

warning "Mise à jour du backend (octobre 2026)" L'inférence GitHub Models a été retirée le 30 juillet 2026. Les références ci-dessous sont conservées uniquement comme contexte historique pendant la migration de ce lab. Utilisez le Lab 013 ? Backends LLM gratuits pour configurer Foundry Local ou Ollama avant de continuer.

Lab 036 : Défense contre l'injection de prompt et sécurité des agents

Niveau : L300 Parcours : Pro Code Durée : ~40 min ? Coût : GitHub Free

Ce que vous apprendrez

Ce qu'est l'injection de prompt et pourquoi elle est dangereuse dans les systèmes agentiques La différence entre les attaques par injection directe et indirecte Les défenses pratiques : assainissement des entrées, validation des sorties, séparation des privilèges Les garde-fous des appels d'outils ? empêcher les agents d'effectuer des actions destructrices Liste de contrôle de sécurité pour les agents en production

Introduction

Une attaque par injection de prompt manipule un agent IA en intégrant des instructions malveillantes dans le contenu que l'agent traite. Contrairement à l'injection SQL (injecter du code dans une requête), l'injection de prompt injecte des instructions dans le contexte du LLM.

Pourquoi les agents sont particulièrement vulnérables : Ils effectuent des actions (envoyer des e-mails, modifier des fichiers, appeler des API). Un agent manipulé ne produit pas seulement du mauvais texte ? il fait de mauvaises choses.

Prérequis

Python 3.11+ pip install openai pydantic GITHUB_TOKEN configuré

Exercice pratique

Étape 1 : Injection directe ? comprendre l'attaque

[diagram or code sample]

Étape 2 : Injection indirecte ? le problème le plus difficile

L'injection indirecte se produit lorsque des instructions malveillantes sont intégrées dans les données que l'agent lit (pages web, documents, e-mails, résultats de bases de données).

[diagram or code sample]

Étape 3 : Défense 1 ? Assainissement des entrées

[diagram or code sample]

Étape 4 : Défense 2 ? Séparation des privilèges

La défense la plus efficace : les agents ne devraient avoir accès qu'à ce dont ils ont besoin.

[diagram or code sample]

Étape 5 : Défense 3 ? Validation des sorties

[diagram or code sample]

Liste de contrôle de sécurité des agents

? | Défense | Implémentation | ? | Assainissement des entrées | Rechercher les motifs d'injection, supprimer les commentaires HTML | ? | Séparation des privilèges | Les agents n'ont que les permissions nécessaires | ? | Liste blanche des appels d'outils | Valider chaque appel d'outil par rapport au rôle | ? | Sortie structurée | Valider la forme de la sortie de l'agent avant d'agir | ? | Humain dans la boucle | Exiger une confirmation pour les actions irréversibles | ? | Journalisation d'audit | Enregistrer tous les appels d'outils avec utilisateur, horodatage, arguments | ? | Limitation de débit | Empêcher les agents d'effectuer trop d'actions trop rapidement | ? | Séparation des données et des instructions | Utiliser des sections de prompt différentes, marquer le contenu utilisateur | ? | Filtrage de contenu | Utiliser les API Azure Content Safety / Responsible AI | ? |

Secrets à moindre privilège | Les clés API de l'agent ont les permissions minimales nécessaires |

Marquage clair du contenu utilisateur

Une technique simple mais efficace : faire en sorte que le LLM distingue le contenu utilisateur des instructions.

[diagram or code sample]

Prochaines étapes

Évaluez la robustesse de votre agent : ? Lab 035 ? Évaluation des agents Fondamentaux de l'IA responsable : ? Lab 008 ? IA responsable pour les agents

Kit d'accompagnement

Le build du site génère des téléchargements PDF et ZIP stables. Le ZIP contient le guide du lab et tous les fichiers d'accompagnement déclarés.

Télécharger le PDF généré Télécharger le ZIP généré

Supporting file: CHEATSHEET.md

Fichiers d'accompagnement: Lab 036 : Défense contre l'injection de prompt et sécurité des agents

Commencez par le parcours local ou gratuit documenté. Consignez les hypothèses et les preuves de validation. Gardez les services hébergés facultatifs sauf exigence explicite du lab.

Supporting file: README.md

Lab 036 : Défense contre l'injection de prompt et sécurité des agents

Utilisez ces fichiers avec le guide du lab et le kit généré.