Hanrong Zhang

(+1)312-479-7822 zhanghr0709@gmail.com 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

Internship

Google **Mountain View, CA, USA**
Student Researcher, Data Synthesis for Multimodal Agents *May 2026 – Aug. 2026*

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

- Build scalable simulation environments for LLM-human multi-turn interactions, covering tool-call scenarios such as retail and airline services under sparse rewards, limited data, and unstable environments.
- Propose a **Tool Dependency Graph** to model tool constraints and generate high-quality multi-turn tool-call dialogues with ground-truth trajectories for verifiable reward computation. Train agents with synthesized trajectories via multi-turn **GRPO**, improving long-horizon tool use in human-agent interactions.

Selected Publications and Preprints

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

[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**.[Paper] [Code] [Website] (A foundational benchmark for LLM Agent Security, 300 citations & Github stars)

[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**. [Paper] [Code] (Machine Learning Safety)

[3] **Hanrong Zhang***, Shicheng Fan*, Henry Peng Zou, Yankai Chen†, Zhenting Wang, Jiayu Zhou, Chengze Li, Wei-Chieh Huang, Yifei Yao, Kening Zheng, Xue Liu, Xiaoxiao Li, Philip S. Yu, *CoEvoSkills: Self-Evolving Agent Skills via Co-Evolutionary Verification*, **Preprint**. [Paper] [Code] [Website] (Agent Skill Self-Evolution)

[4] Zhaofen Wu*, **Hanrong Zhang***†, Fulin Lin, Wujiang Xu, Xinran Xu, Yankai Chen, Henry Peng Zou, Shaowen Chen, Weizhi Zhang, Xue Liu, Philip S. Yu, Hongwei Wang†, *GAM: Hierarchical Graph Memory for LLM-based Agents*, **ACL 2026**; ICLR 2026 Workshop on Memory for LLM-Based Agentic Systems (**Oral**). [Paper] (Agent Memory)

[5] **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, SCI Q1, Top Journal) [Paper] (Continuous Learning)

[6] Yifei Yao*, **Hanrong Zhang***, Fulin Lin*, Ziyan Jin, Xiaoxiao Li, Hongwei Wang, Ying Chi, *From Uncertainty to Clarity: Uncertainty-Guided Class-Incremental Learning for Limited Biomedical Samples via Semantic Expansion*, **IEEE Journal of Biomedical and Health Informatics**. (IF=6.8, SCI Q1, Top Journal) [Paper] (Continuous Learning)

[7] 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, SCI Q1, Top Journal) [Paper] (OOD Distribution Detection)

[8] 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, SCI Q1) [Paper] (Imbalanced Classification)

[9] **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)

[10] Zhiling Yan, Dingjie Song, **Hanrong Zhang**, Wei Liang, Yuxuan Zhang, Yutong Dai, Lifang He, Philip S. Yu, Ran Xu, Xiang Li, Lichao Sun, *OpenSkill: Open-World Self-Evolution for LLM Agents*, **Preprint**. [Paper] [Code] (Agent Self-Evolution)

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
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

Research Experience

Agent Security Bench (ASB) **ICLR 2025, First Author**

- **Motivation:** LLM agents extend LLM reasoning into real-world interaction: by calling tools, updating memory, and executing multi-step plans, they can change external environments, making agent security more critical than standard chatbot safety; however, prior evaluations lacked a unified benchmark across prompts, tools, memory, and planning.
- **Contribution:** Introduce the first LLM agent security benchmark with 10 scenarios, 10 agents, 400+ tools, 27 attack/defense methods, 7 metrics, and Net Resilient Performance (NRP) for balancing utility and security.
- **Results:** Reveal that current LLM agents remain highly vulnerable and existing defenses provide limited protection: attacks reach up to 84.30% success across prompt, tool, and memory stages, while defenses struggle to preserve both utility and security across 13 LLM backbones and 90,000+ test cases.

Towards Self-Evolving Skills in LLM Agents **CoEvoSkills & OpenSkill, Preprint**

- **Motivation:** Skills can boost professional tasks, but human-curated skills are costly and can even hurt performance in some SkillsBench domains. Meanwhile, one-shot self-generation methods are nearly useless, and prior evolution methods are mostly designed for tools and cannot reliably build structured multi-file skill packages.
- **Contribution:** Propose **CoEvoSkills** for closed-world skill evolution and **OpenSkill** for open-world self-evolution, refining multi-file skills through surrogate verification, open-world knowledge retrieval.
- **Results:** **CoEvoSkills** reaches average 71.1% Pass rate on SkillsBench (+40.5pp over no-skill, +17.6pp over human-curated skills baselines); **OpenSkill** reaches 43.6% on Claude Opus 4.6 and 42.1% on GPT 5.2, within 1–3pp of human-curated skills and about +9pp over the best closed-world baselines.

GAM: Hierarchical Graph-based Memory for LLM Agents **ACL 2026, Co-First & Corresponding Author**

- **Motivation:** Long-term agent memory faces a core conflict between fast contextual adaptation and stable knowledge retention: stream-based memories update quickly but lack write isolation, causing memory loss and semantic drift, while static graph memories preserve structure but struggle with fluid open-domain dialogue.
- **Contribution:** Propose **GAM**, a hierarchical graph memory framework that decouples memory encoding from consolidation: an Event Progression Graph isolates ongoing dialogue events, while semantic-boundary triggers consolidate complete episodes into a Topic Associative Network for stable long-term knowledge.
- **Results:** **GAM** achieves SOTA performance among six memory baselines on LoCoMo and LongDialQA.

Open-Source Projects

Dr. Claw [Code] **Core Contributor · ∼1,000 GitHub stars, 100+ forks**

- Build an **Auto Research** workspace that orchestrates agentic pipelines from research ideation and literature/code survey to experiment execution, analysis, paper writing, and presentation generation.
- Integrate multi-source research news tracking, 100+ research skills, project/session management, and multi-agent backends into a full-stack research IDE adopted by the open-source community.

Academic Service

Reviewer: ICLR, ICML, NeurIPS, CVPR, ECCV, KDD, MICCAI; TPAMI, TNNLS, TMLR, Pattern Recognition

Last updated: June 13, 2026