MOHAMMAD RAOUF ABEDINI

AI Security Research · LLM Agent Red-Teaming · Offensive & Defensive Security Engineering

Sydney, Australia · Ready to relocate now to San Francisco, CA · Travel-ready (Washington, DC) · Visa sponsorship required

raoof.r12@gmail.com · linkedin.com/in/mohammad-raouf-abedini-885a9226a

github.com/Raoof128 · raoufabedini.dev · ORCID 0009-0000-6214-258X

PROFESSIONAL SUMMARY

Security researcher building systems that measure and contain the cyber capabilities of frontier AI. Creator of Project Simurgh, a provider-agnostic containment-attestation framework that red-teams LLM agents under an adversarial, dishonest-producer threat model and produces Ed25519-signed, offline-verifiable evidence of what an agent did after a guardrail miss: 138/138 classifier-missed cases contained, live-agent attack success cut 9/140 to 0/140 on AgentDojo, five machine-checked Lean theorems. Evaluated Claude outputs for exploitable code and guardrail circumvention in Anthropic's safety-evaluation program (via Alignerr). Positioned as the defense-in-depth layer complementary to inline classifiers: they govern what a model may say; this attests what an agent was allowed to do. Shipped detection end to end: on-device phishing ML at 87% F1 and real-time intrusion detection at 500K+ packets/sec. Cofounder of an incubator-backed campus-AI startup; 70+ projects shipped.

CORE COMPETENCIES

AI / Agent Security: LLM agent red-teaming · prompt-injection & jailbreak-consequence containment · tool-use authority gating · guardrail & classifier evaluation · adversarial test generation · AgentDojo · Claude Computer Use

Security Research: vulnerability research & responsible disclosure · exploit-chain replay · threat modeling · threat-actor TTP profiling · MITRE ATT&CK · digital forensics · OSINT

Experiments & Data: Python · SQL · automated evaluation pipelines · reproducible benchmarks · signed evidence bundles · large-dataset analysis

Engineering: C/C++17/20 · TypeScript · Bash · Docker · Terraform · PostgreSQL · GitHub Actions CI/CD · Linux · Ed25519 attestation

Frameworks & Tooling: NIST CSF · OWASP Top 10 · Wireshark · Nmap · Burp Suite · libpcap

SELECTED SECURITY RESEARCH & PUBLICATIONS

Project Simurgh — Verifiable Containment Attestation for Agentic AI (Creator) 2026

AGPL-3.0 · github.com/Raoof128/Project-Simurgh · Zenodo preprints

- Built a provider-agnostic framework producing signed, offline-reproducible evidence of what an AI agent did after a guardrail miss, across four containment boundaries: input firewall, context-provenance guard, tool-invocation gate, and output-leakage firewall.
- Designed an adversarial, dishonest-producer threat model with a multi-class escape taxonomy and reproducible red-team campaigns; decision-replay and emission-completeness checks detect falsified or dropped evidence.
- **Guardrail-miss containment:** captured a real Llama Guard 4 (12B) input classifier over a 180-case run-set; contained 138/138 malicious cases the classifier missed (120 downstream-injection cases an input-only classifier structurally cannot see, 18 direct-input misses); combined targeted attack-success 0/150; zero unsafe tool executions or exports.
- **Live-agent containment:** drove a self-hosted Llama-3.3-70B through AgentDojo's workspace suite (140 pre-registered injection cases); the tool-authority gate cut targeted attack success from 9/140 to 0/140 with benign utility held (0/949 on the full four-suite run); containment claimed only within the declared, pre-frozen taxonomy.
- **Complementary to frontier safeguards:** built a frozen Fable-5-era reference corpus and a positioning brief mapping Simurgh as the defense-in-depth layer alongside Anthropic's inline classifier: the classifier governs what a model may say; the attestation governs what an agent was allowed to do (non-claims signed and explicit, including that it would not have caught the June 2026 content-generation bypass itself).
- **Attacked its own proof:** red-team sweep of the attestation core across eight attack classes (tamper, key-swap, canonical-laundering, digest-collision, cross-stage replay, self-proof mutation, policy drift); trust root held on all eight; two detector weaknesses found, versioned into detector-v2, and re-tested; producer-independent witness: zero false accusations, zero missed lies.
- **Formally checked oversight:** approval-gate friction receipts prove an oversight checkpoint preceded every protected authority crossing via a two-key pincer that defeats self-approval and backdating; closed with five machine-checked Lean theorems (fail-closed, friction precedence, no-silent-exemption).
- one-command offline reproduction of a 12-rung signed release ladder; Ed25519-signed byte-reproducible evidence bundles; AgentDojo integration with Claude Computer Use scenarios scoped as documented next step; 3,057 automated tests; preprints state explicit non-claims.

