Der Hamburgische Beauftragte für
Datenschutz und Informationsfreiheit

Diskussionspapier: Large Language Models und personenbezogene Daten

Dieses Diskussionspapier bildet den derzeitigen Wissens- und Erkenntnisstand beim Hamburgischen Beauftragten für Datenschutz und Informationsfreiheit (HmbBfDI) zur Frage der Anwendbarkeit der Datenschutz-Grundverordnung (DSGVO) auf Large Language Models[1] (LLMs) ab. Das Papier ist ein *Debattenimpuls*. Es soll Unternehmen und Behörden dabei unterstützen, datenschutzrechtliche Komplexe besser zu verorten. Zu diesem Zweck werden vorliegend relevante technische Aspekte von LLMs erläutert, vor dem Hintergrund der Rechtsprechung des Gerichtshofs der Europäischen Union (EuGH) zum Personenbezug bewertet und daraus resultierende Folgen für die Praxis beleuchtet. Hieraus lassen sich drei grundlegende Thesen ableiten:

1. **Die bloße Speicherung eines LLMs stellt keine Verarbeitung im Sinne des Art. 4 Nr. 2 DSGVO dar. Denn in LLMs werden keine personenbezogenen Daten gespeichert. Soweit in einem LLM-gestützten KI-System personenbezogene Daten verarbeitet werden, müssen die Verarbeitungsvorgänge den Anforderungen der DSGVO entsprechen. Dies gilt insbesondere für den Output eines solchen KI-Systems.**

2. **Mangels Speicherung personenbezogener Daten im LLM können die Betroffenenrechte der DSGVO nicht das Modell selbst zum Gegenstand haben. Ansprüche auf Auskunft, Löschung oder Berichtigung können sich jedoch zumindest auf Input und Output eines KI-Systems der verantwortlichen Anbieter:in oder Betreiber:in beziehen.**

3. **Das Training von LLMs mit personenbezogenen Daten muss datenschutzkonform erfolgen. Dabei sind auch die Betroffenenrechte zu beachten. Ein ggf. datenschutzwidriges Training wirkt sich aber nicht auf die Rechtmäßigkeit des Einsatzes eines solchen Modells in einem KI-System aus.**

[1] Gemeint sind hierbei allein die Modelle als wichtiger, aber nicht alleiniger Bestandteil eines KI-Systems (z. B. eines LLM-basierten Chatbots).

www.datenschutz-hamburg.de

E-Mail: mailbox@datenschutz.hamburg.de
Ludwig-Erhard-Straße 22 - D-20459 Hamburg - Tel.: 040 - 4 28 54 - 40 40 - Fax: 040 - 4 28 54 – 40 00

Einleitung

Wenn ein LLM als Teil eines KI-Systems[2] Prompts verarbeitet (sog. „Inferenz“), kann es vorkommen, dass der Output des LLMs Angaben über natürliche Personen enthält, insbesondere, wenn durch den Prompt gezielt danach gefragt wird. Dies wirft die Frage auf, ob in einem LLM personenbezogene Daten gespeichert sind.

Zur Beantwortung dieser Frage ist es zunächst wichtig, zwischen einem KI-System und einem ggf. darin enthaltenen LLM zu unterscheiden. Ein KI-System besteht aus mehreren Komponenten. Ein LLM ist eine solche Komponente. Man kann es ohne andere Komponenten, die mit dem Modell gemeinsam ein KI-System bilden, nicht sinnvoll nutzen. Ein bekanntes Beispiel für KI-Systeme sind Chatbots wie etwa ChatGPT. Zu ihren wichtigsten Komponenten zählen die Benutzerschnittstelle[3], Eingangs- und Ausgangsfilter sowie das LLM. Die Nutzereingaben werden vor der Inferenz des LLMs in der Regel zunächst durch weitere Komponenten des KI-Systems verarbeitet. So können die Nutzereingaben („Prompts“) z. B. mit weiteren Informationen aus einer Datenbank, einer Internetsuche oder durch sog. Retrieval Augmented Generation (RAG) angereichert werden. Erst dann verarbeitet das LLM den ggf. modifizierten Prompt. Der unmittelbare Output des LLMs wird danach typischerweise durch Filter weiterverarbeitet, bevor er – in der Regel – über die Benutzerschnittstelle ausgegeben wird.

Die nachfolgenden Ausführungen (II. und III.) bewerten nicht die Verarbeitungstätigkeiten im gesamten KI-System. Sie befassen sich ausschließlich mit der Frage, ob in LLMs personenbezogene Daten gespeichert sind.

