EC-Council
Building A Culture Of Security

C|OASP
Certified Offensive AI Security Professional

CERTIFIED OFFENSIVE AI SECURITY PROFESSIONAL

Master the Tactical Methodology to Hack LLMs and Secure Agentic AI: the Global Command for Offensive Teams

LLMs are vulnerable. Prompt injection can bypass guardrails. Data poisoning can corrupt models.

EC-Council's Certified Offensive AI Security Professional (C|OASP) credential validates that you can red-team AI systems, exploit vulnerabilities in LLMs and agents, and build defenses that survive real-world attacks.

AI Red-Teaming Is a *New Discipline*

AI is transforming products, operations, and decision-making across industries. But when AI systems move into production, they open new attack paths, through models, prompts, data pipelines, agent workflows, APIs, and integrations, creating vulnerabilities adversaries are already targeting.

Traditional pentesting doesn't fully cover LLM vulnerabilities. Prompt injection, data poisoning, and model manipulation require specialized offensive skills. Certified Offensive AI Security Professional is the first credential built specifically for AI red teamers.

The Market Problem

Why organizations are vulnerable to AI attacks:

- Pentesters are not trained to exploit LLMs or AI agents
- There is no standardized methodology for AI red-teaming
- Traditional vulnerability scanners miss AI-specific flaws
- SOC teams struggle to detect AI-powered attacks
- Security architects lack fluency in AI threat models

When AI is attacked, defenses are tested.

Adversaries exploit LLMs and AI agents faster than security teams can assess them. Organizations need clear, repeatable methods to test AI systems against real-world attacks and prove defenses hold up before attackers find the gaps.

Introducing Certified Offensive AI Security Professional C|OASP

Certified Offensive AI Security Professional

C|OASP is a hands-on, practitioner-level credential that validates your ability to **ethically attack AI systems** so you can **defend them with engineering-grade controls.**

C|OASP is not about building AI models or running AI programs. It is about proving you can:

- Think like an attacker inside AI systems
- Uncover weaknesses across models and pipelines
- Validate security controls
- Reduce operational risk before deployment

This is the only credential built for offensive AI security work with outcomes you can demonstrate.

Skills This Program Verifies

The C|OASP credential validates your ability to:

- Execute prompt injection, jailbreaking, and prompt chaining attacks
- Red-team AI agents, including memory corruption, tool misdirection, and checkpoint manipulation
- Apply OWASP LLM Top 10 and MITRE ATLAS frameworks
- Conduct adversarial ML attacks, including data poisoning and model extraction
- Build detection rules and hardening strategies for AI systems

Verifiable Skills You'll Gain with Certified Offensive AI Security Professional C|OASP

C|OASP trains you to perform **end-to-end adversarial testing** and deliver **defensive validation evidence**, including:

Simulate adversarial AI kill chains

reconnaissance → mapping → exploitation → manipulation → exfiltration

Harden AI architectures

secure **system prompts**, context windows, tool integrations, **RAG pipelines**, and agent memory

Conduct AI security assessments

aligned to **MITRE ATLAS, OWASP LLM/ML Top 10, NIST AI RMF,** and **DoD Test & Evaluation** practices

Build SOC-ready capabilities

AI-focused detection logic, incident playbooks, and forensic procedures

Execute prompt injection, adversarial prompting

and **data poisoning** against LLMs/ML systems to identify training-time and inference-time weaknesses

Assess AI supply-chain risk

across models, datasets, dependencies, and third-party integrations using **SBOM/ MBOM** approaches

Implement defensive engineering controls

filtering, sandboxing, rate limiting, anomaly detection, and drift monitoring

Produce assurance and compliance artifacts

mapped to **NIST AI RMF, ISO/IEC 42001**, and emerging AI regulatory expectations

Enterprise Impact of *Verifiable Skills* from Certified Offensive AI Security Professional C|OASP

- Helps organizations identify and neutralize AI-specific threats before attackers do.
- Bridges security, engineering, and data science so controls exist across the full AI life cycle.
- Addresses the lack of standardized AI red-teaming methodology by applying OWASP LLM Top 10 and MITRE ATLAS frameworks.
- Strengthens resilience across plugins, APIs, and vendor ecosystems by exposing third-party risk.
- Improves monitoring and response because defenders understand attacker tactics at the model, application, and system level.
- Reinforces secure and ethical AI deployment, supporting innovation without sacrificing trust.

