Xinhang (Owen) Ma

m.owen@wustl.edu • xhOwenMa.github.io • Google Scholar

Education

Washington University in St. Louis — St. Louis, MO
Ph.D. in Computer Science; Advisor: Prof. Yevgeniy Vorobeychik — *Jan. 2025 – Present*

Washington University in St. Louis — St. Louis, MO
M.S. in Computer Science; Advisor: Prof. Yevgeniy Vorobeychik — *Jan. 2024 – Dec. 2024*

– Thesis: *Towards Verified Vision-Based Neural Network Controllers for Autonomous Lane-Following* [link]

Washington University in St. Louis — St. Louis, MO
B.S. in Computer Science and Mathematics — *Aug. 2020 – Dec. 2023*

Publications

Protecting Language Models Against Unauthorized Distillation through Trace Rewriting
Xinhang Ma, William Yeoh, Ning Zhang, Yevgeniy Vorobeychik
Annual Meeting of the Association for Computational Linguistics (ACL), 2026, Main Conference [arXiv]

Conformal Reachability for Safe Control in Unknown Environments
Xinhang Ma, Junlin Wu, Yiannis Kantaros, Yevgeniy Vorobeychik
International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS), 2026 [arXiv]

Learning Vision-Based Neural Network Controllers with Semi-Probabilistic Safety Guarantees
Xinhang Ma, Junlin Wu, Hussein Sibai, Yiannis Kantaros, Yevgeniy Vorobeychik
AAAI Conference on Artificial Intelligence (AAAI), 2026 [arXiv]

Work in Progress / Under Review

ROPE: Routed Origin Policy Enforcement against Indirect Prompt Injection
Xinhang Ma, Chaowei Xiao, William Yeoh, Ning Zhang, Yevgeniy Vorobeychik
arXiv preprint, 2026 [arXiv]

Robust Context-Aware Detection of Malicious Instructions in Text
Buzhao Liu*, **Xinhang Ma***, Yevgeniy Vorobeychik (* equal contribution)
arXiv preprint, 2026 [arXiv]

AutoDojo: Adaptive Black-Box Attacks Reveal the Limits of IPI Defenses and Task-Specification Effects in LLM Agents
Xinhang Ma, Taoran Li, Chaowei Xiao, Zhiyuan Yu, Ning Zhang, Yevgeniy Vorobeychik
arXiv preprint, 2026 [arXiv]

Talks and Presentations

Protecting Language Models Against Unauthorized Distillation through Trace Rewriting
Oral presentation, Annual Meeting of the Association for Computational Linguistics (ACL) 2026

LLM Security: Jailbreak and Prompt Injection Attacks and Defenses
Guest lecture, Adversarial AI, Washington University in St. Louis 2026

Teaching

Adversarial AI Washington University in St. Louis
Mentored Teaching Experience (guest lectures, course design) *2026*

Data-Driven Privacy and Security Washington University in St. Louis
Teaching Assistant *2025*

Analysis of Algorithms Washington University in St. Louis
Teaching Assistant *2024*

Data Structures and Algorithms Washington University in St. Louis
Teaching Assistant *2022, 2023*

Service

Conference Reviewer: AAAI 2027, NeurIPS 2026, L4DC 2026