I. Technische Bewertung von LLMs

LLMs verarbeiten Sprache, in der Regel sogar mehrere Sprachen.[4] Initial werden sie mit großen Mengen textlichen Inputs in den entsprechenden Sprachen trainiert. Im Output liefern sie wiederum sprachliche Ergebnisse.

[2] Vgl. auch Art. 3 Nr. 1 KI-VO.
[3] So können KI-Systeme über Webseiten oder eigens dafür entwickelte Apps benutzt werden.
[4] Damit sind sowohl verschiedene natürliche Sprachen wie Englisch, Französisch, Deutsch gemeint als auch Computersprachen wie Python, Javascript, Ruby. Für dieses Papier sind i.W. nur die natürlichsprachlichen Aspekte relevant.

1. Das Token als Basiselement der Informationsverarbeitung von LLMs

Von zentraler Bedeutung für die hier zu beantwortende Frage ist ein Verständnis für die Art und Weise der Verarbeitung und Speicherung der sprachlichen Informationen in LLMs. Ein Schlüsselelement ist dabei die sog. Tokenisierung des Inputs. Sämtliche Texte werden in vergleichsweise kleine, vordefinierte Stücke, sog. Tokens, geteilt, bevor sie Eingang in ein LLM finden. Diese Stücke sind in der Regel kleiner als ganze Wörter und größer als einzelne Buchstaben. Die Herausforderung bei der Erstellung von LLMs liegt darin, mit einer überschaubaren Menge (einige zigtausend) von Grundelementen auszukommen, die miteinander in Beziehung gesetzt werden können. Daher finden keine längeren Wörter, Satzteile oder gar ganze Sätze als solche Eingang in ein LLM. Der Satz „Ist ein LLM personenbezogen?“ wird von einem typischen Tokenisierer[5] z. B. wie folgt auf 12 Tokens aufgeteilt: [I][st][e][in][LL][M][person][en][be][z][ogen][?]. Diese Tokens werden auf numerische Werte[6] abgebildet, mit denen im Folgenden innerhalb des Modells ausschließlich gearbeitet wird.

Texte sind dort nicht mehr bzw. nur noch als Fragmente in Form dieser numerischen Tokens und ihrer weiteren Bearbeitung als sog. Embeddings[7] vorhanden. Durch die Embeddings werden die erlernten Zusammenhänge zugänglich, indem sie die Tokens zueinander ins Verhältnis setzen, also nach Wahrscheinlichkeitsgewichten zuordnen. Das umschreibt das eigentliche „Training“ eines LLM. Auf diese mathematische Repräsentation des antrainierten Inputs wird bei der Inferenz eines Prompts zurückgegriffen. Diese stellt sozusagen das erlernte „Wissen“ des LLMs dar.

Entsprechend erfolgt der Output eines LLMs zunächst als Folge von Tokens, die dann zurück in die zugehörigen Buchstabenfolgen gewandelt werden, bevor sie weiterverarbeitet werden.[8] Ein Output wie z. B. „Mia Müller hat gelogen.“ wird vom LLM ebenfalls aus Tokens konstruiert, konkret sind das erste „**M**“ und der Wortteil „**ogen**“ dieselben Token wie in dem obigen Beispielsatz [I][st][e][in][LL][**M**] [person][en][be][z][**ogen**]“.[9] In einem bestimmten und geeigneten Kontext wird nach dem Token „gel“ das Token „ogen“ gewählt, um das Wort „gelogen“ zu erzeugen, während in einem anderen Kontext nach „gel“ z. B. das Token „b“ folgt, um das Wort „gelb“ zu erzeugen.

[5] Hier OpenAI bei GPT-3, siehe https://platform.openai.com/tokenizer.
[6] Im vorliegenden Beispiel sind dies die Werte [40, 301, 304, 259, 27140, 44, 1048, 268, 1350, 89, 6644, 30], also „[I]“ dem Wert 40, „[st]“ dem Wert 301 usw. – die konkreten Werte sind LLM-spezifisch und nicht übergreifend standardisiert.
[7] Hierbei handelt es sich mathematisch um Vektoren in einem vieldimensionalen Vektorraum, z. B. [-0.74, 0.42, -0.53, ..., 0.02].
[8] Z. B. der Output über die Benutzerschnittstelle.
[9] Konkret ist die Tokenisierung hier [M][ia][Mü][ller][hat][gel][ogen][.].

