# Yuhui Wang

Phone: +1 631-943-6953 | Email: yuhui.wang.1@stonybrook.edu
LinkedIn: linkedin.com/in/yuhui-wang-705a41379 | Website: zjuwyh.github.io

## SUMMARY

PhD student in Computer Science, Stony Brook University, USA. Specializing in empirical LLM/Agent safety and alignment, with work spanning jailbreak defenses, adaptive red-teaming, and agentic RL. **3 first-authored papers** in ICLR and IEEE S&P; **$5K fellowship** awardee. Expected graduation 2028.

## EDUCATION

**Stony Brook University** | Computer Science | PhD Student | NY, USA 2024.09 – Present

Advisor: Prof. Ting Wang • Research Interest: Empirical LLM Safety & Alignment, Agent Safety, Jailbreak Defenses

**Shanghai Jiao Tong University** | Industrial Engineering | Master's Degree | Shanghai, China 2021.09 – 2024.03

Advisor: Prof. Di Wang • Research Interest: Multi-dimensional Time-series Classification & Prediction

**Zhejiang University** | Industrial Engineering | Bachelor's Degree | Hangzhou, China 2017.09 – 2021.06

## RESEARCH EXPERIENCE

**Customized RL&SFT for LLM Alignment Enhancement | Project Leader | 2 First-Authored ICLR Papers** 2024.09 – Present

- Found the vulnerabilities of existing safeguards against intensive fine-tuning jailbreak attacks; proposed a **self-destructive** defense with a custom loss coupling benign and harmful gradients, enabling the model to retain ~**5%** initial harmful metric or be **fully destroyed** post-attack; implemented via adapting the HuggingFace Trainer.
- Proposed a perturbation framework to prove and quantify the joint impact of CoT and memory on the final answer. Based on this finding, **suppressed retrieval shortcuts in reinforcement learning** to strengthen model reasoning and ensure trustworthy outputs; improved CoT robustness by **47.8%** and pass@1 accuracy by **5.8%** across diverse benchmarks, implemented through customized adaptation of the **veRL trainer**.

**Shadow-Memory Agent Defense & Adaptive Long-horizon Red-Teaming | 1 ICLR, 1 ACL, 3 Papers Under Review** 2024.09 – Present

- Propose **MAGE** that defends against **adaptive and long-horizon prompt injection** attacks via **shadow memory** that distill the security-aware information throughout long-term interactions; customized the **agentic RL framework (RLLM)** for training the underlying model for memory extraction and risk assessment. Achieved ~**0%** ASR, high BU and 22× cost reduction compared to the SOTA method on AgentDojo and SHADE-Arena.
- Proposed the **adaptive long-horizon prompt injection** in AgentLab Benchmark that uses **evasive** attack plan and **adaptively** refines injected content based on agent tool-call trajectories; achieving up to **93.1%** ASR on AgentDojo environments.
- Participated in developing **AutoRAN**, an **automated jailbreak attack framework** that uses a red-team reasoning model to simulate and refine prompts via target models' intermediate reasoning steps. Achieved ~**100%** success rates on GPT-o3/o4-mini and Gemini-2.5-Flash.

**GraphRAG Data Poisoning | Project Leader | 1 Co-first-Authored IEEE S&P Paper** 2024.09 – 2025.02

- As the **first data poisoning study on GraphRAG** (developed by Microsoft), proposed **GragPoison**, a poisoning attack targeting GraphRAG by leveraging its unique relation and entity retrieval mechanism to **amplify the poisoning effect**. Achieved up to **98%** attack success with **<68%** poisoning text on medical, cybersecurity, and commonsense datasets across multiple GraphRAG variants compared with benchmarks.

## SKILLS

- **Programming:** Python, SQL, C++, Java, C
- **ML/AI Frameworks:** PyTorch, HuggingFace, vLLM, GraphRAG, veRL, Ray, Megatron, FSDP, wandb, scikit-learn, rllm, slime
- **Specialized Skills:** LLM Post-Training (LoRA, RLHF, DPO, PPO, GRPO, DAPO, GSPO), Adversarial ML, RAG, Context Engineering, Jailbreak Defenses, Model Steering, LLM Agent, Prompt Injection, Machine Learning, Software Engineering, Agentic RL, Coding Agent

## AWARDS & SERVICE

Nisha and Vinod K. Singhi Graduate Fellowship (1 PhD student/year in CS Dept., $5,000), Stony Brook University 2024, 2025
Reviewer of NeurIPS 2025, ICLR 2026, ICML 2026, JCS, IEEE CASE 2025 & 2024, IEEE TASE 2024, 2025
Teaching Assistant of CSE352 & CSE582 — Office hours, Grading, Student Communication 2024, 2025

## SELECTED PUBLICATIONS

Google Scholar: scholar.google.com/citations?user=eJgbw-oAAAAJ [135 Citations]

[1] **Yuhui Wang**, Tanqiu Jiang, Jiacheng Liang, Charles Fleming, Ting Wang, "MAGE: Safeguarding LLM Agents against Long-Horizon Threats via Shadow Memory," *under review at **ACM CCS***, 2026.

[2] **Yuhui Wang**, Changjiang Li, Guangke Chen, Jiacheng Liang, Ting Wang, "Reasoning or Retrieval? A Study of Answer Attribution on Large Reasoning Models," *International Conference on Learning Representations (**ICLR**)*, 2026.

[3] **Yuhui Wang**, Rongyi Zhu, Ting Wang, "Self-Destructive Language Model," *International Conference on Learning Representations (**ICLR**)*, 2026.

[4] Tanqiu Jiang, **Yuhui Wang**, Jiacheng Liang, Ting Wang, "AgentLAB: Benchmarking LLM Agents against Long-Horizon Attacks," *under review at **ICML***, 2026.

[5] Jiacheng Liang*, **Yuhui Wang***, Changjiang Li, Rongyi Zhu, Tanqiu Jiang, Neil Gong, Ting Wang, "GraphRAG under Fire," *IEEE Symposium on Security and Privacy (**IEEE S&P**)*, 2026.

[6] Tanqiu Jiang, Zian Wang, Jiacheng Liang, Changjiang Li, **Yuhui Wang**, Ting Wang, "RobustKV: Defending Large Language Models against Jailbreak Attacks via KV Eviction," *International Conference on Learning Representations (**ICLR**)*, 2025.