TrustNLP 2024

The Fourth Workshop on Trustworthy Natural Language Processing

Proceedings of the Workshop (TrustNLP 2024)

June 21, 2024

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

Gold

amazon

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 979-8-89176-113-1

Introduction

Welcome to TrustNLP 2024, the fourth Workshop on Trustworthy Natural Language Processing. Co-located with NAACL 2024, the workshop is scheduled for June 21, 2024. To facilitate the participation of the global NLP community, we conduct this year's workshop in a hybrid format.

The continued evolution of Large Language Models (LLMs) has led to unprecedented growth in Natural Language Processing applications. Incorporating vision capabilities into AI-powered content creation tools, such as Anthropic's Claude 2.0 and OpenAI's ChatGPT 4.0, has ushered in a new era of creative writing and multimodal interaction. The release of new text-to-video models (Sora, Gen-2, Pika) and the integration of text-to-image models into widely adopted tools (DALL-E 3, Firefly) has further expanded the creative possibilities. In the healthcare domain, MedPaLM 2, Google's medical LLM, has demonstrated impressive performance in medical question answering. However, as these advancements continue to shape various aspects of our lives, they also raise pressing concerns about the ethical, social, and technical implications of their widespread adoption. Therefore, as the influence of these technologies grows, so does the need for responsible development and deployment practices.

In response to these challenges, the NLP community has been actively pursuing research on various aspects of trustworthiness, such as fairness, safety, privacy, and transparency. However, these efforts have often been siloed, limiting our understanding of the complex interplay between these objectives. For example, ensuring fairness might necessitate access to sensitive user data, which could compromise privacy. The TrustNLP 2024 workshop aims to foster a more holistic approach to Trustworthy NLP by bringing together researchers working on these interconnected topics and encouraging dialogue on their intersections.

Our agenda features four keynote speeches, a presentation session, and two poster sessions. This year, we were delighted to receive 44 submissions, out of which 40 papers were accepted. Among these, 21 have been included in our proceedings. These papers span a wide array of topics including fairness, robustness, factuality, privacy, explainability, and model analysis in NLP.

We would like to express our gratitude to all the authors, committee members, keynote speakers, and participants and gratefully acknowledge Amazon's generous sponsorship.

Program Committee

Organizers

Anaelia Ovalle, UCLA
Kai-Wei Chang, UCLA, Amazon Visiting Academic
Yang Trista Cao, University of Maryland
Ninareh Mehrabi, Amazon AGI Foundations
Jieyu Zhao, University of Southern California
Jwala Dhamala, Amazon AGI Foundations
Aram Galstyan, USC Information Sciences Institute, Amazon Visiting Academic
Anoop Kumar, Amazon AGI Foundations
Rahul Gupta, Amazon AGI Foundations

Program Committee

Saied Alshahrani, Clarkson University
Nishant Balepur, University of Maryland
Connor Baumler, University of Maryland
Gagan Bhatia, University of British Columbia
Keith Burghardt, USC Information Sciences Institute
Javier Carnerero Cano, IBM Research Europe, Imperial College London
Christina Chance, UCLA
Xinyue Chen, Carnegie Mellon University
Canyu Chen, Illinois Institute of Technology
Jwala Dhamala, Amazon AGI Foundations
Ninareh Mehrabi, Amazon AGI Foundations
Árdís Elíasdóttir, Amazon
Aram Galstyan, USC Information Sciences Institute, Amazon Visiting Academic
Yang Trista Cao, University of Maryland
Usman Gohar, Iowa State University
Zihao He, University of Southern California
Pengfei He, University of Washington
Qian Hu, Amazon
Satyapriya Krishna, Harvard University
Anaelia Ovalle, UCLA
Jooyoung Lee, Penn State University
Yanan Long, University of Chicago
Subho Majumdar, Vijil
Sahil Mishra, IIT Delhi
Isar Nejadgholi, National Research Council Canada
Huy Nghiem, University of Maryland, College Park
Aishwarya Padmakumar, Amazon
Kartik Perisetla, Apple
Salman Rahman, New York University
Chahat Raj, George Mason University
Anthony Rios, University of Texas at San Antonio
Patricia Thaine, University of Toronto
Simon Yu, University Of Edinburgh
Yixin Wan, UCLA

Xinchen Yang, University of Maryland, College Park
Chenyang Zhu, Capital One
Xinlin Zhuang, East China Normal University

Table of Contents

Program

Friday, June 21, 2024

09:00 - 09:10	*Opening Remarks*
09:10 - 09:50	*Keynote 1 (Maria Pacheco)*
09:50 - 10:30	*Keynote 2 (Ahmad Beirami)*
10:30 - 11:10	*Virtual Poster Session + Coffee Break*
11:10 - 11:50	*Keynote 3 (Jieyu Zhao)*
11:50 - 12:30	*Keynote 4 (Prasanna Sattigeri)*
12:30 - 02:00	*Lunch*
02:00 - 03:30	*In-person Poster Session*
03:30 - 04:00	*Coffee Break*
04:00 - 05:20	*Best Paper Presentations + Spotlight Paper Presentations*
05:20 - 05:30	*Closing Remarks*