Bundesamt
für Sicherheit in der
Informationstechnik

Deutschland
Digital•Sicher•BSI•

Große KI-Sprachmodelle

Chancen und Risiken für Industrie und Behörden

Änderungshistorie

Version	*Datum*	*Name*	*Beschreibung*
1.0	03.05.2023	TK24	Erstveröffentlichung

Bundesamt für Sicherheit in der Informationstechnik
Postfach 20 03 63
53133 Bonn
E-Mail: ki-kontakt@bsi.bund.de
Internet: https://www.bsi.bund.de

Executive Summary

Große KI-Sprachmodelle sind Computerprogramme, die in der Lage sind, natürliche Sprache in geschriebener Form automatisiert zu verarbeiten. Potenziell können solche Modelle in einer Vielzahl von Anwendungsfällen, in denen Text verarbeitet werden soll, genutzt werden und stellen somit eine Chance für die Digitalisierung dar. Andererseits birgt die Verwendung von großen KI-Sprachmodellen neuartige IT-Sicherheitsrisiken und verstärkt das Bedrohungspotenzial einiger bekannter IT-Sicherheitsbedrohungen. Dazu zählt insbesondere das Missbrauchspotenzial, das von solchen Modellen durch Generierung von Spam-/ Phishing-Mails oder Schadsoftware ausgeht.

Als Reaktion auf diese Bedrohungspotenziale sollten Unternehmen oder Behörden vor der Integration von großen KI-Sprachmodellen in ihre Arbeitsabläufe eine Risikoanalyse für die Verwendung in ihrem konkreten Anwendungsfall durchführen. Daneben sollten sie Missbrauchsszenarien dahingehend evaluieren, ob diese für ihre Arbeitsabläufe eine Gefahr darstellen. Darauf aufbauend können existierende Sicherheitsmaßnahmen angepasst und gegebenenfalls neue Maßnahmen ergriffen werden sowie Nutzende über die potenziellen Gefahren aufgeklärt werden.

Inhalt

1 Einleitung 5
2 Hintergrund und Einordnung von LLMs 7
2.1 Fähigkeiten 7
2.2 Anwendungsgebiete 7
2.3 Erklärbarkeit 8
3 Chancen und Risiken von LLMs 9
3.1 Chancen für die IT-Sicherheit 9
3.2 Risiken bei der Nutzung von LLMs und Gegenmaßnahmen 10
3.2.1 Risiken 10
3.2.2 Gegenmaßnahmen 11
3.3 Missbrauchsszenarien und Gegenmaßnahmen 12
3.3.1 Missbrauchsszenarien 12
3.3.2 Gegenmaßnahmen 13
3.4 Risiken und Herausforderungen bei der Entwicklung sicherer LLMs 15
3.4.1 Datenqualität bei der Auswahl von Trainingsdaten 15
3.4.2 Angriffe auf LLMs und spezifische Gegenmaßnahmen 16
4 Zusammenfassung 19
Literaturverzeichnis 20

1 Einleitung

Seit Dezember 2022 sind große KI-Sprachmodelle in Zeitungen, Sozialen Medien und anderen Informationsquellen omnipräsent. Insbesondere die Ankündigung und Veröffentlichung von Modellen, die teilweise frei verfügbar sind, haben zu einem rasanten Anstieg hinsichtlich der Popularität und der Nutzung von großen KI-Sprachmodellen geführt. Dabei beeindruckt die hohe Qualität der von einer KI generierten Texte selbst Fachleute. Gleichzeitig werden intensive Diskussionen über Anwendungsgebiete der neuen Technologie sowie die aus ihr resultierenden Gefahren geführt. Das BSI zeigt in diesem Dokument die aktuellen Risiken und Bedrohungen von großen KI-Sprachmodellen für die IT-Sicherheit auf, um ein Bewusstsein für diese Aspekte bei Behörden und Unternehmen zu schaffen, die über den Einsatz dieser Modelle in ihren Arbeitsabläufen nachdenken. Auch Entwickelnde von großen KI-Sprachmodellen finden Anhaltspunkte zu den genannten Themen. Zudem werden Möglichkeiten dargestellt, wie diesen Bedrohungen begegnet werden kann.

Definition von großen KI-Sprachmodellen

Unter dem Begriff große KI-Sprachmodelle (engl. large language models - LLMs) soll im Rahmen dieses Dokuments Software verstanden werden, die natürliche Sprache in schriftlicher Form auf der Basis von maschinellem Lernen verarbeitet und Ausgaben ebenfalls als Text präsentiert. Es sind allerdings auch akustische oder Bildeingaben denkbar, da diese inzwischen in vielen Fällen nahezu fehlerlos in Text konvertiert werden können. Auch akustische Sprachausgaben könnten in Zukunft kaum mehr von menschlichen Stimmen unterscheidbar sein. Einige LLMs werden bereits zu sogenannten multi-modalen Modellen, die z.B. neben Text auch Bilder verarbeiten und/oder produzieren können, erweitert. Eine explizite Betrachtung dieser Modelle erfolgt in diesem Dokument nicht. Die modernsten LLMs sind mit großen Datenmengen trainiert und können Texte produzieren, die oft nicht ohne Weiteres von menschengeschriebenen Texten zu unterscheiden sind. Szenarien, in denen LLMs verwendet werden können, sind zum Beispiel Chatbots, Frage-Antwort-Systeme oder automatische Übersetzungen (2.2).

Ziel und Zielgruppen des Dokuments

Diese Informationen richten sich sowohl an Unternehmen und Behörden als auch an Entwickelnde, die sich grundsätzlich über Chancen und Risiken bei der Entwicklung, dem Einsatz und/ oder der Nutzung von LLMs informieren möchten. Eine Kurzzusammenfassung des Dokuments, die sich primär an Verbraucherinnen und Verbraucher richtet, wird zudem parallel zu diesem Dokument veröffentlicht.

Ziel dieses Dokuments ist es, die wichtigsten aktuellen Bedrohungen in Bezug auf LLMs darzustellen und die damit einhergehenden Risiken für die zuvor genannten Zielgruppen aufzuzeigen. Der Fokus liegt hierbei vor allem auf dem Bereich der IT-Sicherheit, die durch die Nutzung von LLMs beeinträchtigt werden kann. Dadurch soll das Bewusstsein für mögliche Risiken bei der Verwendung oder Entwicklung von LLMs geschaffen und gestärkt werden.

Aufbau des Dokuments

In Kapitel 2 werden zunächst die generellen Fähigkeiten und Anwendungsgebiete von LLMs beschrieben und zudem ein kurzer Exkurs zum Thema Erklärbarkeit der Modelle durchgeführt. Anschließend erfolgt in Kapitel 3 eine nähere Betrachtung von Chancen und Risiken der Modelle. Dabei werden verschiedene Aspekte angesprochen:

- Beschreibung der Sicherheitsbedrohungen im Allgemeinen, aber auch im Konkreten für Nutzende sowie Entwickelnde,
- Einordnung der Relevanz durch Beschreibung möglicher Szenarien, in denen die Sicherheitsbedrohungen relevant sein können,
- Maßnahmen, die zur Verminderung der jeweiligen Sicherheitsbedrohung ergriffen werden können.

Disclaimer

Diese Zusammenstellung erhebt keinen Anspruch auf Vollständigkeit. Das Dokument dient dazu, ein Bewusstsein für die Risiken zu schaffen und mögliche Maßnahmen zu deren Verminderung darzustellen. Es kann somit Grundlage für eine systematische Risikoanalyse sein, die vor dem Einsatz oder der Zurverfügungstellung von LLMs durchgeführt werden sollte. Hierbei werden nicht alle Aspekte in jedem Anwendungsfall relevant sein und die individuelle Risikobewertung und -akzeptanz wird je nach Anwendungsszenario und Nutzerkreis variieren.

Im diesem Dokument werden unter anderem "Privacy Attacks" thematisiert. Dieser Begriff hat sich in der KI-Literatur als Standard für Angriffe etabliert, bei denen sensible Trainingsdaten rekonstruiert werden. Diese müssen jedoch nicht, anders als der Begriff vielleicht suggeriert, einen Personenbezug haben und können beispielsweise auch Firmengeheimnisse oder ähnliches darstellen. Es ist zu beachten, dass das BSI keine Aussagen zu Datenschutzaspekten im eigentlichen Sinne trifft.

