nag, Aman Chadha, Amit Sheth and Amitava Das.....	599