ARMOURINFOSEC

Hands-On Cyber Security Training · Indore, India

AI/ML Penetration Testing

Course Syllabus

Master offensive security testing of AI and machine learning systems including large language models, classifiers, recommendation engines, RAG pipelines, and autonomous agents. Learn prompt injection, jailbreaks, adversarial examples, model extraction, training-data poisoning, supply-chain attacks, and governance assessment aligned with OWASP LLM Top 10, OWASP ML Top 10, NIST AI RMF, and MITRE ATLAS.

// COURSE INFORMATION

Course Information

DURATION	3 Months / 12 Weeks / 90 Hours
LEVEL	Advanced
MODULES	13
FORMAT	Hands-on Labs / Hybrid (Online + Indore Classroom)

// COURSE OVERVIEW

Course Overview

The AI/ML Penetration Testing course is designed for penetration testers, red teamers, application security engineers, ML engineers, and AI security researchers who need to evaluate the security posture of modern AI systems. The curriculum covers the full attack surface of LLM-powered applications, traditional ML pipelines, and agentic systems. You will perform prompt injection and jailbreak attacks, build adversarial example payloads with FGSM, PGD and Carlini-Wagner, execute training-data poisoning and backdoor insertion, attack RAG pipelines and embedding stores, abuse Model Context Protocol (MCP) servers and tool-calling agents, conduct model extraction and inversion attacks, and assess ML supply chains and governance controls against NIST AI RMF and MITRE ATLAS.

// LEARNING OBJECTIVES

Learning Objectives

- Scope and execute end-to-end penetration tests of LLM-powered applications
- Perform direct and indirect prompt injection against production-style systems
- Build adversarial examples using FGSM, PGD, and Carlini-Wagner attacks
- Conduct training-data poisoning and backdoor insertion exercises
- Attack RAG pipelines, vector stores, and embedding-based retrieval
- Abuse MCP servers and tool-calling agents through confused-deputy and tool poisoning

- Execute model extraction, inversion, and membership inference attacks
- Assess ML supply chains for malicious models, pickles, and dependency risks
- Use Garak, PyRIT, ART, Counterfit, and custom harnesses effectively
- Map findings to OWASP LLM Top 10, OWASP ML Top 10, MITRE ATLAS, and NIST AI RMF
- Produce professional AI red team reports with prioritized remediation

// PREREQUISITES

Prerequisites

- Solid penetration testing or application security foundation
- Working Python proficiency for scripting and tooling
- Familiarity with REST APIs, HTTP, and Burp Suite
- Basic understanding of machine learning concepts (training, inference, loss, gradients)
- Comfort with Linux command line and Git
- Laptop with 16GB+ RAM (GPU helpful but not required)

// MODULE BREAKDOWN

Module Breakdown

01 AI/ML Security Landscape & Threat Modeling

- Modern AI/ML deployment patterns
- LLM applications, RAG, and agentic systems
- Traditional ML vs generative AI threat models
- STRIDE applied to ML systems
- OWASP LLM Top 10 walkthrough
- OWASP Machine Learning Top 10 walkthrough
- Mapping attacker goals to ML kill chain

02 MITRE ATLAS & NIST AI RMF

- MITRE ATLAS tactics and techniques
- ATLAS Navigator for engagement planning
- NIST AI Risk Management Framework (AI RMF) overview
- Govern, Map, Measure, Manage functions
- Mapping findings to ATLAS and AI RMF
- Engagement scoping for AI red teaming
- Rules of engagement for production LLMs

03 LLM Prompt Injection

- Direct prompt injection fundamentals
- Indirect prompt injection via documents, URLs, and tools
- System prompt extraction
- Instruction hierarchy bypass
- Multilingual and obfuscated payloads
- Token-level smuggling and Unicode tricks
- Cross-context injection in chat histories
- Building a prompt injection harness in Python

04 Jailbreaks & Guardrail Bypass

- Role-play and persona-based jailbreaks
- Many-shot and crescendo jailbreaks
- Encoding, translation, and cipher-based bypasses
- Refusal classifier evasion
- Bypassing content filters and moderation APIs
- Automated jailbreak discovery with Garak and PyRIT
- Tracking guardrail regressions across model versions

05 RAG Pipeline Attacks

- Anatomy of a RAG system
- Indirect injection through ingested documents
- Embedding inversion and recovery of source text
- Vector store poisoning
- Retrieval ranking manipulation
- Cross-tenant data leakage in multi-tenant RAG
- Metadata and citation forgery
- Hardening retrieval pipelines

06 MCP & Agentic System Abuse

- Model Context Protocol (MCP) architecture
- Malicious MCP server design
- Tool poisoning and tool-description injection
- Confused-deputy attacks on autonomous agents
- Function-calling abuse and argument smuggling
- Privilege escalation through chained tools
- Sandbox escapes in code-execution tools
- Auditing agent action logs

