Kai Chen

kchen035@usc.edu Homepage Google Scholar LinkedIn

RESEARCH INTEREST

Alignment, Data Synthesis, large language models (LLMs) agent, Societal Impacts of LLMs.

EDUCATION

University of Southern California — **Los Angeles, USA**
PhD in Computer Science — *08/2023 - Present*
Advisor: Jonathan May; Kristina Lerman

University of Southern California — **Los Angeles, USA**
MS in Computer Science — *08/2020 - 12/2022*

Zhejiang University of Technology — **Hangzhou, China**
BE in Software Engineering — *09/2016 - 06/2020*

PUBLICATIONS AND PREPRINTS

STEER-BENCH: A Benchmark for Evaluating the Steerability of Large Language Models
Kai Chen, Zihao He, Taiwei Shi, Kristina Lerman
To appear at EMNLP 2025

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

SAFER-INSTRUCT: Aligning Language Models with Automated Preference Data
Taiwei Shi, **Kai Chen**, Jieyu Zhao
NAACL 2024

IsamasRed: A Public Dataset Tracking Reddit Discussions on Israel-Hamas Conflict
Kai Chen, Zihao He, Keith Burghardt, Jingxin Zhang, Kristina Lerman
ICWSM 2024

Large Language Models Reveal Information Operation Goals, Tactics, and Narrative Frames
Keith Burghardt, **Kai Chen**, Kristina Lerman
CySoc @ ICWSM 2024

Anger Breeds Controversy: Analyzing Controversy and Emotions on Reddit
Kai Chen, Zihao He, Rong Ching Chan, Jonathan May, and Kristina Lerman
SBP-BRiMS 2023

EXPERIENCE

Research Assistant @ USC Information Sciences Institute, Los Angeles, USA — 06/2022 - 08/2025
Early Detection of Influence Indicators with Machine Intelligence
Advisor: Kristina Lerman

Research Assistant @ USC Information Sciences Institute, Los Angeles, USA 01/2022 - 12/2022
DARMA: Dialogue Agent for Reducing Malicious Acts
Advisor: Jonathan May, Kristina Lerman

HONORS AND AWARDS

• Best Paper Runner-up at ICLR 2024 Workshop on Secure and Trustworthy Large Language Models (SeT LLM), 2024.
• SBP-BRiMS Scholarship, 2023.
• Provincial Scholarship (top 5%), 2018.
• The Second Prize in the National Competition of Service Outsourcing Innovation and Entrepreneurship Competition, 2018&2019.

SERVICES

• Reviewer: ACL Rolling Review (2024-2025), ICWSM (2024-2025), COLING (2024)

TEACHING EXPERIENCE

• CSCI-360 Artificial Intelligence: Principles and Foundations. Instructor: Mohammad Reza Rajati. Fall 2025
• CSCI-596 Scientific Computing and Visualization. Instructor: Aiichiro Nakano. Fall 2024

Skills

• **Programming languages**: Python, C++, Java
• **Tools and Systems**: Slurm, Docker, Git
• **Libraries and Frameworks**: PyTorch, Transformers, veRL, LLaMA-Factory, NumPy, Pandas