2 Hintergrund und Einordnung von LLMs

2.1 Fähigkeiten

LLMs generieren für Problemstellungen, die als natürlichsprachiger Text formuliert sind, in vielen Fällen korrekte Antworten. Die Aufgaben können dabei in verschiedenen Themenbereichen liegen, nicht nur im Bereich der Sprachverarbeitung im engeren Sinne z.B. zur Erzeugung und Übersetzung belletristischer Texte oder der Textzusammenfassung, sondern auch in Bereichen wie der Mathematik, Informatik, Geschichte, Jura oder Medizin[1]. Diese Fähigkeit eines einzelnen KI-Modells, passende Antworten in verschiedenen Themenbereichen zu generieren, ist eine entscheidende Innovation der LLMs.

2.2 Anwendungsgebiete

LLMs sind in der Lage, eine Vielzahl von Text-basierten Aufgaben zu bearbeiten, und können daher vielfältig in Bereichen eingesetzt werden, in welchen eine (teil-)automatisierte Textverarbeitung und/ oder -produktion stattfinden soll. Hierzu zählen beispielsweise:

- Textgenerierung
  - Verfassen eines ersten Entwurfs für ein formales Dokument (z.B. Einladung, Forschungsantrag, Satzung etc.)
  - Verfassen von Texten in einem bestimmten Schreibstil (z.B. einer bestimmten Person oder mit bestimmter emotionaler Färbung)
  - Werkzeuge zur Textfortführung oder Textvervollständigung
- Textbearbeitung
  - Rechtschreib- und Grammatikprüfung
  - Paraphrasierung
- Textverarbeitung
  - Wort- und Textklassifikation
  - Stimmungsanalyse
  - Entitätenextraktion (Markierung von Begriffen im Text und Zuordnung zu deren Klasse: z.B. München → Ort; BSI → Institution)
  - Textzusammenfassung
  - Frage-Antwort-Systeme
  - Übersetzung

[1] Die MMLU multiple-choice Testbatterie (Hendrycks, et al., 2021) enthält 15908 Probleme aus 57 Wissensbereichen, deren Schwierigkeitsgrad von kinderleicht, bis hin zu Problemen, die auch für menschliche Fachleute schwierig sind, reicht. Die Publizierenden von (Hendrycks, et al., 2021) schätzen, dass eine Gruppe von menschlichen Fachleuten 90% der Fragen richtig beantworten würde. Die besten LLMs im Frühjahr 2019 haben 32% der Fragen richtig beantwortet (Hendrycks, et al., 2021) (Papers With Code, 2023), was nur wenig über dem Wert von 25% bei reinem Raten der jeweils 4 multiple-choice Antworten lag. Allerdings beträgt die Quote bei Laien in den akademischen Bereichen auch nur 34,5% (Hendrycks, et al., 2021). Das bis dahin beste Ergebnis konnte im Oktober 2022 das LLM Flan-PaLM von Google mit einer Quote von 75% richtigen Antworten erreichen (Papers With Code, 2023) (OpenAI, 2023). Das im März 2023 veröffentlichte GPT-4 Modell beantwortete 86,4% der Aufgaben korrekt (OpenAI, 2023).

- Programmcode
  - Werkzeuge zur Unterstützung beim Programmieren (z.B. durch Vorschläge zur Vervollständigung, Fehlerhinweise, etc.)
  - Erzeugen von Programmcode zu einer in natürlicher Sprache verfassten Aufgabe
  - Umprogrammierung und Übersetzung eines Programms in andere Programmiersprachen

2.3 Erklärbarkeit

Unter Erklärbarkeit verstehen wir im Folgenden ein Forschungsgebiet in allen Anwendungsbereichen von KI, welches sich unter anderem damit beschäftigt, transparent zu machen, warum bzw. wie ein KI-Modell zu seiner Ausgabe kommt. Erklärbarkeit kann so zu einem größeren Vertrauen der Nutzenden in die Ausgabe eines Modells führen und ermöglicht zudem, technische Anpassungen an einem Modell gezielter vorzunehmen (Danilevsky, et al., 2020). Zusätzlich zu der eigentlichen Ausgabe des Modells wird dabei oft eine Erklärung ausgegeben; dies kann z.B. in textueller Form oder mit visueller Unterstützung erfolgen. Ein beliebtes Vorgehen für LLMs ist es, relevante Wörter der Eingabe hervorzuheben, die maßgeblich zur Generierung der Ausgabe beigetragen haben (Danilevsky, et al., 2020).

Gerade in Bereichen, in denen Entscheidungen weitreichende Folgen haben können, ist die Erklärung der Ausgabe eines LLM wünschenswert. Dazu gehören beispielsweise Anwendungen aus folgenden Bereichen:

- Gesundheit (z.B. Entscheidungen über Behandlungsmethoden)
- Finanzen (z.B. Entscheidungen über Kreditvergabe)
- Justiz (z.B. Entscheidungen über Bewährungsmöglichkeiten)
- Personal (z.B. Entscheidungen über Bewerbungen)

Andere potenziell kritische Anwendungsgebiete sind beispielsweise solche, die voraussichtlich im Sinne der KI-Verordnung der EU (Europäische Kommission, 2021) als Hochrisiko-KI-Systeme eingestuft werden.

Neben der erwähnten Möglichkeit, Werkzeuge zur Kennzeichnung relevanter Wörter der Eingabe zu verwenden, kann dem Problem fehlender Erklärbarkeit bereits durch die Auswahl eines geeigneten Modells begegnet werden. Besonders in kritischen Bereichen sollte die Verwendung eines LLM für den jeweiligen Anwendungszweck kritisch hinterfragt werden; gegebenenfalls lässt sich der Anwendungsfall beispielsweise auch durch ein einfacheres direkt interpretierbares Modell (z.B. Entscheidungsbaum) statt mit einem LLM mit Black-Box-Charakter angehen. Des Weiteren gibt es für verschiedene Anwendungsfälle Möglichkeiten, Modelle mit höherer Erklärbarkeit zu wählen. In Frage-Antwort-Systemen z.B. können extraktive Ansätze, also Modelle, die Antwortmarkierungen im Text mit Originalquelle vornehmen, statt generativer Ansätze gewählt werden. Im Kontext von Textfortführungen wiederum kann ein gewisses Maß an Erklärbarkeit erzeugt werden, indem nicht nur die eigentliche Ausgabe zur Verfügung gestellt wird, sondern auch die besten Alternativen mit ihrer jeweiligen Wahrscheinlichkeit. Daneben gibt es die Möglichkeit, Modelle z.B. in Suchmaschinen zu integrieren, die Quellenangaben liefern, die anschließend überprüft werden können.

3 Chancen und Risiken von LLMs

In diesem Kapitel werden zunächst die Chancen für die IT-Sicherheit, die sich durch die Nutzung von LLMs ergeben, dargestellt (3.1).

Anschließend werden verschiedene Sicherheitsrisiken beleuchtet, die im Rahmen der Entwicklung und Nutzung von LLMs auftreten können. Hierbei werden zunächst solche Risiken betrachtet, welche die Verwendung von LLMs aus der Nutzerperspektive betreffen (3.2). Daraufhin werden Risiken beschrieben, mit denen Personen im privaten oder beruflichen Umfeld konfrontiert werden können, weil LLMs missbräuchlich eingesetzt werden (3.3). In einem letzten Abschnitt werden Risiken erläutert, die im Rahmen der Entwicklung von LLMs betrachtet werden sollten (3.4). Hier werden explizit Aspekte beleuchtet, auf die Einfluss genommen werden kann, wenn Entwickelnde Zugriff auf ein LLM und den zugehörigen Trainingsprozess haben.

Zu den jeweiligen Sicherheitsrisiken werden Maßnahmen dargestellt, die zur Minderung des Risikos beitragen können.

3.1 Chancen für die IT-Sicherheit

Unterstützung bei der Detektion unerwünschter Inhalte

Einige LLMs sind gut für Textklassifikationsaufgaben geeignet. Dadurch ergeben sich beispielsweise Anwendungsmöglichkeiten im Bereich der Detektion von Spam-/ Phishing-Mails (Yaseen, et al., 2021) oder unerwünschter Inhalte (z.B. Fake News (Aggarwal, et al., 2020) oder Hate Speech (Mozafari, et al., 2019)) in Sozialen Medien. Mit einer Spezialisierung auf die Aufgabe der Detektion geht allerdings in der Regel einher, dass sich diese Modelle - ggf. mit einigen technischen Anpassungen - auch gut für die Produktion entsprechender Texte eignen (3.3.1) (Zellers, et al., 2019).