Certified Offensive AI Security Professional COASP Certification *Modules*

Module 01
Offensive AI and AI System Hacking Methodology

Build a foundation in offensive AI security by learning how AI systems are designed, where they fail, and how adversaries exploit them, using structured hacking methodologies and globally recognized AI security frameworks.

What You will Learn

- Understand AI and machine learning fundamentals from an offensive security perspective
- Identify AI attack surfaces, threat landscapes, and adversary techniques aligned to MITRE ATLAS
- Apply AI system hacking methodologies, frameworks, and risk implications
- Classify AI attack taxonomies and models
- Define offensive AI scoping fundamentals and foundations for securing AI systems
- Provide an overview and mapping of OWASP LLM & ML Top 10 (2025) to AI threats and governance considerations

Module 02
AI Reconnaissance and Attack Surface Mapping

Learn advanced AI-focused OSINT techniques to identify, enumerate, and analyze AI assets, data pipelines, models, APIs, and attack surfaces, and apply exposure mitigation and hardening strategies to support continuous AI security monitoring.

What You will Learn

- Apply OSINT tools and techniques to identify and profile AI assets
- Gather intelligence from AI data sources and training pipelines
- Discover and map AI attack surfaces using publicly available intelligence
- Enumerate AI endpoints, services, APIs, and exposed parameters
- Identify and analyze AI models and vector stores from an attacker’s perspective
- Evaluate OSINT exposure and apply hardening controls to reduce risk
- Use AI threat intelligence to support continuous monitoring and defensive readiness

Certified Offensive AI Security Professional C|OASP Certification *Modules*

Module 03
AI Vulnerability Scanning and Fuzzing

Master AI-specific vulnerability assessment and fuzzing techniques to identify, analyze, and mitigate security weaknesses across modern AI systems and applications.

What You will Learn

- Understand core principles of AI vulnerability assessment and threat discovery
- Use tools and techniques for scanning vulnerabilities in AI models, pipelines, and deployments
- Apply practical fuzzing methods tailored for AI systems and model interfaces
- Integrate scanning and fuzzing into AI security workflows for proactive risk mitigation

Module 04
Prompt Injection and LLM Application Attacks

Analyze and exploit LLM trust boundaries using advanced prompt injection, jailbreaking, and output manipulation techniques, while identifying risks related to sensitive data exposure and insecure LLM application design.

What You will Learn

- LLM architecture, trust boundaries, and associated attack vectors
- Execute prompt injection and jailbreaking techniques in real-world LLM applications
- Identify sensitive information disclosure and system prompt leakage risks
- Evaluate improper output handling vulnerabilities and misinformation threats
- Apply advanced prompt-based attack techniques and exploitation strategies
- Implement secure LLM application design principles and defensive controls

Module 05
Adversarial Machine Learning and Model Privacy Attacks

Execute and analyze adversarial machine learning, privacy, and model extraction attacks to assess AI system robustness, trustworthiness, and risk, and apply defensive strategies to mitigate them.

What You will Learn

- Identify core adversarial machine learning attack classes
- Execute practical adversarial input attacks across data modalities
- Apply privacy, inference, and model extraction attack techniques
- Evaluate robustness, trustworthiness, and risk evaluation methods
- Implement defensive strategies for model privacy and resilience

Certified Offensive AI Security Professional C|OASP Certification *Modules*

Module 06
Data and Training Pipeline Attacks

Compromise AI systems through data poisoning and backdoor insertion targeting training pipelines and model integrity.

