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

Mis à jour le 12/08/2026

S'inscrire

Formation Certification Offensive AI Expert (HTB COAE)

3 jours (21 heures)

Présentation

La formation Certification Offensive AI Expert (HTB COAE) vous prépare à une certification officielle centrée sur l'IA offensive, les attaques adversariales et l'évaluation de systèmes d'IA en conditions réelles. Vous y développez une approche méthodique pour analyser, exploiter et documenter des failles sur des environnements machine learning et LLM.

Cette formation Certification Offensive AI Expert (HTB COAE) vous accompagne dans la préparation de l'examen officiel Certified Offensive AI Expert (HTB COAE), avec un programme orienté validation des compétences et mise en pratique. Vous travaillez les notions essentielles attendues dans un contexte de certification, depuis la compréhension des surfaces d'attaque jusqu'à la rédaction d'un rapport exploitable.

Vous apprendrez à identifier les vulnérabilités liées à l'adversarial machine learning, au prompt injection, au jailbreaking, à l'exploitation des sorties LLM et aux enjeux de confidentialité des données. L'objectif est de vous rendre opérationnel sur des scénarios proches de ceux rencontrés dans l'examen et dans des missions de red teaming.

Au fil de la formation, vous consoliderez une méthode d'attaque structurée, une lecture critique des comportements modèle, ainsi qu'une capacité à prioriser les vecteurs les plus pertinents. Vous saurez également organiser vos constats, qualifier leur impact et produire une restitution claire, attendue dans une certification de niveau intermédiaire à technique élevé.

Cette préparation aborde enfin les exigences pratiques de l'examen, la gestion du temps, la logique de démonstration technique et les bonnes pratiques de reporting. Vous arriverez ainsi avec une vision complète des attendus de Certified Offensive AI Expert (HTB COAE) et une base solide pour réussir la certification.

Comme toutes nos formations, celle-ci vous présentera **la dernière version stable** de la technologie et ses nouveautés.

Objectifs

- Comprendre le périmètre de la certification Certified Offensive AI Expert (HTB COAE) et ses attendus techniques.
- Maîtriser les principales familles d'attaques sur les systèmes IA et LLM.
- Mettre en oeuvre une démarche de red teaming appliquée à des environnements d'IA réalistes.
- Identifier, exploiter et documenter des vulnérabilités liées au prompt injection, à l'évasion et à la fuite d'informations.
- Structurer un rapport technique clair, argumenté et exploitable dans un contexte de certification.
- Se préparer efficacement au format et aux contraintes de l'examen HTB COAE.

Public visé

- Pentester
- Consultant cybersécurité
- Analyste sécurité
- Red teamer
- Ingénieur sécurité IA

Pré-requis

- Connaissances de base en cybersécurité offensive et en méthodologie de test d'intrusion.
- Maîtrise intermédiaire de Python pour lire, adapter ou automatiser des scripts.
- Compréhension des principes de machine learning et des usages courants des LLM.
- Expérience avec les environnements Linux et les outils de ligne de commande.
- Capacité à analyser des comportements applicatifs et à rédiger un compte rendu technique.

Pré-requis techniques

- Un ordinateur récent avec suffisamment de mémoire vive pour exécuter les outils de test et les environnements web.
- Une connexion Internet stable en Wi-Fi ou Ethernet, indispensable pour les labs et les ressources en ligne.
- Python 3.x installé, avec un environnement de travail fonctionnel pour les scripts et automatisations.
- Un navigateur moderne à jour, ainsi qu'un éditeur de code comme Visual Studio Code ou équivalent.
- Un micro-casque si la session est suivie à distance, pour faciliter les échanges et les démonstrations.

Programme de notre formation Offensive AI Expert (HTB COAE)

[Jour 1 - Matin]

Fondations de l’IA offensive

- Comprendre le périmètre de la certification HTB COAE et la structure de l’examen.
- Revoir les notions fondamentales : Machine Learning, LLM, architectures Transformer et surfaces d’attaque.
- Différencier l'approche applicative classique (OWASP Top 10) de l'approche spécifique IA (OWASP for LLM / ML).
- Cartographier les risques majeurs, de la chaîne de collecte de données aux API d'inférence.
- Atelier pratique : prise en main de l'environnement de lab, cartographie d'un système IA cible et identification des points d'entrée.

[Jour 1 - Après-midi]

Prompt injection et jailbreaking

- Étudier la mécanique des prompt injections directes.
- Maîtriser les techniques de jailbreaking et le désalignement des garde-fous (contournement RLHF/DPO).
- Explorer les contournements avancés (encodages, multilinguisme, découpage de contexte, obfuscation).
- Évaluer l’impact métier d’une réponse manipulée ou hors cadre.
- Atelier pratique : réussir une série de contournements graduels sur un assistant conversationnel protégé et documenter la démarche.

[Jour 2 - Matin]

Prompt injection indirecte et exploitation des RAG

- Analyser les vecteurs d'injection indirecte (fichiers piégés, emails, sources web externes).
- Comprendre les vulnérabilités de l'architecture RAG (Retrieval-Augmented Generation).
- Exploiter les faiblesses des bases de données vectorielles (Vector DB poisoning, manipulation d'embeddings).
- Exfiltrer des données confidentielles chargées dynamiquement dans le contexte.
- Atelier pratique : infiltrer un document malveillant dans une base documentaire pour prendre le contrôle des réponses de l'IA.