2. Speicherung von Informationen in LLMs

Nirgends in dem Modell ist der Text oder das Token „Mia Müller“ vorhanden. Die einzelnen Tokens „M“, „ia“, „Mü“ und „ller“ sind für sich genommen reine Sprachfragmente. Die vektoriellen Bezüge der Tokens „Mü“ und „ller“ sind derart, dass vermutlich häufiger (jedenfalls in bestimmten Kontexten) auf „Mü“ das Token „ller“ folgt als z. B. das Token „he“ zur Erzeugung des Worts „Mühe“. Diese Beziehungen der Tokens untereinander – das Embedding – sind gerade die Leistung des Sprachmodells und machen seine Nützlichkeit letztlich aus. Als Kenngröße dient hierbei u. a. die Anzahl der sog. Parameter, die als Ergebnis des Trainingsprozesses die Beziehung der Tokens untereinander bestimmen. Aktuelle LLMs liegen hierbei in der Größenordnung von vielen Milliarden solcher Parameter.[10] Diese Beziehungen sind erlernte Größen eines LLM, können nicht im Einzelnen verstanden werden und lassen sich nicht gezielt anpassen, ohne die Funktion des gesamtes Modells zu gefährden.[11] Sie stellen den „Sinn“ der trainierten Texte dar, ohne die Texte selbst zu repräsentieren.[12] Sind in den Trainingsdaten entsprechend personenbezogene Daten enthalten, durchlaufen sie im Prozess des maschinellen Lernens eine Transformation, bei der sie in abstrakte mathematische Repräsentationen überführt werden. Dieser Abstraktionsprozess führt dazu, dass die konkreten Merkmale und Bezüge zu bestimmten Personen verloren gehen und stattdessen allgemeine Muster und Zusammenhänge erfasst werden, die sich aus der Gesamtheit der Trainingsdaten ergeben.

Dabei ist es eine etwas widersprüchlich erscheinende Eigenschaft solcher Modelle, dass sie einerseits die zum Training verwendeten Texte nicht in ihrer Ursprungsform speichern, sondern so verarbeiten, dass der Trainingsdatensatz in Gänze nie wieder aus dem Modell rekonstruiert werden kann. Andererseits verarbeiten LLMs diese Trainingstexte auf eine sehr spezifische und auf Kontextzusammenhängen basierende Art, die es später ermöglicht, ähnliche und damit oftmals nützliche Output-Texte zu erzeugen. Alles, was LLMs erzeugen, ist jedoch in dem Sinne „ausgedacht“, dass es keine einfache Wiedergabe von etwas Gespeichertem darstellt (wie z. B.

[10] Das LLM Llama 3 gibt es z. B. in einer Ausführung mit 8 Milliarden oder mit 70 Milliarden Parametern.
[11] Dies zeigt auch ein Forschungsprojekt von Anthropic, deren Anpassung des LLM-Features „Golden Gate Bridge“ dazu führte, dass Antworten eines Chatbots sich nur noch um diese Brücke drehten, obwohl der Prompt sie nicht erwähnte, zu Beispielen vgl. https://www.anthropic.com/news/golden-gate-claude; zum Forschungspapier Templeton et. al., 2024, https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html.
[12] In der aktuellen Diskussion wird gelegentlich vorgebracht, die Speicherung von Embeddings sei mit einer kryptografischen Verschlüsselung vergleichbar. Dies trifft nicht zu. Denn Verschlüsselung ist eine bijektive Abbildung zwischen Klar- und Chiffretext. Das bedeutet, jedem Element der Klartextmenge wird genau ein Element der Chiffretextmenge zugeordnet und umgekehrt. Mit dem richtigen Schlüssel lässt sich der Klartext eindeutig aus dem Chiffretext rekonstruieren. Die Informationen bleiben also vollständig erhalten, sind aber ohne Schlüssel nicht lesbar, solange das Verschlüsselungsverfahren nicht gebrochen wurde. Dies ist bzgl. der Trainingsdaten eines LLMs und ihrer abstrakten Repräsentation in den Embeddings nicht der Fall. Einen „Schlüssel“, der die Originaldaten der Trainingsphase vollständig und umfänglich wiederherstellt, gibt es nicht.

ein Eintrag in einer Datenbank oder ein Textdokument), sondern etwas neu Produziertes. Diese probabilistische Generierungsfähigkeit unterscheidet sich grundlegend von herkömmlichen Arten der Datenspeicherung oder des Datenabrufs.