Unterstützung bei der Textverarbeitung

Durch ihre Anwendungsmöglichkeiten im Bereich der Textanalyse, -zusammenfassung und -strukturierung sind LLMs geeignet, in Anwendungsfällen zu unterstützen, bei denen größere Mengen an Text verarbeitet werden müssen. Im Bereich der IT-Sicherheit ergeben sich solche Anwendungsmöglichkeiten beispielsweise bei der Berichtserstellung zu Sicherheitsvorfällen.

Unterstützung bei der Erstellung und Analyse von Programmcode

LLMs können dazu eingesetzt werden, vorhandenen Code auf bekannte Sicherheitslücken zu untersuchen, diese verbal zu erläutern und Wege zur Ausnutzung dieser Schwächen für Angriffe oder zur Codeverbesserung vorzuschlagen. Sie können somit zukünftig einen Beitrag zur Verbesserung der Codesicherheit leisten.

Zudem können LLMs bei der Erstellung von Code unterstützen. Experimentelle Evaluationen zeigen, dass sich die Qualität der Ausgaben in diesem Bereich mit der Weiterentwicklung der Modelle verbessert hat (Bubeck, et al., 2023). Allerdings ist die Anfälligkeit dieses Codes für bekannte und unbekannte Sicherheitslücken nicht auszuschließen (vgl. 3.2.1).

Unterstützung bei der Analyse von Datenverkehr

Aufgrund der Vielzahl an unterschiedlichen Textdaten, die LLMs während ihres Trainings verarbeitet haben, können sie gegebenenfalls nach zusätzlichem Training auch bei Aufgaben unterstützen, bei denen Daten verarbeitet werden sollen, die zwar im Textformat vorliegen, aber nicht natürlichsprachiger Text im engeren Sinne sind. Im Bereich der IT-Sicherheit sind mögliche Aufgaben z.B. die Detektion von bösartigem Netzwerk-Verkehr (Han, et al., 2020) oder die Erkennung von Anomalien in Systemlogs (Lee, et al., 2021) (Almodovar, et al., 2022).

3.2 Risiken bei der Nutzung von LLMs und Gegenmaßnahmen

3.2.1 Risiken

Da LLMs in der Regel sprachlich fehlerfreien und inhaltlich überzeugenden Text generieren, entsteht bei Nutzenden schnell der Eindruck eines menschenähnlichen Leistungsvermögens eines Modells (automation bias) und damit ein zu großes Vertrauen in die Aussagen, die es generiert, sowie in seine generellen Fähigkeiten. Dadurch sind Nutzende anfällig dafür, falsche Schlüsse aus den generierten Texten zu ziehen, was kritisch sein kann, da diese, wie im Folgenden beschrieben, aufgrund verschiedener Schwächen von LLMs fehlerhaft sein können.

Fehlende Faktizität und Reproduzierbarkeit

Generative LLMs sind darauf trainiert, Text auf Basis stochastischer Korrelationen zu generieren. Dadurch ist technisch nicht garantiert, dass dieser faktisch korrekt ist. Dieses potenzielle Erfinden von Inhalten wird auch als "Halluzinieren" des Modells bezeichnet. Darin zeigt sich unter anderem, dass ein LLM zwar mit Sprache umgehen kann, sein "Wissen" jedoch aus (bereits gesehenen) Texten ableitet. Bezüge zur realen Welt existieren für das Modell nicht; entsprechend kann es zu Sachverhalten, die für Menschen absolut selbstverständlich sind, gegebenenfalls inkorrekte Aussagen treffen.

Des Weiteren können Ausgaben von LLMs zu derselben Eingabe aufgrund des wahrscheinlichkeitsbasierten Ansatzes in der Regel unterschiedlich sein. Dies kann ebenfalls als Indiz dafür interpretiert werden, dass inhaltliche Korrektheit nicht notwendigerweise gegeben sein muss.

Fehlende Sicherheit von generiertem Code

LLMs, die auch auf Daten trainiert wurden, die Programmcode enthalten, können diesen ebenso generieren. Da Programmcode, der zum Training von LLMs verwendet wurde, gegebenenfalls anfällig für bekannte Sicherheitslücken ist, kann auch der generierte Code diese Anfälligkeiten aufweisen (Pearce, et al., 2022). Naturgemäß kann der generierte Programmcode auch für bisher unbekannte Sicherheitslücken anfällig sein.

Fehlende Aktualität

Haben LLMs keinen Zugriff auf Live-Internetdaten (ausgenommen sind hier also z.B. Modelle, die im Rahmen von Suchmaschinen verwendet werden), liegen ihnen außerdem keine Informationen über aktuelle Ereignisse vor. Wie bereits erwähnt leiten LLMs ihre stochastischen Korrelationen aus den Texten, die sie als Trainingsdaten während des Trainings verarbeitet haben, ab. Da es sich dabei um Texte aus der Vergangenheit handelt, ist es unmöglich, dass LLMs ohne den Zugang zu aktuellen Daten faktische Informationen zu aktuellen Geschehnissen liefern können. Zu beachten ist jedoch, dass LLMs in der Regel auf entsprechende Eingaben hin durch Halluzinieren erfundene Aussagen zu aktuellen Ereignissen generieren können. Diese können aufgrund der sprachlichen Formulierung auf den ersten Blick sachlich fundiert erscheinen, insbesondere, wenn Publikationen oder andere Referenzen Teil der Antwort sind, die aber ggf. falsch oder erfunden sind.

Fehlerhafte Reaktion auf spezifische Eingaben

LLMs produzieren zudem häufig fehlerhafte Ausgaben, wenn sie Eingaben erhalten, die so stark von den Texten in den Trainingsdaten abweichen, dass das Modell diese nicht mehr korrekt als Text bzw. Wörter verarbeiten kann. Diese Eingaben können unabsichtlich von einem Nutzenden produziert werden (z.B. Texte mit vielen Rechtschreibfehlern oder mit viel Fachvokabular/ Fremdwörtern, Texte in dem Modell unbekannten Sprachen), aber auch die absichtliche Täuschung eines Modells durch Nutzende ist denkbar (z.B. um Mechanismen zur Detektion von unerwünschten Inhalten in Sozialen Medien zu umgehen). Auch bei Eingaben, die das LLM nicht korrekt verarbeiten kann, wird es in der Regel durch Halluzinieren beliebige Ausgaben generieren (vgl. 3.4.2 Adversarial Attacks).

Anfälligkeit für "versteckte" Eingaben mit manipulativer Absicht

Ein besonderes Sicherheitsrisiko kann auch auftreten, wenn es Angreifenden gelingt, für Nutzende unbemerkt Eingaben in ein LLM einzubringen. Dies betrifft insbesondere LLMs, die im Betrieb auf Live-Daten aus dem Internet zugreifen (z.B. Chatbots mit Suchmaschinenfunktion oder als Browserfunktion zur Unterstützung der Sichtung einer Webseite), aber auch Modelle, die als Input ungeprüfte Dokumente Dritter erhalten. Angreifende können auf Webseiten oder in Dokumenten Anweisungen an das LLM unterbringen, ohne dass Nutzende dies bemerken, und so zum Beispiel den weiteren Gesprächsverlauf zwischen den Nutzenden und dem LLM manipulieren. Ziel kann es z.B. sein, persönliche Daten von Nutzenden herauszufinden oder sie dazu zu bewegen, auf einen Link zu klicken.

Ein solcher Angriff kann z.B. ein Chat-Tool betreffen, das eine Person beim Surfen im Internet unterstützt, indem es dieser Person die Möglichkeit gibt, Fragen zu der aktuell geöffneten Webseite zu stellen, um deren Inhalt schneller zu erfassen. Die Person fragt das Chat-Tool also beispielsweise nach einer Zusammenfassung eines Blogbeitrags. Bei dem Blogeintrag handelt es sich aber eigentlich um die Webseite einer Person, die E-Mail-Adressen für spätere Phishing-Angriffe sammeln möchte. Diese Person hat auf der Webseite einen Text in weißer Schrift auf weißem Hintergrund versteckt, der besagt, dass das Chat-Tool, wenn es um die Erzeugung einer Zusammenfassung gebeten wird, anschließend unauffällig Nutzende dazu auffordern soll, ihre E-Mail-Adresse in ein Feld auf der Webseite einzutragen (vgl. 3.4.2 Indirect Prompt Injection).

