Sicherheit im Fokus

Mitte November 2024 wurden von Open Web Application Security Project (OWASP) die Top 10 Sicherheitsrisiken für KI-Sprachmodelle 2025 herausgegeben. Diese Risiken adressieren bekannte, aber auch neue Punkte, die es einzuhalten gilt. Werden bereits im Vorfeld Massnahmen ergriffen, können diese Modelle manipulationssicher betrieben werden.

Bild: Pixabay

Die Sicherheit grosser KI-Modelle ist eine kontinuierliche Herausforderung, die strategische Planung und konsequente Umsetzung erfordert.

Das OWASP ist eine gemeinnützige Initiative, die sich der Sicherheit von Software widmet. Bekannt sind die verschiedenen Top-10-Schwachstellenlisten, die zum Standard bei jeder Web-Entwicklung gehören und ein Verständnis der kritischsten Sicherheitslücken in Anwendungen ermöglichen. Die erste dieser Listen wurde 2003 veröffentlicht und danach regelmässig aktualisiert. Es folgten Mobile Security, API Security und Cloud-Native Application Security. Am 1. August 2023 wurde die Initiative «OWASP Top 10 for LLM Applications» veröffentlicht und durch eine internationale Gruppe von 500 Expertinnen und Experten in kurzer Zeit die Version 1.0 erstellt. Im November 2024 folgte die aktuelle Anpassung/Erweiterung 2025. LLM steht für Large Language Models und meint damit Sprachmodelle, die zur Textgenerierung genutzt werden. Die LLM Top 10 beschreiben jede der Schwachstellen, gefolgt von Beispielen, typischen Angriffsszenarien und empfohlenen Präventionsmassnahmen. Ergänzt werden Quellen für die weitere Beschäftigung mit dem jeweiligen Thema.

ZUM AUTOR

Andreas Wisler, Dipl. Ing FH
goSecurity AG
Schulstrasse 11
CH-8542 Wiesendangen

T +41 (0)52 511 37 37
www.goSecurity.ch
wisler@gosecurity.ch

LLM01 Prompt Injection

Eine Prompt-Injection-Schwachstelle tritt dann auf, wenn Eingabeaufforderungen das Verhalten oder die Ausgabe des LLM auf unbeabsichtigte Weise verändern. Diese Eingaben können sich auf das Modell auswirken, auch wenn sie für den Menschen nicht wahrnehmbar sind.

Daher müssen Eingabeaufforderungen nicht für den Menschen sichtbar/lesbar sein, solange der Inhalt vom Modell analysiert wird. Angriffe wie «Data Poisoning» können dazu führen, dass Modelle falsche oder unerwünschte Ergebnisse liefern.

Lösung
- Implementierung von robusten Validierungsmechanismen.
- Regelmässige Überprüfung der Trainingsdaten.
- Für risikoreiche Aktionen die Zustimmung von Menschen verlangen.

LLM02 Sensitive Information Disclosure

Sensible Informationen können sowohl die LLM als auch ihren Kontext betreffen. Dazu gehören personenbezogene Daten, Finanzdaten, Gesundheitsinformationen, vertrauliche Geschäfts- oder Sicherheitsdaten. LLMs, insbesondere wenn sie in Anwendungen eingebettet sind, bergen das Risiko, dass durch ihre Ausgabe sensible Daten oder vertrauliche Details offengelegt werden. Dies kann zu unbefugtem Datenzugriff, Verletzungen der Privatsphäre und Verstössen gegen das geistige Eigentum führen.

Um dieses Risiko zu verringern, sollten LLM-Anwendungen eine angemessene Datenbereinigung durchführen, um zu verhindern, dass Benutzerdaten in das Trainingsmodell gelangen. Die Eigentümer der Anwendungen sollten ausserdem klare Nutzungsbedingungen bereitstellen, die es den Benutzern ermöglichen, die Aufnahme ihrer Daten in das Trainingsmodell abzulehnen.

Lösung
- Regelmässige Datenbereinigung
- Eingeschränkte Zugriffskontrolle.
- Training der Benutzer.

LLM03 Supply Chain

LLM-Lieferketten sind anfällig für verschiedene Schwachstellen, die sich auf die Integrität von Trainingsdaten, Modellen und Plattformen auswirken können. Diese Risiken können zu verzerrten Ergebnissen, Sicherheitsverletzungen oder Systemausfällen führen.