II. Speicherung personenbezogener Daten im LLM

Der Begriff des personenbezogenen Datums in Art. 4 Nr. 1 DSGVO ist ein durch die EuGH-Rechtsprechung konkretisierter Rechtsbegriff. Er ist nicht gleichzusetzen mit dem allgemeinen Verständnis des Zusammenhangs eines Datums zu einer Person. Er setzt voraus, dass einem Datum eine Information über eine Person zugrunde liegt, die identifiziert oder identifizierbar ist. Dies wäre zum Beispiel bei einem Bibliotheksausweis der Fall, auch wenn dieser lediglich eine auf den Inhaber bezogene Nummer ausweist. Der Personenbezug kann nicht deshalb abgelehnt werden, weil die Bibliotheksausweisnummer bloß aus Zahlen besteht. Wenn die hinter der Bibliotheksausweisnummer stehende Person über weitere Hilfsmittel ausfindig gemacht werden kann, z. B. über das Bibliothekssystem, würde der EuGH für einen Zugriffsberechtigten annehmen, dass es sich bei der Zahlenfolge um ein personenbezogenes Datum handelt. Für die Identifizierung dürfen nach Auffassung des EuGH jedoch nur solche Mittel in Betracht kommen, die nicht gesetzlich verboten sind und deren Einsatz keinen praktisch unverhältnismäßigen Aufwand voraussetzt.[13]

Der EuGH hat zur Speicherung von personenbezogenen Daten in LLMs oder vergleichbaren Technologien noch keine Entscheidung getroffen. Unter Berücksichtigung der bisherigen EuGH-Rechtsprechung und der bisher bekannten Angriffsformen auf LLMs kommt der HmbBfDI zu dem Ergebnis, dass ein LLM keine personenbezogenen Daten im Sinne des Art. 4 Nr. 1, 2 DSGVO i. V. m. Erwägungsgrund 26 speichert.[14] Obwohl beobachtet wird, dass vereinzelt feingetunte LLMs durch Privacy-Attacken dazu gebracht werden, Trainingsdaten wiederzugeben, ist zweifelhaft, ob derartige Attacken den rechtlichen Schluss zulassen, dass im LLM personenbezogene Daten gespeichert werden.

[13] EuGH, Urt. v. 19.10.2016, C-582/14, Rn. 46.
[14] Nach Auffassung der dänischen Datenschutzaufsichtsbehörde enthalte ein KI-Modell als solches keine personenbezogenen Daten. Es sei lediglich das Ergebnis der Verarbeitung personenbezogener Daten. Dies folge daraus, dass ein statistischer Bericht ebenfalls nicht als personenbezogene Daten gelte, wenn der Bericht nur Schlussfolgerungen und aggregierte Daten enthalte, die das Ergebnis der statistischen Analyse seien (Leitfaden der dänischen Datenschutzaufsichtsbehörde zum Einsatz künstlicher Intelligenz, S.7, veröffentlicht im Oktober 2023, abrufbar unter: https://www.datatilsynet.dk/Media/638321084132236143/Offentlige%20myndigheders%20brug%20af%20kunstig%20intelligens%20-%20Inden%20I%20g%C3%A5r%20i%20gang.pdf).

1. Anknüpfung an (Embedded) Tokens im LLM

Die vom EuGH behandelten Konstellationen zum Personenbezug betrafen insbesondere IP-Adressen, Prüfungsantworten, behördliche Vermerke, Fahrzeugidentifizierungsnummern oder andere codierte Zeichenketten wie den TC-String.[15] Diese weisen einen Bezug zur Identifizierung einer bestimmten Person oder zu Personen zugeordneten Objekten aus. Es handelt sich um Kennungen, Identifier oder – nach dem EuGH – um „Informationen ‚über' die in Rede stehende Person".[16]

Dieser Bezug ergibt sich aus der Funktion dieser Kennungen sowie den darin enthaltenen Informationen.[17] IP-Adressen dienen der Zuordnung eines Gerätes, damit dessen Nutzende im Internet Daten senden und empfangen können. Sie haben eine Identifizierungsfunktion, indem sie eine Verbindung zwischen einer Online-Aktivität und einer physischen Person herstellen.[18] Prüfungsantworten und Prüferanmerkungen sollen eine konkret zu identifizierende Person und deren fachliche Kompetenz bewerten.[19] Ein behördlicher Vermerk über die Erfüllung von Aufenthaltsvoraussetzungen gibt detailliert Auskunft über die persönlichen Umstände und Eigenschaften von Antragstellenden, wie beispielsweise die Staatsangehörigkeit, den Familienstand oder die berufliche Situation.[20] Die Fahrzeugidentifizierungsnummer ermöglicht durch die Kodierung fahrzeugspezifischer Merkmale wie Hersteller, Typ, Ausstattung, Produktionsort und Herstellungsjahr die Identifizierung eines konkreten Fahrzeugs und mittelbar dessen Halter:in.[21] Mit dem TC-String werden konkrete Informationen über die Präferenzen und das Verhalten des Nutzenden kommuniziert, etwa für welche Verarbeitungszwecke und Anbieter:innen eine Person ihre Einwilligung erteilt oder verweigert hat.[22] All diese Identifier fungieren somit als Platzhalter für die Identität, Eigenschaften oder Verhältnisse einer bestimmten Person oder Sache einer Person; ihre Funktion ist die gezielte Zuordnung.

