+33 (0)1 48 10 09 46
formation@ambient-it.net
140 Avenue Jean Lolive 93500 Pantin - France

Updated on August 12, 2026

Sign Up

Offensive AI Expert Certification Training (HTB COAE)

5 days (35 hours)

Overview

The Offensive AI Expert (HTB COAE) Certification Training prepares you for an official certification focused on offensive AI, adversarial attacks, and the evaluation of AI systems under real-world conditions. You'll develop a methodical approach to analyzing, exploiting, and documenting vulnerabilities in machine learning and LLM environments.

This Offensive AI Expert Certification (HTB COAE) training course guides you through preparation for the official Certified Offensive AI Expert (HTB COAE) exam, with a curriculum focused on skills validation and hands-on application. You'll work on the essential concepts required for certification, ranging from understanding attack surfaces to writing a actionable report.

You will learn to identify vulnerabilities related to adversarial machine learning, prompt injection, jailbreaking, the exploitation of LLM outputs, and data privacy issues. The goal is to make you operational in scenarios similar to those encountered in the exam and in red teaming missions.

Throughout the training, you will develop a structured attack methodology, a critical understanding of model behavior, and the ability to prioritize the most relevant attack vectors. You will also learn to organize your findings, assess their impact, and produce a clear report—as expected for an intermediate-to-advanced-level certification.

Finally, this preparation covers the practical requirements of the exam, time management, the logic behind technical demonstrations, and best practices for reporting. You'll thus arrive with a comprehensive understanding of the Certified Offensive AI Expert (HTB COAE) requirements and a solid foundation for passing the certification.

Like all our training courses, this one will introduce you to **the latest stable version** of the technology and its new features.

Objectives

- Understand the scope of the Certified Offensive AI Expert (HTB COAE) certification and its technical requirements.
- Master the main categories of attacks on AI and LLM systems.
- Implement a red teaming approach applied to realistic AI environments.
- Identify, exploit, and document vulnerabilities related to prompt injection, escape, and information leakage.
- Structure a clear, well-reasoned, and actionable technical report in the context of certification.
- Effectively prepare for the format and requirements of the HTB COAE exam.

Target Audience

- Penetration Tester
- Cybersecurity consultant
- Security analyst
- Red teamer
- AI Security Engineer

Prerequisites

- Basic knowledge of offensive cybersecurity and penetration testing methodologies.
- Intermediate proficiency in Python to read, adapt, or automate scripts.
- Understanding of machine learning principles and common uses of LLMs.
- Experience with Linux environments and command-line tools.
- Ability to analyze application behavior and write a technical report.

Technical Requirements

- A recent computer with sufficient RAM to run testing tools and web environments.
- A stable Wi-Fi or Ethernet internet connection, essential for labs and
  .
- Python 3.x installed, with a functional development environment for scripts and automation.
- A modern, up-to-date browser, as well as a code editor such as Visual Studio Code or equivalent.
- A headset if attending remotely, to facilitate communication and demonstrations.

Curriculum for our Offensive AI Expert (HTB COAE) Training Course

[Day 1 - Morning]

Foundations of Offensive AI

- Understand the scope of the HTB COAE certification and the exam structure.
- Review fundamental concepts: Machine Learning, LLM, Transformer architectures, and attack surfaces.
- Distinguish between the traditional application-based approach (OWASP Top 10) and the AI-specific approach
  (OWASP for LLM/ML).
- Map major risks, from the data collection chain to inference APIs.
- Hands-on workshop: getting started with the lab environment, mapping a target AI system, and identifying entry points.

[Day 1 – Afternoon]

Prompt Injection and Jailbreaking

- Examining the mechanics of direct prompt injections.
- Master jailbreaking techniques and safeguard misalignment (RLHF/DPO bypass).
- Explore advanced bypass techniques (encodings, multilingualism, context segmentation, obfuscation).
- Assess the business impact of a manipulated or out-of-bounds response.
- Hands-on workshop: Successfully perform a series of incremental bypasses on a protected chatbot and document the process.

[Day 2 - Morning]

Indirect prompt injection and exploitation of RAGs

- Analyze indirect injection vectors (trapped files, emails, external web sources).
- Understand the vulnerabilities of the RAG (Retrieval-Augmented Generation) architecture.
- Exploit weaknesses in vector databases (vector DB poisoning, embedding manipulation).
- Exfiltrate confidential data dynamically loaded into the context.
- Hands-on workshop: infiltrate a malicious document into a document repository to take control of AI responses.

