Rathinam Trainers & Consultants Private Limited

The AI Engineering Series · Program 07 · 2026

Master AI Security

Red-team and then harden production LLM & AI-agent systems — by attacking and defending RedVault, a deliberately-vulnerable enterprise AI assistant — in 16 sessions.

4	16	32	18
MODULES	SESSIONS	HOURS	SEEDED VULNS

THE PROBLEM

Who Secures the AI Systems Everyone Is Shipping?

Organizations ship LLM and agent applications faster than they can secure them. Prompt injection and MCP supply-chain attacks are the #1 AI incidents of 2026 — yet the experts who can **both attack and defend** AI systems are critically scarce. Every team is wiring up RAG pipelines, tool-using agents, and MCP servers. Almost none of them have a person who can break those systems and then re-architect them secure-by-design.

> **The skill gap isn't building AI apps — it's securing them.**
> This program closes it by red-teaming and hardening a real AI system end-to-end.

WHAT YOU WILL BE ABLE TO DO

Learning Outcomes

- ✓ Threat-model AI systems with OWASP LLM Top-10 v2 and MITRE ATLAS
- ✓ Exploit agentic systems — excessive agency, confused-deputy, memory poisoning
- ✓ Apply adversarial ML — model extraction, membership inference, embedding inversion
- ✓ Engineer guardrails and I/O defenses with Llama Guard and NeMo Guardrails
- ✓ Secure RAG and data layers and harden the supply chain — MCP signing, scanning
- ✓ Execute prompt-injection attack chains — direct, indirect, multi-hop exfiltration
- ✓ Attack the AI supply chain — poisoned MCP servers, tool-description injection, lateral movement
- ✓ Build a custom automated red-team harness
- ✓ Architect least-privilege, sandboxed, egress-controlled agents
- ✓ Build detection + CI security gates, run AI incident response, map to NIST AI RMF & the EU AI Act

WHY THIS PROGRAM

Five Things No Other Program Does

01

Attack AND Defend the Same Real App

You exploit RedVault in Module 2, then harden the exact same weaknesses in Module 3 — and replay every attack to prove it now fails. Not slideware.

02

Expert Depth, Single-Stack

One toolchain taken to the core. The scope is covered exhaustively, nothing left out — not a shallow multi-tool tour that never goes deep.

03

Framework-Complete Coverage

Full OWASP LLM Top-10 v2.0, MITRE ATLAS, NIST AI RMF, and EU AI Act / ISO 42001 mapping — every attack and defense is provably mapped.

04

Graded Adversarial Capstone

A live engagement: attack → harden → defend under the instructor's attacks. You prove mastery in real time, not on a quiz.

05

Free & Local, $0

No vendor lock-in. The whole course runs on Ollama on a 16 GB laptop with open-source tools — Promptfoo, PyRIT, Garak, Llama Guard, NeMo Guardrails. Cloud is optional. You own every piece of the stack.

PROVABLE COVERAGE

Four Frameworks. Complete Coverage.

Every attack and defense is mapped to the industry frameworks — so coverage is provable, not anecdotal.

OWASP LLM Top-10 v2.0

The canonical AI-app risk list. Every category — from LLM01 Prompt Injection to LLM10 Unbounded Consumption — is attacked and then defended against RedVault.

MITRE ATLAS

Adversarial-ML & agent TTPs. Each attack maps to an ATLAS technique ID, so you speak the same language as threat intel and red teams worldwide.

NIST AI RMF

Govern, Map, Measure, Manage. The RMF (incl. the GenAI Profile) becomes concrete controls you implement and verify against RedVault — not a poster on the wall.

EU AI Act + ISO/IEC 42001

Security obligations as policy-as-code. Map regulatory and management-system requirements to OPA/Rego controls and auto-generate a compliance report.

WHO SHOULD ATTEND

Built for AI Security Practitioners

From security pros learning how AI breaks to builders learning how to defend what they ship — this program delivers hands-on value whether you come from attack, defense, or engineering.

🛡️

AI Security Engineers

Own the full attack-and-defend lifecycle for LLM and agent systems end-to-end.

🔐

AppSec Engineers

Extend application security into the AI surface — prompts, context, tools, supply chain.

