Zihao He

zihaoh@usc.edu | LinkedIn | Google Scholar | Homepage | Los Angeles, CA

RESEARCH INTERESTS

Natural Language Processing, Computational Social Science, Societal Impacts of LLMs, Safety and Alignment

EDUCATION

University of Southern California — Los Angeles, CA, USA
Doctor of Philosophy, Computer Science; ***GPA: 3.91/4.00*** — *Aug 2019 – present (Dec 2024 expected)*
Advisor: Prof. Kristina Lerman

Tsinghua University — Beijing, China
Master of Engineering, Computer Engineering (Transferred) — *Sep 2018 – Jun 2019*

Beijing University of Posts and Telecommunications — Beijing, China
Bachelor of Engineering, Communications Engineering; ***GPA: 90.38/100, Rank: 7/556*** — *Sep 2014 – Jun 2018*

SELECTED PUBLICATIONS AND PREPRINTS

LLM Personalization

COMMUNITY-CROSS-INSTRUCT: Unsupervised Instruction Generation for Aligning Large Language Models to Online Communities
Zihao He, Rebecca Dorn, Siyi Guo, Minh Duc Chu, Kristina Lerman
Preprint, 2024.

Improving and Assessing the Fidelity of Large Language Models Alignment to Online Communities
Minh Duc Chu, Zihao He, Rebecca Dorn, Kristina Lerman
Preprint, 2024.

Reading Between the Tweets: Reading between the tweets: Deciphering ideological stances of interconnected mixed-ideology communities
Zihao He, Ashwin Rao, Siyi Guo, Negar Mokhberian, Kristina Lerman
In *Findings of EACL, 2024.*

Safety and Alignment

How Susceptible are Large Language Models to Ideological Manipulation?
Kai Chen, Zihao He, Jun Yan, Taiwei Shi, Kristina Lerman
Preprint, 2024. Best Paper Runner-up at SeT LLM @ ICLR 2024

Whose Emotions and Moral Sentiments Do Language Models Reflect?
Zihao He, Siyi Guo, Ashwin Rao, Kristina Lerman
In *Findings of ACL, 2024.*

Multimodal Machine Learning

ALCAP: Alignment-Augmented Music Captioner
Zihao He, Weituo Hao, Wei-Tsung Lu, Changyou Chen, Kristina Lerman, Xuchen Song
In *Proceedings of ACL, 2024.*

Representation Learning

Infusing Knowledge from Wikipedia to Enhance Stance Detection
Zihao He, Negar Mokhberian, Kristina Lerman
In *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis, 2022.*

Speaker Turn Modeling for Dialogue Act Classification
Zihao He, Leili Tavabi, Kristina Lerman, Mohammad Soleymani
In *Findings of EMNLP, 2021.*

Detecting Polarized Topics Using Partisanship-aware Contextualized Topic Embeddings
Zihao He, Negar Mokhberian, António Câmara, Andrés Abeliuk, Kristina Lerman
In *Findings of EMNLP, 2021.*

Graph Embedding with Personalized Context Distribution
Zihao He, Di Huang, Yuzhong Huang, Kexuan Sun, Sami Abu-El-Haija, Bryan Perozzi, Kristina Lerman, Fred Morstatter, Aram Galstyan
In *Companion Proceedings of the Web Conference, 2020.*

INDUSTRIAL EXPERIENCE

Research Intern @ Tiktok — San Jose, CA, USA
Mentors: Weituo Hao, Xuchen song — *May 2022 – Aug 2022, May 2023 – Aug 2023*
Project: Automatic Generation of Music Interpretation (music2text)

Applied Science Intern @ Amazon — Seattle, WA, USA
Mentor: Matthew Butler — *May 2021 – Aug 2021*
Project: Text Embedding-based Book Personalization

Algorithms Development Intern @ Didi Global — Beijing, China
Mentor: Chao Feng — *Apr 2019 – July 2019*
Project: Multi-task Learning for Multi-label Driver Misbehavior Classification

ACADEMIC SERVICES

Reviewer/PC: ACL Rolling Review, EMNLP, ACL, NAACL, SIGKDD, WSDM, WWW, CSS

TEACHING EXPERIENCE

DSCI-531: Fairness in Artificial Intelligence. Instructor: Kristina Lerman. Spring 2022
CSCI-566: Deep Learning and its Applications. Instructor: Yue Zhao. Spring 2024

HONORS & AWARDS

Graduate School Travel Award, University of Southern California, 2024
Annenberg Fellowship, University of Southern California, 2019
National Scholarship for Excellent Academic Performance (top 1.2%), 2018
Meritorious Winner for the Interdisciplinary Contest in Modeling (top 10%), 2017
Second Prize for the Chinese Mathematics Competitions (top 7.5%), 2017
First Prize for National English Contest for College Students (top 0.6%), 2017

SKILLS

Programming Languages: Python, C++, Java, HTML
Machine Learning Libraries: PyTorch, Transformers, Scikit-Learn, NumPy, Pandas