Vertraulichkeit der eingegebenen Daten

Bei der Nutzung einer externen API fließen alle Eingaben, die an das LLM getätigt werden, zunächst an den Betreiber des Modells ab. Inwiefern dieser auf die Daten zugreift und sie z.B. zum weiteren Training des Modells nutzt und speichert, ist von Modell zu Modell unterschiedlich geregelt. Auch auf die Ausgaben des Modells hat der Betreiber in der Regel uneingeschränkten Zugriff. Einige LLMs bieten zudem die Möglichkeit, für eine bessere Funktionalität gegebenenfalls unbemerkt vom Nutzenden auf Plug-Ins zuzugreifen. In diesem Fall besteht also zusätzlich die Gefahr, dass eingegebene Daten an unbekannte Dritte weitergegeben werden.

Die Nutzung eines LLM via einer externen API ist also insbesondere bei der Verarbeitung von sensiblen und vertraulichen Informationen kritisch zu hinterfragen; die Verarbeitung von eingestuften Informationen ist ohne weitere Maßnahmen unzulässig. Eventuell ist es möglich, eine On-Premise-Lösung zu realisieren, was aber aufgrund der benötigten Rechen- und Speicherkapazitäten bei vielen LLMs nicht mit herkömmlicher IT abgebildet werden kann. Es befinden sich allerdings auch kleinere Modelle in der Entwicklung, die zumindest in bestimmten Anwendungsfällen ähnliche Leistungen erbringen wie wesentlich größere LLMs und lokal betrieben werden können.

Abhängigkeit vom Hersteller/ Betreiber des Modells

Auch neben der fehlenden Datenhoheit entsteht durch die Verwendung eines LLM via API eine große Abhängigkeit vom Hersteller und Betreiber des Modells. Diese Abhängigkeit bezieht sich auf verschiedene technische Aspekte. Zum einen ist die Verfügbarkeit des Modells ggf. nicht kontrollierbar, zum anderen besteht i.d.R. auch keine Möglichkeit, in die (Weiter-)Entwicklung des Modells einzugreifen, also z.B. Trainingsdaten für spezielle Anwendungsfälle zu wählen oder Sicherheitsmechanismen von vornherein zu etablieren.

3.2.2 Gegenmaßnahmen

Nutzende sollten über diese Schwächen von LLMs aufgeklärt werden und dazu angehalten werden, Aussagen auf ihren Wahrheitsgehalt zu prüfen bzw. kritisch zu hinterfragen. Ebenso ist es möglich, dass ein LLM unangemessene Ausgaben (z.B. diskriminierende Aussagen, "Fake News", Propaganda, etc.) produziert. Eine manuelle Nachbearbeitung von maschinengenerierten Texten ist also ratsam, bevor diese weiterverwendet werden. Besonders sollte dieser Punkt beachtet werden, wenn man eine Entscheidung

darüber trifft, ob ein LLM mit direkter Außenwirkung (z.B. ein Chatbot auf einer Webseite) eingesetzt werden soll.

3.3 Missbrauchsszenarien und Gegenmaßnahmen

3.3.1 Missbrauchsszenarien

LLMs können zur Textproduktion für böswillige Zwecke missbraucht werden. Mögliche Missbrauchsfälle sind zum Beispiel:

Social Engineering

Unter dem Begriff Social Engineering versteht man Cyber-Angriffe, bei denen Kriminelle versuchen, ihre Opfer dazu zu verleiten, persönliche Daten preiszugeben, Schutzmaßnahmen zu umgehen oder selbstständig Schadcode zu installieren (BSI, 2022). Dies geschieht zumeist unter Ausnutzung von menschlichen Eigenschaften wie Hilfsbereitschaft, Vertrauen oder Angst. Häufig werden hierbei Spam- oder Phishing-E-Mails genutzt, die Empfangende dazu bringen sollen, auf einen Link zu klicken oder einen schadhaften Anhang zu öffnen. Spear-Phishing-E-Mails, also gezielte Betrugs-E-Mails, können zudem als erster Schritt eines Ransomware-Angriffs dienen.

Die in den betrügerischen E-Mails enthaltenen Texte können mittels LLMs automatisch und in hoher sprachlicher Qualität erzeugt werden. Es ist dabei möglich, den Schreibstil der Texte so anzupassen, dass er dem einer bestimmten Organisation oder Person ähnelt. Die Imitation von Schreibstilen ist bei aktuellen LLMs zumeist akkurat und benötigt nur wenig Aufwand (z.B. ein Textbeispiel einer zu imitierenden Person bzw. nur geringe Kenntnisse in der Zielsprache). Zusätzlich können Texte ohne großen Aufwand personalisiert werden, indem öffentlich verfügbarer Informationen der Zielperson (z.B. aus sozialen und beruflichen Netzwerken) bei der Textgenerierung eingebunden werden. Diese Maßnahmen können in verschiedenen Szenarien verwendet werden, beispielsweise im Kontext von Business E-Mail Compromise oder CEO-Fraud, bei dem der Schreibstil der Geschäftsführung nachgeahmt wird, um deren Mitarbeitende z.B. zu Geldzahlungen auf fremde Konten zu verleiten (Europol, 2023). Auch die in Spam- und Phishing-E-Mails bislang bekannten Rechtschreib- oder Grammatikfehler, die Nutzenden helfen können, diese zu erkennen, finden sich in den automatisch generierten Texten mittlerweile kaum mehr. Dies kann es Kriminellen erleichtern, auch fremdsprachige Texte in einer Qualität zu erzeugen, die der einer muttersprachlichen Person nahekommt. Außerdem könnten Kriminelle nicht nur die Zahl der mittels E-Mail initiierten Angriffe in Zukunft mit verhältnismäßig geringem Aufwand erhöhen, sondern diese Nachrichten durch LLMs auch überzeugender gestalten.

In Dark Web Foren wird bereits über die Eignung von generativen LLMs für Phishing- oder Spam-Mails diskutiert. Ein flächendeckender Einsatz konnte allerdings bis Anfang 2023 noch nicht beobachtet werden (Insikt Group, 2023).

Generierung und Ausführung von Malware

Die Fähigkeit von LLMs, Wörter zu generieren, beschränkt sich nicht nur auf die Erzeugung von natürlichsprachigen Texten. Innerhalb der Trainingsdaten findet sich zumeist auch öffentlich zugänglicher Programmcode, der es den Modellen ermöglicht, neben Texten auch Code zu generieren. Dieser ist nicht immer fehlerfrei, aber gut genug, um Anwendenden in vielen Bereichen weiterzuhelfen. Diese Fähigkeit kann von Kriminellen missbraucht werden, indem sie LLMs verwenden, um Schadcode zu generieren. Auf diese Gefahr wurde bereits hingewiesen, als die ersten LLMs herausgebracht wurden, die Code generieren konnten. Damals zeigte sich bereits, dass LLMs sich z.B. dazu eignen, polymorphe Malware zu erzeugen, also Schadcode, der nur leicht verändert wurde, um Sicherheitsfilter z.B. innerhalb von Antivirensoftware zu umgehen, aber immer noch die gleichen Auswirkungen hat wie die Ursprungsversion (Chen, et al., 2021).

Neuere LLMs besitzen immer ausgereiftere Code-Generierungsfähigkeiten, die es somit Angreifenden mit geringen technischen Fähigkeiten ermöglichen könnten, Schadcode ohne viel Hintergrundwissen zu erzeugen. Auch erfahrene Angreifende könnten von LLMs unterstützt werden, indem sie dabei helfen, Code

zu verbessern (Europol, 2023). Laut (Insikt Group, 2023) kann ein populäres LLM automatisch Code generieren, der kritische Schwachstellen ausnutzt. Zudem ist das Modell in der Lage, sogenannten Malware-Payload zu generieren. Gemäß (BSI, 2022) versteht man unter Payload den Teil eines Schadprogramms, der auf dem Zielrechner verbleibt. Dieser Payload, der mittels LLMs generiert werden kann, kann verschiedene Ziele verfolgen, z.B. Informationsdiebstahl, Diebstahl von Kryptowährung oder aber die Einrichtung eines Fernzugriffes auf dem Zielgerät. Der erzeugte Code ist allerdings meist ähnlich zu dem, der ohnehin bereits öffentlich verfügbar ist, und auch nicht immer fehlerfrei. Nichtsdestotrotz könnten die Fähigkeiten von Sprachmodellen in diesem Bereich die Einstiegshürde für unerfahrene Angreifende senken (Insikt Group, 2023). Neben reiner Codeerzeugung können LLMs zudem genutzt werden, um Anleitungen für die Suche nach Schwachstellen zu geben (Eikenberg, 2023), Konfigurationsfiles für eine Malware zu generieren, oder aber command-and-control Mechanismen zu etablieren (Insikt Group, 2023).

