Yuxin Xiao

Research Interest

My current research focuses on ethical and deployable large language models (LLMs) for healthcare. I am particularly interested in evaluating and enhancing LLM alignment in safety and faithfulness.

Education

Massachusetts Institute of Technology (MIT)
Ph.D. in Social & Engineering Systems and Statistics
Advised by Prof. Marzyeh Ghassemi. GPA: 5.00/5.00
Affiliated with Institute for Data, Systems, and Society (IDSS) and Laboratory for Information and Decision Systems (LIDS)

Carnegie Mellon University (CMU)
M.S. in Machine Learning
Advised by Prof. Eric Xing and Prof. Louis-Philippe Morency. GPA: 4.12/4.33

University of Illinois at Urbana-Champaign (UIUC)
B.S. in Computer Science; B.S. in Statistics, Mathematics
Advised by Prof. Jiawei Han and Prof. Hari Sundaram. GPA: 3.93/4.00

Awards & Honors

2020 C. W. Gear Outstanding Undergraduate Award, UIUC (2 at UIUC)
2020 CRA Outstanding Undergraduate Researcher Award (Honorable Mention), CRA (4 at UIUC)

Publications

(* indicates equal contribution)

Speak Easy: Eliciting Harmful Jailbreaks from LLMs with Simple Interactions
Yik Siu Chan*, Narutatsu R*, YUXIN XIAO*, Marzyeh Ghassemi
Preprint 2025 [Paper], [Code]

SFTMix: Elevating Language Model Instruction Tuning with Mixup Recipe
YUXIN XIAO*, Shujian Zhang, Wenxuan Zhou, Marzyeh Ghassemi, Sanqiang Zhao
Preprint 2024 [Paper], [Code]

In the Name of Fairness: Assessing the Bias in Clinical Record De-identification
YUXIN XIAO*, Shuhammitte Lim*, Tom Joseph Pollard, Marzyeh Ghassemi
FAccT 2023 [oral presentation]

Uncertainty Quantification with Pre-trained Language Models: A Large-Scale Empirical Analysis
YUXIN XIAO, Paul Pu Liang, Umang Bhatt, Willie Neiswanger, Ruslan Salakhutdinov, Louis-Philippe Morency
EMNLP 2022 [findings] [Paper], [Code]

SAIS: Supervising and Augmenting Intermediate Steps for Document-Level Relation Extraction
YUXIN XIAO, Zecheng Zhang, Yuning Mao, Carl Yang, Jiawei Han
NAACL 2022 [oral presentation] [Paper], [Code]

Amortized Auto-Tuning: Cost-Efficient Bayesian Transfer Optimization for Hyperparameter Recommendation
YUXIN XIAO, Eric P. Xing, Willie Neiswanger
Preprint 2021 [Paper], [Code]

Heterogeneous Network Representation Learning: A Unified Framework with Survey and Benchmark
CARL YANG*, YUXIN XIAO*, Yu Zhang*, Yizhou Sun, Jiawei Han
IEEE BigData 2019 [oral presentation]

Discovering Strategic Behaviors for Collaborative Content-Production in Social Networks
YUXIN XIAO, Adit Krishnan, Hari Sundaram
WWW 2020 [Paper], [Code]

Non-local Attention Learning on Large Heterogeneous Information Networks
YUXIN XIAO*, Zecheng Zhang*, Carl Yang, Chengxiang Zhai
IEEE BigData 2019 [oral presentation]

Industry Experience

Zoom  RESEARCH INTERN, GenAI R&D Team
San Jose, CA 06/2024 - 08/2024

Bosch Center for AI  RESEARCH INTERN, NLP Team
Renningen, Germany 06/2023 - 08/2023

Cars.com  SOFTWARE ENGINEER INTERN, Mobile Development Team
Chicago, IL 01/2018 - 08/2018

Teaching Experience

MIT  14.310X DATA SCIENCE FOR SOCIAL SCIENTISTS, Teaching Assistant & Session Instructor
Summer 2022

UIUC  CS446 MACHINE LEARNING, Course Assistant
Fall 2019

UIUC  CS410 TEXT INFORMATION SYSTEMS, Course Assistant
Fall 2019

UIUC  CS125 INTRO TO COMPUTER SCIENCE, Course Assistant
Spring 2017