Bundesamt
für Sicherheit in der
Informationstechnik

Nationales
IT-Lagezentrum
BSI

SCHWACHSTELLE | GEFÄHRDUNG | VORFALL | IT-ASSETS

Indirect Prompt Injections - Intrinsische Schwachstelle in anwendungsintegrierten KI-Sprachmodellen

CSW-Nr. 2023-249034-1032, Version 1.0, 18.07.2023

IT-Bedrohungslage*: 1 / Grau

Achtung: Für die schriftliche und mündliche Weitergabe dieses Dokumentes und der darin enthaltenen Informationen gelten gemäß dem Traffic Light Protokoll (TLP) die folgenden Einschränkungen:

TLP:CLEAR: Unbegrenzte Weitergabe

Abgesehen von urheberrechtlichen Aspekten, die das TLP explizit nicht adressiert, dürfen Informationen der Stufe TLP:CLEAR ohne Einschränkungen frei weitergegeben werden.

Das Dokument ist durch den Empfänger entsprechend den vereinbarten „Ausführungsbestimmungen zum sicheren Informationsaustausch mit TLP“ zu verarbeiten, aufzubewahren, weiterzugeben und zu vernichten. Weitere Informationen zum TLP finden Sie am Ende dieses Dokumentes.

Sachverhalt

Große KI-Sprachmodelle (engl. Large Language Model (LLM)) erfreuen sich zunehmender Beliebtheit und werden beispielsweise eingesetzt, um Textdokumente automatisiert zu verarbeiten und Anwenderinnen sowie Anwendern mittels Chatbots und autonomer Agenten zu assistieren. Hierbei wird die Funktionalität fortlaufend erweitert. So ist es zum Beispiel Chatbots mittlerweile möglich, mittels Plugins Internetseiten oder Dokumente automatisiert auszuwerten sowie auf Programmierumgebungen oder E-Mail-Postfächer zuzugreifen. Bei vielen der antizipierten Anwendungsfälle werden ungeprüfte Daten aus unsicheren Quellen verarbeitet.

In diesem Fall sind LLMs anfällig für sogenannte *Indirect Prompt Injections*: Angreifende können die Daten in diesen Quellen manipulieren und dort unerwünschte Anweisungen für LLMs platzieren. Greifen LLMs auf diese Daten zu, werden die unerwünschten Befehle unter Umständen ausgeführt. Angreifende können dadurch das Verhalten der LLMs gezielt manipulieren. Die potentiell schadhaften Befehle können kodiert oder versteckt sein und sind für Anwenderinnen sowie Anwender unter Umständen nicht erkennbar.

* 1 / Grau: Die IT-Bedrohungslage ist ohne wesentliche Auffälligkeiten auf anhaltend hohem Niveau.
2 / Gelb IT-Bedrohungslage mit verstärkter Beobachtung von Auffälligkeiten unter temporärer Beeinträchtigung des Regelbetriebs.
3 / Orange Die IT-Bedrohungslage ist geschäftskritisch. Massive Beeinträchtigung des Regelbetriebs.
4 / Rot Die IT-Bedrohungslage ist extrem kritisch. Ausfall vieler Dienste, der Regelbetrieb kann nicht aufrecht erhalten werden.

In einfachen Fällen könnte dies zum Beispiel ein Text auf einer Webseite mit Schriftgröße Null oder ein versteckter Text im Transkript eines Videos sein. Darüber hinaus ist es jedoch auch möglich, Anweisungen zu kodieren, sodass diese von LLMs weiterhin problemlos interpretiert werden, von Menschen jedoch nur schwer lesbar sind (z.B. unter Verwendung von ASCII-Code oder ähnlichem). Eine weitere Möglichkeit ist, dass Anfragen von Chatbots durch den Webserver aufgrund anderer Aufrufparameter mit anderen Inhalten beliefert werden als sie menschliche Nutzerinnen und Nutzer durch die Browseranfragen erhalten.

