KI-Sprachmodelle und Tücken

ChatGPT und die OWASP-Rangfolge der Sicherheitsgefährdungen

(BS/Oliver Wege) Derzeit gibt es einen großen Hype um ChatGPT, dabei ist dies nur die bekannteste Variante der KI-Sprachmodelle (LLMs, Large Language Models). Auch andere Tech-Konzerne investieren in diese neue Technik. Dazu wurde von OWASP aktuell eine Top-Ten-Liste der Schwachstellen für die KI-Sprachmodelle veröffentlicht.

Neben ChatGPT von OpenAI hat Microsoft zwischenzeitlich das KI-Programm in seine Suchmaschine Bing integriert (MS BingAI). Aber auch die anderen US-Tech-Konzerne unternehmen große Anstrengungen in diese Richtung. So arbeitet Google an seinem KI-Programm „Bard" und Meta an „Llamla2", einer Open-Source-Variante auf der Azure Cloud von Microsoft. Apple, das mit seiner „Siri"-Software zur Erkennung und Verarbeitung von natürlich gesprochener Sprache lange Zeit führend war, will mit dem sogenannten „AppleGPT" nun auch wieder aufholen. Selbst *Elon Musk* will mit einem neuen Unternehmen „xAI" und der Software „TruthAI" in den Markt einsteigen.

Auch bisher nicht so bekannte Namen tauchen hier auf. Anthropic A.I. ist ein amerikanisches Start-up und gemeinnütziges Unternehmen, das von ehemaligen Mitgliedern von OpenAI gegründet wurde. Bis Juli 2023 hatte Anthropic 1,5 Mrd. US-Dollar an Finanzmitteln aufgebracht. Google investierte in das Unternehmen 300 Mio. US-Dollar für einen Anteil von 10 Prozent, wobei Anthropic Rechenressourcen von Google Cloud nutzt. Inflection AI ist ein weiteres solches Start-up, das mit vornehmlich Nvinda-Prozessoren einen Supercomputer baut. Darauf trainiert die unter anderem von *Bill Gates, Eric Schmidt* und Nvidia finanzierte Firma ihren KI-Chatbot „Pi". Inflection AI kooperiert dabei mit der Azure Cloud von Microsoft, Nvidia und dem US-Cloud-Dienstleister CoreWeave und hat bis Ende Juni 1,3 Milliarden US-Dollar an frischem Kapital eingesammelt.

Zum Vergleich: In Deutschland will die Bundesregierung bis 2025 insgesamt fünf Milliarden Euro für die Umsetzung der KI-Strategie bereitstellen, die – neben dem Sprachmodell-Bereich – weitere KI-Bereiche in insgesamt zwölf Handlungsfeldern umfasst.

Dabei ist das GPT-Grundprinzip immer das Gleiche, kurz gesagt: GPT-Programme formulieren einen Satz, indem sie Wort für Wort abschätzen, wie er weitergehen könnte. Ein Nachteil des Prinzips ist, dass das Programm kein Verständnis für die Inhalte hat. Deshalb kann es auch überzeugend Informationen ausgeben, die völlig falsch sind, es „halluziniert". Dieses Problem ist besonders aus Datenschutzsicht relevant, da solche falschen Informationen aus der KI quasi nicht mehr weg zu bekommen sind.

Intrinsische Schwachstellen nutzen

Weil es keine klare Trennung zwischen Daten und Anweisungen gibt, werden neue sogenannte Prompt-Injection-Angriffe möglich. Im Einzelfall ist es über sogenannte Prompts möglich, die KI zu bewegen, Dinge auszugeben, die nicht vorgesehen sind. Die Programmierer geben zwar Regeln mit, um kritische Fragen abzublocken, aber durch geschickte Fragekonstruktionen und Anweisungen kann man diese umgehen. „Da dies eine intrinsische Schwachstelle der derzeitigen Technologie ist, sind Angriffe dieser Art grundsätzlich schwierig zu verhindern", so das Bundesamt für Sicherheit in der Informationstechnik (BSI). OWASP (Open Worldwide Application Security Project), bekannt für die Auflistung der Top-Ten-Schwachstellen von Web-Applikationen, hat nun auch eine Top Ten für die KI-Sprachmodelle veröffentlicht. Prompt-Injection-Angriffe wurden auf Platz eins gesetzt, und zwar in Korrespondenz mit Platz sechs „Offenlegen sensibler Informationen", wenn sensible Daten über solche Prompts ausgegeben werden. Auf Platz zwei folgt „Unsicheres Output-Handling" (man denke nur an Cross-Site-Scripting etc.) und auf Platz drei das „Vergiften der Trainingsdaten", um die KI zu falschen Aussagen zu bewegen. Platz vier und fünf sowie sieben und acht sind klassische Sicherheitsthemen jeder Software wie Denial-of-Service-Angriffe, mögliche Schwachstellen in der Software-Lieferkette, unsichere Plugins auch außerhalb der KI, um entfernten Programmcode auszuführen (Remote Code Execution) sowie zu viele Berechtigungen eines Chatbots. Erst auf Platz neun folgt das „übermäßige Vertrauen" in die KI-Ausgaben im Zusammenhang mit dem Fake-News-Problem. Platz zehn wird dann vom Kopieren und damit Stehlen ganzer Modelle eingenommen, wobei dem Unternehmen wirtschaftliche Verluste entstehen können. Nicht genannt wird dagegen die Verwendung der GPTs zu automatisierten Phishing- und Malware-Attacken.

Kriminelle GPT-Varianten

Allerdings gibt es zwischenzeitlich auch schon kriminelle GPT-Varianten. „Fraud GPT" kann Phishing-Mails schreiben, Cracking-Tools entwickeln und hilft, Opfer zu finden, die besonders leicht zu betrügen sind. FraudGPT sieht dabei aus wie eine dunkle Version des bekannten ChatGPT, wobei die Beschränkungen beispielsweise nach einem Text für betrügerische SMS oder Mails entfernt wurden. WormGPT" ist ein weiterer Chatbot, der für kriminelle Zwecke trainiert wurde. Er soll auf einer älteren Open-Source-Variante von GPT-3 (GPT-J) basieren, die Barrieren eingebaut hatte, um nicht für kriminelle Zwecke ausgenutzt zu werden. Dieses wurde bei WormGPT offensichtlich ebenfalls ausgehebelt.