Mengxiao Wang

Research Focus: Specialized in targeted attacks and defenses for domain-specific LLM agents, with emphasis on financial trading systems. Research covers prompt injection, jailbreak, and prompt leaking attacks against LLM-based autonomous agents in high-stakes domains.

Education

Texas A&M University — College Station, TX
Ph.D. in Computer Science (*Advisor*: Dr. Nitesh Saxena) — Aug 2023 – Present
GPA: 4.0/4.0

Carnegie Mellon University — Pittsburgh, PA
Master of Science in Electrical and Computer Engineering — Sep 2021 – May 2023
GPA: 3.78/4.0

University of Minnesota, Twin Cities — Minneapolis, MN
Bachelor of Arts in Computer Science, Minor in Statistics — Jan 2018 – May 2021
GPA: 3.815/4.00

Skills

Languages: C/C++, Python, Go, Java, JavaScript/TypeScript, X86 Assembly, SQL, Rust
Technologies: Docker, Linux, MySQL, Git, Kali Linux (Burp Suite, Nmap, Metasploit), Design Patterns

Work Experience

Inspur — Jinan, China
LLM Security Research Intern — May 2024 – July 2024

- Assisted in developing a large language model (LLM) using LangChain and the ChatGLM model with API implementations. Utilized Python, FastAPI, and Flask to create a local deployment system for clients, enabling document analysis and summarization. The client could select different models to adjust LLM autonomy, with parameters like temperature and max tokens being configurable.
- Developed a fuzzing tool to test LLM vulnerabilities (e.g., prompt injection, unauthorized access), generating 1,000+ test cases. Discovered three novel methods to bypass content restrictions: context manipulation through subtle prompt phrasing, recursive question embedding, and hidden token injection via encoding tricks.
- Implemented an initial input-output filter with a high false positive rate (22.3%), which was reduced to 4% after integrating semantic analysis. Also applied strict file permissions and access control to mitigate LLM unauthorized file access.

Carnegie Mellon University — Pittsburgh, PA
Research Assistant (CyLab) — July 2021 – May 2023

- Led the effort to analyze TypeScript code flows using taint analysis, contributing to a crawl of over 2.1 million npm packages and improving the tool's detection of command injection vulnerabilities by 40%.
- Helped the team submit 30+ new CVEs related to code execution and command injection vulnerabilities in npm packages by refining detection mechanisms and enhancing vulnerability discovery.
- Manually crafted complex input flows, identifying vulnerabilities missed by automated tools, and improved the tool to handle edge cases such as type-specific inputs, further enhancing its detection capabilities.

Inspur — Jinan, China
Full Stack Intern — May 2018 – August 2018

- Scraped province-wide maps via Amap API with Python, integrating external data to boost coverage by 50% and improve MongoDB accuracy. Built the frontend in React and backend in Java (Spring

Boot), reducing map query load times by 25%.
- Performed security tests with Burp Suite Pro, fixing SQL injection and XSS. Developed proof-of-concept exploits using Python and Bash, reducing attack surface by 15% and exploit risk by 30%.

Publications

SoK: Trading Agents or Market Crashers? Dissecting Robustness and Security Failures in Academic Financial LLM Trading Schemes M Wang, N Saxena. Under review, 2026

Demystifying Progressive Web Application Permission Systems Wang, M., & Gu, G. In Proc. of the 56th IEEE/IFIP International Conference on Dependable Systems and Networks (DSN), 2026.
Cutting through the Confusion: A Measurement Study of Homograph Domains in Ethereum Name Service J Huang, SR Chintapalli, M Wang, G Gu. In Proceedings of The Web Conference (WWW), 2025
Application of Computer Vision in Logistics Warehouse Safety Management Wang, M. Software, 41(10), 180-183. ISSN:1003-6970

Awards & Grants

OpenAI Researcher Access Program Grant 2025

Professional Service

- **Program Committee,** ACM The Web Conference (WWW), 2025
- **External Reviewer,** Network and Distributed System Security Symposium (NDSS), 2023, 2024
- **External Reviewer,** USENIX Security Symposium (USENIX Security), 2024, 2025
- **External Reviewer,** IEEE Symposium on Security and Privacy (IEEE S&P), 2023
- **External Reviewer,** ACM Conference on Computer and Communications Security (CCS), 2023–2025

Student Supervising/Mentoring

Landry Taylor [Bachelor of Business Administration] Spring 2026 - Present
Texas A&M University, College Station, TX, USA