

Miao Yu

EDUCATION

The University of Hong Kong (HKU)	Sep. 2026 - Jun. 2030 (Expected)
Hong Kong, China	Ph.D. in Computer Science
Supervisor: Zuming Jiang	
University of Science and Technology of China (USTC)	Sep. 2021 - Jun. 2025
Hefei, Anhui, China	B.E. in Computer Science

RESEARCH EXPERIENCES

LLM Engineer	Mar. 2026 – Aug. 2026
Tencent (Qingyun Plan), Shenzhen, China	Supervisor: Shengyuan Chen
Research Assistant	Oct. 2025 - Jan. 2026
Shanghai Artificial Intelligence Laboratory, Shanghai, China	Supervisor: Yafu Li & Yu Cheng
Research Assistant	Mar. 2025 - Jun. 2025
Agency for Science, Technology and Research (A*STAR), Singapore	Supervisor: Xingrui Yu & Qing Guo
Lab Assistant	Mar. 2024 - Jun. 2024
University of Science and Technology of China, Hefei, China	Supervisor: Yan Song

PUBLICATIONS

Google Scholar Citation: 1100+ * Equal Contribution

[1] SafeSeek: Universal Attribution of Safety Circuits in Language Models

Accepted at ICML 2026 (Co-first Author, CCF-A)

Authors: **Miao Yu***, Siyuan Fu*, Moayad Aloqaily, Zhenhong Zhou, Safa Otoum, Kun Wang, Yufei Guo, Qingsong Wen

[2] Uncovering Hidden Triggers: Backdoor Attribution in Language Models

Accepted at ICML 2026 (Co-first Author, CCF-A)

Authors: **Miao Yu***, Zhenhong Zhou*, Moayad Aloqaily, Kun Wang, Biwei Huang, Stephen Wang, Yueming Jin, Qingsong Wen

[3] A Survey on Trustworthy LLM Agents: Threats and Countermeasures

Accepted at KDD 2025 Tutorial (Co-first Author)

Authors: **Miao Yu***, Fanci Meng*, Xinyun Zhou, Shilong Wang, Junyuan Mao, Linsey Pang, Tianlong Chen, Kun Wang, Xinfeng Li, Yongfeng Zhang, Bo An, Qingsong Wen

[4] NetSafe: Exploring the Topological Safety of Multi-agent System

Accepted at ACL 2025 Findings (Co-first Author)

Authors: **Miao Yu***, Shilong Wang*, Guibin Zhang, Junyuan Mao, Chenlong Yin, Qijiong Liu, Qingsong Wen, Kun Wang, Yang Wang

[5] Backdoor Collapse: Eliminating Unknown Threats via Known Backdoor Aggregation in Language Models

Accepted at ACL 2026 (Co-first Author, CCF-A)

Authors: Liang Lin*, **Miao Yu***, Moayad Aloqaily, Zhenhong Zhou, Kun Wang, Linsey Pang, Prakhar Mehrotra, Qingsong Wen

[6] Hidden in the Noise: Unveiling Backdoors in Audio LLMs Alignment through Latent Acoustic Pattern Triggers

Accepted at AAAI 2026 Oral (Co-first Author, CCF-A)

Authors: Liang Lin*, **Miao Yu***, Kaiwen Luo*, Yibo Zhang, Lilan Peng, Dexian Wang, Xuehai Tang, Yuanhe Zhang, Xikang Yang, Zhenhong Zhou, Kun Wang, Yang Liu

[7] G-Safeguard: A Topology-Guided Security Lens and Treatment on LLM-based Multi-agent Systems

Accepted at ACL 2025 (Contributor, CCF-A)

Authors: Shilong Wang*, Guibin Zhang*, **Miao Yu**, Guancheng Wan, Fanci Meng, Chongye Guo, Kun Wang, Yang Wang

[8] G-Designer: Architecting Multi-agent Communication Topologies via Graph Neural Networks

Accepted at ICML 2025 Spotlight (Contributor, CCF-A)

Authors: Guibin Zhang*, Yanwei Yue*, Xiangguo Sun, Guancheng Wan, **Miao Yu**, Junfeng Fang, Kun Wang, Tianlong Chen, Dawei Cheng

[9] G-Memory: Tracing Hierarchical Memory for Multi-Agent Systems

Accepted at NeurIPS 2025 Spotlight (Contributor, CCF-A)

Authors: Guibin Zhang, Muxin Fu, Kun Wang, Guancheng Wan, **Miao Yu**, Shuicheng Yan

[10] OrthAlign: Orthogonal Subspace Decomposition for Non-Interfering Multi-Objective Alignment

Accepted at ICLR 2026 (Contributor, CCF-A)

Authors: Liang Lin*, Zhihao Xu*, Junhao Dong, Jian Zhao, Yuchen Yuan, Guibin Zhang, **Miao Yu**, Yiming Zhang, Zhengtao Yao, Huahui Yi, Dongrui Liu, Xinfeng Li, Kun Wang

[11] AgentAsk: Multi-Agent Systems Need to Ask

Accepted at ACL 2026 (Contributor, CCF-A)

Authors: Bohan Lin*, Kuo Yang*, Zelin Tan, Yingchuan Lai, Chen Zhang, Guibin Zhang, Xinlei Yu, **Miao Yu**, Xu Wang, Yudong Zhang, Yang Wang