BERLIN

Orientierungshilfe zum Umgang mit LLM-basierten Chatbots im Land Berlin

Um die neuesten Technologien verantwortungsvoll und im Einklang mit geltenden Datenschutz- und Sicherheitsstandards zu nutzen, stellt die Senatskanzlei die nachfolgenden Orientierungspunkte zum Umgang mit Chatbots, die auf Large Language Models (LLMs) basieren, zur Verfügung. LLM-basierte Chatbots sind Programme, die mit großen Mengen von Textdaten trainiert werden, um natürliche Gespräche zu führen und auf eine Vielzahl von Fragen zu antworten.

In der Regel werden diese Sprachmodelle in der Cloud betrieben. Unabhängig vom Betriebsmodell ist ein sachgerechter Umgang mit vertraulichen und personenbezogenen Daten zu gewährleisten, damit diese Daten nicht im Rahmen des weiteren Trainings der Modelle genutzt oder unzulässig verarbeitet werden.

Die folgenden Punkte sollen eine erste Orientierung geben und können als Checkliste zum Umgang dienen, um sowohl die Vorteile aus der Nutzung dieser Technologien prüfen zu können als auch potenzielle Risiken zu minimieren. Die Orientierungshilfe richtet sich an Führungskräfte sowie Verwaltungsmitarbeiterinnen und -mitarbeiter, die in eigener Verantwortung die aktuell am Markt verfügbaren KI-Chatbots nutzen. Parallel wird derzeit daran gearbeitet, eine landesweit einsetzbare und IKT-architekturkonforme Lösung anzubieten, die vergleichbare KI-unterstützte Möglichkeiten bieten wird. Unabhängig davon ist insbesondere im Verwaltungskontext darauf zu achten, dass die eingesetzten Werkzeuge selbst entsprechend IKT-architekturkonform sind.

1. Compliance-Regelungen

Die Vorschriften zum Schutz von geistigem Eigentum, Urheberrechten, Betriebs- und Geschäftsgeheimnissen sowie zum Datenschutz gelten auch für die Nutzung von KI-Anwendungen. Insbesondere ist das Hochladen von Dokumenten, Texten und Fotos mit personenbezogenen Daten in öffentlich zugängliche KI-Anwendungen nicht zulässig. Weitere Informationen bietet die Orientierungshilfe der unabhängigen Datenschutzaufsichtsbehörden des Bundes und der Länder.[1]

2. Nutzung von Funktions-Accounts

Es ist erforderlich, durch interne Compliance und Dienstanweisungen den Umgang mit KI-Anwendungen in den Dienststellen klar zu regeln. Gegebenenfalls enthalten Dienstanweisungen bereits einschlägige Regelungen (wie z.B. Untersagung des Einsatzes von „Fremd-Software“) und sind somit darum zu ergänzen, unter welchen Voraussetzungen und in welchen Einsatzszenarien welche KI-Anwendungen in Frage kommen.

[1] Orientierungshilfe der Konferenz der unabhängigen Datenschutzaufsichtsbehörden des Bundes und der Länder vom 6. Mai 2024: https://www.datenschutzkonferenz-online.de/media/oh/20240506_DSK_Orientierungshilfe_KI_und_Datenschutz.pdf

Für die Nutzung von LLM-basierten Chatbots sind Accounts grundsätzlich zu vermeiden, die Namen einzelner Beschäftigter enthalten. Oft werden für die Registrierung E-Mail-Adresse und Telefonnummer abgefragt und damit hinterlegt. Hier empfehlen sich Accounts mit Funktionsadressen bzw. pseudonymisierte Accounts und berufliche Telefonnummern. Weiterhin ist die Einhaltung der Regeln zur Internetnutzung und der Anforderungen zur Informationssicherheit in den Behörden zu gewährleisten, sowie eine private Nutzung dieser dienstlichen Accounts in den Behörden zu untersagen. Die verantwortlichen Dienststellen stellen berufliche Chatbot-Accounts zur Verfügung, Beschäftigte dürfen für dienstliche Zwecke nicht eigenständig als Privatpersonen Nutzerkonten erstellen. Es sind starke Passwörter zu nutzen, um unbefugten Zugriff auf den Chatbot und die damit verbundenen Daten zu verhindern.