Im Gegensatz hierzu weisen einzelne Tokens als Sprachfragmente („M", „ia", „ Mü" und „ller") keinen individuellen Informationsgehalt auf und fungieren auch nicht als Platzhalter hierfür.

[15] Der Transparency and Consent String dient in der Online-Werbewirtschaft dazu, die Einwilligungen und Präferenzen von Nutzenden bezüglich der Datenverarbeitung zu Werbezwecken zu kodieren und an die beteiligten Parteien zu übermitteln.
[16] EuGH, Urt. v. 20.12.2017 – C-434/16, Rn. 34.
[17] EuGH, Urt. v. 20.12.2017 – C-434/16, Rn. 35; vgl. EuGH, Urt. v. 4.5.2023 – C-487/21, Rn. 24, EuGH, Urt. v. 8.12.2022 – C180/21, Rn. 70.
[18] Vgl. EuGH, Urt. v. 19.10.2016, C-582/14, Rn. 36.
[19] EuGH, Urt. v. 20.12.2017 – C-434/16, Rn. 38.
[20] Vgl. EuGH, Urt. v. 17.7.2014 – C-141/12, C-372/12, Rn. 48.
[21] EuGH, Urt. v. 9.11.2023 – C-319/22, Rn. 49.
[22] EuGH Urt. v. 7.3.2024 – C-604/22, Rn. 21.

Auch in der Beziehung dieser Tokens zueinander, den Embeddings, sind lediglich mathematische Repräsentationen des antrainierten Inputs gespeichert. Dass die Tokens „Mü“ und „ller“ in bestimmten Kontexten häufiger in Beziehung gesetzt werden als „Mü“ und „he“ ist noch keine Information über die Person Mia Müller, sondern über die sprachliche Funktion der Fragmente zueinander. Gespeichert sind damit hochgradig abstrahierte und aggregierte Datenpunkte aus den Trainingsdaten sowie deren Verhältnisse zueinander ohne konkrete Merkmale oder Bezüge „über“ natürliche Personen. Weder den einzelnen Tokens noch ihrem Embedding lassen sich ggf. im Trainingsdatensatz enthaltene Einzelinformationen über natürliche Personen entnehmen, die vergleichbar mit den Kennungen aus der EuGH-Rechtsprechung dazu dienen, diese Personen zu identifizieren. Daher kommt es für die Frage der Speicherung personenbezogener Daten in LLMs nach dem Maßstab der EuGH-Rechtsprechung nicht auf die Mittel eines etwaigen Verantwortlichen an. Es fehlt im LLM bereits an der Speicherung des notwendigen Identifiers im Sinne einer gezielten individuellen Zuordnung von Informationen, welche personenbezogene Daten in der Rechtsprechung des EuGH charakterisiert: der Information „über“ eine natürliche Person.

2. Auswirkungen von Privacy Attacks und PII Extraction

Bei LLMs, die durch Fine-Tuning für spezifische Aufgaben optimiert wurden, wird derzeit beobachtet, dass sie unter besonderen Umständen Trainingsdaten wiedergeben – auch solche über natürliche Personen.[23] Aus dieser Wiedergabe von Trainingsinhalten wird geschlossen, dass diese Informationen über natürliche Personen im LLM – trotz der Darstellung in Form von Tokens bzw. Embeddings – gespeichert („memorisiert“) sein müssten. Allerdings ist zweifelhaft, ob diese Art der Extrahierung auch den rechtlichen Schluss zulässt, dass im LLM personenbezogene Daten gespeichert werden.

Zunächst ist das bloße Vorhandensein plausibler Personeninformationen im LLM-Output kein zwingender Nachweis, dass personenbezogene Daten memorisiert worden sind, da LLMs in der Lage sind, zufällig mit Trainingsdaten übereinstimmende Texte zu generieren.

[23] Zum Komplex, vgl. Das et al., 2024, https://arxiv.org/html/2402.00888v1; solche Extraktionsattacken sind abzugrenzen von Model Inversion oder Membership Inference Attacken, die versuchen durch gezielte Anfragen an das LLM Rückschlüsse auf statistische Eigenschaften bzw. die Zugehörigkeit zu einem Trainingsdatensatz zu ziehen, ohne die Datensätze als solche zu rekonstruieren. Sie treffen keine Aussage über die etwaige Speicherung von Daten im LLM selbst, sondern über die Wahrscheinlichkeit, dass Daten mit bestimmten Eigenschaften zum Training bzw. ein bestimmtes Datenset verwendet worden sind, vgl. Maini et. al. 2024, https://arxiv.org/abs/2406.06443; zum Unterschied von Model Inversion Attacken und Membership Inference Attacken, vgl. Yang et. al. 2020, https://arxiv.org/pdf/2005.03915.