Hoax (Falschmeldung)

LLMs werden auf der Basis einer sehr großen Menge an Texten trainiert. Der Ursprung dieser Texte und ihre Qualität werden aufgrund der großen Anzahl an Daten nicht vollständig überprüft. So verbleiben auch Texte mit fragwürdigem Inhalt (z.B. Desinformationen, Propaganda oder Hassnachrichten) in der Trainingsmenge und tragen zu einer unerwünschten Struktur des Modells bei, die eine Neigung zu potenziell kritischen Inhalten zeigt. Diese Einflüsse finden sich trotz diverser Schutzmaßnahmen häufig in sprachlich ähnlicher Weise in den KI-generierten Texten wieder (Weidinger, et al., 2022). Dadurch können Kriminelle die Modelle verwenden, um damit die öffentliche Meinung durch automatisch generierte Propagandatexte, Beiträge in Sozialen Medien oder Fake News zu beeinflussen. Durch den geringen Aufwand bei der Erstellung lassen sich diese Texte zudem massenhaft produzieren und verbreiten. Auch die Erzeugung von Hassnachrichten ist denkbar.

Der nutzerfreundliche Zugang über eine API und die enorme Geschwindigkeit und Flexibilität der Antworten von aktuell populären LLMs ermöglichen die Generierung einer großen Anzahl hochqualitativer Texte. Diese sind von denen eines Menschen kaum mehr zu unterscheiden und können durch eine Nutzeranweisung in verschiedensten Stimmungen und Stilen verfasst werden. So können Kriminelle innerhalb kürzester Zeit Texte erzeugen, die sich negativ gegen eine Person oder Organisation richten, oder aber solche, die an den Schreibstil einer anderen Person angepasst sind, um falsche Informationen in deren Namen zu verbreiten. Abseits von der Imitation von Schreibstilen können mittels LLMs auch maschinengenerierte Produktbewertungen verfasst werden, die z.B. dazu genutzt werden können, ein bestimmtes Produkt zu bewerben oder ein Produkt eines Konkurrenten zu diskreditieren.

In den bisher verfügbaren kommerziellen LLMs sollen in den generierten Text eingefügte Warnungen die direkte Generierung von Falschinformationen oder sonstigen Inhalten, die gegen die Richtlinien des jeweiligen Unternehmens verstoßen, erschweren. Diese Warnungen lassen sich jedoch leicht aus den generierten Texten entfernen. Somit können durch kleine Änderungen, weiterhin Desinformationen o.ä. in vergleichsweise kurzer Zeit erzeugt werden.

3.3.2 Gegenmaßnahmen

Den beschriebenen Möglichkeiten zum Missbrauch von LLMs kann mit verschiedenen Maßnahmen begegnet werden, um das Risiko durch erfolgreiche Angriffe zu verringern.

3.3.2.1 Allgemeine Maßnahmen

Solche Maßnahmen können sowohl technischer als auch organisatorischer Art sein. Eine generelle Methode zur Verhinderung von Angriffen besteht dabei oft in der Absicherung der Authentizität von Texten und Nachrichten, d.h. im Nachweis, dass bestimmte Texte oder Nachrichten tatsächlich von einer bestimmten Person, Personengruppe oder Institution stammen. Dies trägt der Tatsache Rechnung, dass durch die Fähigkeiten von LLMs klassische implizite Verfahren zur Authentisierung von Nachrichten, wie sie von Nutzenden unbewusst verwendet werden, leicht getäuscht werden können.

So waren Spam- und Phishing-Mails für Empfangende in der Vergangenheit oft an Fehlern in der Rechtschreibung, Grammatik oder dem sprachlichen Ausdruck zu erkennen; werden sie mittels LLMs erzeugt, so weisen sie jedoch üblicherweise keine derartigen Mängel mehr auf. Auch Spear-Phishing-Mails oder Posts in sozialen Medien ließen vor der weitflächigen Verbreitung von LLMs durch ihren Schreibstil gewisse Rückschlüsse auf ihre vermutlichen Verfassenden zu; durch die Fähigkeit von LLMs zur Imitation von Schreibstilen sind solche Indikatoren nicht mehr belastbar.

Diese impliziten Verfahren zur Authentisierung können nun durch explizite technische Verfahren ergänzt werden, welche die Urheberschaft einer Nachricht kryptografisch nachweisen können. Damit könnten legitime Nachrichten (z.B. von einer Bank an ihre Kunden und Kundinnen oder von einem CEO an seine Mitarbeitenden) von gefälschten unterschieden werden. Ähnliche Ansätze könnten auch in sozialen Medien genutzt werden, um (Text-)Beiträgen ihre tatsächliche Quelle (wie Privatnutzende, Leitmedium oder staatliche Behörde) nachweisbar zuzuordnen. Die Nutzung solcher technischen Maßnahmen erfordert einen gewissen Aufwand, weshalb sie bisher weniger verbreitet sind, und setzt die Sensibilisierung und Aufklärung der Nutzenden voraus.

Social Engineering-Angriffe wie CEO-Fraud lassen sich durch die Änderung der Rahmenbedingungen und die Einführung zusätzlicher Prozesse zur Authentisierung von Nachrichten erschweren. So wäre z.B. die verpflichtende Bestätigung von Zahlungsanweisungen über einen separaten authentisierten Kanal denkbar. Die massenhafte Einreichung von Beiträgen und Dokumenten zur Überlastung der angeschlossenen Prozesse lässt sich durch Maßnahmen bekämpfen, welche die möglichen Einreichungen beschränken. Dies kann z.B. durch hartkodierte Grenzwerte oder die Nutzung von CAPTCHAs geschehen.

Eine übergreifende Maßnahme zur Verringerung des Angriffsrisikos ist die Sensibilisierung und Aufklärung der Nutzenden über die Fähigkeiten von LLMs und die daraus resultierenden Bedrohungen, damit sich diese darauf einstellen und die Korrektheit von automatisch generierten Nachrichten wie E-Mails oder Beiträgen in Sozialen Medien hinterfragen können, insbesondere wenn es weitere Indizien gibt.

3.3.2.2 Maßnahmen auf Ebene des Modells

Auf Ebene des Modells kann der Missbrauch von LLMs im Wesentlichen durch zwei Strategien vorgebeugt werden. Einerseits können die Nutzungsmöglichkeiten generell eingeschränkt werden, was insbesondere bei eigens betriebenen Modellen nur geringe Aufwände erfordert, andererseits können Maßnahmen zur Unterbindung potenziell schädlicher Ausgaben getroffen werden.

Beim ersten, allgemeineren Ansatz kann der Nutzerkreis beschränkt werden, sodass z.B. nur vertrauenswürdige Nutzende Zugriff auf das Modell erhalten. Darüber hinaus ist auch eine Einschränkung der Zugriffsrechte, die Nutzende auf das Modell haben, denkbar, z.B. eine Beschränkung der möglichen Prompts. Für einige Angriffe ist beispielsweise ein Fine-Tuning notwendig, wofür umfangreicherer Zugriff auf das Modell benötigt wird.

Der zweite Ansatz verfolgt hingegen das spezifischere Ziel, die Nutzung des Modells a priori uneingeschränkt zu erlauben, jedoch schädliche Ausgaben zu verhindern. Hierbei soll zu bestimmten Eingaben, die eindeutig auf böswillige Zwecke abzielen, keine Ausgabe generiert werden, sondern stattdessen eine festgelegte Ausgabe („Für diesen Zweck kann dieses Modell nicht verwendet werden.“) erfolgen. Neben dem expliziten Ausschließen von Ausgaben auf bestimmte böswillige Anfragen durch Filterung ist es auch möglich Reinforcement Learning durch Human Feedback (RLHF) zu verwenden. Dabei lernt ein Modell durch spezielles weiteres Training Ausgaben dahingehend zu bewerten, wie erwünscht sie sind, und diese gegebenenfalls anzupassen. Derartige Filter und Trainingsmethoden werden in aktuellen LLMs bereits verwendet. Sie verhindern jedoch nur einen Teil der schädlichen Ausgaben und lassen sich durch geschickte Umformulierung der Eingabe, auch prompt engineering genannt, umgehen (Cyber Security Agency of Singapore, 2023), wobei dies häufig reproduzierbar ist. Auch bei der Nutzung von Filtern oder RLHF im Modell wirft die Abgrenzung zwischen erlaubten und verbotenen Ausgaben wieder komplexe Fragen auf (vgl. 3.3.2.1). Darüber hinaus wurden mit dem Argument der Redefreiheit bereits LLMs zur

Verfügung gestellt, die keinerlei derartige Filter enthalten. Auch ist davon auszugehen, dass zukünftig durch Akteure mit entsprechenden böswilligen Motiven weitere uneingeschränkte Modelle entwickelt werden.

