PRACTICAL AI SECURITY: ATTACKS, DEFENSES, AND APPLICATIONS

Expert-Led Cybersecurity Training · Beginner to Advanced

COURSE OVERVIEW

This course gives you a practical, hands-on path into AI security with a strong focus on LLM-powered applications. You'll learn how modern AI systems are built, how they fail, and how to secure them against real threats. Everything is structured around doing rather than theory, so you immediately apply what you learn.

You begin with a solid technical foundation. Through hands-on labs, you'll build working LLM applications using the Hugging Face Transformers ecosystem, implement RAG pipelines with LangChain, LlamaIndex, and FAISS, and explore how tokenization, embeddings, and context windows shape model behavior. You'll also learn advanced prompt engineering patterns and build your own MCP servers to automate security tasks and integrate AI into real workflows.

The security phase takes you deep into offensive and defensive techniques. You'll practice prompt injection, multimodal exploitation, and workflow manipulation against agents and AI-generated ("vibe-coded") applications. You'll map these risks to Google's Secure AI Framework and learn how to threat model, and harden RAG systems, agent logic, and custom MCP servers with proper authentication and validation.

8KSEC

You finish by learning how to use AI to accelerate your own work. You'll use tools like Fabric AI, OpenRouter, and Perplexity to automate threat intelligence, research, and analysis, giving you a repeatable process to move faster with better accuracy.

By the end of the course, you'll be able to design, assess, and secure AI-powered systems with confidence.

You'll also be prepared for the **Certified AI Security Researcher (CAISR)** exam, with one exam attempt included.

WHY SHOULD YOU TAKE THIS COURSE?

This is a completely hands-on course designed for beginners to intermediate professionals seeking to master AI security. Instead of just theory, students actively exploit vulnerabilities and build security tools guided by expert instructors.
Both on-site and virtual attendees get cloud-based lab environments to perform exercises with LLMs, vector databases, and AI frameworks without complex setup.
A dedicated Slack channel ensures pre-course readiness and post-course collaboration for continuous learning.

WHAT WILL THE STUDENT GET?

- Certificate of completion
- Complete course materials (slides, lab guides)
- Source code for all vulnerable AI applications used in class
- Source code for exploit PoCs used in assessments
- All Python scripts and tools developed during the course
- Cloud instances for the duration of the course
- Access to pre-configured lab environment with required tools
- Slack access for collaboration and AI security discussions
- MCP server templates and custom tools repository
- An attempt at the Certified AI Security Researcher (CAISR) exam

WHO SHOULD ATTEND?

This course is designed for AI enthusiasts, Penetration testers, Security researchers, Security engineers, DevSecOps professionals, and anyone interested in learning AI security and understanding both offensive and defensive aspects of LLM security

8KSEC

KEY LEARNING OBJECTIVES

- Understand the core concepts distinguishing AI, Machine Learning, and LLMs, including supervised vs unsupervised learning, neural networks, Generative AI, diffusion models, and the complete ML model training lifecycle from data preprocessing to deployment.

- Master the fundamentals of Large Language Models, including Transformer architecture, tokenization mechanisms (BPE), context windows, embeddings, and the differences between foundational and fine-tuned models like GPT vs BERT architectures.

- Become proficient in Prompt Engineering techniques including system vs user prompts, prompt templates, leaked system prompts analysis, and controlling model output via sampling parameters (Temperature, Top-k, Top-p) for security-focused workflows like threat modeling assistants.

- Learn to use essential AI development tools including Hugging Face Transformers, LangChain (with memory and tool integration), LlamaIndex (multi-file processing), OpenWebUI for local LLM deployment, vector databases like FAISS for RAG implementations, and fine-tuning workflows.

- Build and deploy production-ready AI applications, including custom RAG (Retrieval-Augmented Generation) systems with vector storage, conversational agents with short and long-term memory, AI-powered security tools with proper rule-based and advanced guardrails, and FastAPI-based scanners.

- Master Model Context Protocol (MCP) servers for integrating AI with security tools to understand MCP vs traditional connectors, build custom MCP servers, and leverage them for reverse engineering, Mobile malware analysis, and automated penetration testing workflows. Configure MCP with Cursor and Claude for enhanced AI-assisted security research.