Während sich herkömmliche Software-Schwachstellen auf Probleme wie Codefehler und System-Abhängigkeiten konzentrieren, erstrecken sich diese Risiken auch auf vorab trainierte Modelle und Daten von Drittanbietern.

Lösung
- Lieferanten und Datenquellen sorgfältig prüfen.
- Nur Modelle aus überprüfbaren Quellen verwenden.
- Führen eines Inventars der eingesetzten Komponenten.

LLM04 Data and Model Poisoning

Eine Verfälschung der Daten tritt auf, wenn diese vor dem Training, der Feinabstimmung oder der Einbettung in ein System manipuliert werden, um Schwachstellen, Hintertüren oder Verzerrungen einzuführen. Diese Manipulation kann die Sicherheit, Leistung oder das ethische Verhalten des Modells beeinträchtigen und zu schädlichen Ergebnissen oder eingeschränkten Fähigkeiten führen. KI-Modelle können damit für schädliche Zwecke missbraucht werden, wie zum Beispiel zur Erstellung von Fake News, Deepfakes oder schädlichem Code.

Lösung
- Implementierung von Missbrauchserkennungsmechanismen.
- Begrenzung der Modellfähigkeiten durch regelbasierte Filter
- Sensibilisierung der Nutzer für ethische Anwendungen.

LLM05 Improper Output Handling

Unsachgemässe Handhabung von Ausgaben bezieht sich speziell auf unzureichende Validierung, Bereinigung und Handhabung der von einem LLM generierten Ausgaben, bevor sie an nachgelagerte Komponenten und Systeme weitergeleitet werden. Damit ist das Modell anfällig für indirekte Prompt-Injection-Angriffe, die es einem Angreifer ermöglichen könnten, privilegierten Zugriff auf die Umgebung eines Zielbenutzers zu erlangen.

Lösung
- Kein Vertrauen in das Modell, jede Eingabe muss validiert werden.
- Keine direkte Code-Ausgabe.
- Implementieren von Protokollierungs- und Überwachungssystemen.

LLM06 Excessive Agency

Einem LLM-basierten System wird von seinem Entwickler oft ein gewisses Mass an Handlungsfähigkeit eingeräumt – die Fähigkeit, Funktionen aufzurufen oder über Erweiterungen (von verschiedenen Anbietern manchmal als Tools, Skills oder Plugins be-

zeichnet) mit anderen Systemen zu kommunizieren, um auf eine Eingabe hin Aktionen auszuführen. Agentenbasierte Systeme rufen in der Regel wiederholt eine LLM auf und verwenden dabei die Ausgabe früherer Aufrufe, um nachfolgende Aufrufe zu begründen und zu steuern. Die Sprachmodelle können sich dabei nicht selbst kontrollieren oder einschränken.

Excessive Agency ist die Schwachstelle, die es ermöglicht, schädliche Aktionen als Reaktion auf unerwartete, mehrdeutige oder manipulierte Ausgaben einer LLM auszuführen, unabhängig davon, was die Fehlfunktion der LLM verursacht.

Lösung
- LLM-Agenten auf das erforderliche Minimum einschränken.
- Funktionen der LLM-Erweiterungen begrenzen (zum Beispiel bei Zugriff auf ein Mail-Postfach, nur Lesen, kein Schreiben).
- Keine offenen Erweiterungen zulassen.

LLM07 System Prompt Leakage

Die Anfälligkeit für Systemaufforderungen in LLMs bezieht sich auf das Risiko, dass die Systemaufforderungen oder Anweisungen, die zur Steuerung des Modellverhaltens verwendet werden, auch sensible Informationen enthalten können, die nicht offengelegt werden sollten.

Wenn diese Informationen entdeckt werden, können sie zur Erleichterung anderer Angriffe verwendet werden. Systemeingabeaufforderung dürfen weder als Geheimnis betrachtet noch als Sicherheitskontrolle verwendet werden. Dementsprechend sollten sensible Informationen wie Anmeldedaten, Passwörter, Verbindungen zu Datenbanken usw. nicht in der Eingabeaufforderung enthalten sein.

