

Jiaxu Zhao

📍 Eindhoven, NL ✉ zjx19960405@gmail.com ☎ +31 687 286 408

🔗 <https://jiaxu-zhao.github.io/> 🎓 Google Scholar

Research Interests

Large Language Models, Fairness and Bias, AI Safety, Causal Reasoning, Sparse Neural Network, Graph Neural Network, Robustness, Adversarial Attack

Education

Eindhoven University of Technology
PhD in Mathematics & Computer Science

Sep 2021 – Sep 2025

- Research Group: Data Mining
- Supervisors: Mykola Pechenizkiy, Meng Fang, Yulong Pei
- Research Focus: Understanding and Mitigating Unwanted Biases in Generative Language Models

University of Electronic Science and Technology of China
Master in Computer Technology

Sep 2018 – July 2021

- Research Group: CI Lab
- Supervisors: Xiaobin Wang, Hong Qu
- Thesis: *The Research and Solution of Exposure Bias in Neural Machine Translation*

Professional Experience

Research Scientist
OpenML

Eindhoven, The Netherlands
June 2024 – Aug 2024

- Developed and implemented a Retrieval-Augmented Generation (RAG) framework integrated with Large Language Models to enhance the dataset recommendation system for the OpenML platform.

Program Mentor
Responsible AI Fall 2023 Research Program
(New York University & Ukrainian Catholic University)

New York, United States
Aug 2023 – Nov 2024

- Students: Anastasiia Mazur, Bohdan Turbal
- Project: This project investigates how transfer learning affects the adversarial robustness of large language models (LLMs). Although transfer learning improves standard accuracy, the study finds it often increases vulnerability to adversarial attacks—especially in smaller models—highlighting a critical trade-off between performance and robustness.

AI Researcher (Intern)
Digital Brain Laboratory

Shanghai, China
March 2023 – June 2023

- Responsibility: Evaluated fairness and bias in Chinese LLMs using Statistical Parity, KL divergence, and group-wise accuracy across demographic groups. Applied Proximal Policy Optimization (PPO) for Reinforcement Learning from Human Feedback (RLHF) to guide behavior alignment. Focused on Chinese Large Language Models development and evaluation.

Selected Publications

Jiaxu Zhao, Meng Fang, Fanghua Ye, Ke Xu, Qin Zhang, Joey Tianyi Zhou, Mykola Pechenizkiy. **Understanding Large Language Model Vulnerabilities to Social Bias Attacks.** *ACL Oral 2025*

Jiaxu Zhao, Meng Fang, Kun Zhang, Mykola Pechenizkiy. **Unmasking Style Sensitivity: A Causal Analysis of Bias Evaluation Instability in Large Language Models.** *ACL 2025*

Jiaxu Zhao, Tianjin Huang, Shiwei Liu, Jie Yin, Yulong Pei, Meng Fang, Mykola Pechenizkiy. **FS-GNN: Improving Fairness in Graph Neural Networks via Joint Sparsification.** *Neurocomputing 2025*

Xiao Xiao, **Jiaxu Zhao**, Terry Payne, Meng Fang. **Empirical Study of Social Bias in Medical Question Answering via Large Language Models.** *AiIH 2025*

Qin Zhang, Sihan Cai, **Jiaxu Zhao**, Mykola Pechenizkiy, Meng Fang. **CHAmbi: A New Benchmark on Chinese Ambiguity Challenges for Large Language Models.** *EMNLP findings 2024*

Jiaxu Zhao, Zijing Shi, Yitong Li, Yulong Pei, Ling Chen, Meng Fang, Mykola Pechenizkiy. **More than Minorities and Majorities: Understanding Multilateral Bias in Language Generation.** *ACL findings 2024*

Jiaxu Zhao, Meng Fang, Zijing Shi, Yitong Li, Ling Chen, Mykola Pechenizkiy. **CHBias: Bias Evaluation and Mitigation of Chinese Conversational Language Models.** *ACL 2023*

Jiaxu Zhao, Lu Yin, Shiwei Liu, Meng Fang, Mykola Pechenizkiy. **Rest: Enhancing group robustness in DNNs through reweighted sparse training.** *ECML PKDD 2023*

Bohdan Turbal, Anastasiia Mazur, **Jiaxu Zhao**, Mykola Pechenizkiy. **On Adversarial Robustness of Language Models in Transfer Learning.** *NeurIPS SoLaR 2024*

Wei Wu, Junjie Xiao, **Jiaxu Zhao**, Jianxin Wang, Meng Fang. **Enhancing Long-Form Question Answering via Reflection with Question Decomposition.** *Information Processing and Management 2025*

Tianjin Huang, Tianlong Chen, Meng Fang, Vlado Menkovski, **Jiaxu Zhao**, Lu Yin, Yulong Pei, Decebal Constantin Mocanu, Zhangyang Wang, Mykola Pechenizkiy, Shiwei Liu. **You Can Have Better Graph Neural Networks by Not Training Weights at All: Finding Untrained GNNs Tickets.** *LoG (Best Paper Award) 2022*

Jiaxu Zhao, Meng Fang, Shirui Pan, Wenpeng Yin, Mykola Pechenizkiy **GPTbias: A comprehensive framework for evaluating bias in large language models** *arXiv preprint arXiv:2312.06315 2023*

Shenghui Li, Fanghua Ye, Meng Fang, **Jiaxu Zhao**, Yun-Hin Chan, Edith C-H Ngai, Thiemo Voigt **Synergizing foundation models and federated learning: A survey** *arXiv preprint arXiv:2406.12844 2024*

Supervision Activities

Master Projects

- Large Language Model-Based Chatbots and Their Applications. (2025.07 - Now, Co-supervised with Prof. Meng Fang) Student: Dan Liu
- Persuasiveness and Bias in LLM: Investigating the Impact of Persuasiveness and Reinforcement of Bias in Language Models. (2024.4 - 2025.7, Co-supervised with Prof. Mykola Pechenizkiy) Student: Saumya Roy
- Exploring the synergy within a context-aware and domain-flexible pipeline for neural machine translation. (2022.11 - 2023.7, Co-supervised with Prof. Meng Fang) Student: Luc Geven

Teaching

Assistant Lecturer

Eindhoven University of Technology

Eindhoven, The Netherlands

Oct 2024 – Jan 2025

- Professor: Sibylle Hess
- Course: 2HIG0 Data Mining and Machine Learning Course

Service

Journal Reviewer

- Artificial Intelligence Research (JAIR), Neurocomputing.

Conference Reviewer

- ACL 2023, ICML 2024, ACL ARR 2024, EMNLP 2024, ICLR 2025, KDD 2025 (Outstanding Reviewer), ACL ARR 2025, EWAF 2025 Workshop.

Volunteer

- EWAF 2025 Workshop.