Yuping Lin

linyupin@msu.edu ◇ +1 (413) 512-9746

Education

Michigan State University

Ph.D. in Computer Science 09/2023 - Present

- Advisor: Dr. Jiliang Tang
- Research Area: Safety in Large Language Models, Trustworthy AI

University of Massachusetts, Amherst

M.S. in Electrical and Computer Engineering 01/2022 - 08/2023

- Advisor: Dr. Tongping Liu
- Research Area: Machine Learning Model Compression

University of Electronic Science and Technology of China

B.Eng in Information Security 09/2017 - 07/2021

Research Experience

Fake Safety Alignment of Large Language Models (LLMs) Ongoing

- Current safety alignment benchmarks are vulnerable to backdoor attacks: a malicious LLM can be engineered to deliver harmful outputs when triggered, while still passing safety checks in benign scenarios.
- Our ongoing investigation provides early evidence that these vulnerabilities exist, with preliminary findings indicating that Mixture-of-Expert (MoE) architectures are particularly vulnerable because of the flexibility of their designs.
- We are exploring countermeasures to enhance the reliability of safety alignment benchmarks and mitigate these risks.

Memory Management in LLM-based Agents Submitted to ICML 2025

- The memory module is a crucial component of LLM-based agents, yet there is no consensus on its optimal design.
- We investigate the general principles of memory management for LLM agents by examining the impact of two core operations—the addition and deletion of memories—on LLM agent performance.
- We examine various memory addition and deletion strategies, and find that with the optimal memory addition strategy, the agent's execution success rate can increase from 13% to 38%. Additionally, the optimal memory deletion strategy reduces the memory size from 1000 to 250, highlighting the critical importance of effective memory management.
- Our manuscript, titled *Towards Optimal Memory Management: Investigating Experience-Following Behavior of Large Language Model Agents*, is currently under submission to **ICML 2025**.

Understanding Jailbreak Attacks in LLMs EMNLP 2024 (Main)

- Although various jailbreak attacks have been developed in the literature, a comprehensive understanding of their underlying mechanisms remains elusive. This gap hinders the advancement of robust jailbreak attacks as well as the development of effective alignment and defense strategies.
- We propose an analytical framework that leverages the hidden representations of LLMs to examine the behavior of jailbreak attacks. Our analysis reveals a clear separation between harmful and benign prompts in the hidden space, with successful jailbreak attacks shifting harmful prompt representations toward the benign direction.
- We demonstrate that steering the representation of jailbreak attacks toward the benign region increases the attack success rate from 26% to 62%, thereby exposing a critical vulnerability in current LLM architectures.
- Our work, titled *Towards Understanding Jailbreak Attacks in LLMs: A Representation Space Analysis*, has been accepted for poster presentation at **EMNLP 2024 (Main Conference)**.

Publications and Preprints

1. Shenglai Zeng, Jiankun Zhang, Bingheng Li, **Yuping Lin**, Tianqi Zheng, Dante Everaert, Hanqing Lu, Hui Liu, Yue Xing, Monica Xiao Cheng, Jiliang Tang. *Towards Knowledge Checking in Retrieval-augmented Generation: A Representation Perspective*. [**NAACL 2025**]

2. Yingqian Cui, Jie Ren, **Yuping Lin**, Han Xu, Pengfei He, Yue Xing, Lingjuan Lyu, Wenqi Fan, Hui Liu, Jiliang Tang. *FT-Shield: A Watermark Against Unauthorized Fine-tuning in Text-to-Image Diffusion Models*. [**SIGKDD Explorations Newsletter**, 2025]

3. Pengfei He, **Yuping Lin**, Shen Dong, Han Xu, Yue Xing, Hui Liu. *Red-Teaming LLM Multi-Agent Systems via Communication Attacks*. **[Submitted to ACL 2025]**

4. **Yuping Lin***, Zidi Xiong*, Wenya Xie*, Pengfei He, Jiliang Tang, Himabindu Lakkaraju, Zhen Xiang. *Towards Optimal Memory Management: Investigating Experience-Following Behavior of Large Language Model Agents*. **[Submitted to ICML 2025]**

5. Kaiqi Yang, Hang Li, Yucheng Chu, **Yuping Lin**, Tai-Quan Peng, Hui Liu. *Unpacking Political Bias in Large Language Models: Insights Across Topic Polarization*. **[Submitted to NAACL 2024]**

6. **Yuping Lin***, Pengfei He*, Han Xu, Yue Xing, Makoto Yamada, Hui Liu, Jiliang Tang. *Towards Understanding Jailbreak Attacks in LLMs: A Representation Space Analysis*. **[EMNLP 2024]**

Teaching

Teaching Assistant Spring 2025
CSE 335 Object-oriented Software Development

Teaching Assistant Fall 2024
CSE 482 Big Data Analysis

Awards

- Travel grant, 2024, Michigan State University
- Outstanding Student Scholarship in the academic year of 2018, 2019, UESTC
- Third prize in the 16th UESTC ACM Programming Competition, 2018
- CCF National Olympiad in Informatics 2015 (2nd), 2016 (2nd)
- Second prize in Computer Production Activities for Primary and Secondary schools in Shandong Province, 2016

Professional Skills

- **Programming Languages**: Python, C, C++, Pascal, JavaScript, Java, HTML/CSS, MATLAB
- **Softwires and Systems**: PyTorch, HuggingFace, Numpy, Pandas, Scikit-Learn, Git, Docker, Linux, Vue, SpringBoot
- **AI skills**: Large Language Models (LLaMA, DeepSeek, Qwen, Gemma, Mixtral, Vicuna, GPT), Computer Vision (Diffusion models), Data Science, Data Visualization