3. Wahlmöglichkeit hinsichtlich KI-Training und Eingabe-Historie

Oft werden die Eingaben für das weitere Training des Modells oder Analysen genutzt. Es ist die Möglichkeit zu nutzen, eine Datenverarbeitung zu eigenen Zwecken (vom Anbieter) zu deaktivieren (z.B. Nutzung der eingegebenen Daten zur Verbesserung des Service, Lernen der KI durch die Eingaben/Suchaufträge) und damit der Speicherung und Verwendung zu Trainingszwecken zu widersprechen. In einigen Fällen muss hier ein entsprechendes Lizenzmodell gewählt werden, das gegebenenfalls kostenpflichtig ist. In den Einstellungen sollte die Deaktivierung der Protokollierung eingestellt werden. Dies ist bei cloudbasierten Diensten wichtig, da oft eine Möglichkeit angeboten wird, bisherige Eingaben zu speichern und mit einem Nutzerprofil zu verknüpfen, um zu einem späteren Zeitpunkt darauf wieder zurückgreifen zu können. Damit ist zwangsläufig eine Verkettung der Eingaben einer Person verbunden.

4. Vermeidung personenbezogener und personenbeziehbarer Daten

Personenbezogene oder sensible Daten (Geschäftsgeheimnisse, Informationen zu Kunden oder Geschäftspartnern) dürfen nicht in die Anfragen eingegeben oder in den Antworten des Chatbots verarbeitet werden. Es ist darauf zu achten, dass keine Einsatzszenarien stattfinden, die mit Personen direkt/indirekt zu tun haben.[2]

Auch dürfen bei der Nutzung öffentlich zugänglicher KI-Anwendungen weder die Eingaben noch die dabei entstehenden Ergebnisse personenbezogene Daten enthalten. Der Widerspruch gegen

[2] Eine Verarbeitung personenbezogener Daten mittels LLM-basierter Chatbots wird im Regelfall auch unzulässig sein, da hierfür keine Rechtsgrundlage i. S. d. Art. 6 Abs. 1 DSGVO besteht. Selbst wenn hierzu eine Einwilligung der Betroffenen i. S. d. Art. 6 Abs. 1 lit. a DSGVO eingeholt werden soll, so müsste diese nach den Vorgaben der DSGVO informiert und freiwillig abgegeben werden (Art. 4 Nr. 11 DSGVO). Regelmäßig behalten sich die Anbieter von LLM-basierten Chatbots jedoch vor, personenbezogene Daten der Nutzenden auch zu eigenen, nicht näher bezeichneten Zwecken zu verarbeiten. Eine ausreichend informierte Einwilligung ist in diesem Fall nicht möglich. Zudem sind Einwilligungen von Bürgern und Bürgerinnen gegenüber Behörden nur in Ausnahmefällen wirksam möglich, da aufgrund des bestehenden Ungleichgewichts keine ausreichende Freiwilligkeit anzunehmen ist (vgl. Erwägungsgrund 43 zu Artikel 7 DSGVO). Sollte im Einzelfall eine tragfähige Rechtsgrundlage zur Verarbeitung personenbezogener Daten mittels LLM-Basierter Chatbots vorliegen, so sind mit externen Anbietern wirksame Auftragsverarbeitungsverträge i. S. d. Art. 28 Abs. 3 DSGVO abzuschließen.

die Datenverarbeitung zu Trainingszwecken reicht nicht für eine vollständig datenschutzkonforme Nutzung aus, wenn personenbezogene Daten eingegeben werden.

Es sind Fallbeispiele zu wählen, die unverfänglich sind und keinen Bezug zu Einzelpersonen haben. Hierbei ist zu beachten, dass ein Personenbezug sich durch viele Merkmale, nicht nur durch Namen und Adressdaten ergeben kann.