🔎

SOC / Blue-Team Analysts

Build detection, telemetry, and incident response for AI abuse and agentic incidents.

🎯

Pen Testers / Red-Teamers

Add the OWASP GenAI / MITRE ATLAS offensive playbook to your engagement toolkit.

🤖

AI/ML Engineers & Builders

Upskill into security — learn how your RAG, agents, and MCP tools actually get broken.

🏗️

Security Architects & Leads

Design secure-by-design AI architecture and map controls to NIST AI RMF and the EU AI Act.

Prerequisites: Builder/developer fluency (Python, Git, APIs) **or** professional security experience, plus gated pre-work (a Security Primer for builders, an LLM/Agent/MCP Primer for security pros). 16 GB laptop.

CURRICULUM

Four Modules. 16 Sessions. One Real AI System.

The whole course is one arc — attack RedVault, harden it, then defend it live. Each module advances RedVault from vulnerable to secure-by-design.

MODULE 1

Sessions 1-4

Foundations & AI Threat Modeling

AI attack surface, OWASP LLM Top-10 v2, MITRE ATLAS, STRIDE-for-LLM, AI-BOM, and red-team lab setup (Promptfoo / PyRIT / Garak).

MODULE 2

Sessions 5-8

Offensive: Attacking AI Systems

Prompt injection, agentic exploitation, MCP and supply-chain attacks, adversarial ML — plus a learner-authored custom red-team harness in CI.

MODULE 3

Sessions 9-12

Defensive: Hardening & Architecture

Guardrails, secure least-privilege agent architecture, RAG and data-layer defense, supply-chain hardening — every Module-2 attack replayed and blocked.

MODULE 4

Sessions 13-16

Detection, Operations, Governance & Capstone

Detection engineering, AI incident response, compliance-as-code (OPA), and a graded live red-vs-blue capstone.

THE FULL TECH STACK

30+ Tools — Offensive & Defensive, All Free/OSS

Every tool below is used hands-on. We keep the repo-wide distinction strict: **AI tools** have a core AI/ML capability (they probe, classify, or orchestrate attacks); **platform tools** are execution infrastructure (they host, store, and enforce).

Offensive / Red-Team (AI tools)

Tool	What It Does	Free Tier
Promptfoo	Breadth red-teaming — sweep all OWASP LLM Top-10 attack plugins against RedVault (OpenAI-acquired Mar 2026, remains OSS)	✓ OSS
PyRIT	Depth red-teaming — multi-turn / multi-modal attack orchestration (Microsoft)	✓ OSS
Garak	LLM vulnerability probe scanner — 37+ probe families, reports which jailbreaks succeed (NVIDIA)	✓ OSS

Defensive (AI tools)

Tool	What It Does	Free Tier
Llama Guard	Open-weight safety classifier for input/output moderation; runs in Ollama	✓ open weights
NeMo Guardrails	Programmable rails (input / dialog / retrieval / execution / output) (NVIDIA)	✓ OSS
Guardrails AI	Structured output validation via validators / Guardrails Hub	✓ OSS

Target App / Platform Tools

Tool	What It Does	Free Tier
Python 3.13	Language runtime for RedVault and all tooling	✓ OSS
FastAPI	RedVault API and agent-loop host	✓ OSS
SQLAlchemy 2.0	ORM / data-access layer	✓ OSS
PostgreSQL 17 + pgvector	RAG vector store and relational data	✓ OSS
MCP server	Exposes agent tools over Model Context Protocol — the agency attack surface	✓ OSS
React 19 + TypeScript	RedVault chat UI and admin frontend	✓ OSS
Docker	One-command reproducible lab environment	✓ OSS
Ollama	Local model runtime — the $0 engine (Llama 3.x + Llama Guard)	✓ OSS

Ollama runs everything locally — $0 API cost during labs. No paid frontier model is needed to demonstrate every OWASP LLM Top-10 attack.

