Hanrong Zhang

Tel: (+1)312-479-7822 | Email: zhanghr0709@gmail.com | WeChat: henry_zhang0709 | Homepage | Google Scholar

Education

University of Illinois Chicago — **Aug. 2025 – Present**
Ph.D. Student in Computer Science — *Chicago, USA*
Big Data and Social Computing Lab; Supervisor: Prof. Philip S. Yu

Zhejiang University — **Sep. 2022 – Mar. 2025**
MEng. Computer Engineering — *Hangzhou, China*
Ranking: 1/82; GPA: 93/100; Supervisor: Prof. Hongwei Wang

University of Leeds — **Sep. 2018 – Jun. 2022**
BSc. Computer Science - First Class Honors Degree — *Leeds, United Kingdom*

Southwest Jiaotong University — **Sep. 2018 – Jun. 2022**
BEng. Computer Science and Technology — *Chengdu, China*
Ranking: 1/74; GPA: 3.81/4.0, 92/100; Supervisor: Prof. Tianrui Li

Selected Papers

*Full publication list is available on Google Scholar. * denotes Equal Contribution.*

Trustworthy Machine Learning:

[1] **Hanrong Zhang**, Jingyuan Huang, Kai Mei, Yifei Yao, Zhenting Wang, Chenlu Zhan, Hongwei Wang, Yongfeng Zhang, *Agent Security Bench (ASB): Formalizing and Benchmarking Attacks and Defenses in LLM-based Agents*, **ICLR 2025**, Singapore, Apr. 2025. [Paper] [Code] [Website] (LLM Agent Security)

[2] **Hanrong Zhang**, Zhenting Wang, Tingxu Han, Mingyu Jin, Chenlu Zhan, Mengnan Du, Hongwei Wang, Shiqing Ma, *Invisible Backdoor Attack in Self-supervised Learning*, **CVPR 2025**, Nashville, USA, Jun. 2025. [Paper] [Code] (Machine Learning Safety)

Machine Learning and Data Mining:

[1] **Hanrong Zhang**, Yifei Yao, Zixuan Wang, Jiayuan Su, Mengxuan Li, Peng Peng, Hongwei Wang, *Class Incremental Fault Diagnosis under Limited Fault Data via Supervised Contrastive Knowledge Distillation*, **IEEE Transactions on Industrial Informatics**. (IF=12.3, JCR Q1 SCI) [Paper] (Continuous Learning)

[2] Xingyue Wang*, **Hanrong Zhang***, Xinlong Qiao, Ke Ma, Shuting Tao, Peng Peng, Hongwei Wang, *Generalized Out-of-distribution Fault Diagnosis (GOOFD) via Internal Contrastive Learning*, **IEEE Transactions on Industrial Informatics**. (IF=12.3, JCR Q1 SCI) [Paper] (OOD Distribution Detection)

[3] Peng Peng*, **Hanrong Zhang***, Xinyue Wang, Wanqiu Huang, Hongwei Wang, *Imbalanced Chemical Process Fault Diagnosis Using Balancing GAN With Active Sample Selection*, **IEEE Sensors Journal**. (IF=4.3, JCR Q1 SCI) [Paper] (Imbalanced Classification)

[4] Wanqiu Huang, **Hanrong Zhang**, Peng Peng, Hongwei Wang, *Multi-gate Mixture-of-Expert Combined with Synthetic Minority Over-sampling Technique for Multimode Imbalanced Fault Diagnosis*, **IEEE International Conference on Computer Supported Cooperative Work in Design 2023**. **(Best Paper Award Finalist)** [Paper] (Imbalanced Classification)

[5] **Hanrong Zhang**, Xinyue Wang, Jiabao Pan, Hongwei Wang, *SAKA: an intelligent platform for semi-automated knowledge graph construction and application*, **Service Oriented Computing and Applications**. [Paper] (Knowledge Graph)

[6] **Hanrong Zhang**, Xinyue Wang, Bo Qin, Hongwei Wang, *An Intelligent System for Semantic Information Extraction and Knowledge Graph Construction from Multi-Type Data Sources*, **IEEE ICEBE 2022**. [Paper] (Knowledge Graph)

Internship

RL for Human-Agent Multi-turn Interaction — **Alibaba Group**
Research Intern — *Hangzhou, May 2025 – Aug. 2025*

- To address the challenges of sparse rewards, limited data, and environment instability in LLM agent-human multi-turn interactions, I develop scalable simulation environments with various tool-call scenarios (e.g., retail, airline) to facilitate multi-turn interactions.
- I propose a novel Tool Dependency Graph structure to model tool dependencies, enabling high-quality, multi-turn tool-call dialogues and generating ground truth trajectories, which is used for verifiable reward signal computation in multi-turn conversations.
- I leverage this data to conduct multi-turn GRPO training, resulting in significant improvements in the agent's performance in tool usage during human-agent multi-turn interactions.