Außerdem erfolgt die Wiedergabe regelhaft nur durch gezielte Attacken auf LLMs,[24] etwa sog. „Privacy Attacks“ oder „PII Extraction“.[25] Nach der EuGH-Rechtsprechung können Daten aber nur dann als personenbezogen eingestuft werden, wenn die Identifizierung mit Mitteln des Verantwortlichen oder Dritter keinem gesetzlichen Verbot unterliegt bzw. nicht bloß mit einem unverhältnismäßigen Aufwand an Zeit, Kosten und Arbeitskräften möglich ist.[26] Die Durchführung solcher Attacken könnte dabei als praktisch unverhältnismäßiger Aufwand und ggf. gesetzlich verbotenes Mittel zur Ermittlung des Personenbezugs von im LLM gespeicherten Daten betrachtet werden.

Die Konzeption und Durchführung effektiver Privacy-Attacks auf LLMs erfordern erhebliche technische Expertise und zeitliche Ressourcen, über die der durchschnittliche Nutzende nicht verfügt.[27] Um zu verifizieren, ob die von einem LLM generierten Texte tatsächlich aus den Trainingsdaten extrahiert oder halluziniert worden sind, benötigen bisherige Privacy-Attacken Zugriff auf die ursprünglichen Trainingsdaten.[28] Erst durch den direkten Vergleich des LLM-Outputs mit diesen Daten lässt sich argumentieren, dass ggf. personenbezogene Daten aus dem Trainingsdatensatz gespeichert worden sind.[29] Allerdings sind Trainingsdatensätze von LLMs in der Regel nicht vollständig öffentlich zugänglich.[30] Gleichzeitig sind immense Mengen von Trainingsdaten erforderlich, um die Masse an möglichen Output-Varianten eines LLMs auf ihre Übereinstimmung mit dem Trainingsdatensatz zu überprüfen. Diese Grundvoraussetzungen einer Attacke sind dabei nicht nur tatsächlich aufwendig, sondern auch kostenintensiv,[31] sodass es sich um einen praktisch unverhältnismäßigen Aufwand handelt, der gegenwärtig nur zur

[24] Zu übrigen Angriff-Varianten vgl. Das et al., 2024, https://arxiv.org/html/2402.00888v1.
[25] Wobei das Akronym „PII“ Personable Identifiable Information einen in unterschiedlichen US-Gesetzen verwendeten Begriff meint, der nicht mit dem Begriff des personenbezogenen Datums nach der DSGVO gleichgesetzt werden kann.
[26] EuGH, Urt. v. 19.10.2016, C-582/14, Rn. 46.
[27] Das gilt umso mehr, als dass bekannte Angriffe aus der Vergangenheit regelhaft von Entwickler:innen aufgegriffen werden, um weitere Schutzmaßnahmen zu ergreifen und es dadurch erhebliche Expertise braucht, um neue Angriffe zu entwickeln, vgl. hierzu auch weiter unten.
[28] Vgl. Das et al., 2024, https://arxiv.org/html/2402.00888v1.
[29] Solche Attacken, welche ohne Vergleich mit den Trainingsdaten erfolgen, sind ebenfalls aufwendig – nicht zuletzt, weil Modellentwickler:innen Schutzvorkehrungen implementieren – und stellen u. U. ein nach deutschem Recht verbotenes Mittel dar.
[30] Shi et al, 2024, „Although large language models (LLMs) are widely deployed, the data used to train them is rarely disclosed”, https://arxiv.org/html/2310.16789v3, auch bei als „Open Source” bezeichneten LLMs werden nicht alle Trainingsdaten offengelegt: „for Llama, the corporate preprint notes that fine-tuning was done based on “a large dataset of over 1 million binary comparisons based on humans applying our specified guidelines, which we refer to as Meta reward modeling data”, and which remains undisclosed“, Liesenfeld/Dingemanse, Rethinking open source generative AI, open-washing and the EU AI Act, https://pure.mpg.de/rest/items/item_3588217_2/component/file_3588218/content.
[31] Vgl. Das et al., 2024, https://arxiv.org/html/2402.00888v1; die Aussage eines anderes Forschungsprojekts, eine PII-Attacke habe nur Kosten von 200 USD ausgelöst, betrifft letztlich nur die Kosten der immerhin 25 Millionen Queries, nicht aber die fachlichen Expertise, den Zeitaufwand, die hierfür erforderliche Vergütung sowie den Vergleich der extrahierten Daten mit neun Terabyte an Teil-Trainingsdatensets, vgl. Nasr et al, 2023, https://arxiv.org/pdf/2311.17035.

