TAIWEI SHI

taiweish@usc.edu ◇ linkedin.com/in/maksimstw/ ◇ maksimstw.github.io

EDUCATION

University of Southern California, Ph.D. in Computer Science Aug 2023 - Present
Advisor: Prof. Jieyu Zhao

Georgia Institute of Technology, B.S. in Computer Science Aug 2020 - May 2023
Thesis Advisor: Prof. Diyi Yang, Prof. Mark Riedl
GPA: 3.96. Major GPA: 4.0. Highest Honors

George School, High School Diploma Aug 2017 - May 2020
Head of School's List. Honor Roll

SELECTED PUBLICATIONS AND PROJECTS (* EQUAL CONTRIBUTION)

CoAct-1: Computer-using Agents with Coding as Actions
Linxin Song, Yutong Dai, Viraj Prabhu, Jieyu Zhang, Taiwei Shi, Li Li, Junnan Li, Silvio Savarese, Zeyuan Chen, Jieyu Zhao, Ran Xu, Caiming Xiong
Preprint 2025

Efficient Reinforcement Finetuning via Adaptive Curriculum Learning
Taiwei Shi, Yiynag Wu, Linxin Song, Tianyi Zhou, Jieyu Zhao
In Submission to NeurIPS 2025

The Hallucination Tax of Reinforcement Finetuning
Linxin Song, Taiwei Shi*, Jieyu Zhao*
EMNLP 2025 Findings

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

Discovering Knowledge Deficiencies of Language Models on Massive Knowledge Base
Linxin Song, Xuwei Ding, Jieyu Zhang, Taiwei Shi, Ryotaro Shimizu, Rahul Gupta, Yang Liu, Jian Kang, Jieyu Zhao
COLM 2025

WildFeedback: Aligning LLMs with In-situ User Interactions and Feedback
Taiwei Shi, Zhuoer Wang, Longqi Yang, Ying-Chun Lin, Zexue He, Mengting Wan, Pei Zhou, Sujay Jauhar, Sihao Chen, Shan Xia, Hongfei Zhang, Jieyu Zhao, Xiaofeng Xu, Xia Song, Jennifer Neville
NeurIPS 2024 Workshop on Behavioral Machine Learning

Detecting and Filtering Unsafe Training Data via Data Attribution
Yijun Pan, Taiwei Shi, Jieyu Zhao, Jiaqi Ma
Preprint 2025

On the Trustworthiness of Generative Foundation Models
Yue Huang, Chujie Gao, Siyuan Wu, Haoran Wang, Xiangqi Wang, Yujun Zhou, Yanbo Wang, Jiayi Ye, Jiawen Shi, Qihui Zhang, Yuan Li, Han Bao, Zhaoyi Liu, Tianrui Guan, Dongping Chen, Ruoxi Chen, Kehan Guo, Andy Zou, Bryan Hooi Kuen-Yew, Caiming Xiong, Elias Stengel-Eskin, Hongyang Zhang, Hongzhi Yin, Huan Zhang, Huaxiu Yao, Jaehong Yoon, Jieyu Zhang, Kai Shu, Kaijie Zhu, Mohit Bansal, Ranjay Krishna, Swabha Swayamdipta, Taiwei Shi, Weijia Shi, Xiang Li, Yiwei Li, Yuexing Hao, Zhihao Jia, Zhize Li, Xiuying Chen, Zhengzhong Tu, Xiyang Hu, Tianyi Zhou, Jieyu Zhao, Lichao Sun, Furong Huang, Or Cohen Sasson, Prasanna Sattigeri, Anka Reuel, Max Lamparth, Yue Zhao, Nouha Dziri, Yu Su, Huan Sun, Heng Ji, Chaowei Xiao, Nitesh V. Chawla, Jian Pei, Jianfeng Gao, Michael Backes, Philip S. Yu, Neil Zhenqiang Gong, Pin-Yu Chen, Bo Li, Xiangliang Zhang
Preprint 2025

How Susceptible Are Large Language Models to Ideological Manipulation?
Kai Chen, Zihao He, Jun Yan, Taiwei Shi, Kristina Lerman
EMNLP 2024
🏆 *Best Paper Runner-up* at ICLR 2024 Workshop on Secure and Trustworthy Large Language Model

