# ZHI YANG

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

Research area/interest: Trustworthy Machine Learning (ML), Large Language Models (LLMs).

## EDUCATION

- **Southern University of Science and Technology (SUSTech)** *Sep. 2023 - Present*
  *Master's degree in Computer Science and Technology (expected)*
  - Supervisor: Changwu Huang, Ke Tang, 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: Advanced Mathematics (92), Discrete Mathematics (94), Probability and Statistics (95), Computer Networks (95), Software Engineering (93), 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

- **Adaptive Personalized LLM for Multi-Turn Conversations**
  *Ongoing*
  - Aims to develop a novel personalized LLM capable of dynamically adapting its responses to a user's demographic characteristics, evolving preferences, background knowledge, emotional state, communication style, and personality across multi-turn dialogues.

- **Multi-Agent Systems with LLMs**
  *Partially ongoing*
  - Investigated the impact of emotional prompts on agent behaviors in a multi-agent LLMs system across diverse network topologies, concluding that modern LLMs are robust to such stimuli, with no consistent topology-dependent variance.
  - Explored distilling multi-agent collaboration into a single LLM, aiming to enable efficient collective reasoning within a unified model.

- **Fair Privacy [S.1]** *[]*
  *Master's Thesis*
  - Summary: While fairness in conventional machine learning and differentially private machine learning (DPML) has been extensively studied, **the fairness of privacy protection** across groups remains underexplored. Existing methods for assessing group privacy risks are either insufficiently accurate or computationally expensive. To address these limitations, we propose a novel membership inference game that efficiently approximates worst-case privacy risks, enabling **stricter and more reliable assessment of the disparity in group privacy risks**. To further enhance fairness, we introduce a **group-specific adaptive gradient clipping strategy** for DP-SGD, which effectively reduces disparities in group privacy risks.

- **Fairness & Privacy [C.1]** *[]*
  *Master's Thesis*
  - Summary: Recent research has highlighted the challenges of integrating differential privacy (DP) with group fairness. One line of work focuses on mitigating accuracy disparities across sensitive groups introduced by DP mechanisms, while another seeks to preserve outcome fairness in DP-trained models. However, **these two objectives often conflict**, and existing methods typically address them in isolation. To tackle both problems simultaneously, we propose a group-specific DP-SGD training framework combined with classification threshold optimization. Our approach jointly reduces accuracy disparity and achieves outcome fairness.

- **Explainability & Fairness [C.2]**
  *Bachelor's Thesis*
  - Summary: In this work, we propose an **explainable feature selection (ExFS) method** to improve the fairness of ML by recursively eliminating features that contribute to unfairness based on the **feature attribution explanations (FAE) of the model's predictions**.

## Honors and Awards

- 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 laboratory sessions by preparing instructional materials and slides, explaining core concepts, guiding hands-on exercises, and answering student questions; designed assignments/quizzes/exams to assess student learning and provide 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.