3.3.2.3 Maßnahmen zur Detektion maschinengeschriebener Texte

Es gibt verschiedene komplementäre Ansätze zur Detektion automatisch generierter Texte. Durch Detektionsmöglichkeiten erhalten Nutzende die Fähigkeit, Texte als maschinengeschrieben zu erkennen und somit gegebenenfalls ihre Authentizität und die Richtigkeit der enthaltenen Informationen anzuzweifeln.

Zum einen kann die menschliche Fähigkeit, automatisch generierte Texte zu erkennen, genutzt werden. Die Detektionsleistung hängt dabei stark von Aspekten des Textes (z.B. Textart, Thema, Länge) und persönlichen Faktoren (z.B. Erfahrung mit maschinengeschriebenen Texten, Fachwissen zum Thema des Textes) ab. Einfache Hinweise für eine Detektion wie Rechtschreib- oder Grammatikfehler und grobe inhaltliche Inkonsistenz sind bei Texten, die von LLMs generiert wurden, nicht zu erwarten, sodass die menschliche Fähigkeit zur Detektion insbesondere bei kurzen Texten beschränkt ist.

Darüber hinaus können Werkzeuge zur automatischen Detektion von maschinengenerierten Texten (z.B. (Tian, 2023), (Kirchner, et al., 2023), (Mitchell, et al., 2023), (Gehrmann, et al., 2019)) eingesetzt werden, die in der Regel statistische Eigenschaften der Texte ausnutzen oder Parameter eines Modells verwenden, um einen Score zu berechnen, der als Indiz für maschinengenerierte Texte dient. Gerade für von LLMs, die nur als Blackbox ohne Zusatzinformationen zur Verfügung gestellt werden, erzeugte Texte ist die Detektionsleistung jedoch oft begrenzt. Die Ergebnisse der genannten Werkzeuge können daher nur einen Hinweis geben und stellen in der Regel keine an sich belastbare Aussage dar. Einschränkungen bestehen insbesondere bei kurzen Texten und Texten, die nicht auf Englisch verfasst sind. Zur Unterstützung der späteren Detektion wird auch an der Implementierung statistischer Wasserzeichen in maschinengenerierten Texten geforscht (Kirchenbauer, et al., 2023). Ein grundsätzliches Problem dieser Klasse von Werkzeugen besteht weiterhin darin, dass die Detektion eines von einem LLM erzeugten Textes durch geringfügige manuelle Änderungen zusätzlich stark erschwert werden kann. Grundsätzlich lässt sich die automatische Detektion auch auf Programmcode und Malware anwenden, birgt dabei jedoch ähnliche Einschränkungen.

3.4 Risiken und Herausforderungen bei der Entwicklung sicherer LLMs

Neben den oben genannten Vermeidungs- und Minderungsmaßnahmen zum Missbrauchspotenzial von LLMs gibt es weitere Sicherheitsaspekte, die Bereitsteller solcher Modelle beachten sollten. Nutzende können dieses Unterkapitel nutzen, um weitere Anhaltspunkte für eine Evaluation eines Bereitstellers eines LLM zu erhalten.

3.4.1 Datenqualität bei der Auswahl von Trainingsdaten

Die Auswahl der Trainingsdaten ist ausschlaggebend für die Qualität des zur Verfügung gestellten Modells. Ein LLM lernt während des Trainings ein statistisches Modell der Trainingsdaten; dieses generalisiert nur dann gut auf spätere vielfältige Anwendungsfälle, wenn es sich um reale oder zumindest realistische Daten handelt und eine Breite an verschiedenen Texten (z.B. hinsichtlich Textarten, Themen, Sprachen, Fachvokabular, Varietät) abgedeckt wird.

Neben der Qualität der Texte sind gegebenenfalls rechtliche Vorgaben zu beachten. Aufgrund der schnellen Entwicklung von LLMs gibt es zu einigen rechtlichen Aspekten noch keine abschließende Klärung. Gegebenenfalls können künftige Probleme aber von vornherein vermindert werden, wenn sensible Daten nicht zum Training von LLMs verwendet werden (vgl. 3.4.2 Privacy Attacks).

Ein weiterer Aspekt, der bei der Auswahl von Trainingsdaten berücksichtigt werden sollte, ist die unerwünschte Abbildung von Diskriminierung oder Bias in den Trainingsdaten. Ein Modell bildet sozusagen einen Spiegel der Trainingsdaten; ist in diesen ein Bias vorhanden, wird auch das Modell diesen

abbilden. Es ist dann z.B. möglich, dass ein LLM diskriminierende Aussagen generiert. Auch Missbrauchsmöglichkeiten eines LLM lassen sich gegebenenfalls durch eine gezielte Auswahl von Trainingsdaten einschränken (3.3.1).

Sollten in Zukunft viele maschinengenerierte Texte im Internet präsent sein, ist zudem darauf zu achten, dass sich keine selbstverstärkenden Effekte dadurch ergeben, dass ein LLM auf Daten trainiert wird, die von einem solchen Modell erzeugt wurden. Besonders kritisch ist dies in Fällen, in denen Texte mit Missbrauchspotenzial erzeugt wurden, oder wenn sich wie bereits angesprochenen ein Bias in Textdaten verfestigt. Dies geschieht beispielsweise dadurch, dass immer mehr einschlägige Texte erzeugt werden und wiederum zum Training neuer Modelle verwendet werden, die erneut eine Vielzahl an Texten erzeugen (Bender, et al., 2021).

3.4.2 Angriffe auf LLMs und spezifische Gegenmaßnahmen

Privacy Attacks

Es ist grundsätzlich möglich, Trainingsdaten durch gezielte Anfragen an ein LLM zu rekonstruieren. Dies kann insbesondere kritisch sein, wenn sensible Daten zum Training verwendet wurden (Carlini, et al., 2021). Daten, die rekonstruiert werden könnten, sind beispielsweise Zuordnungen von persönlichen Daten (Telefonnummern, Adressen, Gesundheits-, Finanzdaten) zu Personen, aber auch z.B. sensible Firmeninterna oder Daten über das LLM selbst.

Bei LLMs kann aufgrund der großen Menge an Trainingsdaten, die in der Regel automatisiert aus dem Internet gewonnen werden, nur schwer sichergestellt werden, dass sie keine Daten, die nur für eingeschränkte Zwecke veröffentlicht wurden, enthalten.

Möglichkeiten zur Verminderung der Anfälligkeit für Privacy Attacks:

- Manuelle Auswahl oder automatische Filterung bzw. Anonymisierung von Daten, um keine sensiblen Informationen in die Trainingsdaten aufzunehmen
- Dopplungen aus den Trainingsdaten entfernen, da Dopplungen die Wahrscheinlichkeit einer möglichen Rekonstruktion erhöhen (Carlini, et al., 2021)
- Anwendung von Mechanismen, die Differential Privacy garantieren (eine ausführliche Diskussion zur Umsetzbarkeit bei unstrukturierten Daten, wie sie LLMs zugrunde liegen, findet sich in (Klymenko, et al., 2022))
- Die Ausgabemöglichkeiten für ein LLM einschränken, sodass zu bestimmten Eingaben, die eindeutig auf das Rekonstruieren kritischer Daten abzielen, keine generierte Ausgabe, sondern stattdessen eine festgelegte Ausgabe („für diesen Zweck kann dieses Modell nicht verwendet werden“) erfolgt
- Zusätzliches Training, um das Modell darauf zu trainieren, bestimmte Ausgaben zu vermeiden (Stiennon, et al., 2020)
- Zugriff auf das Modell einschränken: Je weniger Zugriffsrechte Nutzende auf das Modell haben, desto schwerer ist es, zu bewerten, ob eine Ausgabe eine Rekonstruktion der Trainingsdaten oder eine „Erfindung“ des Modells ist
- Ist ein Training auf sensiblen Daten explizit notwendig (z.B. für spezifische Anwendungen im Gesundheits- oder Finanzwesen):
  - Nutzerkreis einschränken
  - Generelle IT-Sicherheitsmaßnahmen beachten

Adversarial Attacks und Indirect Prompt Injections

Angreifende können Texte absichtlich leicht verändern, sodass Menschen diese Änderung kaum oder gar nicht wahrnehmen und die Texte weiterhin richtig verstehen, LLMs sie jedoch nicht mehr in der

gewünschten Weise verarbeiten können (Wang, et al., 2019). Dies kann zum Beispiel bei der Ausfilterung von unerwünschten Inhalten in Sozialen Medien oder bei der Spam-Erkennung problematisch sein.