What You will Learn

- Understand AI data and training pipeline architecture and threat surfaces
- Execute practical data poisoning techniques and attack scenarios
- Apply backdoor and trojan insertion during model training
- Implement security measures to safeguard data and training pipelines

Module 07
Agentic AI and Model-to-Model Attacks

Analyze and exploit autonomous AI agents and multi-model architectures by targeting excessive agency, cross-LLM interactions, orchestration workflows, and unbounded resource consumption, while understanding defensive strategies to secure agentic systems.

What You will Learn

- Understand agentic AI architecture and attack surface
- Apply excessive agency and autonomy exploitation techniques
- Identify cross-LLM and model-to-model attack vectors
- Asses denial-of-wallet risks and unbounded resource consumption
- Execute attacks targeting AI workflows and orchestration layers
- Implement defensive strategies for securing agentic AI applications

Module 08
AI Infrastructure and Supply Chain Attacks

Explore offensive techniques targeting AI infrastructure, system integrations, and third-party dependencies, while learning how to identify, exploit, and harden AI supply chain weaknesses.

What You will Learn

- Understand AI infrastructure components and system integration architectures
- Identify vulnerabilities in AI systems, frameworks, and deployment pipelines
- Analyze abuse of tools, plugins, and APIs in AI-enabled applications
- Assess AI supply chain threats and dependency risks (deep dive)
- Implement hardening strategies for AI infrastructure and supply chains

Certified Offensive AI Security Professional C|OASP Certification *Modules*

Module 09
AI Security Testing, Evaluation, and Hardening

Apply structured AI security testing and evaluation methodologies to assess risk, validate controls, and implement hardening best practices across enterprise AI systems.

What You will Learn

- Understand AI security testing methodologies and evaluation techniques
- Apply red team frameworks for offensive AI assessment
- Identify, validate, and report AI vulnerabilities and risk
- Implement security hardening and mitigation best practices for AI systems

Module 10
AI Incident Response and Forensics

Master AI-specific incident response and forensics, concluding with hands-on engagement in AI red team activities.

What You will Learn

- Detect and respond to AI-specific security incidents
- Collect and analyze AI logs, telemetry, and digital evidence
- Analyze root causes in post-incident analysis

20+ Hands-on AI Offensive Security Techniques Covered in the *Program*

- OSINT-Driven Supply Chain Assessments
- AI Model Vulnerability Research
- AI Reconnaissance via Model Fingerprinting
- Multi-Protocol Reconnaissance
- Telemetry Analysis to Map AI Decision Boundaries
- API Reconnaissance
- Model Artifact Exfiltration on AI-Powered Infrastructure
- AI Model Robustness Evaluation using Coverage-Guided Fuzzing
- Stateful AI Vulnerability Scanner using PyRIT
- Prompt Injection and Persona-Based Privilege Escalation in AI Chatbots
- Automating LLM Filter Evasion with Prompt Orchestration
- PAIR, TAP, and GCG Orchestration for Automating Advanced Jailbreaks
- Natural Language to SQL (NL2SQL) Vulnerabilities in AI-Powered Search Interfaces
- Remote Code Execution (RCE) Vulnerabilities in DocsGPT
- Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD) Attacks on Image Classifiers
- Transfer, Boundary, and Noise Attacks on AI Models
- Projected Gradient Descent (PGD) Attacks on Audio Classification and Transcription Models
- Gradient-Based and Heuristic Attacks on NLP Classifiers
- API Reconnaissance and Model Extraction on AI Endpoints
- RAG Poisoning and Contextual Prompt Injection Attacks on AI Systems
- Training Data Poisoning via Label Flipping
- Supply Chain Data Poisoning Attacks
- Privilege Escalation and Command Injection Attacks on Autonomous AI Agents
- Multi-Stage Prompt Injection and Denial of Service Attacks on AI Assistants
- AI Model Supply-Chain Poisoning Attacks on Access Control Systems
- Remote Code Execution (RCE) via Insecure Deserialization in MLflow
- Forensic Signal Extraction and Telemetry Analysis on Obfuscated Datasets
- AI Systems from Training Data Corruption

Key *Alignments*

- MITRE ATLAS Framework
- OWASP LLM Top 10 (2025)
- OWASP ML Security Top 10 (2025)
- OWASP Top 10 for Agentic Applications
- DoD AI Test & Evaluation Specialist (672) Framework
- NIST AI Risk Management Framework