wissenschaftlichen Erforschung der Funktionsweise von LLMs getätigt wird.[32] Außerdem sind derartige Attacken auch häufig Anlass für LLM-Entwickler:innen, an der Verbesserung von Schutzmaßnahmen zu arbeiten, um solche Extrahierungen zu erkennen und zu unterbinden.[33] Werden derartige technische Schutzmaßnahmen bewusst überwunden, ist zudem nicht auszuschließen, dass es sich nach deutschem Recht um verbotene Mittel handeln könnte.[34]

III. Praktische Folgen

Die These, dass LLMs keine personenbezogenen Daten speichern, kann weitreichende Folgen für die Praxis haben. Das illustrieren folgende Fallbeispiele:

1. Folgen eines rechtswidrigen Trainings von LLMs

Ein Unternehmen oder eine Behörde setzt ein von einem Dritten entwickeltes LLM ein. Später stellt sich heraus, dass der Dritte beim Training des Modells personenbezogene Daten verwendet hat, ohne dass hierfür eine Rechtsgrundlage gegeben war.

Eine eventuelle Rechtswidrigkeit[35] beim Training eines Modells wirkt sich nicht auf die Rechtmäßigkeit des Einsatzes eines solchen Modells aus. Datenschutzverstöße beim Training eines LLMs sind nicht dem Verantwortlichen zuzurechnen, welcher das LLM einsetzt, sondern ausschließlich der Entwickler:in des Modells. Letzterer hat ebenso wie Unternehmen und Behörden, welche LLMs einsetzen und nachtrainieren wollen, datenschutzrechtliche Vorgaben zu beachten.

2. Bedeutung für Betroffene

Eine Person gibt ihren Namen in den LLM-basierten Chatbot eines Unternehmens oder einer Behörde ein. Der LLM-basierte Chatbot gibt falsche Informationen über sie aus. Welche Betroffenenrechte kann sie in Bezug zu welchem Gegenstand geltend machen?

[32] Auch in der Forschung wird konstatiert: „Memorization Can Be Hard to Discover.", Carlini et al., 2021, https://www.usenix.org/system/files/sec21-carlini-extracting.pdf.
[33] Vgl. zu geeigneten Praktiken des Red Teamings als Methode zur (Fort-)Entwicklung von Schutzmaßnahmen, Feffer et al https://arxiv.org/pdf/2401.15897; zu gegenwärtigen Limitationen solcher Privacy Attacks durch Schutzvorkehrungen, Das et. al., 2024, https://arxiv.org/html/2402.00888v1.
[34] Vgl. § 202a StGB; zu den rechtlichen Risiken sowie ethischen Vorbehalten von Privacy Attacks, vgl. Carlini et. al, 2021, Extracting Training Data from Large Language Models, https://arxiv.org/abs/2311.17035; Nasr et al., 2023, https://arxiv.org/abs/2311.17035.
[35] Z. B. aufgrund fehlender Rechtsgrundlage für die damit verbundene Verarbeitung personenbezogener Daten in den Trainingsdaten.

Unternehmen und Behörden müssen sicherstellen, dass bei der Verarbeitung personenbezogener Daten die Anforderungen der DSGVO erfüllt werden. Da in einem LLM keine personenbezogenen Daten gespeichert sind, kann es nicht selbst Gegenstand der Betroffenenrechte aus Artt. 12 ff. DSGVO sein.[36] Wenn in einem KI-System im Übrigen personenbezogene Daten verarbeitet werden, insbesondere bei dessen Output oder etwaigen Datenbankabfragen, hat der Verantwortliche die Betroffenenrechte zu erfüllen.
Dies bedeutet für den oben skizzierten Fall, dass die betroffene Person von dem Unternehmen oder der Behörde jedenfalls in Bezug auf den Input und den Output des LLM-basierten Chatbots verlangen kann,

- dass ihr eine Auskunft gemäß Art. 15 DSGVO erteilt wird,
- dass die sie betreffenden personenbezogenen Daten gemäß Art. 16 DSGVO berichtigt werden,
- ggf., dass die sie betreffenden personenbezogenen Daten gemäß Art. 17 DSGVO gelöscht werden.

