

OUTSELL PLAYBOOK

OUTSELL JOURNAL — BESCHLUSSBESCHREIBUNG

Prompt Injection: Quellen und Anweisungen auseinanderhalten

FRAGE

Was kann schief gehen, wenn ein Marketing-Agent eine externe Seite liest?

LEITFAHRZEUGE

Externe Texte können Anweisungen enthalten, die einen Agenten umlenken sollen. Behandeln Sie abgerufene Inhalte als Daten und setzen Sie erlaubte Aktionen außerhalb des Modells durch.

HIER FREIST

Externe Eingaben beschriften.

DEINE LAGE

Account oder Kohorte

Entscheidung zu treffen

Ausgangszustand und Ausschlüsse

Evidenzquellen, Daten und Grenzen

Unbekannt zur Validierung

ARBEITEN DURCH DEN LEITFAHRT

1. Externe Inputs auf Etiketten

Nachweis oder Eingabe

Output oder Entscheidung

Eigentümer und Überprüfungsstelle

2. Beschränken Sie Daten- und Werkzeugberechtigungen

Nachweis oder Eingabe

Output oder Entscheidung

Eigentümer und Überprüfungsstelle

3. Testen Sie einen synthetischen Angriff

Nachweis oder Eingabe**Output oder Entscheidung****Eigentümer und Überprüfungsstelle**

4. Ablehnung der Ausführung

Nachweis oder Eingabe**Output oder Entscheidung****Eigentümer und Überprüfungsstelle**

5. Rehearse Pause und Erholung

Nachweis oder Eingabe**Output oder Entscheidung****Eigentümer und Überprüfungsstelle****VERGLEICH DER OPTIONEN**

Ausgangstext

Nützlich, wenn: Öffentliche Forschungsergebnisse

Einschränkung: Kann kontradiktorische Anweisungen enthalten

Vorgeschlagene Maßnahme: Behandeln Sie als nicht vertrauenswürdige Daten

Warum das zu unserer Situation passt**Vor dem Handeln erforderliche Beweise**

Sofortige Anweisung

Nützlich, wenn: Modellverhalten führen

Einschränkung: Keine Zugangskontrollgrenze

Vorgeschlagene Maßnahmen: Hinzufügen externer Durchsetzung

Warum das zu unserer Situation passt**Vor dem Handeln erforderliche Beweise**

Eingeschränktes Werkzeug

Nützlich, wenn: Ein begrenztes Lesen oder Schreiben

Einschränkung: Benötigt gepflegte Berechtigungen

Vorgeschlagene Maßnahme: Zulässige Zielorte prüfen

Warum das zu unserer Situation passt

Vor dem Handeln erforderliche Beweise

Sanierungsplan

Nützlich, wenn: Eine unerwartete Aktion auftritt

Einschränkung: Unversöhnlicher Zustand kann bestehen bleiben

Empfohlene Aktion: Anhalten und die Spur überprüfen

Warum das zu unserer Situation passt

Vor dem Handeln erforderliche Beweise

ZUSAMMENFASSUNG ZU EINEM ABGEBOTENEN NÄCHSTEN SCHRITT

Ausgewählte Maßnahme

Verantwortlicher Eigentümer

Erforderliche Genehmigung

Erfolgsmessung und -definition

Datum der Überprüfung

Kriterien fortsetzen, überarbeiten oder stoppen

Leitfaden: <https://outsell.com/de/blogs/prompt-injection-in-marketing-agents/>

Veröffentlicher und Evidenzpolitik: <https://outsell.com/de/editorial-policy/>