Auch der Hersteller OpenAI weist im Zusammenhang mit der Nutzung von *Plugins* beim Produkt *ChatGPT*auf diese Schwachstelle hin: "However, there are still open research questions. For example, a proof-of-concept exploit illustrates how untrusted data from a tool's output can instruct the model to perform unintended actions." [OAI23].

Welches Risiko aus dem Angriffsvektor folgt, hängt stark von dem konkreten Anwendungsfall und den Einsatzbedingungen des LLMs ab, wie zum Beispiel den Aktionsmöglichkeiten oder Rechten. Es werden Beispiele aufgezählt, um die Risiken für verschiedene Anwendungsfälle einordnen zu können. Alle Beispiele basieren auf konkreten *Proof of Concepts (PoCs)*:

- Verwendung eines LLM zur Zusammenfassung oder Analyse von Texten aus externen Quellen
  - › Angreifende könnten das Ergebnis gezielt manipulieren
- Verwendung eines Chatbots, der auf modifizierte Internet-Seiten zugreift
  - › Ergebnisse von Anfragen könnten gezielt manipuliert werden
  - › Der Chatbot könnte ein unerwünschtes Verhalten aufweisen und beispielsweise rechtlich bedenklich oder unerwünschte Aussagen treffen
  - › Der Chatbot könnte Nutzende dazu motivieren, einen (bösartigen) Link aufzurufen
  - › Der Chatbot könnte versuchen, sensitive Informationen von Nutzenden zu erlangen (z.B. Kreditkarteninformationen)
  - › Angreifende könnten (unbemerkt) sensitive Informationen aus dem Chatverlauf extrahieren, falls beispielsweise die Möglichkeit zum Aufrufen von URLs oder dem Anzeigen externer Bilder existiert
  - › Der Chatbot könnte selbst weitere Plugins aufrufen und damit unerwünschte Aktionen ausführen, wie zum Beispiel:
    - Zugriff auf das E-Mail-Konto, Zusammenfassung der letzten E-Mails und Extraktion der Informationen
    - Veröffentlichung privater Quellcode-Repositories
- Autonomer Agent der lokal in einem Docker Container läuft und auf ein LLM via API zugreift:
  - › Angreifende könnten aus dem Container ausbrechen und root-Rechte auf dem Zielsystem erlangen