Detection & Ops

Tool	What It Does	Free Tier
Langfuse	LLM observability / tracing — score and detect attack patterns in traces	✓ OSS (self-host)
GitHub Actions	CI security gate — block merge on red-team regression	✓ free (public repos)
Prometheus	Metrics for attack/defense telemetry	✓ OSS
Grafana	Dashboards for abuse and defense signals	✓ OSS

Threat Modeling & Compliance

Tool	What It Does	Free Tier
OWASP Threat Dragon	Threat-model the GenAI architecture — STRIDE-for-LLM, DFDs, attack trees	✓ OSS
OPA / Rego	Compliance-as-code — policy gate on dangerous MCP tool calls; map EU AI Act / NIST AI RMF / ISO 42001 to controls	✓ OSS

The only optional paid resource is one instructor-side GCP A100 VM for heavier adversarial-ML demos — learners pay $0 and replicate on smaller local models.

THE PROJECT

Break & Defend RedVault — A Real AI System

RedVault is a deliberately-vulnerable, multi-tenant enterprise AI assistant — chat + RAG + an agent with MCP tools — shipped in **vulnerable** and **hardened** modes. You attack the vulnerable build, then re-architect it secure-by-design and prove every attack now fails.

> **Your final deliverable:** a hardened, monitored AI application + a professional red-team report + a secure-by-design architecture — a portfolio piece proving you can both break and defend AI systems.

Subsystem	What It Does	Curriculum Area
Chat Interface & Model Router	Streaming chat over local models via Ollama	Foundations / Guardrails
RAG Pipeline + pgvector	Retrieve and ground answers from a vector store	Data-layer defense
Agent + MCP Tools	Tool-calling agent exposed over Model Context Protocol	Agentic exploitation / Secure architecture
Red-Team Harness	Automated Promptfoo / PyRIT attack suite, custom probes	Offensive AI
Guardrails Engine	Llama Guard + NeMo Guardrails on the I/O path	Defensive AI
Detection & Telemetry	Langfuse tracing + GitHub Actions CI security gates	Detection engineering
Compliance-as-Code	OPA/Rego policy gates and auto-generated reports	Governance
Incident-Response Runbook	IR playbook + forensic log analysis for AI incidents	Operations

PROGRAM DETAILS

Format & Structure

This is a deep, immersive program. Each session pairs concept coverage (~40 min) with a live hands-on demo (~80 min) the instructor performs and learners replicate on their own laptop during the week.

FORMAT	Instructor-led — concepts (~40 min) + live hands-on demo (~80 min), replicated during the week
DELIVERY	Virtual (live, instructor-led)
DURATION	16 weeks (one 2-hour session per week)
SESSION	Every Saturday, 10:00 AM – 12:00 noon IST (batch runs Jul–Oct 2026)
LAB SETUP	Personal laptop (16 GB RAM) with Ollama for local models — $0; optional GCP for a few sessions
LANGUAGES	Python 3.13 (backend) · TypeScript / React 19 (frontend)
TAKEAWAYS	All attack/defense lab code, the hardened RedVault, and a reference guide

Session Time by Timezone

Region	Timezone	Session Time
India	IST (UTC+5:30)	**Saturday 10:00 AM – 12:00 noon**
USA (East Coast)	EDT (UTC-4)	Saturday 12:30 AM – 2:30 AM
USA (West Coast)	PDT (UTC-7)	Friday 9:30 PM – 11:30 PM
UK / Europe	BST (UTC+1)	Saturday 5:30 AM – 7:30 AM
UAE / Middle East	GST (UTC+4)	Saturday 8:30 AM – 10:30 AM
Singapore / East Asia	SGT (UTC+8)	Saturday 12:30 PM – 2:30 PM
Australia (Sydney)	AEST (UTC+10)	Saturday 2:30 PM – 4:30 PM

IMPORTANT

Note on Free Tiers & Costs

> **The course runs $0 locally.** Every hands-on exercise is reproducible on a 16 GB laptop with Ollama and open-source tools — verified as of June 2026. **Students are responsible for any optional cloud, GPU, or API costs** they choose to incur. Free tiers and tool licenses (e.g. Promptfoo post-acquisition) may change, reduce, or be removed at any time without notice — re-verify before each cohort.

Ready to Break — and Defend — AI Systems?

Contact us for batch schedules, corporate training, and custom program design.

Rathinam Trainers & Consultants Private Limited

We train engineers, not just tool users.