AYI NEDJIMI Consultants

Pentest IA : Méthodologie d'Audit des Systèr Artificielle

30 April 2026 • Mis à jour le 30 April 2026 • 47 min de lecture •

Méthodologie pentest IA : OWASP ML Top 10, MITRE ATLAS, prompt inject poisoning. Outils Garak, PromptFoo, ART. Conformité AI Act.

Le **pentest IA** et l'**audit de sécurité des systèmes d'intelligence artificielle** émerg critiques face à la prolifération des modèles de langage (LLM), des systèmes de m des architectures RAG (Retrieval-Augmented Generation) dans les infrastructures l'application des méthodologies offensives traditionnelles aux systèmes d'IA — co de la **prompt injection** avancée à l'extraction de modèles, en passant par l'empois attaques adversariales sur le machine learning, les vulnérabilités des API de servi des modèles (Hugging Face, registres de modèles). Les référentiels **OWASP ML T** **RMF** (Risk Management Framework) fournissent les cadres méthodologiques pou que des outils comme **Garak**, **PromptFoo**, **ART** (Adversarial Robustness Toolbox) d'intrusion spécifiques à l'IA. Avec l'entrée en application de l'**AI Act** européen et l associées, l'audit de sécurité IA n'est plus optionnel — il devient une obligation ré haut risque. Ce guide expert couvre l'ensemble de la méthodologie de pentest LL

fondamentaux théoriques aux techniques d'exploitation avancées, en intégrant les
matière de détection et de défense.

À RETENIR

Points clés de cet article :

- Le **pentest IA** applique les méthodologies offensives aux systèmes d'intelli
  classique, RAG, agents autonomes
- Les référentiels **OWASP ML Top 10**, **MITRE ATLAS** et **NIST AI RMF** structure
  d'évaluation
- La **prompt injection** (directe et indirecte) reste la vulnérabilité la plus critiqu
  production
- L'**extraction de modèle** (model stealing) et l'**inversion de données d'entraî**
  intellectuelle et la confidentialité
- Les outils **Garak**, **PromptFoo**, **ART** et **Counterfit** automatisent les tests de s
- L'**AI Act** européen impose des évaluations de sécurité obligatoires pour les
  partir d'août 2026
- La **supply chain IA** (modèles pré-entraînés, datasets, frameworks ML) cons
  croissante et sous-évaluée

Cadres méthodologiques pour le pentest IA

L'évaluation de sécurité des systèmes d'IA nécessite des cadres méthodologiques
spécifiques au machine learning et aux modèles de langage, au-delà des vulnérab
Trois référentiels majeurs structurent aujourd'hui la discipline : l'OWASP ML Top 10