KAITLYN ZHOU

katezhou@stanford.edu
(425) 373 - 6037

EDUCATION

Stanford University
Ph.D. Computer Science, 2019-
Advisor: Professor Dan Jurafsky

University of Washington
M.S. Computer Science, 2019
B.S. Computer Science, 2018
B.S.E. Human Centered Design and Engineering, 2018
Advisor: Professor Kate Starbird

INDUSTRY EXPERIENCE

Allen Institute for AI Seattle, Washington
Research Intern, Summer 2023
Conducted research on the use of epistemic markers by large language models. Findings illustrate the impacts of how RLHF and finetuning have resulted in overconfidence in language generation.
Mentors: Xiang Ren, Maarten Sap, and Jena Hwang

Microsoft Research Montréal, Canada
Research Intern, Summer 2021
Fairness, Accountability, Transparency, and Ethics in AI (FATE) Team
Led a research study on the unnamed practices and assumptions of natural language generation evaluation and its impacts on fairness and inclusion.
Mentors: Alexandra Olteanu, Su Lin Blodgett, Adam Trischler, Hal Daumé III

Docugami Seattle, Washington
Product Manager and Data Science Intern, 2019
Led the design of the flagship product, conducted user research studies, and ran natural language processing experiments at an eight-person start-up.

AWARDS & GRANTS

Microsoft Accelerating Foundation Models Research Grant
Multi-lingual Expressions of Uncertainty ($20,000)

IBM Research Grant
Safety Risks of Language Model Overconfidence ($85,000)

Stanford University
Stanford Graduate Fellowship, 2019-2022 - Full tuition and stipend

University of Washington
College of Engineering Dean's Medal of Academic Excellence, 2018

Phi Beta Kappa
Inducted as a Junior, 2017

PUBLICATIONS

12. REL-A.I.: An In Situ Framework for Measuring Human-LM Reliance
Kaitlyn Zhou, Jena Hwang, Nouha Dziri, Xiang Ren, Dan Jurafsky, Maarten Sap
(Under Review)

11. Relying on the Unreliable: The Impact of Language Models' Reluctance to Express Uncertainty
Kaitlyn Zhou, Jena Hwang, Xiang Ren, Maarten Sap
ACL 2024

10. Navigating the Grey Area: How Expressions of Uncertainty and Overconfidence Affect Language Models
Kaitlyn Zhou, Dan Jurafsky, Tatsunori Hashimoto
EMNLP 2023 (Oral)

9. Spotlight Tweets: Attention Dynamics Within Online Sensemaking During Crisis Events.
Kaitlyn Zhou, Tom Wilson, Kate Starbird, Emma Spiro.
Transactions of Social Computing (Journal - 2023), IC2S2 (oral - 2024)

8. Deconstructing NLG Evaluation: Evaluation Practices, Assumptions, and Their Implications
Kaitlyn Zhou, Su Lin Blodgett, Adam Trischler, Hal Daumé III, Kaheer Suleman, Alexandra Olteanu
NAACL 2022 (Oral)

7. Problems with Cosine as a Measure of Embedding Similarity for High Frequency Words
Kaitlyn Zhou, Kawin Ethayarajh, Dallas Card, Dan Jurafsky
ACL 2022

6. Richer Countries and Richer Representations
Kaitlyn Zhou, Kawin Ethayarajh, Dan Jurafsky
ACL 2022 - Findings

5. On the Opportunities and Risks of Foundation Models
Rishi Bommasani, Drew A. Hudson ... **Kaitlyn Zhou**, Percy Liang et al.
arXiv:2108.07258, 2021.

4. The Disagreement Deconvolution: Bringing Machine Learning Performance Metrics In Line With Reality
Mitchell L. Gordon, **Kaitlyn Zhou,** Kayur Patel, Tatsunori Hashimoto, Michael S. Bernstein. In Proceedings of the International ACM Conference on Human Factors in Computing Systems.
CHI 2021

3. Assembling Strategic Narratives: Information Operations as Collaborative Work Within an Online Community.
Tom Wilson, **Kaitlyn Zhou**, and Kate Starbird. Proceedings of the ACM on Human-Computer Interaction.
CSCW 2018

2. Centralized, Parallel, and Distributed Information Processing During Collective Sensemaking.
Peter Krafft, **Kaitlyn Zhou**, Isabelle Edwards, Kate Starbird, and Emma S. Spiro. In Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems.
CHI 2017

1. Could This be True? I Think So! Expressed Uncertainty in Online Rumoring.
Kate Starbird, Emma Spiro, Isabelle Edwards, **Kaitlyn Zhou**, Jim Maddock, and Sindhuja Narasimhan. In Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems.
CHI 2016

LEADERSHIP & SERVICE

Stanford University - CS Ph.D. Student Advisory Council
Founder and Chair, 2020-2022
Founder and chair of the Ph.D. student advisory council. Collaborating with student leaders and department leadership on a variety of student needs such as diversity and inclusion and student well-being.

University of Washington - Board of Regents
Regent, 2018-2019
Appointed by Washington State Governor, Jay Inslee, to serve as the student regent. Oversaw performance of three campuses, an 8 billion dollar operating budget, and four regional hospitals.

MENTORING

Sally Gao (Master's student, Stanford)
Zouberou Sayibou (Senior, Stanford)
Michael Brockman (Junior, Stanford)
Sofia Kim (Junior, Stanford)
Sarah Chen (Junior, Stanford)
Caeley Woo (Sophomore, Stanford)
Neil Rathi (Sophomore, Stanford)
Bolu Aminu (Freshmen, Stanford)

SERVICE

CS Diversity Committee 2019-2020
School of Engineering's Inaugural Dean's Advisory Council 2019-2021
CS Rebuilding Community Working Group 2021
Co-author of the PhD Climate Survey 2021
Reviewer for the Application Assistance Program 2021 -
Mentor for STEM Fellows 2022-
Speaker at the Faculty Retreat - 2021
Speaker at the DEI Town Hall Meeting 2021

TEACHING

Stanford University, Stanford, California
NLP for Computational Social Science - Professor: Diyi Yang
Lectured, mentored students on quarter-long projects, graded presentations and designed assignments.

University of Washington, Seattle, Washington
Data Visualization - Professor: Jeffrey Heer
Teaching assistant for upper level data visualization course. Lectured, advised student projects, and graded final assignments and homework.

University of Washington, Seattle, Washington
Foundations of Computing II- Professor: Anup Rao
Teaching assistant for introductory computing class in probability and statistics. Led sections, graded homework and finals. Rewrote homework problems to incorporate inclusive and gender-neutral language.

LANGUAGES

English - Native, Fluent
French - Fluent, Diplôme approfondi de langue française (DALF) - C1
Mandarin - Native