15+ MITRE ATLAS Techniques in *Program*

- AI Model Inference API Access
- AI-Enabled Product or Service
- Active Scanning
- Craft Adversarial Data
- Develop Capabilities
- Discover AI Artifacts
- Discover ML Model Ontology
- Exfiltration via ML Inference
- Exploiting Public-Facing Applications
- Extract LLM System Prompt
- Extract ML Model
- Full AI Model Access
- Gather Victim Identity Information
- LLM Data Leakage
- LLM Jailbreak
- LLM Prompt Injection
- Physical Environment Access
- Retrieval Content Crafting
- Reverse Shell
- Search Application Repositories
- Search Open AI Vulnerability Analysis
- Search Victim-Owned Websites
- Verify Attack

20+ Key Tools Covered in the *Program*

- Garak (LLM vulnerability scanner)
- PyRIT (Microsoft's AI red team tool)
- Burp Suite for AI APIs
- OWASP ZAP for web-based AI services
- Atheris
- AFL (American Fuzzy Lop)
- CleverHans
- ART (Adversarial Robustness Toolbox)
- Foolbox
- Alibi Detect
- TensorFlow Data Validation (TFDV)
- Fairlearn
- IBM AI Fairness 360
- Prompt Fuzzer
- ToolFuzz
- Tensorfuzz
- FuzzyAI Fuzzer
- Wfuzz
- TruffleHog
- Gitleaks
- GitRob
- Mindgard
- Promptfoo
- Giskard
- TextAttack
- HiddenLayer
- AutoRTAI

Offensive *AI Security*

Methodology

From reconnaissance to exploitation and from testing to hardening, there is a systematic approach to securing AI systems against adversarial threats.

This framework equips you to think like an attacker and defend like an expert.

RECON 01

Map AI system architectures, enumerate exposed endpoints, and build threat models. Review training pipelines, data flows, and inference APIs to identify where defenses are weakest.

EXPLOIT 02

Execute prompt injection, jailbreaking, data poisoning, and model extraction attacks to validate AI system weaknesses and document exploitable gaps.

DEFEND 03

Implement guardrails, detection mechanisms, and incident response procedures to harden AI systems and ensure resilient, secure deployments.

Who C|OASP Is Designed For

C|OASP is designed for security professionals who wish to master offensive and defensive AI security techniques.

Offensive Security

- Penetration Tester/Ethical Hacker
- Red Team Operator/Red Team Lead
- Offensive Security Engineer
- Adversary Emulation/Purple Team Specialist

Threat Intelligence

- Malware Analyst/Threat Researcher
- Cyber Threat Intelligence (CTI) Analyst – AI Focus
- Fraud/Abuse Detection Analyst (AI-enabled threats)

Security Engineering

- DevSecOps/Secure DevOps Specialist
- Application Security Engineer (LLM Apps/APIs)
- Product Security Engineer/AI Product Security

Defensive Security

- SOC Analyst (Tier 2/3)/Detection Engineer
- Blue Team Engineer/Threat Detection Engineer
- Incident Responder (IR)/DFIR Analyst
- Security Operations Manager (SOC Lead)

AI/ML Engineering

- ML Engineer/Applied AI Engineer
- GenAI Engineer (RAG/Agents)
- AI/LLM Application Developer
- MLOps/AI Platform Engineer

AI Security Architecture

- Secure AI Engineer/AI Security Architect
- LLM Systems Engineer

Offensive AI Security Across *Industries*

C|OASP applies wherever AI is deployed in high-impact environments:

Finance

model security, adversarial testing, regulatory alignment

Healthcare

protect diagnostics, patient data, inference integrity

Government

defend AI in defense and public-sector operations

Manufacturing

secure automation, supply chains, agent workflows

Technology

harden LLM apps, agentic AI, platforms, and integrations

Job Roles Enabled by Certified Offensive AI Security Professional C|OASP Credential

Verifiable skills from C|OASP align with high-impact AI and cyber roles.