- Develop Offensive AI capabilities, including building autonomous AI agents and workflows for vulnerability scanning, CVE finding, reconnaissance, IAM policy analysis, threat intelligence gathering, and exploit development assistance using frameworks like LangChain.

- Execute advanced attacks against AI systems, including Prompt Injection variants (direct, indirect, multimodal attacks on CV screeners, meeting summarizers, image analyzers), jailbreaking techniques, data exfiltration through prompt manipulation, and exploiting MCP server vulnerabilities (Confused Deputy attacks, information disclosure, bruteforcing, arbitrary file read/write).

- Implement Defensive AI strategies, including securing AI-powered applications against prompt injection, analyzing vulnerabilities in "vibe-coded" AI-generated applications, securing MCP servers with proper authentication and authorization, and applying pre-launch security checklists for AI-assisted apps.

8KSEC

- Deploy and configure AI Gateways to secure production LLM applications and learn to migrate existing apps behind AI Gateways, implement multi-layered guardrails for input/output validation, configure rate limiting policies, and leverage analytics and comprehensive logging for monitoring, compliance, and cost optimization.
- Master AI-powered Threat Modeling using STRIDE methodology and understand the engineering logic of systematic threat modeling, leverage LLMs to identify threats across Spoofing, Tampering, Repudiation, Information Disclosure, Denial of Service, and Elevation of Privilege categories, and develop practical mitigations with AI assistance.
- Apply AI to enhance Security Operations and Reverse Engineering workflows using Fabric AI for knowledge mining, log parsing, email header analysis, threat intelligence processing, video knowledge extraction, breaking language barriers in security research, and integrating AI into tools like Ghidra and JADX for automated malware analysis.
- Understand and implement enterprise AI security frameworks, including comprehensive coverage of Google's Secure AI Framework (SAIF) with all 14 security risks (data poisoning, unauthorized training data, model tampering, prompt injection, model evasion, sensitive data disclosure, etc.).
- Debug, intercept, and secure MCP implementations using MCP Inspector for debugging, Burp Suite for traffic interception and modification, and apply the comprehensive MCP Server Security Cheatsheet for identifying and remediating common vulnerabilities in both custom and third-party MCP servers.
- Secure AI supply chains by pinning dependencies, verifying model signatures, understanding format risks, and detecting tampering or backdoors.
- Earn the Certified AI Security Researcher (CAISR) certification by demonstrating mastery across all course modules from foundational AI/ML concepts through advanced offensive and defensive AI security techniques in real-world scenarios.

HARDWARE/SOFTWARE REQUIREMENT

- Laptop with 16+ GB RAM (32 GB recommended) and 100 GB free space
- Access to Linux cloud instances (provided)
- API keys for LLM providers (provided for exercises)
- Administrative access on your local system
- Setup instructions and Slack details sent prior to course start

PREREQUISITE KNOWLEDGE

- Working knowledge of cybersecurity and testing fundamentals
- Basic Python programming skills
- Understanding of APIs and web services
- Familiarity with command-line interfaces
- Basic knowledge of authentication, authorization, and encryption
- Basic web application security knowledge (recommended, not required)

8KSEC

COURSE SYLLABUS

Module 1: Before the Prompt – Rise of the Transformers

- Understanding Transformer Architecture
- "Attention Is All You Need" paper breakdown
- How attention mechanisms work
- Evolution from RNNs to Transformers
- Installing dependencies and setting up environment
- Exploring transformer architectures through hands-on exercises

Module 2: Foundations of AI, ML, and LLMs

- Introduction to Artificial Intelligence concepts
- AI vs ML vs LLM distinctions
- Supervised vs Unsupervised Learning paradigms
- Neural Network architecture and principles
- Generative AI fundamentals
- Diffusion Models explained
- Building Spam Detector (supervised)
- Building Customer Clustering model (unsupervised)
- Implementing Generative AI examples
- StableDiffusion for image generation

Module 3: Inside Large Language Models (LLMs)

- What is a Large Language Model?
- Understanding GPT, BERT, and related architectures
- Interactive LLM experiments
- ChatGPT vs Grok comparison
- From Words to Tokens: text splitting
- Building a simple tokenizer
- Embeddings: giving meaning to numbers
- Visualizing vectors and embeddings
- The Context Window: understanding memory limitations
- Temperature and parameter tuning exercises