The Invisible Window — Vulnerability Research & Responsible Disclosure 2026

IEEE-format paper · DOI 10.5281/zenodo.20376495

- Independently discovered and responsibly disclosed a cross-platform screen-capture evasion class exploiting OS-level display-affinity APIs (Windows/macOS) that defeats browser-based capture and AI-vision pipelines.

Aion-BibleQA — RAG Faithfulness & Adversarial Robustness 2026

DOI 10.5281/zenodo.20522874

- Authored an 8-page preprint and 40-question benchmark for citation faithfulness and false-premise robustness; R@5 = 0.941, zero unsupported citations, 6/6 false-premise refusals.

Selected Additional Engineering (70+ shipped projects)

- Project Zurvan: local-first LLM knowledge engine extracting structured claims, decisions, and entities from raw documents, exposed to AI agents over an MCP stdio server (Python, SQLite FTS5 + embeddings, 218 tests).
- Cofounder of Syllabus-Sync, an agent-readable campus-AI platform accepted into the Macquarie University startup incubator as a formal startup (Next.js, Supabase, 503 tests across 92 files, live at syllabus-sync.app); NanoMatch, a C++20 limit-order-book engine at 1M+ orders/sec with O(1) cancellation; eBPF kernel runtime-monitoring research; technical write-ups at raoufabedini.dev.
- Offensive capability for defensive ends: hands-on PoCs in Windows hook-chaining (Rust), raw-socket SYN-scan detection without libpcap (Python), and LSB steganography (Go), each built to understand a detection or evasion class from the attacker's side.

PROFESSIONAL EXPERIENCE

AI Safety Evaluator, Claude — Alignerr Jan 2026 – Mar 2026

Anthropic AI safety-evaluation program · Remote (Sydney)

- Assessed Claude outputs for security vulnerabilities and harmful patterns, identifying where models produce dangerous or exploitable code.
- Analyzed how safety guardrails can be circumvented, building practical knowledge of LLM misuse vectors; delivered structured, rubric-based findings (continues earlier 2024 Claude Code evaluation work).

Freelance Security Engineer & Full-Stack Developer — Self-Employed Jan 2024 – Present

Sydney

- Architected the security backend for a production, multi-frontend student platform on Supabase: WebAuthn passkeys, TOTP/SMS MFA, zero-trust edge middleware, distributed rate limiting, strict Postgres Row-Level Security, and tamper-resistant audit logging via SECURITY DEFINER RPCs.
- Built SentinelFlow, a real-time C++ intrusion detection system parsing 500K+ packets/sec with signature and stateful detection of port scans, SYN floods, and DNS tunnelling.
- Built Mehr Guard, an offline phishing and QR/URL threat scanner (Kotlin Multiplatform, 5 targets) flagging brand impersonation, homograph attacks, and malicious redirects on-device via an ensemble ML model plus 25+ heuristics: 87% F1, sub-5ms latency, 1,248+ tests, and a built-in red-team suite of 19 curated attack scenarios.
- Engineered CI/CD with 500+ automated security tests (GitHub Actions), cutting deployment vulnerabilities ~40%; shipped security-first apps for 1,000+ users with OWASP Top 10 controls, OAuth 2.0, and cryptographic receipt validation.
- Sustained a 5.0-star rating across 31 client reviews and 35 completed tasks (95% completion, ID-verified) on Airtasker, delivering web development, Python, API integration, and IT/security work for paying clients.

IT Manager & Security Operations — Iran Pharmacy 2019 – 2024

Multi-site organization

- Investigated unauthorized-access incidents and built automated detection workflows across a 50+ staff multi-site organization, reducing manual triage ~30%.
- Enforced RBAC and least-privilege policies; maintained 99% uptime through monitoring, incident investigation, and rapid response.

EDUCATION

Bachelor of Cyber Security — Macquarie University Expected Nov 2026

Coursework: Digital Forensics, Network Security, Systems Security, NLP & Machine Learning, Privacy-Preserving Data Analysis, Cloud Computing

Diploma of Information Technology — Macquarie University 2023 – 2024

CERTIFICATIONS & CREDENTIALS

Anthropic (2026): Building with the Claude API · Claude Code in Action · Introduction to Claude Cowork · Claude Platform 101 · AI Fluency: Framework & Foundations · Claude Code 101 · Claude 101

Macquarie University (2026, grades 93–100%): Security of AI · Application of AI · Applied Cryptography · Data Security & Information Privacy · Digital Forensics · DevSecOps · Identity & Access Management · Mobile Security · GRC I (Governance) & II (Risk & Compliance) · Essentials for Managers & Leaders

LEADERSHIP & VOLUNTEER

Cofounder, **Macquarie Persian Students Society** (2026): founded and help run the university's Persian student community (events, peer support, cultural programming).

LANGUAGES

English (Professional Working) · Persian / Farsi (Native) · Japanese (Elementary)