Yuxin Xiao

yuxin102@mit.edu | xiaoyuxin1002.github.io | xiaoyuxin1002 | xiaoyuxin1002 | @YuxinXiao6

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 terms of safety, faithfulness, and fairness.

Education

Massachusetts Institute of Technology (MIT) — Cambridge, MA

PH.D. IN SOCIAL & ENGINEERING SYSTEMS AND STATISTICS — 09/2022 - Present

- 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) — Pittsburgh, PA

M.S. IN MACHINE LEARNING — 08/2020 - 12/2021

- Advised by **Prof. Eric Xing** and **Prof. Louis-Philippe Morency**; GPA: 4.12/4.33

University of Illinois at Urbana-Champaign (UIUC) — Urbana, IL

B.S. IN COMPUTER SCIENCE; B.S. IN STATISTICS, MATHEMATICS — 08/2016 - 05/2020

- 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)

SFTMix: Elevating Language Model Instruction Tuning with Mixup Recipe — Preprint 2024
YUXIN XIAO, SHUJIAN ZHANG, WENXUAN ZHOU, MARZYEH GHASSEMI, SANQIANG ZHAO — [Paper], [Code]

HarmScore and UserAttack: Towards Harmful LLM Jailbreak From a User Perspective — Preprint 2024
YIK SIU CHAN*, NARUTATSU RI*, YUXIN XIAO*, MARZYEH GHASSEMI — [Paper], [Code]

In the Name of Fairness: Assessing the Bias in Clinical Record De-identification — FAccT 2023
YUXIN XIAO*, SHULAMMITE LIM*, TOM JOSEPH POLLARD, MARZYEH GHASSEMI *(oral presentation)* — [Paper], [Code]

Uncertainty Quantification with Pre-trained Language Models: A Large-Scale Empirical Analysis — EMNLP 2022
YUXIN XIAO, PAUL PU LIANG, UMANG BHATT, WILLIE NEISWANGER, RUSLAN SALAKHUTDINOV, LOUIS-PHILIPPE MORENCY *(findings)* — [Paper], [Code]

SAIS: Supervising and Augmenting Intermediate Steps for Document-Level Relation Extraction — NAACL 2022
YUXIN XIAO, ZECHENG ZHANG, YUNING MAO, CARL YANG, JIAWEI HAN *(oral presentation)* — [Paper], [Code]

Amortized Auto-Tuning: Cost-Efficient Bayesian Transfer Optimization for Hyperparameter Recommendation — Preprint 2021
YUXIN XIAO, ERIC P. XING, WILLIE NEISWANGER — [Paper], [Code]

Heterogeneous Network Representation Learning: A Unified Framework with Survey and Benchmark — TKDE 2020
CARL YANG*, YUXIN XIAO*, YU ZHANG*, YIZHOU SUN, JIAWEI HAN *(330+ citations, 300+ GitHub stars and forks)* — [Paper], [Code]

Discovering Strategic Behaviors for Collaborative Content-Production in Social Networks — WWW 2020
YUXIN XIAO, ADIT KRISHNAN, HARI SUNDARAM *(oral presentation)* — [Paper], [Code]

Non-local Attention Learning on Large Heterogeneous Information Networks — IEEE BigData 2019
YUXIN XIAO*, ZECHENG ZHANG*, CARL YANG, CHENGXIANG ZHAI *(oral presentation)* — [Paper], [Code]

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