

OUTSELL PLAYBOOK

JOURNAL EXTÉRIEUR — DECISION

Injection de prompt : distinguer les sources des instructions

QUESTION

Que peut-il se passer quand un agent de marketing lit une page externe?

PENDANCE DU GUIDE

Un texte externe peut contenir des instructions visant à détourner l'agent. Traitez les contenus récupérés comme des données et contrôlez les actions autorisées en dehors du modèle.

Commencez ici.

Étiquette les entrées externes.

VOTRE SITUATION

Compte ou cohorte

Décision de prendre

État de départ et exclusions

Sources de données, dates et limites

Inconnus à valider

TRAVAILLEZ PAR LE GUIDE

1. Entrées externes de l'étiquette

Preuves ou commentaires

Produit ou décision

Propriétaire et point d'examen

2. Limiter les autorisations de données et d'outils

Preuves ou commentaires

Produit ou décision

Propriétaire et point d'examen

3. Tester une attaque synthétique

Preuves ou commentaires

Produit ou décision

Propriétaire et point d'examen

4. Vérifier le rejet de l'exécution

Preuves ou commentaires

Produit ou décision

Propriétaire et point d'examen

5. Répéter la pause et la récupération

Preuves ou commentaires

Produit ou décision

Propriétaire et point d'examen

COMPARER LES OPTIONS

Texte source

Utile quand : Preuves de recherche publique

Limitation: Peut contenir des instructions contradictoires

Mesures suggérées : Traiter comme des données non fiables

Pourquoi cela correspond à notre situation

Preuves nécessaires avant d'agir

Instructions rapides

Utile quand: comportement de modèle directeur

Limitation: Pas une limite de contrôle d'accès

Mesures suggérées : Ajouter une application externe

Pourquoi cela correspond à notre situation

Preuves nécessaires avant d'agir

Outil restreint

Utile quand : Une lecture ou une écriture limitée

Limitation: Besoins de permissions maintenues

Mesure suggérée : Vérifier les destinations autorisées

Pourquoi cela correspond à notre situation

Preuves nécessaires avant d'agir

Plan de redressement

Utile quand : Une action inattendue se produit

Limitation : Un état non réconcilié peut persister

Mesures suggérées : Laisser tomber et examiner la trace

Pourquoi cela correspond à notre situation

Preuves nécessaires avant d'agir

ENGAGEMENT À UNE PROCHAINE ÉTAPE

Action choisie

Propriétaire responsable

Approbation requise

Mesure de succès et définition

Date de révision

Continuer, réviser ou arrêter les critères

Guide : <https://outsell.com/fr/blogs/prompt-injection-in-marketing-agents/>

Politique en matière d'édition et de preuve : <https://outsell-ai-agency.marketplace-0455.chatgpt.site/editorial-policy/>