

TrustNLP 2022

**The 2nd Workshop on Trustworthy Natural Language
Processing**

Proceedings of the Workshop

July 14, 2022

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

Sponsor



©2022 Association for Computational Linguistics

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
209 N. Eighth Street
Stroudsburg, PA 18360
USA
Tel: +1-570-476-8006
Fax: +1-570-476-0860
acl@aclweb.org

ISBN 978-1-955917-78-0

Organizing Committee

Organizing Committee

Yada Pruksachatkun, Infinitus Systems

Apurv Verma, Amazon Alexa

Jwala Dhamala, Amazon Alexa

Yang Trista Cao, University of Maryland College Park

Kai Wei Chang, University of California, Los Angeles

Aram Galstyan, University of Southern California

Program Committee

Program Committee

Rahul Gupta, Amazon
Naveen Kumar, Disney Research
Tianlu Wang, Facebook
Joe Near, Vermont University
Sunipa Dev, University of California Los Angeles
Jieyu Zhao, University of California Los Angeles
David Darais, Galois
Paul Pu Liang, Carnegie Mellon University
Hila Gonen, Bar-Ilan University
Ninareh Mehrabi, University of Southern California
Arjun Subramonian, University of California Los Angeles
Emily Sheng, Twitter
Isar Nejadgholi, IMRSV Data Labs
Eric W. Davis, Galois
Anthony Rios, University of Texas at San Antonio
Jamie Hayes, University College London
Hitesh Sapkota, Rochester Institute of Technology
Anirudh Raju, Amazon
Umang Gupta, University of Southern California
Krishna Somandepalli, Google
Caleb Zeims, Georgia Institute of Technology
Varun Kumar, Amazon
Robik Shrestha, Rochester Institute of Technology
Griffin Adams, Columbia University
Walt Woods, Galois
Jialu Wang, University of California Santa Cruz

Invited Speakers

Fei Wang, Cornell University
Subho Majumdar, Splunk
Diyi Yang, Georgia Institute of Technology

Table of Contents

<i>An Encoder Attribution Analysis for Dense Passage Retriever in Open-Domain Question Answering</i> Minghan Li, Xueguang Ma and Jimmy Lin	1
<i>Attributing Fair Decisions with Attention Interventions</i> Ninareh Mehrabi, Umang Gupta, Fred Morstatter, Greg Ver Steeg and Aram Galstyan	12
<i>Does Moral Code have a Moral Code? Probing Delphi’s Moral Philosophy</i> Kathleen C. Fraser, Svetlana Kiritchenko and Esma Balkir	26
<i>The Cycle of Trust and Responsibility in Outsourced AI</i> Maximilian Castelli and Linda C. Moreau, Ph.D.	43
<i>Explaining Neural NLP Models for the Joint Analysis of Open-and-Closed-Ended Survey Answers</i> Edoardo Mosca, Katharina Harmann, Tobias Eder and Georg Groh.....	49
<i>The Irrationality of Neural Rationale Models</i> Yiming Zheng, Serena Booth, Julie Shah and Yilun Zhou	64
<i>An Empirical Study on Pseudo-log-likelihood Bias Measures for Masked Language Models Using Paraphrased Sentences</i> Bum Chul Kwon and Nandana Mihindukulasooriya	74
<i>Challenges in Applying Explainability Methods to Improve the Fairness of NLP Models</i> Esma Balkir, Svetlana Kiritchenko, Isar Nejadgholi and Kathleen Fraser.....	80
<i>ER-TEST Evaluating Explanation Regularization Methods for NLP Models</i> Brihi Joshi, Aaron Chan, Ziyi Liu and Xiang Ren.....	93