3. Anforderungen an Fine-Tuning von LLMs

Ein Unternehmen oder eine Behörde möchte das von einem Dritten entwickelte LLM mit eigenen Trainingsdaten für einen konkreten Nutzungszweck nachtrainieren.

Für das Unternehmen oder die Behörde bedeutet das insbesondere:

- Es ist empfehlenswert, dass in den Trainingsdaten möglichst keine personenbezogenen Daten enthalten sind. Synthetische Daten sind – soweit sie zum Training geeignet sind – zu bevorzugen.[37]
- Soweit dennoch Daten mit Personenbezug für das Fine-Tuning verwendet werden sollen, müssen Unternehmen und Behörden dies auf eine Rechtsgrundlage[38] stützen und sicherstellen, dass die Betroffenenrechte erfüllt werden können.

[36] Ebenso sind die Verarbeitungsgrundsätze des Art. 5 DSGVO, insbesondere der Grundsatz der Datenrichtigkeit, gemäß Art. 5 Abs. 1 lit. d DSGVO, auf ein LLM selbst nicht anwendbar.
[37] Dies leitet der Europäische Datenschutzbeauftragte aus dem Grundsatz der datenschutzgerechten Technikgestaltung („Privacy by Design“) her, https://www.edps.europa.eu/press-publications/publications/techsonar/synthetic-data_en, vgl. auch Art. 25 Abs. 1 DSGVO.
[38] Zu den möglichen Rechtsgrundlagen vgl. das Diskussionspapier „Rechtsgrundlagen im Datenschutz beim Einsatz von Künstlicher Intelligenz“ des Landesbeauftragten für Datenschutz und Informationsfreiheit Baden-Württemberg, abrufbar unter: https://www.baden-wuerttemberg.datenschutz.de/rechtsgrundlagen-datenschutz-ki/.

4. Anforderungen an den lokalen LLM-Betrieb

Ein Unternehmen oder eine Behörde möchte ein lokal betriebenes LLM einsetzen, um über ein Webinterface ein internes Wissensmanagement-Tool einzusetzen.

Für das Unternehmen oder die Behörde bedeutet das insbesondere:

1. Die Speicherung eines LLMs auf dem Server eines Unternehmens oder einer Behörde ist datenschutzrechtlich nicht relevant.
2. Allerdings muss das einzusetzende KI-System jedenfalls in Bezug auf seinen Input und Output die Erfüllung von Betroffenenrechten ermöglichen.
3. Verantwortliche sollten sicherstellen, dass Extrahierungen, etwa durch Privacy Attacks und PII Extraction unterbunden werden.[39]
    a. Die von der LLM-Entwickler:in bereitgestellten Schutzmaßnahmen („safeguards“) sollten umgesetzt werden.
    b. Neben den Schutzmaßnahmen der Entwickler:in sollte der Verantwortliche eigene Maßnahmen, wie z. B. Filter, ergreifen, um Privacy Attacks und PII Extraction zu verhindern.

5. Anforderungen an den Betrieb von Drittanbieter-LLMs

Ein Unternehmen oder eine Behörde schließt einen Vertrag mit einer Drittanbieter:in über die Zurverfügungstellung eines LLM etwa über eine Programmierschnittstelle (API), um Mitarbeitern über ein Webinterface Textzusammenfassungen zu ermöglichen.

Für das Unternehmen oder die Behörde bedeutet das insbesondere:

1. Das einzusetzende KI-System muss jedenfalls in Bezug auf seinen Input und Output die Erfüllung von Betroffenenrechten ermöglichen.

[39] Aber auch hier bleibt zweifelhaft, ob ggf. extrahierbare Datensätze personenbezogene Daten darstellen, vgl. oben III. 2.

E-Mail: mailbox@datenschutz.hamburg.de
Ludwig-Erhard-Straße 22 - D-20459 Hamburg - Tel.: 040 - 4 28 54 - 40 40 - Fax: 040 - 4 28 54 – 40 00

2. Bei der Auswahl der Anbieter:in sollte darauf geachtet werden, dass Schutzmaßnahmen zur Verhinderung von Privacy Attacks und PII Extraction vorhanden sind.
3. Vor der Inbetriebnahme des LLMs sind die Verantwortlichkeiten zu klären (Auftragsverarbeitung, gemeinsame Verantwortlichkeit oder selbstständige Verantwortlichkeiten).[40]

[40] Vgl. hierzu DSK Orientierungshilfe v. 06.05.2024, „Künstliche Intelligenz und Datenschutz“, Rn. 32 ff., https://www.datenschutzkonferenz-online.de/media/oh/20240506_DSK_Orientierungshilfe_KI_und_Datenschutz.pdf.