Selbst wenn der genaue Wortlaut nicht offengelegt wird, können Angreifer, die mit dem System interagieren, mit ziemlicher Sicherheit viele der Schutzmassnahmen ermitteln.

Lösung
- Sensible Daten von Systemaufforderungen trennen.
- Sicherstellen, dass Sicherheitskontrollen unabhängig vom LLM durchgesetzt werden.
- Nicht auf die Systemaufforderungen verlassen.

LLM08 Vector and Embedding Weaknesses

Schwachstellen bei Vektoren und Einbettungen in Systeme stellen erhebliche Sicherheitsrisiken dar, die Retrieval Augmented Generation (RAG) verwenden. Bei RAG-Modellen werden nicht nur die eigenen Quellen abgefragt, sondern auch Datenbanken und Webseiten. Schwachstellen bei der Generierung, Speicherung oder Abfrage können durch böswillige Handlungen (absichtlich oder unabsichtlich) ausgenutzt werden, um schädliche Inhalte einzuschleusen, Modellausgaben zu manipulieren oder auf sensible Informationen zuzugreifen.

Lösung
- Umfassende Berechtigung und Zugriffskontrolle.
- Alle Daten und Quellen müssen validiert werden
- Nutzung von Loggin.g-Systemen zur Nachverfolgung von Anomalien.

LLM09 Misinformation

Fehlinformationen treten auf, wenn LLMs falsche oder irreführende Informationen produzieren, die glaubwürdig erscheinen. Diese Schwachstelle kann zu Sicherheitsverletzungen, Rufschädigung und rechtlicher Haftung führen. Eine der Hauptursachen für Fehlinformationen sind Halluzinationen – wenn das LLM Inhalte generiert, die zwar korrekt erscheinen, aber erfunden sind. Halluzinationen treten auf, wenn LLMs Lücken in ihren Trainingsdaten mit statistischen Mustern füllen, ohne den Inhalt wirklich zu verstehen.

Infolgedessen kann das Modell Antworten liefern, die zwar korrekt klingen, aber völlig unbegründet sind. Halluzinationen sind zwar eine Hauptquelle für Fehlinformationen, aber nicht die einzige Ursache; auch durch die Trainingsdaten eingeführte Vorurteile und unvollständige Informationen können dazu beitragen. Ein damit zusammenhängendes Problem ist das übermässige Vertrauen.

Lösung
- Antworten mit vertrauenswürdigen Quellen abgleichen.
- Nutzung von automatischen Validierungsmechanismen.
- Schulung und Sensibilisierung der Benutzer.

LLM10 Unbounded Consumption

Unbounded Consumption bezieht sich auf den Prozess, bei dem ein LLM basierend auf Eingabeabfragen oder -aufforderungen Ausgaben generiert. Angriffe, die darauf abzielen, den Dienst zu stören (Denial-of-Service, kurz DoS), hohe Kosten zu erzeugen oder sogar geistiges Eigentum zu stehlen, indem das Verhalten eines Modells geklont wird. Dies trifft dann ein, wenn es Benutzern möglich ist, übermässige und unkontrollierte Schlussfolgerungen zu ziehen. Die hohen Rechenanforderungen von LLMs, insbesondere in Cloud-Umgebungen, sind ein hohes Risiko für den sicheren Betrieb des Systems.

Lösung
- Setzen von klaren Begrenzungen von Ressourcen und Abfragen.
- Festlegen von Timeouts oder Drosselung der Verarbeitung.
- Einfügen von Wasserzeichen, um die unbefugte Nutzung zu erkennen.

Fazit

Die Sicherheit grosser KI-Modelle ist eine kontinuierliche Herausforderung, die strategische Planung und konsequente Umsetzung erfordert. Die OWASP Top 10 für LLMs bieten einen Leitfaden, um Sicherheitsrisiken zu identifizieren und zu minimieren. Die vorgeschlagenen Präventionsstrategien sind praxisnah und unterstützen Entwickler, Sicherheitsfachleute und Unternehmen dabei, diese Risiken zu minimieren. Durch eine Kombination aus technologischen, organisatorischen und proaktiven Massnahmen können Entwickler und Unternehmen sicherstellen, dass ihre KI-Systeme sicher und verantwortungsvoll eingesetzt werden.

Die aktuelle Version kann kostenlos mittels QR-Code heruntergeladen werden.