AYI NEDJIMI Consultants

Red Teaming LLM on-premise : mét

16 mai 2026 • Mis à jour le 17 mai 2026 • 18 min de lecture • 3439 mots •

Découvrez la méthodologie complète du red teaming LLM on-premise : pr
2025, outils garak et PyRIT pour auditer vos modèles IA internes.

À RETENIR

A retenir - Red Teaming LLM on-premise

Le **red teaming LLM on-premise** est devenu une discipline incontournable po
langage en environnement d'entreprise. Les vecteurs d'attaque specifiques
exfiltration de donnees -- necessitent des outils et une methodologie adapte
L OWASP LLM Top 10 2025 structure desormais ces risques et fournit un cad
securite doivent imperativement maitriser garak, PyRIT et promptfoo pour au

Le deploiement massif de modeles de langage (LLM) en environnement on-premis
surface d'attaque radicalement nouvelle que les equ ent imperati
s'impose comme la reponse methodologique structu x risques :

des applications classiques, mais d'explorer les comportements emergents, les fa
intrinseques aux architectures transformer. En 2026, avec l'acceleration des deplo
ou Qwen 2.5 en interne, comprendre comment un attaquant peut detourner, extrai
LLM est devenu une competence fondamentale pour tout RSSI ou equipe red team
complete, des outils eprouves et des cas concrets pour auditer efficacement vos
jusqu'au reporting de remediation. Nous couvrons les techniques d'attaque docum
reference, la taxonomie des jailbreaks et les procedures de verification post-audit

Comprendre la surface d'attaque specifique aux LLM on-premise

Contrairement aux applications web traditionnelles, un LLM deploye en entreprise
qui s'etend bien au-dela des interfaces reseau classiques. Le modele lui-meme --
d'entrainement fine-tune -- constitue une cible a part entiere. Les **vecteurs d'atta**
attaques sur l'entree (prompt manipulation), les attaques sur le contexte (system p
sur les sorties (data exfiltration via generation, model inversion).

Pour un LLM on-premise typique, la surface inclut l'API d'inference (REST ou gRPC
les connecteurs de donnees (RAG, base documentaire, outils externes), et les pipe
Chacune de ces couches merite une analyse specifique et systematique.

Les organisations qui deploient des LLM sans red teaming prealable s'exposent a
confidentielles presentes dans le contexte systeme, contournement des politiques
manipulation des agents IA pour executer des actions non autorisees. Une etude d
d'entreprise testes etaient vulnerables a au moins une forme de prompt injection s

La classification des modeles on-premise par niveau de risque est une premiere e
reseau presente un profil de risque bien different d'un agent autonome avec acces
Cette cartographie conditionne la profondeur et les p ming.

Réponse sous 24h

Devis gratuit →