Jiaxu Zhao

PhD Candidate,
TU Eindhoven

j.zhao@tue.nl
+31 687 286 408
zhao1402072392.github.io

Languages

Chinese ●●●●●
English ●●●●○
Dutch ●●○○○

Skills

Coding: Python, C

Frameworks: Keras, Pytorch, Matlab

Research Interests

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

About Me

I am passionate about fairness and robustness in NLP, with a strong interest in interdisciplinary collaboration. I value diverse perspectives and believe that responsible AI starts from inclusive design and transparent evaluation.

Beyond the academic adventure, I also dive into photography, horse riding, squash, and badminton.

Looking For

Actively seeking research scientist or applied scientist roles focusing on NLP, LLMs, and AI Safety, in either industry research labs or academia.

Education

2021.9 – 2025.9 **PhD - Eindhoven University of Technology**
Research Group: Mathematics & Computer Science, Data Mining
Supervisors: Mykola Pechenizkiy, Meng Fang, Yulong Pei
Research Focus: Fairness and Bias in Generative Language Models

2018.9 – 2021.7 **Master - University of Electronic Science and Technology of China**
Research Group: Computer Technology, CI Lab
Supervisors: Xiaobin Wang, Hong Qu
Thesis: The Research and Solution of Exposure Bias in Neural Machine Translation.

Selected Publications

Jiaxu Zhao, et al. Understanding Large Language Model Vulnerabilities to Social Bias Attacks ***ACL 2025***

Jiaxu Zhao, et al. Unmasking Style Sensitivity: A Causal Analysis of Bias Evaluation Instability in Large Language Models ***ACL 2025***

Jiaxu Zhao, et al. FS-GNN: Improving Fairness in Graph Neural Networks via Joint Sparsification ***Neurocomputing 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***

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

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***

Jiaxu Zhao, et al. More than Minorities and Majorities: Understanding Multilateral Bias in Language Generation. ***ACL findings 2024***

Jiaxu Zhao, et al. CHBias: Bias Evaluation and Mitigation of Chinese Conversational Language Models. ***ACL 2023***

Jiaxu Zhao, et al. Gptbias: A comprehensive framework for evaluating bias in large language models. ***arXiv preprint arXiv:2312.06315***

Jiaxu Zhao, et al. Rest: Enhancing group robustness in DNNs through reweighted sparse training. ***ECML PKDD 2023***

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 2022 (Oral and Best Paper Award)***

Internship Experience

2023.3 – 2023.6 **Digital Brain Laboratory, Shanghai** Student Researcher
Project: Chinese Large Language Models.
Responsibilities: 1) Evaluated fairness and bias in Chinese LLMs using Statistical Parity, KL divergence, and group-wise accuracy across demographic groups. 2) Applied Proximal Policy Optimization (PPO) for Reinforcement Learning from Human Feedback (RLHF) to guide behavior alignment.

Selected Projects

2023.9 – 2024.8 **Responible AI Fall 2023 Research Program**
Led by Prof. Julia Stoyanovich, New York University
Co-supervise with Prof. Mykola Pechenizkiy
Research Project: Investigated robustness of LLMs under adversarial attacks post transfer learning. Found that larger models are more resilient to adversarial attacks, revealing trade-offs between adaptation and security.

2024.7 – 2024.8 **OpenML AI Search Project**
Led by OpenML and Prof. Joaquin Vanschoren
Research Project: Developed a retrieval-augmented generation (RAG) framework for dataset recommendation. Integrated semantic similarity and query reformulation modules to enhance recommendation quality.

2022.11 – 2023.7 **Supervised Project**
Co-supervise with Prof. Meng Fang
Research Project: Exploring the synergy within a context-aware and domain-flexible pipeline for neural machine translation.