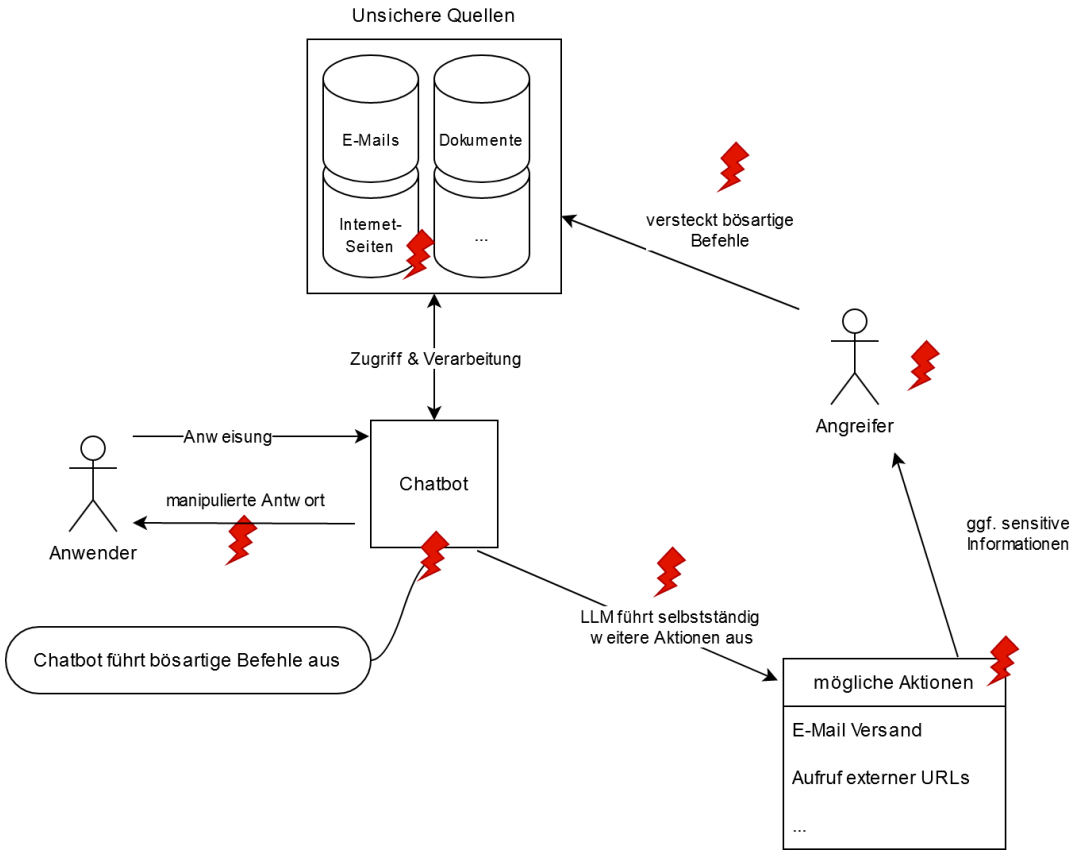

Abb. 1: Schematischer Ablauf einer Indirect Prompt Injection mit moeglichen Folgen am Beispiel eines Chatbots

Nachdem diese neue Schwachstellenklasse im Februar 2023 das erste Mal in der Wissenschaft diskutiert wurde [GRE23], hat das BSI den Angriffsvektor bereits in dem Positionspapier "*Große KI-Sprachmodelle - Chancen und Risikenfür Industrie und Behörden"* thematisiert [BSI23]. Da in den letzten Monaten vermehrt die Darstellung und Diskussion konkreter PoCs zur Ausnutzung der Schwachstelle im Internet beobachtet werden konnten und gleichzeitig die Integration von Sprachmodellen in Anwendungen rasant voranschreitet, sensibilisiert das BSI mit dieser Meldung noch einmal verstärkt für diese neue Schwachstellenklasse.

Bewertung

Die Risiken durch Indirect Prompt Injections sind ernst zu nehmen und entstehen bei der Verarbeitung von Informationen aus unsicheren Quellen durch LLMs. Die genannten Beispiele sind glaubwürdig und wurden durch das BSI zum Teil selbst nachvollzogen. Die Auswirkungen der Schwachstelle sind abhängig vom konkreten Einsatzszenario und den Aktionsmöglichkeiten (bzw. aktivierten Plugins) des LLMs, wie zum Beispiel dem Zugriff auf sensitive Daten. Je nach Szenario können die Auswirkungen eines Angriffs beachtlich sein. Mit welchem Aufwand die Schwachstelle auszunutzen ist und welche Auswirkungen bestehen, kann jedoch nur im konkreten Einzelfall im Rahmen einer systematischen Risikoanalyse eingeschätzt werden.

Texte können in der menschlichen Kommunikation sowohl Informationen übermitteln, als auch Befehle erteilen. Diese Ambiguität wird nun auch in die IT-Sphäre übertragen: Auch bei LLMs existiert keine klare Trennung zwischen Daten und Anweisungen. Da dies eine intrinsische Schwachstelle der derzeitigen Technologie ist, sind Angriffe dieser Art grundsätzlich schwierig zu verhindern. Derzeit ist keine zuverlässige und nachhaltig sichere Mitigationsmaßnahme bekannt, die nicht auch die Funktionalität deutlich einschränkt. Die möglichen Auswirkungen der Schwachwestelle werden verstärkt, wenn das LLM als (teil)autonomes System eingesetzt wird, welches selbständig folgenreiche Aktionen ausführen kann.

Anwendende haben in der Regel keine Möglichkeit, einen solchen Angriff durch Inspektion der Quellen selbst zu erkennen, da die Befehle sowohl versteckt als auch kodiert sein können.

Maßnahmen

Bei der Integration von LLMs in Anwendungen sollte eine systematische Risikoanalyse durchgeführt werden, bei der das Risiko durch Indirect Prompt Injections explizit bewertet wird.

