TAIWEI SHI

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

EDUCATION

University of Southern California, Ph.D. in Computer Science — Aug 2023 - May 2027 (Expected)
Advisor: Prof. Jieyu Zhao
GPA: 4.0

Georgia Institute of Technology, Bachelor of 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

How Susceptible are Large Language Models to Ideological Manipulation?
Kai Chen, Zihao He, Jun Yan, Taiwei Shi, Kristina Lerman
🏆 *Best Paper Runner-up*
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 Ethics Lab — Aug 2023 - Present
Research Assistant for Prof. Jieyu Zhao — *Los Angeles, CA*

- Working on alignment and safety of large language models.

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

- Explored human-AI collaboration on data annotation.
- Worked on uncertainty estimation and multi-objective optimization.

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

- Extracted commonsense knowledge from large language models for ending-guided story generation.
- Explored chain-of-thoughts reasoning and prompt engineering for explainable question-answering calibration.

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.
- Improved blenderbot and GPT-style models' performance on content moderation.
- Conducted human evaluation on dialogue models.

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

- Explored the robustness of QA and deep generative models on different dialects.
- Created the AAVE version of SQuAD and CoQA datasets.
- Revealed a drop of up to 20% in model performance on SQuAD and CoQA.

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

- Contextualized hate speech classifiers and mitigated spurious relationship in deep learning models.
- Mitigated transformer-based model's over-reliance on spurious relationships.

AWARDS

Convergence Innovation Competition Runner-Up Nov 2022
Runner-up (top 2) in the Convergence Innovation Competition at Georgia Tech.

Faculty Honors and Dean's List May 2021 - Dec 2022
Georgia Tech's Faculty Honors and Dean's List recognize a student's commitment to academic excellence.

TALKS

Improving Moderation via Nonviolent Communication Aug 2022
USC Information Sciences Institute

SKILLS

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