Kai Chen

kchen035@usc.edu LinkedIn Google Scholar Homepage

RESEARCH INTEREST

Alignment, Safety, Data Synthesis, Societal impacts.

EDUCATION

University of Southern California — **Los Angeles, USA**
PhD in Computer Science — *08/2023 - Present*
Advisor: Prof. 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

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 - present
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