Das Risiko kann verringert werden, indem der Zugriff auf unsichere Quellen ausgeschlossen wird oder vor der Ausführung von potentiell kritischen Aktionen des LLMs eine menschliche Kontrolle und Autorisierung erfolgt. Dies ist auch deshalb ratsam, da LLMs derzeit auch ohne Angriffe halluzinieren und falsche Entscheidungen treffen können. Um die Auswirkungen eines möglichen Angriffs zu reduzieren, können Aktionen unter Umständen auch so eingeschränkt werden, dass diese reversibel sind oder in einer abgetrennten Umgebung ("Sandbox") ausgeführt werden. In jedem Fall sollten die möglichen Aktionen (bzw. Plugins) von LLMs auf ein für den Anwendungsfall benötigtes Minimum beschränkt werden. Das Durchführen von zielgerichteten Penetrationstests (Red Teaming) kann helfen, die Risiken für ein konkretes Einsatzszenario besser bewerten zu können.

Auch der Hersteller OpenAI zählt im Kontext der Nutzung von Plugins in ChatGPT folgende Gegenmaßnahmen auf: "Developers can protect their applications by only consuming information from trusted tools and by including user confirmation steps before performing actions with real-world impact,such as sending an email, posting online, or making a purchase." [OAI23]

Bekannte Angriffe können von den Betreibern und Herstellern durch Filter blockiert werden. Allerdings ist es schwierig, Variationen zu erkennen. Derzeit werden Mitigationsmaßnahmen von Forschenden, Entwickelnden und Herstellern diskutiert sowie erprobt, um die Ausnutzung der Schwachstelle zu erschweren, wie zum Beispiel die Filterung und Validierung von Eingaben oder das Einführen von Rollen für Chatbots, um Anweisungen und Informationen besser trennen zu können [OWA23]. Insgesamt ist jedoch zu betonen, dass es sich um eine relativ neue Art von Schwachstelle handelt und in diesem Bereich derzeit noch keine Security-Best-Practices existieren.

Es ist wichtig, Anwenderinnen und Anwender sowie Entwicklerinnen und Entwickler von LLMs für mögliche Risiken bzw. Limitierungen der Technologie zu sensibilisieren. Ein Grund ist, dass ein manipulierter Chatbot, ähnlich wie bei einem Social Engineering Angriff, unter Umständen glaubhaft argumentieren kann, warum eine (schadhafte) Aktionen unbedingt autorisiert werden muss.

Die BSI-Publikation "Große KI-Sprachmodelle - Chancen und Risiken für Industrie und Behörden" bietet einen kompakten Überblick der Chancen und Risiken von LLMs [BSI23].

Links

[OAI23] https://openai.com/blog/function-calling-and-other-api-updates

[GRE23] https://arxiv.org/abs/2302.12173

[BSI23] https://www.bsi.bund.de/SharedDocs/Downloads/DE/BSI/KI/Grosse_KI_Sprachmodelle.html

[OWA23] https://owasp.org/www-project-top-10-for-large-language-model-applications/assets/PDF/OWASP-Top-10-for-LLMs-2023-v05.pdf

Anlagen

Kontakt

Bitte wenden Sie sich bei allen Rückfragen zu diesem Dokument an denjenigen Kontakt, der Ihnen das Dokument zugesendet hat. Dadurch bleibt der Informationsfluss kanalisiert. Die Single Points of Contact (SPOCs), welche das Dokument direkt vom Nationalen IT-Lagezentrum des BSI erhalten haben, können sich direkt an die bekannten Kontaktdaten des Nationalen IT-Lagezentrums im BSI wenden.

Erklärungen zum Traffic Light Protokoll (TLP)

Dieses Dokument und die darin enthaltenen Informationen sind gemäß dem TLP eingestuft:

1) Was ist das Traffic Light Protokoll?

