YIAN WANG

+12176488679 ⋄ yian3@illinois.edu

RESEARCH INTERESTS

Machine unlearning, causal representation learning, robustness and safety of LLMs, interpretability, and adversarial ML.

EDUCATION

University of Illinois Urbana Champaign, US *August 2023 - Present*
PhD in Computer Science

University of Illinois Urbana Champaign, US *August 2021 - May 2023*
Master of Science in Computer Science
Grade Point Average: 4.0/4.0

University of Science and Technology of China Hefei, China *September 2016 - June 2020*
Bachelor of Science in physics and minor in Computer Science
Grade Point Average: 3.64/4.3 in physics, 4.17/4.3 in Computer Science
TOEFL iBT: Total:104, Reading: 27, Listening: 26, Speaking: 23, Writing: 28
GRE: Verbal: 158, Quantitative: 169, Analytical Writing: 3.5

PUBLICATIONS

- **Yian Wang**, Yuen Chen, Agam Goyal, Hari Sundaram. *CausalDetox: Causal Head Selection and Intervention for Language Model Detoxification.* In Findings of the Association for Computational Linguistics: ACL 2026.
- **Yian Wang**, Mukhil Shankar, Eshwar Chandrasekharan, Hari Sundaram. The Chilling: Identifying Strategic Antisocial Behavior Online and Examining the Impact on Journalists. In Proceedings of the 19th ACM conference on computer-supported cooperative work & social computing (CSCW), 2025.
- **Yian Wang**, Ali Ebrahimpour-Boroojeny, Hari Sundaram. On the Necessity of Output Distribution Reweighting for Effective Class Unlearning. arXiv:2506.20893, 2025. (Under review).
- **Yian Wang**, Jian Kang, Yinglong Xia, Jiebo Luo, Hanghang Tong. iFiG: Individually Fair Multi-view Graph Clustering. In 2022 IEEE International Conference on Big Data (Big Data), 2022.
- Agam Goyal*, **Yian Wang***, Eshwar Chandrasekharan, Hari Sundaram. *From Plausible to Causal: Counterfactual Semantics for Policy Evaluation in Simulated Online Communities.* In PoliSim Workshop at CHI 2026. **(Best Paper Nomination, Top 5 Papers)**.
- Hongtao Liu,**Yian Wang**, Qiyao Peng, Fangzhao Wu, Lin Gan, Lin Pan, Pengfei Jiao. Hybrid neural recommendation with joint deep representation learning of ratings and reviews. *Neurocomputing*, 374, 77-85, 2020.
- Hongtao Liu,**Yian Wang**, Fangzhao Wu, Pengfei Jiao, Hongyan Xu, Xing Xie. REKER: relation extraction with knowledge of entity and relation. In Natural Language Processing and Chinese Computing: 8th CCF International Conference, NLPCC, 2019.
- Agam Goyal, Vedant Rathi, William Yeh, **Yian Wang**, Yuen Chen, Hari Sundaram. Breaking Bad Tokens: Detoxification of LLMs Using Sparse Autoencoders. In Actionable Interpretability Workshop at ICML 2025.

*Equal contribution.

RESEARCH EXPERIENCE

University of Illinois Urbana Champaign Aug. 2023 - Present
Mentor: Dr.Hari Sundaram

· Use Transformer based model to do coordination detection and strategy discovery.
· Robust Class Unlearning: Design efficient method to forget entire classes from trained models without retraining, with strong empirical privacy and utility guarantees.
· Causally Sound LLM Unlearning: Design an intervention-based method with causal attribution to remove toxic behaviors from LLMs without harming general performance.

University of Illinois Urbana Champaign Aug. 2021 - May. 2023
Mentor: Dr.Hanghang Tong

· Conducted research on individually fair multi-view graph clustering.
· Conducted research on a fair graph learning model that is robust.

Microsoft Research Asia May. - Aug. 2021
Mentor: Dr.Fangzhao Wu

· Conducted a large-scale data collection.
· Conducted research and experiments on large-scale fair recommender systems for news recommendation.

University of Science and Technology of China Sep. 2018 - May. 2020
Mentor: Dr.Qi Liu

· Conducted research on explainable recommender systems.
· Conducted research on representation learning for hybrid recommender systems.

WORKING EXPERIENCE

Pinterest Seattle, WA
Machine Learning Internship May.2025 - Aug.2025
- Designed and implemented novel candidate models to improve L1 ad ranking performance while keeping low latency.
- Conducted utility-based model tuning to balance user engagement and advertiser value.

TEACHING EXPERIENCE

University of Illinois Urbana Champaign Urbana, IL
Teaching assistant Aug.2021 - Dec.2021
- Teaching assistant of CS105: Intro Computing: Non-Tech.
- Supported students for Q&A and course logistics.
- Lead lab sessions.
- Instructor: Dr. Craig Zilles

University of Illinois Urbana Champaign Urbana, IL
Teaching assistant Jan.2023 - May.2023
- Teaching assistant of CS412: Introduction to Data Mining.
- Assist with grading assignments and exams.
- Supervise students during office hours.
- Instructor: Dr. Hanghang Tong