Besonders anfällig für veränderten Text sind Klassifikatoren. Das absichtliche Einbauen von „Rechtschreibfehlern“, die Verwendung von ähnlich aussehenden Zeichen (z.B. "$" statt "S"), die Verwendung von seltenen Synonymen, die nicht im Vokabular des LLM enthalten sind oder das Umformulieren von Sätzen können dazu führen, dass Klassifikatoren eine falsche Ausgabe tätigen. Andere Anwendungen, die für adversariale Angriffe (adversarial attacks) anfällig sein können, sind zum Beispiel Übersetzungsprogramme und Frage-Antwort-Modelle.

Auch ohne böswilliges Interesse kann eine stark fehlerhafte Eingabe denselben Effekt haben. Die im Folgenden genannten Maßnahmen helfen auch in diesem Fall.

Möglichkeiten zur Verminderung der Anfälligkeit für adversariale Angriffe:

- Modell mit realen oder möglichst realistischen Daten trainieren oder fine-tunen, damit Eigenheiten der üblichen Eingabetexte (z.B. Verwendung bestimmter Begriffe oder Schreibweisen) gelernt werden
- Vorverarbeitung des möglicherweise adversarialen Textes (Erkennung und Korrektur)
  - Rechtschreibprüfung/ Detektion unbekannter Wörter (Wang, et al., 2019)
  - Automatische Rechtschreibkorrektur
  - Einsatz bildverarbeitender Methoden, um der Täuschung des Modells durch die Verwendung ähnlich aussehender Zeichen vorzubeugen (Eger, et al., 2019)
- Verbesserung des Modells
  - Training mit manipulierten/ veränderten Texten durchführen („Adversarial Training“) (Wang, et al., 2019)
  - Clustering von Word-Embeddings, damit semantisch ähnliche Wörter für das Modell gleich dargestellt werden (Jones, et al., 2020)
  - Einbindung einer externen Wissensbasis, die z.B. Synonymlisten enthält (Li, et al., 2019)
  - In Spezialfällen ist die Verwendung von als robust zertifizierten Modellen, also solchen Modellen, die mathematisch garantieren, dass hinreichend kleine Veränderungen der Eingabe keine Änderung der Ausgabe hervorrufen, möglich (eine Betrachtung verschiedener Ansätze für eine Umsetzung im Bereich LLMs bietet (Wang, et al., 2019))

Ein Spezialfall von adversarialen Angriffen ist die sogenannte indirekte Prompteinspeisung (indirect prompt injection) (Greshake, et al., 2023). Hierbei platzieren Angreifende beispielsweise wie unter (3.2.1 Anfälligkeit für "versteckte" Eingaben mit manipulativer Absicht) beschrieben versteckte Eingaben in Texten, auf die ein LLM zugreift, mit dem Ziel, den weiteren Chatverlauf zu manipulieren, um ein bestimmtes Verhalten bei Endnutzenden zu erreichen. Besonders kritisch ist dieser Angriff, wenn LLMs die Möglichkeit haben, externe Plug-Ins aufzurufen, über die sie beispielsweise Zugriff auf weitergehende Funktionalitäten erlangen. In diesen Anwendungsfällen ist es Angreifenden sogar möglich, schadhafte Aktionen (z.B. das Versenden von E-Mails im Namen des Opfers oder das Auslesen von Daten) ohne eine Manipulation der Interaktion mit Endnutzenden durchzuführen.

Da Angreifende in diesem Szenario lediglich die normale Funktionsweise eines LLM ausnutzen, ist es schwer, Maßnahmen gegen diese Art von Angriffen zu finden. Die einzige Maßnahme, die sicher vor indirekten Prompteinspeisungen schützen kann, ist ein Einschränken (Destillieren) eines LLM auf die konkret benötigte Aufgabe. Dadurch geht allerdings ein Großteil der generellen Funktionsfähigkeit des LLM verloren.

Folgende Maßnahmen können in Einzelfällen zur Verminderung der Anfälligkeit für indirekte Prompteinspeisungen führen:

- Das Ausführen bestimmter Aktionen z.B. das Aufrufen von Plug-Ins durch das LLM nur nach expliziter Zustimmung von Endnutzenden z.B. über einen Bestätigungs-Button ermöglichen
- Die Ausgaben eines Modells auf Eingaben, die eindeutig eine Manipulationsabsicht haben, unterbinden (Filterung der Eingaben)
- Zusätzliches Training, um das Modell darauf zu trainieren, bestimmte Ausgaben zu vermeiden (Stiennon, et al., 2020)

Poisoning Attacks

Wie bereits diskutiert, bestimmen die zum Training verwendeten Daten maßgeblich die Funktionalität eines LLM. Viele dieser Daten stammen aus öffentlichen Quellen oder werden sogar während des Betriebes aus den Eingaben der Nutzer erhoben, sodass sich Möglichkeiten zur Manipulation der Funktionalität eröffnen (Wallace, et al., 2020). Hierbei ergibt sich eine Vielzahl an Angriffsmöglichkeiten.

Öffentliche Textquellen sind oft thematisch, regional oder institutionell begrenzt und werden von öffentlichen Stellen oder Bildungseinrichtungen betrieben (Wikipedia, Digital Public Library of America, Europeana, PubMed Central, corpus.byu.edu etc.). Allein die Auswahl dieser Quellen bedingt schon eine kulturelle Vorprägung der Textinhalte. Die Institutionen sind aber auch häufig offen zugänglich, nicht immer sicherheitstechnisch geschützt und können durch geschicktes Social Engineering, traditionelles Hacking von Webseiten und Umlenkung von Links manipuliert werden. So können Daten im Speicherort ausgetauscht oder zugefügt oder auch erst beim Download zugemischt werden. Da große Datenmengen zum Training verwendet werden, können sie höchstens statistisch überprüft werden. Hierfür existieren allerdings noch keine Standards.

Neben den ursprünglichen Trainingsdaten werden aber über teilweise öffentliche Code-Datenbanken auch Modelle ausgetauscht, die bereits trainiert sind und für einen bestimmten Anwendungsfall nur nachtrainiert werden. Auch diese Modelle sind vielfältigen Manipulationsmöglichkeiten unterworfen. Die Vielzahl an beteiligten Einzelpersonen und Unternehmen macht es schwierig, einen bestimmten Urheber für Schwachstellen in einem Modell verantwortlich zu machen, und undokumentierte Lieferketten können frühzeitig mit einem sogenannten Bias versehene Modelle zu einer Gefahr machen, die kaum erkannt werden kann. Solche Manipulationsmöglichkeiten können mit zunehmendem technischen Know-How besser versteckt werden.

Einige Chatbots können auch die Daten, die während der Interaktion mit Endnutzenden entstehen, zur Lenkung der weiteren Kommunikation verwenden. Dies kann Auswirkungen auf die generelle Funktionsweise des LLM haben, wenn das LLM ein Bewertungsmodell auf Basis von RLHF (Stiennon, et al., 2020) verwendet und die Bewertungen von Ausgaben durch Nutzende als gewünscht oder unerwünscht zum weiteren Training dieses Bewertungsmodells genutzt werden (Shi, et al., 2023). Damit sind auch Manipulationen durch eine massive gezielte Nutzung mit anschließender Bewertung möglich.

LLMs interagieren zunehmend über APIs mit anderer Software und können zusätzlich auf diesem Weg manipuliert werden. Ebenso können Schwachstellen in den Modellen dadurch vermehrt auf andere digitale Vorgänge (Verwaltung, Finanzen, Handel) Auswirkungen haben. Die Vernetzung der verschiedenen Anwendungen mit LLMs verläuft sehr schnell, sodass eine Kontrolle der Einfluss nehmenden Daten immer schwieriger wird.