[Day 2 - Afternoon]

AI Agent Security and Tool Calling

- Examine how autonomous AI agents work (ReAct pattern, Function/Tool Calling).
- Identify the risks of tool hijacking: RCE, SSRF, SQL queries, privilege escalation.
- Manipulate the agent’s decision-making cycle to force the execution of unauthorized actions.
- Assess the security of APIs and plugins connected to models.

- Hands-on workshop: Compromising an AI agent connected to a system environment to execute commands remotely.

[Day 3 - Morning]

Adversarial Machine Learning and Data Tampering

- Understand evasion and poisoning attacks on traditional ML and deep learning models.
- Generate adversarial examples to fool classifiers.
- Identify vulnerabilities in the MLOps pipeline (deserialization, dataset poisoning, dependencies).
- Hands-on workshop: Manipulate the prediction of a risk analysis model using custom-built adversarial inputs.

[Day 3 - Afternoon]

Model and Secret Extraction

- Examine model inversion and membership inference attacks.
- Implement model stealing techniques via repeated API requests.
- Analyze secret leaks: extraction of system prompts, API keys, and training data.
- Exploit conversational memory and caching mechanisms.
- Hands-on workshop: Retrieve a hidden System Prompt and extract access keys hidden in the AI’s memory.

[Day 4 - Morning]

Automated Red Teaming and Offensive Tools

- Using modern AI Red Teaming frameworks (Garak, PyRIT, Promptfoo, etc.).
- Automate prompt fuzzing, jailbreak detection, and security regression testing.
- Integrate AI vulnerability detection into a continuous assessment pipeline.
- Hands-on workshop: Configure and orchestrate an automated Red Teaming campaign on a target application.

[Day 4 - Afternoon]

Defense, Guardrails, and Hardening

- Review current countermeasures (NeMo Guardrails, Llama Guard, filters, sanitization).
- Audit the effectiveness of guardrails and understand their limitations against advanced attacks.

- Develop a remediation strategy: the principle of least privilege, isolation, and output sanitization.
- Hands-on workshop: analyze a system protected by guardrails, identify its vulnerability, and then propose an architectural fix.

[Day 5 - Morning]

Practical exam simulation (Live-Action CTF)

- Real-world scenario in a comprehensive environment integrating multiple components (LLM, RAG, Agent, ML Model).
- Independent application of offensive methodology: reconnaissance, exploitation, and exfiltration.
- Time management, methodical note-taking, and rigorous collection of proof-of-concept (PoC) evidence.
- Hands-on workshop: timed exercise simulating the conditions of the HTB COAE exam lab.

[Day 5 - Afternoon]

Exam methodology, reporting, and post-mortem

- Analyze the requirements of the official HTB COAE certification report.
- Structure a clear, compelling, and business-focused technical and executive audit report.
- Write flawless reproduction steps and define priority recommendations.
- Group debriefing of the simulation and organizational tips for the 7-day exam.
- Hands-on workshop: drafting and finalizing the technical report based on the morning’s simulation.

Target Companies

This training is intended for both individuals and companies—large or small—seeking to train their teams in new, advanced IT technologies or to acquire specific business knowledge or modern methods.

Entry-Level Assessment

The pre-training assessment complies with Qualiopi quality standards. Upon final registration, the learner receives a self-assessment questionnaire that allows us to evaluate their estimated proficiency level across various types of technologies, as well as their expectations and personal goals for the upcoming training, within the limits imposed by the selected format. This questionnaire also allows us to anticipate certain connection or internal security issues within the company (intra-company or virtual classroom) that could pose challenges for monitoring and ensuring the smooth operation of the training session.

Teaching Methods

Hands-On Training: 60% Practical, 40% Theoretical. Training materials distributed in

digital format to all participants.

Organization

The course alternates between theoretical input from the instructor—supported by examples and reflection sessions—and group work.

Assessment

At the end of the session, a multiple-choice quiz is administered to verify that participants have successfully acquired the skills.

Certification

A certificate will be issued to each trainee who has completed the entire training program.