AYI NEDJIMI Consultants

Indirect Prompt Injection 2026 : Empoisonner u LLM

16 mai 2026 • Mis à jour le 17 mai 2026 • 20 min de lecture • 3762 mots •

L'indirect prompt injection (IPI) injecte une instruction adversariale dans ur document du corpus RAG. ASR > 60% sur LLM 2026 sans defense.

À RETENIR

A retenir — Indirect Prompt Injection

IPI (Greshake et al., 2023) injecte une instruction adversariale dans un doc corpus RAG. ASR 74% sur GPT-5 en 2026.

Vecteurs courants : commentaires HTML, metadata PDF, alt-text image, co base64, white-on-white CSS, ZWS Unicode steganographique.

Impact business : exfiltration de PII, manipulation de reponses chatbot clien propagation par *worm prompts* (CompromptM , 2024).

Defenses 2026 : Spotlighting (Hines et al., Microsoft 2024), StruQ (Chen e
2024), Constitutional Classifiers, isolation prompts/data.

OWASP LLM01 classe IPI comme attaque prioritaire. AI Act annexe IV exige
documentation des tests IPI.

L'**indirect prompt injection** (IPI) est l'attaque la plus pernicieuse de l'ecosysteme L
Contrairement aux jailbreaks frontaux (GCG Adversarial Suffix, Multi-Turn Jailbrea
Crescendo), l'IPI exploite une propriete structurelle des architectures RAG (Retriev
Augmented Generation) : un LLM ne distingue pas, dans son contexte, ce qui est i
utilisateur et ce qui est donnee retrieved. Un attaquant qui controle un document i
RAG d'une entreprise peut injecter des instructions adversariales que le LLM exec
s'il s'agissait du prompt utilisateur. Cet article presente le code Python d'attaque,
de payload (HTML, metadata, steganographie), les defenses 2026 (Spotlighting, S
mapping conformite AI Act / OWASP LLM Top 10. Pour les architectes RAG 2026, l
du statut de risque emergent a celui de menace operationnelle confirmee, exigean
approche defense-in-depth coordonnee avec les regulateurs (CNIL, ENISA, NIS2)

1. Genese et etat de l'art

Le concept d'IPI est formalise par Kai Greshake et al. en avril 2023 dans *Not what
signed up for: Compromising Real-World LLM-Integrated Applications with Indirec
Injection*. La demo est saisissante : un site web piege contient une instruction cach
lorsqu'un assistant IA navigue le site (Bing Chat a l'epoque), prend le controle de l
conversation utilisateur.

Réponse sous 24h

Devis gratuit →