[Jour 2 - Après-midi]

Sécurité des Agents IA et Tool Calling

- Étudier le fonctionnement des Agents IA autonomes (ReAct pattern, Function/Tool Calling).
- Identifier les risques de détournement d'outils : RCE, SSRF, requêtes SQL, élévation de privilèges.
- Manipuler le cycle de décision de l'agent pour forcer l'exécution d'actions non autorisées.
- Évaluer la sécurité des API et plugins connectés aux modèles.

- Atelier pratique : compromettre un Agent IA connecté à un environnement système pour exécuter des commandes à distance.

[Jour 3 - Matin]

Adversarial Machine Learning et altération de données

- Comprendre les attaques par evasion et poisoning sur les modèles de ML classiques et Deep Learning.
- Générer des exemples adversariaux pour tromper les classifieurs.
- Identifier les vulnérabilités de la chaîne MLOps (désérialisation, empoisonnement de jeux de données, dépendances).
- Atelier pratique : altérer la prédiction d'un modèle d’analyse de risques à l'aide d'entrées adversariales construites sur mesure.

[Jour 3 - Après-midi]

Extraction de modèles et de secrets

- Étudier les attaques d'inversion de modèle et d'inférence d'appartenance.
- Mettre en œuvre des techniques de vol de modèle (Model Stealing) via requêtes API répétées.
- Analyser les fuites de secrets : extraction de System Prompts, clés API et données d’entraînement.
- Exploiter la mémoire conversationnelle et les mécanismes de cache.
- Atelier pratique : récupérer un System Prompt masqué et extraire des clés d'accès cachées dans la mémoire de l'IA.

[Jour 4 - Matin]

Red Teaming automatisé et outillage offensif

- Utiliser les frameworks modernes de Red Teaming IA (Garak, PyRIT, Promptfoo, etc.).
- Automatiser le fuzzing de prompts, la recherche de jailbreaks et les tests de régression de sécurité.
- Intégrer la détection de vulnérabilités IA dans un pipeline d'évaluation continue.
- Atelier pratique : configurer et orchestrer une campagne de Red Teaming automatisée sur une application cible.

[Jour 4 - Après-midi]

Défense, garde-fous et hardening

- Passer en revue les contre-mesures actuelles (NeMo Guardrails, Llama Guard, filtres, sanitisation).
- Auditer l'efficacité des garde-fous et comprendre leurs limites face aux attaques avancées.

- Structurer une stratégie de remédiation : principe du moindre privilège, isolation, Output Sanitization.
- Atelier pratique : analyser un système protégé par des garde-fous, en identifier la faille, puis proposer le correctif d'architecture.

[Jour 5 - Matin]

Simulation d'examen pratique (CTF Grandeur Nature)

- Mise en situation réelle sur un environnement complet intégrant plusieurs briques (LLM, RAG, Agent, Modèle ML).
- Application autonome de la méthodologie offensive : reconnaissance, exploitation, escalade d'impact et exfiltration.
- Gestion du temps, prise de notes méthodique et collecte rigoureuse des preuves de concept (PoC).
- Atelier pratique : épreuve chronométrée simulant les conditions du lab de l'examen HTB COAE.

[Jour 5 - Après-midi]

Méthodologie d'examen, reporting et post-mortem

- Analyser les exigences du rapport officiel de la certification HTB COAE.
- Structurer un rapport d'audit technique et exécutif clair, probant et orienté correction métier.
- Rédiger des étapes de reproduction irréprochables et définir les recommandations prioritaires.
- Débriefing collectif de la simulation et conseils d'organisation pour l'épreuve de 7 jours.
- Atelier pratique : rédaction et finalisation du rapport technique à partir de la simulation du matin.

Sociétés concernées

Cette formation s’adresse à la fois aux particuliers ainsi qu’aux entreprises, petites ou grandes, souhaitant former ses équipes à une nouvelle technologie informatique avancée ou bien à acquérir des connaissances métiers spécifiques ou des méthodes modernes.

Positionnement à l'entrée en formation

Le positionnement à l'entrée en formation respecte les critères qualité Qualiopi. Dès son inscription définitive, l'apprenant reçoit un questionnaire d’auto-évaluation nous permettant d’apprécier son niveau estimé sur différents types de technologies, ses attentes et objectifs personnels quant à la formation à venir, dans les limites imposées par le format sélectionné. Ce questionnaire nous permet également d’anticiper certaines difficultés de connexion ou de sécurité interne en entreprise (intraentreprise ou classe virtuelle) qui pourraient être problématiques pour le suivi et le bon déroulement de la session de formation.

Méthodes pédagogiques

Stage Pratique : 60% Pratique, 40% Théorie. Support de la formation distribué au format

numérique à tous les participants.

Organisation

Le cours alterne les apports théoriques du formateur soutenus par des exemples et des séances de réflexions, et de travail en groupe.

Validation

À la fin de la session, un questionnaire à choix multiples permet de vérifier l’acquisition correcte des compétences.

Sanction

Une attestation sera remise à chaque stagiaire qui aura suivi la totalité de la formation.