C|OASP validates skills to identify, exploit, and defend AI systems, supporting roles such as:

- AI Red Team Specialist/Adversarial AI Engineer
- Offensive Security Engineer (AI/LLM)
- Adversarial AI Security Analyst/AI Threat Hunter
- AI Incident Response Engineer/AI Forensics Analyst
- Secure AI Engineer/AI Security Architect
- MLOps/AIOps Security Specialist
- AI Model Risk/AI Risk & Assurance Specialist
- LLM Systems Engineer
- AI Product Security Manager/AI Security Program Manager
- CTI Analyst (AI Focus)/AI Risk Advisor

Key Features of Certified Offensive AI Security Professional C|OASP Certification

Certified Offensive AI Security Professional

Hands-on Labs First:

30 practical exercises across modules covering prompt injection, model extraction, adversarial attacks, and red vs. blue scenarios.

DCWF-aligned Learning Path:

Labs map to DoD/ NICE DCWF KSAs such as reconnaissance, exploitation, extraction, and threat hunting.

Full Adversarial Life Cycle Coverage:

Progress from recon and prompt exploitation to memory manipulation, supply-chain compromise, and risk reporting.

AI security success isn't defined by models alone. It is defined by the ability to anticipate threats, simulate attacks, validate controls, and harden systems across the AI life cycle.

C|OASP certifies practitioners who can demonstrate offensive AI security skills, emulating adversaries, validating defenses, and leading red-team/blue-team exercises to keep AI resilient, reliable, and auditable.

Certified Offensive AI Security Professional C|OASP *Training and Exam*

Training

Title of the Course: Certified Offensive AI Security Professional (C|OASP)

Version: 1

Duration: 5 days

Prerequisite: 3 Years of Cybersecurity Experience

Delivery Mode:

- **Instructor-led Training (ILT)**
  In-person training where you can collaborate with your peers.
- **iLearn (Asynchronous Online Learning)**
  An asynchronous, self-study experience in a video-streaming format.
- **iWeek (Synchronous Online Learning)**
  A live, online course led by an instructor.

Exam

Exam Title: Certified Offensive AI Security Professional (COASP)

Exam Code: 312-52

Number of Questions: 70

Duration: 6 hours

Availability: ECC Exam Portal

Passing Score: 70–80%

Test Format: Multiple Choice Questions and Performance-Based Questions

Certified Offensive AI Security Professional C|OASP

Where AI offense becomes enterprise defense.

About EC-Council

EC-Council's mission is to build and redefine the cybersecurity profession globally

through education, certification, and workforce development, with an expanded focus on **AI-enabled and AI-secure enterprises**.

Through industry-recognized certifications and training, EC-Council prepares professionals to address challenges across **cybersecurity, AI governance, and the secure adoption of AI technologies**. Its certification portfolio includes programs such as Certified Ethical Hacker (C|EH $^{\text{AI}}$), Certified Penetration Testing Professional (C|PENT $^{\text{AI}}$), Computer Hacking Forensic Investigator (C|HFI), Certified Network Defender (C|ND), Certified SOC Analyst (C|SA), Certified Threat Intelligence Analyst (C|TIA), Certified Incident Handler (E|CIH), and the Certified Chief Information Security Officer (C|CISO).

EC-Council also provides cybersecurity services to some of the largest businesses globally. Trusted by 7 of the Fortune 10, 47 of the Fortune 100, the U.S. Department of Defense (DoD), the intelligence community, NATO, and more than 2,000 universities, colleges, and training companies, with certified professionals in more than 140 countries.

EC-Council is an ANAB ISO/IEC 17024-accredited organization and has recognition under **DoD Directive 8140/8570**, as well as recognition in the U.K. by **GCHQ** and other bodies.

Founded in 2001, EC-Council employs more than 400 individuals worldwide and operates 10 global offices in the U.S., U.K., Malaysia, Singapore, India, and Indonesia, including U.S. offices in Albuquerque, New Mexico, and Tampa, Florida.

Learn more at eccouncil.org.