Miao Yu

ymzgkxjsdx@mail.ustc.edu.cn | (+86) 184-5653-4040
University of Science and Technology of China, Hefei, China
Github | Google Scholar

EDUCATION

University of Science and Technology of China (USTC), Hefei, China — Sept 2021 – Jul 2025 (Expected)

Bachelor of Computer Science

GPA: 84.42/100

TOEFL: 102/120

Core Courses: *Text Representation Learning 97/100, Introduction to Deep Learning 92/100, Fundamentals of Artificial Intelligence 90/100, Python and Fundamentals of DL 90/100, Web Information Processing and Applications 90/100, Quantum Computing and Machine Learning 85/100*

Honors and Awards:

1st in USTC-Huawei HarmonyOS Elite Class Internship, USTC/Huawei
Winner of Scholarship for Huaxia CS Elite Class, USTC
Winner of Wanglaoji Scholarship, USTC

RESEARCH EXPERIENCES

Research Assistant — March 2025 – June 2025 (Expected)

Agency for Science, Technology and Research (A*STAR), Singapore — *Supervisor: Xingrui Yu & Qing Guo*

Research Assistant — March 2024 – June 2024

University of Science and Technology of China, Hefei, China — *Supervisor: Yan Song*

PUBLICATIONS

- A Survey on Trustworthy LLM Agents: Threats and Countermeasures, submitted to KDD Tutorial 2025 *(First Author), arxiv.org/abs/2503.09648*
- NetSafe: Exploring the Topological Safety of Multi-agent Networks, cited by 9, submitted to ACL ARR February 2025 *(First Author), arxiv.org/pdf/2410.15686*
- LLM-Virus: Evolutionary Jailbreak Attack on Large Language Models, cited by 2, submitted to TEVC 2025 *(First Author), arxiv.org/pdf/2410.15686*
- Mind Scramble: Unveiling Large Language Model Psychology Via Typoglycemia, cited by 2 (First Author)
- Teaching Multi-agent Systems via Evolutionary Algorithms for Complex Tasks Modeling, submitted to TEVC 2025 *(Co-first Author)*
- G-Safeguard: A Topology-Guided Security Lens and Treatment on LLM-based Multi-agent Systems, submitted to ACL ARR February 2025 *(Second Author), arxiv.org/pdf/2502.11127*
- AgentSafe: Safeguarding Large Language Model-based Multi-agent Systems via Hierarchical Data Management, submitted to ACL ARR February 2025 *(Third Author), arxiv.org/pdf/2503.04392*
- G-Designer: Architecting Multi-agent Communication Topologies via Graph Neural Networks, cited by 9, submitted to ICML 2025 *(Fourth Author), arxiv.org/pdf/2410.11782*

MAIN PUBLICATION DETAILS

1. A Survey on Trustworthy LLM Agents: Threats and Countermeasures

Research Assistant | **Advisor***:* Qingsong Wen, Squirrel AI; Postdoc. Xinfeng Li, NTU; Postdoc. Kun Wang, NTU

Brief Introduction: We present a comprehensive review of the trustworthiness challenges in agents based on large language models (LLMs) and multi-agent systems (MAS). To this end, we introduce the TrustAgent framework, which systematically categorizes trustworthiness into intrinsic (brain, memory, tools) and extrinsic (user, agent, environment) aspects. Through our analysis, we identify and examine emerging threats, defense strategies, and evaluation methodologies, with particular emphasis on critical issues including adversarial attacks and privacy risks.

2. NetSafe: Exploring the Topological Safety of Multi-agent Networks

Research Assistant | **Advisor:** Yang Wang, USTC; Postdoc. Kun Wang, NTU

Brief Introduction: We investigate the security of MAS from a topological perspective. Specifically, we introduce NetSafe, a comprehensive framework designed to evaluate network safety against misinformation, bias, and harmful content propagation. Through our extensive experimental analysis, we uncover critical insights, particularly the vulnerability of highly connected networks to adversarial attacks. Additionally, we identify two novel phenomena: Agent Hallucination and Aggregation Safety. These discoveries not only advance our understanding of MAS vulnerabilities but also establish a solid foundation for designing safer and more resilient multi-agent architectures.

3. LLM-Virus: Evolutionary Jailbreak Attack on Large Language Models

Research Assistant | **Advisor:** Qingsong Wen, Squirrel AI; Postdoc. Kun Wang, NTU

Brief Introduction: We introduce a novel jailbreak attack method leveraging evolutionary algorithms to bypass the safety mechanisms of LLMs. Inspired by biological virus evolution, the LLM-Virus framework treats jailbreak templates as evolving entities, optimizing their effectiveness through LLM-assisted mutation, crossover, and fitness evaluation. The approach enhances attack efficiency, transferability, and reduces computational cost while outperforming existing jailbreak strategies on multiple safety benchmarks. This research highlights critical security vulnerabilities in LLMs and contributes to the development of more robust defense mechanisms.

4. Teaching Multi-agent Systems via Evolutionary Algorithms for Complex Tasks Modeling

Research Assistant | **Advisor:** Qingsong Wen, Squirrel AI; Postdoc. Kun Wang, NTU; Prof. Shirui Pan, GU

Brief Introduction: We explore the integration of LLMs with evolutionary algorithms to address complex tasks, particularly those that require dynamic adaptation and combinatorial optimization. The proposed framework, Dual-Branch Evolutionary Agent Optimization (DBE-AO), combines the strengths of LLM-based agents and traditional evolutionary algorithms to enhance task-solving capabilities.

5. G-Safeguard: A Topology-Guided Security Lens and Treatment on LLM-based Multi-agent Systems

Research Assistant | **Advisor:** Yang Wang, USTC; Postdoc. Kun Wang, NTU

Brief Introduction: We present a novel security framework for multi-agent systems leveraging LLMs. In this work, we introduce a topology-guided approach specifically designed to enhance the security and robustness of these systems while addressing their potential vulnerabilities. Through our comprehensive analysis, we demonstrate significant improvements in system integrity and resilience against various cyber threats. Our framework provides a systematic methodology for strengthening the security posture of LLM-powered multi-agent systems, offering both theoretical insights and practical solutions for mitigating emerging risks in this domain.

6. G-Designer: Architecting Multi-agent Communication Topologies via Graph Neural Networks

Research Assistant | **Advisor:** Dawei Cheng, TJU; Postdoc. Kun Wang, NTU; Tianlong Chen, UNC-Chapel Hill

Brief Introduction: We propose an innovative solution for dynamically designing task-aware communication topologies in multi-agent systems powered by LLMs. By leveraging a variational graph auto-encoder, G-Designer encodes agents and task-specific information to decode high-performing and robust communication graphs. Extensive experiments show that G-Designer outperforms state-of-the-art baselines in performance, adaptiveness, and adversarial robustness, making it a practical and scalable approach for automating collaborative MAS design.

SKILLS

Professional Skills: Python, Pytorch, C, C++, Latex, Linux, Matlab, Verilog

Hobbies: Reading, Writing, Poem, Photography, Guitar, Table Tennis