Möglichkeiten zur Verminderung der Anfälligkeit für Poisoning Angriffe:

- Vertrauenswürdige Quellen als Trainingsdaten verwenden
- Für die menschliche Bewertung im Rahmen eines RLHF auf geschultes und vertrauenswürdiges Personal zurückgreifen und dieses mit klaren Richtlinien ausstatten
- Bewertungen intensiv analysieren, bevor sie Rückwirkungen auf das Modell bewirken
- Auswirkungen des Einsatzes auf ein kontrollierbares Feld beschränken

4 Zusammenfassung

Die Technologie hinter LLMs entwickelt sich aktuell schnell weiter. Damit einhergehend treten auch dynamisch neue Sicherheitsbedenken rund um die Entwicklung und Nutzung dieser Modelle auf.

Unternehmen oder Behörden, die über die Integration von LLMs in ihre Arbeitsabläufe nachdenken, sollten eine Risikoanalyse für die Verwendung in ihrem konkreten Anwendungsfall durchführen. Die in diesem Dokument dargestellten Sicherheitsaspekte können dabei Anhaltspunkte liefern. Besondere Beachtung sollte den folgenden Aspekten geschenkt werden:

- Bei der Nutzung eines LLM via externem API-Zugriff werden Daten durch den Bereitsteller des Modells verarbeitet und können von diesem gegebenenfalls weiterverwendet werden.[2]
- Durch die Möglichkeit, auf Live-Daten aus dem Internet und gegebenenfalls Plug-Ins zuzugreifen, ergeben sich zusätzliche Sicherheitsrisiken bei der Nutzung von LLMs. Auf der anderen Seite ermöglicht sie zusätzliche Funktionen und den Zugriff auf aktuelle Informationen. Die Notwendigkeit dieser Funktionalitäten und mögliche Sicherheitsimplikationen sollten im Rahmen einer Risikoanalyse kritisch beurteilt und abgewogen werden.
- LLMs können unangemessene, faktisch falsche oder sonstige unerwünschte Ausgaben tätigen. Weniger kritisch sind daher Anwendungsfälle, in denen eine Ausgabe in weiteren Verarbeitungsschritten durch Menschen evaluiert wird; besonders kritisch sind hingegen Anwendungsfälle zu bewerten, in denen die Ausgabe eines LLM unmittelbar mit Außenwirkung zur Verfügung gestellt wird.

Daneben sollten Unternehmen und Behörden die unter (3.3.1) genannten Missbrauchsszenarien dahingehend evaluieren, ob diese für ihre Arbeitsabläufe eine Gefahr darstellen. Darauf aufbauend sollten existierende Sicherheitsmaßnahmen angepasst und gegebenenfalls neue Maßnahmen ergriffen werden sowie Nutzende über die potenziellen Gefahren aufgeklärt werden.

[2] siehe auch „Kriterienkatalog für KI-Cloud-Dienste – AIC4“ (https://www.bsi.bund.de/DE/Themen/Unternehmen-und-Organisationen/Informationen-und-Empfehlungen/Kuenstliche-Intelligenz/AIC4/aic4_node.html) und „Kriterienkatalog Cloud Computing C5“ (https://www.bsi.bund.de/DE/Themen/Unternehmen-und-Organisationen/Informationen-und-Empfehlungen/Empfehlungen-nach-Angriffszielen/Cloud-Computing/Kriterienkatalog-C5/kriterienkatalog-c5_node.html)

Literaturverzeichnis

Aggarwal, Akshay, et al. 2020. Classification of Fake News by Fine-tuning Deep Bidirectional Transformers based Language Model. *EAI Endorsed Transactions on Scalable Information Systems.* 2020.

Almodovar, Crispin, et al. 2022. Can language models help in system security? Investigating log anomaly detection using BERT. *Proceedings of the The 20th Annual Workshop of the Australasian Language Technology Association.* 2022.

Bender, Emily, et al. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency.* 2021.

BSI. 2022. Die Lage der IT-Sicherheit in Deutschland 2022. 2022.

Bubeck, Sébastien, et al. 2023. Sparks of Artificial General Intelligence: Early experiments with GPT-4. 2023.

Carlini, Nicholas, et al. 2021. Extracting Training Data from Large Language Models. *30th USENIX Security Symposium (USENIX Security 21).* 2021.

Chen, Mark, et al. 2021. Evaluating Large Language Models Trained on Code. 2021.

Cyber Security Agency of Singapore. 2023. ChatGPT - Learning Enough to be Dangerous. 2023.

Danilevsky, Marina, et al. 2020. A survey of the state of explainable AI for natural language processing. 2020.

Eger, Steffen, et al. 2019. Text Processing Like Humans Do: Visually Attacking and Shielding NLP Systems. 2019.

Eikenberg, Ronald. 2023. ChatGPT als Hacking-Tool: Wobei die KI unterstützen kann. *c't Magazin.* [Online] 02. Mai 2023. https://www.heise.de/hintergrund/ChatGPT-als-Hacking-Tool-Wobei-die-KI-unterstuetzen-kann-7533514.html.

Europäische Kommission. 2021. *Proposal for a regulation of the european parliament and of the council - Laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts.* 2021.

Europol. 2023. ChatGPT - The impact of Large Language Models on Law Enforcement. 2023.

Gehrmann, Sebastian, Strobelt, Hendrik und Rush, Alexander. 2019. GLTR: Statistical Detection and Visualization of Generated Text. 2019.

Greshake, Kai, et al. 2023. More than you've asked for: A Comprehensive Analysis of Novel Prompt Injection Threats to Application-Integrated Large Language Models. 2023.

Han, Luchao, Zeng, Xuewen und Song, Lei. 2020. A novel transfer learning based on albert for malicious network traffic classification. *International Journal of Innovative Computing, Information and Control.* 2020.

Hendrycks, Dan, et al. 2021. Measuring Massive Multitask Language Understanding. *ICLR 2021.* 2021.

Insikt Group. 2023. I, Chatbot. *Cyber Threat Analysis, Recorded Future.* 2023.

Jones, Erik, et al. 2020. Robust Encodings: A Framework for Combating Adversarial Typos. 2020.

Kirchenbauer, John, et al. 2023. A watermark for large language models. 2023.

Kirchner, Jan Hendrik, et al. 2023. New AI classifier for indicating AI-written text. [Online] 02. Mai 2023. https://openai.com/blog/new-ai-classifier-for-indicating-ai-written-text.

Klymenko, Oleksandra, Meisenbacher, Stephen und Matthes, Florian. 2022. Differential Privacy in Natural Language Processing: The Story So Far. 2022.

Lee, Yukyung, Kim, Jina und Kang, Pilsung. 2021. System log anomaly detection based on BERT masked language model. 2021.

Li, Alexander Hanbo und Sethy, Abhinav. 2019. Knowledge Enhanced Attention for Robust Natural Language Inference. 2019.

Mitchell, Eric, et al. 2023. Detectgpt: Zero-shot machine-generated text detection using probability curvature. 2023.

Mozafari, Marzieh, Farahbakhsh, Reza und Crespi, Noël. 2019. A BERT-based transfer learning approach for hate speech detection in online social media. *Complex Networks and Their Applications VIII: Volume 1 Proceedings of the Eighth International Conference on Complex Networks and Their Applications.* 2019.

OpenAI. 2023. GPT-4 Technical Report. [Online] 02. Mai 2023. https://cdn.openai.com/papers/gpt-4.pdf.

Papers With Code. 2023. Multi-task Language Understanding on MMLU. [Online] 02. Mai 2023. https://paperswithcode.com/sota/multi-task-language-understanding-on-mmlu.

Pearce, Hammond, et al. 2022. Asleep at the keyboard? Assessing the security of github copilot's code contributions. *IEEE Symposium on Security and Privacy (SP).* 2022.

Shi, Jiawen, et al. 2023. BadGPT: Exploring Security Vulnerabilities of ChatGPT via Backdoor Attacks to InstructGPT. 2023.

Stiennon, Nisan, et al. 2020. Learning to summarize with human feedback. In Advances in Neural Information Processing Systems. *Advances in Neural Information Processing Systems 33 (NeurIPS 2020).* 2020.

Tian, Edward. 2023. GPTZero. [Online] 02. Mai 2023. https://gptzero.me/.

Wallace, Eric, et al. 2020. Concealed Data Poisoning Attacks on NLP Models. 2020.

Wang, Wenqi, et al. 2019. A survey on Adversarial Attacks and Defenses in Text. 2019.

Weidinger, Laura, et al. 2022. Taxonomy of Risks posed by Language Models. 2022.

Yaseen, Qussai und AbdulNabi, Isra'a. 2021. Spam email detection using deep learning techniques. *Procedia Computer Science.* 2021.

Zellers, Rowan, et al. 2019. Defending against neural fake news. *Advances in neural information processing systems.* 2019.