07 Adversarial Examples for ML Classifiers

- Adversarial example theory and threat models
- Fast Gradient Sign Method (FGSM)
- Projected Gradient Descent (PGD)
- Carlini-Wagner (CW) attacks
- Transferability across models
- Black-box query-based attacks
- Physical-world adversarial patches
- Using Adversarial Robustness Toolbox (ART)

08 Training-Data Poisoning & Backdoors

- Clean-label vs dirty-label poisoning
- Backdoor trigger design
- BadNets and label-flipping attacks
- Poisoning fine-tuning datasets
- Poisoning RLHF preference data
- Detection via spectral signatures and activation clustering

- Supply-chain implications of public datasets

09 Model Extraction & Model Inversion

- Query-based model extraction
- Functional equivalence vs fidelity extraction
- Extracting decision boundaries
- Model inversion to reconstruct training data
- Membership inference attacks
- Attribute inference attacks
- Privacy budget and differential privacy basics
- Defensive rate limiting and watermarking

10 ML Supply Chain Attacks

- Hugging Face Hub threat surface
- Malicious pickle and safetensors payloads
- Compromised model cards and configs
- Dependency confusion in ML tooling
- Notebook and pipeline poisoning
- Container and base-image risks for ML workloads
- Signing, attestation, and SBOMs for models

11 LLM Application Security Testing

- Testing LLM-backed web APIs with Burp Suite
- Authentication and authorization in LLM apps
- Output handling: XSS, SSRF, and command injection via LLM responses
- Insecure plugin design (OWASP LLM07)
- Excessive agency (OWASP LLM08)
- Sensitive information disclosure
- Denial-of-wallet and cost-exhaustion attacks

12 Tooling Deep Dive

- Garak vulnerability scanner for LLMs
- Microsoft PyRIT for automated red teaming
- Adversarial Robustness Toolbox (ART)
- Counterfit for ML attack automation
- MITRE ATLAS Navigator workflows
- Building custom prompt-injection harnesses in Python
- HuggingFace transformers for offline attack research
- Integrating tools with Burp and CI pipelines

13 Reporting, Governance & AI Red Team Operations

- Structuring AI red team findings
- Severity scoring for AI/ML vulnerabilities
- Mapping findings to OWASP, ATLAS, and NIST AI RMF
- Governance review: model cards, data sheets, evaluation suites
- Continuous AI red teaming programs
- Communicating risk to ML and product teams

- Building safe disclosure and remediation workflows

// TOOLS & HANDS-ON LABS

Tools & Hands-On Labs

- Self-hosted vulnerable LLM chatbot with system prompts and tools
- RAG pipeline lab with poisonable document store and vector DB
- Agentic system lab built on Model Context Protocol (MCP)
- Image and tabular classifier lab for adversarial example exercises
- Training-data poisoning sandbox with reproducible fine-tuning jobs
- Garak and PyRIT preconfigured against local and API targets
- Adversarial Robustness Toolbox (ART) and Counterfit ready to run
- Burp Suite with custom extensions for LLM API testing
- Python + HuggingFace transformers environment with sample models
- MITRE ATLAS Navigator workspace for engagement tracking

// TRAINING MODE

Training Mode

Every Armour Infosec course runs as a unified programme delivered in two parallel modes — the same curriculum, the same trainers, the same certification, regardless of how you join.

- ✓ Online Live Classes — real-time, instructor-led, fully interactive sessions
- ✓ On-Premise Classroom Training — in-person at our Indore centre (Sudama Nagar)
- ✓ Both modes run concurrently in every batch; switch between them as your schedule needs
- ✓ Same syllabus, lab access, and certification track for online and on-premise students

// CERTIFICATIONS & CAREER OUTCOMES

Certifications & Career Outcomes

This course aligns with industry-recognised certifications and prepares graduates for offensive-security, application-security, and infrastructure-security roles.

- OSCP+ (Offensive Security Certified Professional+) — Advanced Offensive Security Certification
- CEH (Certified Ethical Hacker)
- OWASP LLM Top 10 alignment
- OWASP Machine Learning Top 10 alignment
- MITRE ATLAS-aligned AI red team methodology
- NIST AI Risk Management Framework readiness
- Generic AI red-teaming and AI security assessor preparation

// ENROL WITH ARMOUR INFOSEC

Enrol With Armour Infosec

Reach out to discuss enrolment, batch schedule, and lab access. Our Indore training centre runs both in-person and live online cohorts with placement assistance.

PHONE +91 99777 47168

EMAIL info@armourinfosec.com

ADDRESS	674, Sudama Dwar, Narendra Tiwari Marg, Sudama Nagar, Indore, Madhya Pradesh 452009
WEBSITE	https://armourinfosec.com

Scan to View Course Online

https://www.armourinfosec.com/training/ai-ml-penetration-testing/