WEIDI LUO

The University of Georgia, Herty Dr, Athens, GA 30602
(+1)614.477.3643 (+86)18990600198 luo.1455@uga.edu www.eddyluo.com

RESEARCH INTERESTS

Trustworthy AI & AI Safety: Using interpretability methods, discover security vulnerabilities in AI systems, including foundation models and AI agents, and develop corresponding algorithms, including safety alignment strategies.
AI in Security: Leverage AI to drive defense and attack strategies on systems, including web system and operating system.
Lifelong AI Algorithms: Develop lifelong learning AI frameworks and defense systems by utilizing reinforcement learning, cognitive science, bio-inspired algorithms, active learning, and so on.

EDUCATION

School of Computing, University of Georgia Aug 2025-Present
P.HD. Student in Computer Science
- Advisor: Prof. Zhen Xiang

Art and Science College, The Ohio State University Aug 2022-Jul 2025
B.A. in Computer and Information Science
- Advisor: Prof. Chaowei Xiao, Dr.Yu Li, Prof.Ruixiang Tang, Prof.Yu Su

WORK EXPERIENCES

Large Language Model Engineer | International Digital Economy Academy Jun 2024-Aug 2024
- Advisor: Yu Li, Principal Researcher

PROFESSIONAL EXPERIENCES

Research Intern | University of Wisconsin-Madison Dec 2023-Apr 2025
- Advisor: Chaowei Xiao, Assistant Professor at Information School

Research Intern | Rutgers University Oct 2024-Dec 2024
- Advisor: Ruixiang Tang, Assistant Professor at Department of Computer Science

Research Assistant | ICICLE Institute | The Ohio State University Aug 2022-Dec 2023
- Advisor: Yu Su, Distinguished Assistant Professor at College of Engineering

REVIEWER EXPERIENCES

ACL'2025, **EMNLP**'2025

PUBLICATION

Weidi Luo, Shenghong Dai, Xiaogeng Liu, Suman Banerjee, Huan Sun, Muhao Chen, Chaowei Xiao, "*AGrail: A Lifelong Agent Guardrail with Effective and Adaptive Safety Detection"*, Accepted by ACL'2025.

Weidi Luo*, Siyuan Ma*, Xiaogeng Liu*, Xiaoyu Guo, Chaowei Xiao, "*JailBreakV-28K: A Benchmark for Assessing the Robustness of MultiModal Large Language Models against Jailbreak Attacks*", Accepted by COLM'2024.

Weidi Luo*, He Cao*, Yu Wang, Zijing Liu, Aidan Wong, Bin Feng, Yuan Yao, Yu Li, "*Guide for Defense (G4D): Dynamic Guidance for Robust and Balanced Defense in Large Language Models.*", Accepted by NAACL'2025.

Mingyu Jin, **Weidi Luo**, Sitao Cheng, Xinyi Wang, Wenyue Hua, Ruixiang Tang, William Yang Wang, Yongfeng Zhang "*Disentangling Memory and Reasoning Ability in Large Language Models*", Accepted by ACL'2025.

Vardaan Pahuja, **Weidi Luo**, Yu Gu, Cheng-Hao Tu, Hong-You Chen, Tanya Berger-Wolf, Charles Stewart, Song Gao, Wei-Lun Chao, Yu Su, "*Bringing Back the Context: Image Classification as Link Prediction on Multimodal Knowledge Graphs*", Accepted by CIKM'2024.

PRE-PRINT

Weidi Luo*, Qiming Zhang*, Tianyu Lu*, Xiaogeng Liu, Yue Zhao, Zhen Xiang, Chaowei Xiao, "*Doxing via the Lens:*

Revealing Privacy Leakage in Image Geolocation for Agentic Multi-Modal Large Reasoning Model", on Arxiv.

Siyuan Ma*****, **Weidi Luo***, Yu Wang, Xiaogeng Liu, Muhao Chen, Bo Li, Chaowei Xiao, "*Visual-RolePlay: Universal Jailbreak Attack on MultiModal Large Language Models via Role-playing Image Character*", on Arxiv.

Zeru Shi, Zhenting Wang, Yongyue Su, **Weidi Luo**, Fan Yang, Yongfeng Zhang, "*Robustness-aware Automatic Prompt Optimization"*, on Arxiv.

AWARD

- $20,000 SafeBench Award from Center for AI Safety, 2025