Das vom BSI verwendete TLP basiert auf der Definition der TLP Version 2.0 des „Forum of Incident Response and Security Team“ (FIRST). Es dient der Schaffung von Vertrauen in Bezug auf den Schutz ausgetauschter Informationen durch Regelungen der Weitergabe. Eine unbefugte Weitergabe kann eine Verletzung der Vertraulichkeit, eine Rufschädigung, eine Beeinträchtigung der Geschäftstätigkeit oder datenschutzrechtliche Belange zur Folge haben. Im Zweifelsfall ist immer in Absprache mit dem Informationsersteller zu handeln.

2) Welche Einstufungen existieren?

- **TLP:CLEAR: Unbegrenzte Weitergabe**
  Abgesehen von urheberrechtlichen Aspekten dürfen Informationen der Stufe TLP:CLEAR ohne Einschränkungen frei weitergegeben werden.

- **TLP:GREEN: Organisationsübergreifende Weitergabe**
  Informationen dieser Stufe dürfen innerhalb der Organisationen und an deren Partner weitergegeben werden. Die Informationen dürfen jedoch nicht veröffentlicht werden. Eine Weitergabe von den Partnerorganisationen an weitere Personen oder Organisationen ist solange zulässig, wie diese weiteren Empfänger derselben Nutzergruppe (bspw. Angehörige der Cybersecurity-Community) angehören.

- **TLP:AMBER: Eingeschränkte interne und organisationsübergreifende Weitergabe**
  Der Empfänger darf die Informationen, welche als TLP:AMBER gekennzeichnet sind, an seine Partner weitergeben, soweit diese die Informationen zur Schadensreduktion oder dem eigenen Schutz benötigen. Eine Weitergabe von den Partnern an Dritte ist nicht erlaubt und auch innerhalb der Partnerorganisationen gilt das Prinzip „Kenntnis nur, wenn nötig“. Der Informationsersteller kann weitergehende oder zusätzliche Einschränkungen der Informationsweitergabe festlegen. Diese müssen eingehalten werden.

  TLP:AMBER+STRICT: Eingeschränkte interne Weitergabe
  Die Einstufung von Informationen als TLP:AMBER+STRICT beschränkt die Weitergabe ausschließlich auf die Organisation des Empfängers. Jegliche Weitergabe darüber hinaus ist untersagt. Es gilt „Kenntnis nur, wenn nötig“. Der Informationsersteller kann weitergehende oder zusätzliche Einschränkungen der Informationsweitergabe festlegen. Diese müssen eingehalten werden.

- **TLP:RED: Persönlich, nur für benannte Empfänger**
  Informationen dieser Stufe sind auf den Kreis der Anwesenden in einer Besprechung oder Video-/Audiokonferenz bzw. auf die direkten Empfänger bei schriftlicher Korrespondenz beschränkt. Eine Weitergabe ist untersagt. TLP:RED eingestufte Informationen sollten möglichst mündlich oder persönlich übergeben werden.

3) Was mache ich, wenn ich das Dokument an jemanden außerhalb des im TLP vorgegebenen Informationsverbundes weitergeben will?

Sollte eine Weitergabe an einen nicht durch die Einstufung genehmigten Empfängerkreis notwendig werden, so ist diese vor einer eventuellen Weitergabe durch den Informationsersteller nachvollziehbar zu genehmigen. Bei ausnahmsweiser Weitergabe im Rahmen einer bestehenden gesetzlichen Verpflichtung ist der Informationsersteller – nach Möglichkeit vorab – zu informieren.

4) Was passiert, wenn ich die Einstufung nicht beachte?

Bei Verstoß gegen die Regeln zur Weitergabe von Informationen erhält der Verpflichtete zukünftig nur noch TLP:CLEAR eingestufte Informationen aus dem Kreis der Verpflichteten.

Hinweis zu Upload-, Prüf- und Übersetzungsdiensten

TLP-eingestufte Dokumente (außer TLP:CLEAR) dürfen nicht auf Plattformen Dritter (wie Virustotal, Übersetzer, etc.) hochgeladen werden, da die Dokumente dort ggf. Dritten zugänglich gemacht werden.