

Lowell, Ma
[Linkedin](#)

Andrew Adiletta

(978) 735 9666
ajadiletta@wpi.edu

Senior AI safety researcher and systems security engineer with a focus on adversarial prompt techniques and holistic system security. Developing guardrail techniques to protect alignment in frontier open source models.

Work Experience

Senior AI Safety PhD Researcher AI Safety and System Security SME	WPI Worcester, Ma	Aug. 2021 - Present Grad. Jan 2026
• Studying mechanistic interpretability on GPT-OSS, Llama4, Phi, Gemma, and other open source models • Optimization based adversarial prompt engineering – developing new capabilities for lab to red team LLMs • Studying novel detection mechanisms for Jailbreak and Prompt Injection attacks on GPT-OSS MoE models		
Senior AI Systems Security Researcher Project Lead	MITRE Bedford, Ma	Aug. 2023 - Present Secret Clearance
• Researched fault injection and side channel risks in GPU and embedded hardware • Collaborated across teams to quantify security features such as secure boot and key protections into hardware • Delivered technical reports and training on GPU security architectures and best practices to engineering teams		
Pre-silicon Validation Engineer Fault Tolerant Validation Utility	Intel Hudson M	2019-22 (22 months)
• Expanded graph-based validation checker, improving runtime and resource efficiency • Produced internal white paper on benefits of graph-based validation, leading to widespread adoption • Implemented validation for various power systems related flows for SoC servers		

Publications

[Impact] Super Suffixes: Bypassing Text Generation Alignment and Guard Models Simultaneously (Target: S&P 2026)

- Discovered novel joint prompt optimization approach to bypass Guard models and text generation models like Microsoft Phi simultaneously. Solves challenges with non-differentiable tokenization barriers
- Developed countermeasure *DeltaGuard* utilizing the cosine similarity to high level “jailbreak” direction in embedding space – outperforms Meta Prompt Guard in Jailbreak detection
- Paper: <https://arxiv.org/abs/2512.11783> Press Coverage: [\[Link\]](#) [\[Link\]](#)

Spill The Beans: Exploiting CPU Cache Side Channels to Leak Tokens from LLMs

- Discovered novel side channel targeting LLMs enabling token leakage in cloud via a CPU cache sidechannel
- Paper: <https://arxiv.org/abs/2505.00817> Press Coverage: [\[Link\]](#) [\[Link\]](#)

Rubber Mallet: A Study of High Frequency Localized Bit Flips and Their Impact on Security (DRAMSeC 2025)

- Discussed bypassing prompt guardrails via fault injection against GGUF model formats
- Paper: <https://arxiv.org/abs/2505.01518>

LeapFrog: The Rowhammer Instruction Skip Attack (EuroS&P 2025)

- Developed control flow subversion attack with Rowhammer, TLS & OpenSSL, and ML classification attacks
- Paper: <https://arxiv.org/abs/2404.07878> Press Coverage: [\[Link\]](#)

Mayhem: Targeted Corruption of Register and Stack Variables (AsiaCCS, 2024)

- Developed technique attacking stack, register variables using Rowhammer (SUDO, OpenSSH, OpenSSL)
- Paper: <https://arxiv.org/abs/2309.02545>

Don't Knock! Rowhammer at the Backdoor of DNN Models (DSN, 2023)

- Coauthored research paper on backdoor injection attacks on machine learning algorithms using fault injection
- Paper: <https://arxiv.org/abs/2110.07683>

Education and Certifications

• PhD ECE (GPA 4.0), Worcester Polytechnic Institute	2021-(Exp. Jan 2026)
• MS ECE (GPA 4.0), Worcester Polytechnic Institute	2021-2023
• BS ECE (GPA 4.0), Worcester Polytechnic Institute	2019-2022