(+1) 919-672-7127
zedian.shao@outlook.com
asympsec.com
LinkedIn
GitHub
Google Scholar

Zedian Shao

Education

Georgia Institute of Technology, School of Cybersecurity and Privacy 2025 – 2030 (expected)
Ph.D. in Computer Science Atlanta, GA
Advisor: Prof. Teodora Baluta

Duke University, Pratt School of Engineering 2023 – 2025
M.Sc. in Electrical and Computer Engineering Durham, NC
Advisor: Prof. Neil Zhenqiang Gong, GPA: 4.0/4.0

Duke Kunshan University 2019 – 2023
B.S. in Data Science Suzhou, China
GPA: 3.8/4.0

Research Interests

My research focuses on AI security and safety, especially attacks and defenses for generative AI. I am currently interested in mechanistic interpretability as tools for understanding, evaluating, and hardening AI systems against security and safety failures.

Selected Publications

- **Zedian Shao**, Charles Fleming, Teodora Baluta. "Cordyceps: Covert Control Attacks on LLMs via Data Poisoning". In *USENIX Security Symposium (USENIX Security)*, 2026. **[Arxiv] [Github]**
- **Zedian Shao***, Hongbin Liu*, Yuepeng Hu, Neil Zhenqiang Gong. "Leave My Images Alone: Preventing Multi-Modal Large Language Models from Analyzing Images via Visual Prompt Injection". In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2026. **[Arxiv] [Github]**
- Yuqi Jia, Yupei Liu, **Zedian Shao**, Jinyuan Jia, and Neil Zhenqiang Gong. "PromptLocate: Localizing Prompt Injection Attacks". In *IEEE Symposium on Security and Privacy (S&P)*, 2026. **[Github]**
- Yuqi Jia, **Zedian Shao**, Yupei Liu, Jinyuan Jia, Dawn Song, Neil Zhenqiang Gong. "A Critical Evaluation of Defenses against Prompt Injection Attacks". In *ACM Symposium on Access Control Models and Technologies (ACM SACMAT)*, 2026.
- **Zedian Shao***, Hongbin Liu*, Jaden Mu, Neil Zhenqiang Gong. "Enhancing Prompt Injection Attacks to LLMs via Poisoning Alignment". In *ACM Workshop on Artificial Intelligence and Security (AISec)*, 2025. **[Arxiv] [Github]**
- Xilong Wang, John Bloch, **Zedian Shao**, Yuepeng Hu, Shuyan Zhou, and Neil Zhenqiang Gong. "WebInject: Prompt Injection Attack to Web Agents". In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2025.

Experience

Georgia Institute of Technology, College of Computing Atlanta, GA
Teaching Assistant | CS 8803: Machine Learning Security 08/2025 – 12/2025

- Designed assignments on adversarial examples, backdoor and poisoning attacks, watermarking, and formal-methods topics for a graduate machine learning security course
- Graded written and programming assignments, developed feedback rubrics, and held office hours to help students reason through attacks, defenses, and implementation details

Duke University, The Fuqua School of Business Durham, NC

Teaching Assistant | DECISION 521Q: Decision Analytics and Modeling 01/2023 – 04/2023

- Evaluated analytics and modeling assignments, clarified optimization and decision-analysis concepts, and supported students through office hours

Lexikat Pte. Ltd. Singapore

Software Development Engineer 09/2022 – 06/2023

- Built the core architecture of an iOS privacy-protection application in SwiftUI, integrating PhotosUI and UIKit for media selection, preview, and editing workflows
- Implemented efficient image and video processing pipelines with PhotoKit and Image I/O to support privacy-preserving transformations on user media
- Developed OpenCV-based text-image transformation features designed to reduce OCR readability while preserving human-readable content

Projects

AI Security Research Knowledge Base Telegram Bot

Python, FastAPI, LangGraph, Qdrant, PostgreSQL

- Built a personal Telegram bot based AI-security research assistant with **agent harness** to coordinate tool execution, workflow state management, retrieval pipelines, and multi-step research reasoning
- Designed a **multi-agent workflow** for topic classification, query expansion, source retrieval, grounded synthesis, and conversation summarization
- Implemented a **RAG knowledge-base pipeline** for arXiv PDFs, conference papers, GitHub READMEs, and personal notes using semantic chunking, metadata filtering, hybrid search, and reranking

Toy Rootkit: Stealth Kernel Module for Process Concealment [Github]

C, Linux Kernel Programming, System Call Hooking

- Built a **Linux kernel module** that intercepts system calls to support **process hiding**, **module hiding**, and **file access redirection**
- Modified the `sys_call_table` to hook `getdents64`, `openat`, and `read`, concealing target processes and kernel modules from user-space utilities such as `ls`, `ps`, and `cat /proc/modules`

Daemonized Multithreaded HTTP Caching Proxy [Github]

C++, Boost, Web Programming, Multithreads, Daemon

- Implemented a daemonized **C++/Boost** HTTP caching proxy that parses client requests, forwards upstream traffic, and stores reusable responses
- Designed an **LRU-based** cache with HTTP revalidation logic to reuse cached resources while preserving response freshness
- Added **multithreaded** request handling to serve concurrent client connections and isolate long-running network operations

Honors and Awards

- Usenix Security Student Grant and GREPSEC VII Travel Award 2026
- Dean's List with Distinction (top 10%) 2022
- Merit-based Scholarship 2019
- Second Prize, Chinese Chemistry Olympiad (CChO) 2018

Invited Talks

- "Covert Control Attacks on LLM" @ OpenSSF AI/ML Security WG 07/2023

Service

- Reviewer, NeurIPS 06/2026
- Reviewer, ACL ARR 05/2025

Skills

Programming Languages: Python, C/C++, Verilog
Tools and Frameworks: Git, LaTeX, PyTorch, Docker, Django, PostgreSQL