Selected Honors

Outstanding Graduate in Zhejiang Province & Zhejiang University	Zhejiang University, 2025
National Scholarship for Graduate Students (Top 0.2%)	Zhejiang University, 2023 - 2024
National Scholarship for Graduate Students (Top 0.2%)	Zhejiang University, 2022 - 2023
National Scholarship for Undergraduate Students (Top 0.2%)	Southwest Jiaotong University, 2019 - 2020
Outstanding Graduate in Sichuan Province	Southwest Jiaotong University, 2022
Pacemaker to Merit Student	Southwest Jiaotong University, 2018 - 2021
Best Student in Computer Science (1/75)	University of Leeds, 2020 - 2021
First-class full-ride Scholarship (1/75)	University of Leeds, 2020 - 2021
Best Student Overall (1/300, 4 majors)	University of Leeds, 2018 - 2019

Selected Competiton Awards

Mathematical Modeling Contest for College Students | *National Second Prize* (Top 0.5% of 45,000 teams)	Sep. 2020
Students Service Outsourcing Innovation and Entrepreneurship Competition | *National Second Prize* (Top 3%)	Aug. 2020
MathorCup College Mathematical Modeling Competition | *First Prize* (Top 3%)	May 2020
May Day Mathematical Modeling Competition | *First Prize* (Top 3%)	May 2020
Asia-Pacific Mathematical Modeling Contest | *First Prize* (Top 3%)	Nov. 2019
Mathematical Modeling Competition in Southwest Jiaotong University | *First Prize* (Top 3%)	Nov. 2019
American College Student Mathematical Modeling Contest | *Honorable Mention* (Top 10%)	Jan. 2020

Research Experience

AgentSecurityBench (ASB): Formalizing and Benchmarking Attacks and Defenses in LLM-based Agents

Research Intern@WISE Lab of Prof. Yongfeng Zhang. ***ICLR 2025 First Author*** *Rutgers University*

- Introduce Agent Security Bench (ASB), a comprehensive framework designed to formalize, benchmark, and evaluate the attacks and defenses of LLM agents, including 10 scenarios (e.g., e-commerce, autonomous driving, finance), 10 agents targeting the scenarios, over 400 tools, 23 different types of attack/defense methods, and 7 evaluation metrics.
- Benchmark 10 prompt injection attacks, a memory poisoning attack, a novel Plan-of-Thought backdoor attack, and 11 corresponding defenses across 13 LLM backends with over 90,000 testing cases in total.
- Reveal critical vulnerabilities in different stages of agent operation, including system prompt, user prompt handling, tool usage, and memory retrieval, with the highest average attack success rate of 84.30%, but limited effectiveness shown in current defenses.

Towards Imperceptible Backdoor Attack in Self-supervised Learning

Research Intern@Lab of Prof. Shiqing Ma. ***CVPR 2025 First Author*** *UMass Amherst*

- Observe that existing imperceptible triggers designed for supervised classifiers have limited effectiveness in SSL, and current backdoor attacks on SSL, like BadEncoder, achieve high ASRs but rely on visible triggers.
- Find that the reason behind such ineffectiveness is the coupling feature-space distributions for the backdoor samples and augmented samples in the SSL models.
- Propose an imperceptible and effective backdoor attack in SSL by disentangling the distribution of backdoor samples and augmented samples in SSL, while constraining the stealthiness of the triggers during the optimization process.
- Extensive experiments on five datasets and six SSL algorithms with different augmentation ways demonstrate our attack is effective and stealthy, and can also be resilient to current SOTA backdoor defense methods.

Few-shot Class Incremental Learning

M.S. student. ***IEEE Trans. on Industrial Informatics First Author*** *Zhejiang University*

- Introduce a novel framework for class-incremental fault diagnosis under limited fault data. It addresses key challenges such as class imbalance, long-tailed distributions, and catastrophic forgetting.
- Propose supervised contrastive knowledge distillation to improve feature extraction from limited fault data while minimizing the forgetting of previously learned fault classes as new ones are introduced.
- Propose marginal exemplar selection, which prioritizes hard-to-classify edge-case samples for storage, helping the model recognize class boundaries and improve generalization.

Academic Service

Conference Reviewer: ICLR, CVPR	2025 - Present
Journal Reviewer: IEEE TNNLS, TMLR, IEEE Trans. on Reliability, Pattern Recognition	2025 - Present

Last updated: November 7, 2025