ZHI YANG

+86-138-2874-1098 | 12332454@mail.sustech.edu.cn | ruayz.github.io

Research area: Trustworthy Machine Learning (ML) and Large Language Models (LLMs).

Education

- **Southern University of Science and Technology (SUSTech)** *Sep. 2023 - Present*
  Master's degree in Computer Science and Technology
  - Supervisor: Changwu Huang, Xin Yao
  - GPA: 3.5/4.0
  - Core courses: Advanced Artificial Intelligence (B), Advanced Algorithms (A-), Numerical Methods(B), Human Brain Intelligence and Machine Intelligence (A-), etc.
- **Yunnan University (YNU)** *Sep. 2019 - Jun. 2023*
  Bachelor's degree in Software Engineering
  - **GPA: 3.8/4.0 (Top 2%)**
  - Core courses: Calculus B(1) (92), Discrete Mathematics (94), Probability Statistics (95), The Design Practice of Python Language (94), Computer Networking (95), etc.

Publications

C=Conference, S=In Submission

[S.1] Zhi Yang, Changwu Huang, Ke Tang, Xin Yao. **On the Fairness of Privacy Protection: Measuring and Mitigating the Disparity of Group Privacy Risk for Differentially Private Machine Learning**. arXiv preprint arXiv:2510.09114.

[C.1] Zhi Yang, Changwu Huang, Xin Yao. **Towards Private and Fair Machine Learning: Group-Specific Differentially Private Stochastic Gradient Descent with Threshold Optimization**. *The International Conference on Neural Information Processing (ICONIP)*. Singapore: Springer Nature Singapore, 2024: 66-80.

[C.2] Zhi Yang, Ziming Wang, Changwu Huang, Xin Yao. **An Explainable Feature Selection Approach for Fair Machine Learning**. *International Conference on Artificial Neural Networks (ICANN)*. Cham: Springer Nature Switzerland, 2023: 75-86.

Research Experience

- **Impact of Visual Redundancy on Multimodal Reasoning** *Jan. 2025 - Present*
  Ongoing Project
  - Investigated the phenomenon of **information redundancy** in Vision-Language Models (VLMs), specifically how excessive visual tokens can degrade performance in cross-modality question answering tasks.
- **Adaptive and Continual LLM Personalization via Interaction** *Oct. 2025 - Present*
  Ongoing Project
  - Developed a framework for **rapid adaptation** to user backgrounds using activation engineering, effectively mitigating the cold-start problem without the need for resource-intensive fine-tuning; Engineered **a closed-loop feedback mechanism to facilitate continual learning**, enabling the model to dynamically evolve and refine its alignment through multi-turn user interactions.
- **Emotion Influence on LLM-Based Multi-Agent Systems** *Jun. 2025 - Sep. 2025*
  Concluded Project[]
  - Investigated how emotional stimuli influence agent behavior in multi-agent LLMs systems across diverse network topologies, finding no consistent sensitivity patterns across node positions or propagation structures, and showing that modern LLMs remain robust against emotion-induced perturbations.
- **Fair Privacy [S.1]** *Nov. 2024 - May. 2025*
  Master's Thesis []
  - Summary: We address the question of whether AI systems expose different subgroups to unequal privacy risks by proposing **a stricter measurement for group privacy risk and an effective mitigation strategy**. Existing methods are either inaccurate or computationally expensive. We introduce a novel membership inference game to efficiently estimate worst-case privacy risks, enabling a more reliable assessment of disparities across groups; To enhance fairness, we develop a group-specific adaptive gradient clipping method for DP-SGD that reduces differences in privacy risks among subgroups.
- **Fairness & Privacy [C.1]** *Mar. 2024 - Jul. 2024*
  Master's Thesis []

  - Summary: Recent research highlights the challenges of integrating differential privacy (DP) with group fairness. Existing methods typically address either the amplified accuracy disparities across sensitive groups caused by DP or the potential outcome unfairness in DP-trained models, often in isolation. We propose a group-specific DP-SGD training framework combined with classification threshold optimization, which **jointly reduces accuracy disparities while achieving outcome fairness**.

- **Explainability & Fairness [C.2]** *Feb. 2023 - Jun. 2023*
  Bachelor's Thesis
  - Summary: In this work, we propose an **explainable feature selection (ExFS) method** to enhance model outcome fairness by recursively removing features that contribute to unfairness, guided by the **feature attribution explanations (FAE) of the model's predictions**.

Honors and Awards

• Top Academic Scholarship (Top 20% in Dept. of CS)	*Oct. 2025*
• **Outstanding Graduate Teaching Assistant**	*Sep. 2024*
• Top Research Assistance Scholarship (Top 20% in Dept. of CS)	*Sep. 2024*
• **National Scholarship (Top 1% in School of Software)**	*Dec. 2022*
• National Silver Award, 8th Internet+ Innovation and Entrepreneurship Competition	*Nov. 2022*
• University-level First-class Scholarship	*Dec. 2021*
• University-level First-class Scholarship	*Dec. 2020*

Other Experience

- **Teaching Assistant** *Sep. 2023 - Jun. 2025*
  SUSTech
  - Overview: Served as a teaching assistant for four semesters in *the Fundamentals of Python Programming* course offered by the Department of Computer Science and Technology.
  - Responsibilities: Led lab sessions by preparing instructional materials and slides, explaining core concepts, guiding hands-on exercises, and addressing student questions. Designed assignments, quizzes, and exams to evaluate student learning and provide constructive feedback.

- **Project Experience**
  SUSTech/YNU
  - SUSTech Trusted Intelligent Systems Innovation Laboratory AI Transparency Project: Conducted a comprehensive review and analysis of existing transparency concepts by synthesizing insights from academic literature on AI ethics, societal impacts, and regulatory frameworks (including the EU AI Act and Ethics Guidelines for Trustworthy AI).
  - Knowledge Graph–Based Integrated Health Knowledge System: Designed and implemented an intelligent Q&A platform covering traditional Chinese medicine and mental health knowledge; built the web application using CSS, HTML, and JavaScript, implemented the backend with a Neo4j-based database, and applied the Aho–Corasick string matching algorithm as a core technique.
  - Facial Recognition System: Integrated the Hongsoft facial recognition interface and developed the system using Spring Boot, MyBatis-Plus, and Angular frameworks.
  - Huan Yan software: Utilized the open-source GFPGAN algorithm to restore old and blurred photos; developed and deployed as a WeChat Mini Program using WXML, WXS, and JavaScript.