Module 4: Building with LLMs

- Introduction to the Transformers library
- Building NLP pipelines for tasks
- Creating a News Summarizer
- Building a Q&A tool from scratch
- Building an Unsafe Text Detector
- Running LLMs locally and via OpenWebUI
- Comparing local vs cloud deployment

8KSEC

Module 5: RAG, LangChain & LlamaIndex

- Fundamentals of Retrieval Augmented Generation
- Vector storage and semantic search
- RAG demo with FAISS and SentenceTransformers
- Secure API key management
- LangChain introduction and conversational memory
- Custom prompts and external tool connections
- LlamaIndex for document processing
- Fine-tuning LLMs for security use cases

Module 6: Prompt Engineering

- Fundamentals and best practices
- System vs User Prompts
- Analyzing real-world leaked system prompts
- Prompt security and attack surfaces
- Prompt template design for consistency
- Creating a Threat Model Assistant
- Adding UI for prompt-based tools
- Advanced prompt engineering techniques

Module 7: MCP Servers

- Model Context Protocol (MCP) fundamentals
- Architecture, setup, and configuration
- Connectors vs MCP Servers
- Reverse Engineering with MCP Servers
- Android Malware Analysis examples
- Automated Pentesting using MCP Servers
- Creating your first MCP Server
- Security implications and threat modeling

Module 8: AI Agents and Workflows

- Fundamentals of AI Agents
- Agent architecture design
- Building CVE Finder and IAM Policy Analyzer
- Creating Recon and CVE Analyzer Agents
- Building a FastAPI-based vulnerability scanner
- Workflow automation strategies

Module 9: Google Secure AI Framework (SAIF)

- SAIF fundamentals and pillars 1–6
- Implementation checklist and risks overview
- 14 major AI risks including:
  - Data Poisoning
  - Model Tampering
  - Prompt Injection
  - Model Exfiltration
  - Evasion and Output Disclosure
- Applying SAIF to real-world AI systems

Module 10: Human Augmentation Using Fabric AI

- Introduction to Fabric AI
- Installation and configuration
- Knowledge Mining & Distillation
- AI-powered translation and OSINT
- Threat Intelligence automation
- Video and log analysis using Fabric

8KSEC

Module 11: Attacking LLMs – Prompt Injection & Multimodal Attacks

- Prompt Injection taxonomy and classification
- Exploiting AI-powered CV screeners, meeting summarizers, and image analyzers
- Securing vulnerable systems
- Multimodal attack strategies and defenses

Module 12: Vulnerabilities in AI-Generated Applications

- "Vibe-coded" app vulnerabilities
- Common misconfigurations and exploits
- Security checklists and code review strategies
- Identifying AI-specific flaw patterns

Module 13: Exploiting and Securing MCP Servers

- Debugging MCP Servers using MCP Inspector
- Configuring Cursor & Claude for Interacting with MCP Servers
- Intercepting and Modifying MCP Server Traffic using Burp Suite
- Arbitrary file r/w in MCP Servers
- Exploiting Confused Deputy Attack via Delegation in MCP Servers
- Exploiting Information Disclosure in MCP Servers
- Bruteforcing Accounts in MCP Servers
- Securing MCP Servers against Attacks
- MCP Server Cheatsheet

Module 14: AI-powered Threat Modeling with STRIDE methodology

- STRIDE methodology for Threat Modeling
- Practical Threat Modeling using LLMs'

Module 15: Securing Apps Using AI Gateways

- Understanding the role of AI Gateways in modern applications
- Migrating an existing app behind an AI Gateway
- Adding guardrails and enforcing security policies
- Using analytics, logs, and rate limiting for monitoring and protection

Module 16: Supply Chain Integrity in AI Applications

- Understanding supply chain risks in AI systems
- Dependency pinning and version control strategies
- How model signing works and why it matters
- Signing and verifying models in practice
- Model file formats and their security considerations
- Exploiting insecure serialization (Pickle code injection)
- Detecting backdoored or tampered models

About the company

8kSec is a foremost cybersecurity research company offering exceptional training and consulting services to aid clients in enhancing their security stance. Our experts possess extensive experience in delivering specialized cybersecurity training and consulting to multiple commercial and defense organizations across the United States, Europe, and the Middle East and North Africa region.

Get in touch

8kSec.io
info@8ksec.io