Safer-Instruct: Aligning Language Models with Automated Preference Data
Taiwei Shi, Kai Chen, Jieyu Zhao
NAACL 2024

Can Language Model Moderators Improve the Health of Online Discourse?
Hyundong Cho, Shuai Liu, Taiwei Shi, Darpan Jain, Basem Rizk, Yuyang Huang, Zixun Lu, Nuan Wen, Jonathan Gratch, Emilio Ferrara, Jonathan May
NAACL 2024

CoAnnotating: Uncertainty-Guided Work Allocation between Human and Large Language Models for Data Annotation
Minzhi Li, Taiwei Shi, Caleb Ziems, Min-Yen Kan, Nancy F. Chen, Zhengyuan Liu, Zhengyuan Liu, Diyi Yang
EMNLP 2023

Neural Story Planner
Anbang Ye, Christopher Zhang Cui, Taiwei Shi, Mark Riedl
AAAI 2023 Workshop on Creative AI Across Modalities

Investigating African American Vernacular English in Question Answering Systems
Taiwei Shi
Georgia Tech Undergraduate Thesis

EXPERIENCE

USC Language, Intelligence, and Model Evaluation Lab Aug 2023 - Present
Research Assistant for Prof. Jieyu Zhao *Los Angeles, CA*

- Working on alignment, synthetic data, and reinforcement finetuning (RFT).

Microsoft Office of Applied Research Aug 2024 - Present
Research Intern for Dr. Sihao Chen and Dr. Longqi Yang *Redmond, WA*

- Working on reinforcement finetuning, human-AI collaboration, and AI privacy.

Microsoft Research May 2024 - August 2024
Research Intern for Prof. Jennifer Neville and Dr. Longqi Yang *Redmond, WA*

- Worked on aligning large language models with in-situ user interactions and feedback.

Stanford NLP Group April 2023 - Aug 2023
Research Assistant for Prof. Diyi Yang *Remote*

- Worked on human-AI collaboration on data annotation.

Georgia Tech Entertainment Intelligence Lab Aug 2022 - May 2023
Research Assistant for Prof. Mark Riedl *Atlanta, GA*

- Worked on commonsense reasoning and ending-guided story generation.

USC Information Sciences Institute May 2022 - Dec. 2022
Research Intern for Prof. Jonathon May, Xuezhe Ma *Marina del Rey, CA*

- Worked on combating norm violation in social media using nonviolent communication and large language models.

Georgia Tech Social and Language Technologies Lab Aug 2021 - Aug 2022
Research Assistant for Prof. Diyi Yang *Atlanta, GA*

- Worked on the robustness of language models on different dialects.

Nanyang Technological University NLP Group Jun 2021 - Mar 2022
Research Assistant for Prof. Luu Anh Tuan *Remote*

- Worked on contextualized hate speech classifiers.

AWARDS

Best Paper Runner-up at ICLR 2024 Workshop on SeT LLM May 2024
Best Paper Runner-up at ICLR 2024 Workshop on Secure and Trustworthy Large Language Model

Faculty Honors and Dean's List May 2021 - May 2023
Georgia Tech's Faculty Honors and Dean's List recognize a student's commitment to academic excellence.

Convergence Innovation Competition Runner-Up Nov 2022
Runner-up in the Convergence Innovation Competition at Georgia Tech.

TALKS

Improving Moderation via Nonviolent Communication Aug 2022
USC Information Sciences Institute

MENTORSHIP

Wendy Wu, B.S. in Computer Science, University of Southern California Aug 2024 - Present
Working on Reinforcement Finetuning (RFT).

Yijun Pan, B.S. in Computer Science, University of Michigan May 2024 - Present
Working on detecting unsafe training data via data attribution.

SERVICE

Reviewer	NAACL 2024, COLM 2024, ACL 2025, NeurIPS 2025
Volunteer	NAACL 2024

SKILLS

Programming	Python, C/C++, Java, Javascript, CSS, HTML, SQL, Rust
Framework & Tools	PyTorch, Hugging Face, NumPy, pandas, Azure, PySpark
Languages	Chinese (native), English (fluent), French (basic)
Other Interests	Philosophy, Table Tennis, Chess, Cooking