STEPHANIE DING

San Francisco, CA hello@sdi.ng +1 626 650 3010 github.com/onionymous

WORK EXPERIENCE

Member of Technical Staff Apr. 2026 – Present
Thinking Machines Lab *San Francisco, CA*

- Working on safety and application security

AI Research Engineer Jul. 2024 – Apr. 2026
Meta (Meta Superintelligence Labs, Trust & Safety) *Pasadena, CA (Remote)*

- Led agentic and youth safety workstreams; conducting detailed safety analyses required for all first-party product and open-source model releases, including building datasets and state-of-the-art auto-raters for judging model responses, running evaluations and making recommendations according to internal policies
- Developed and open-sourced evaluations for multimodal and agentic prompt injection
- Developed open-source guardrails, including a model for jailbreak detection, a chain-of-thought monitor for agent misalignment, and an agentic guardrail orchestration framework
- Developed a synthetic audio detection model, released as part of the Llama Defenders Program

Software Engineer Aug. 2023 – Jul. 2024
Meta (Product Security, Mobile) *Pasadena, CA (Remote)*

- Architected, developed and deployed multiple Android and iOS frameworks (iOS URL signing and deeplink access controls, Android intent authentication based on app signatures, and file format and path validation frameworks), enabling secure-by-default development for thousands of Meta's mobile engineers
- Led pilot implementation of the Token Binding protocol on Instagram for Android to cryptographically link access tokens to devices and mitigate critical token theft and replay attacks

Security Engineer Sep. 2020 – Aug. 2023
Meta (Product Security, Web) *New York, NY*

- Developed and productionized scalable, automated detections for web application vulnerabilities using taint-flow static analysis
- Designed and built secret scanning system preventing credential leakage in internal infrastructure

RESEARCH EXPERIENCE

ML Alignment & Theory Scholars (MATS) Fellow Jan. 2026 – Mar. 2026
In collaboration with Kang Lab, University of Illinois Urbana-Champaign *Advisor: Daniel Kang*

- Designed a protocol for verifying and deploying software in Trusted Execution Environments (TEEs), enabling verifiable third-party audits of confidential AI without the disclosure of confidential source code. We develop an end-to-end implementation that runs in real TEEs, along with a benchmark of vulnerabilities in confidential AI and a reference agent implementation to audit for these vulnerabilities.

Supervised Program for Alignment Research (SPAR) Fellow Sep. 2025 – Dec. 2025
In collaboration with FAR.AI's Integrity Team *Advisor: Kellin Pelrine*

- Worked on jailbreaking and prompt injection research, including building an agentic prompt injection demo, investigating a new clustering-based technique for more efficient token-level adversarial optimization, and building a general framework for adversarial prompt optimization attacks

Summer Undergraduate Research Fellow Jun. 2017 – Sep. 2017
California Institute of Technology, Center for Data-Driven Discovery (CD3) *Advisor: Julian Bunn*

- Developed a two-stage, network-based botnet detection method using machine learning on network flows generated over time-limited intervals

EDUCATION

California Institute of Technology Sep. 2016 – Jun. 2020
Bachelor's in Computer Science (GPA: 4.0/4.3)

PUBLICATIONS

- LlamaFirewall: An open source guardrail system for building secure AI agents (Meta AI Security, arXiv:2505.03574) May 2025
- CYBERSECEVAL 3: Advancing the Evaluation of Cybersecurity Risks and Capabilities in Large Language Models (Meta AI Security, arXiv:2408.01605) Aug. 2024
- Machine Learning for Cybersecurity: Network-based Botnet Detection Using Time-Limited Flows (Caltech Undergraduate Research Journal, 2018) Jul. 2017

SPEAKING

- LlamaFirewall: Guardrails for Controlling Agentic AI Systems (Arsenal, Black Hat USA 2025) Aug. 2025
- LlamaFirewall: An open source guardrail system for building secure AI agents (Track 1, The Diana Initiative 2025) Aug. 2025
- High Stakes Open Questions in AI Security at Meta (Workshop, Paris AI Security Forum '25) Feb. 2025
- Influence the Next Generation of Meta's Open-Source LLM Security Tooling (Workshop, Vegas AI Security Forum '24) Aug. 2024

INTERNSHIPS

Security Engineer Intern Jul. 2019 – Sep. 2019
Meta (Product Security, Web) *Menlo Park, CA*

- Developed DOM XSS prevention frameworks using Trusted Types, conducted security reviews and assisted with triage of external bug bounty reports

Software Engineer Intern Dec. 2018 – Jan. 2019
MemVerge *Shanghai, China*

- Developed benchmarking suite for a distributed persistent memory store and an improved custom memory allocation scheme to reduce fragmentation compared to the default allocator in Intel's PMDK

Software Engineer, Tools & Infrastructure Intern Jun. 2018 – Sep. 2018
Google *Venice, CA*

- Developed a full-stack health analysis tool for an internal rule-based production monitoring and alerting platform, enabling hundreds of engineers to monitor the effectiveness of their configured rules

SKILLS

Languages	Python, Hack, C/C++, Rust, Java, Kotlin, Swift, C#, OCaml, HTML/JavaScript
Security	static analysis, web/mobile application security, LLM security (prompt injection, jailbreaking)
ML & AI	PyTorch, Transformers, evals, AI safety, red teaming