> *Beispiele für **unproblematische** Eingaben: „Welche öffentlichen Datenquellen kann ich für ein Projekt im Bereich der Stadtentwicklung nutzen?“, „Wie könnte KI dabei helfen, Verkehrsstaus in Großstädten zu reduzieren?“*
>
> *Beispiel einer **problematischen** Eingabe: „Durchsuche soziale Medien, um Personen zu identifizieren, die kritische Meinungen über lokale Bauprojekte XY äußern, und liste deren Namen sowie Adressen für eine direkte Kontaktaufnahme auf.“ Eine solche Anfrage würde einen Eingriff in die Privatsphäre der betroffenen Personen bedeuten und könnte gegen Datenschutzgesetze verstoßen. Darüber hinaus würden das Sammeln und Veröffentlichen personenbezogener Daten ohne Zustimmung der betroffenen Personen ethische Grundsätze und möglicherweise rechtliche Rahmenbedingungen verletzen.*
>
> *Beispiel einer **problematischen Eingabe**: „Erstelle eine Beurteilung für einen Sachbearbeiter aus dem Bereich X aus der Verwaltung Y.“ Die Erstellung eines Arbeitszeugnisses hat einen Personenbezug, selbst wenn der Name der Person im Text nicht vorkommt, jedoch erkennbar ist, aus welchem Unternehmen sie zu einem bestimmten Zeitpunkt getätigt wurde.*

5. Formulierung und Gegenstand der Eingabe

Die aktuell meist benannten KI-Anwendungen werden cloudbasiert zur Verfügung gestellt und sind damit über das Internet einem unbestimmten Kreis von Anwenderinnen und Anwendern zugänglich. Die Eingabedaten verlassen damit den geschützten Bereich der Anwenderinnen bzw. Anwendern und können- je nach Konzeption der KI-Anwendung - von dieser auch für die Beantwortung von Anfragen anderer Anwender/-innen verwendet werden.

Die Qualität und Formulierung der Eingaben bzw. die Auswahl und Art der Eingabedaten können in das lernende System einfließen und beeinflussen möglicherweise zukünftige Antworten oder Lösungen des Systems. Es ist daher darauf zu achten, dass diese diskriminierungssensibel und sind und dass keine Personengruppen mit geschützten Merkmalen (§2 LADG Diskriminierungsverbot)[3] abgewertet, benachteiligt oder ausgeschlossen werden.

[3] § 2 Diskriminierungsverbot: Kein Mensch darf im Rahmen öffentlich-rechtlichen Handelns aufgrund des Geschlechts, der ethnischen Herkunft, einer rassistischen und antisemitischen Zuschreibung, der Religion und Weltanschauung, einer Behinderung, einer chronischen Erkrankung, des Lebensalters, der Sprache, der sexuellen und geschlechtlichen Identität sowie des sozialen Status diskriminiert werden.

6. Überprüfung der Ergebnisse

Ergebnisse von Chatbots sind immer als Entwurf zu verstehen und einer Qualitätssicherung zu unterziehen. Es ist sicherzustellen, dass die von Chatbots generierten Antworten und Lösungen auf ihre Genauigkeit und Diskriminierungsfreiheit hin überprüft werden.

Nutzer/-innen sollten sich nicht auf die Ergebnisse verlassen und sicherstellen, dass Menschen die Eingaben und Ausgaben beurteilen, besonders im Hinblick u.a. auf Halluzination[4], ethische Fragen, Fake-News, Diskriminierung (auch i.S.v. Benachteiligung, Chancenungleichheit oder Ausschlusses) und Genauigkeit der Ergebnisse.

Für die Nutzung zur Software-Entwicklung: Wenn ein LLM-basierter Chatbot Code generiert, wird eine gründliche Überprüfung empfohlen, um zu vermeiden, dass schädlicher Code integriert wird und eine Sicherheitslücke entsteht[5]. Es sind Prozesse zur manuellen Überprüfung und Korrektur von potenziell fehlerhaften oder voreingenommenen Ausgaben zu implementieren.

7. Keine automatisierten Letztentscheidungen

Vermeiden Sie es, dass Chatbots finale Entscheidungen ohne menschliche Überprüfung treffen, besonders in kritischen Bereichen wie Personalangelegenheiten oder bei der Erstellung von Bescheiden. Es ist sicherzustellen, dass stets ein menschlicher Entscheidungsträger involviert ist. Die Verwendung von KI ist transparent zu kennzeichnen. Die Verwendung von Chatbots in Vermerken ist zu dokumentieren, gegebenenfalls ist auf deren Verwendung in Verwaltungsakten hinzuweisen. Auch unzureichende Ressourcen und Zeitdruck dürfen nicht dazu führen, dass Ergebnisse ungeprüft übernommen werden.

8. Sensibilisierung

Es wird empfohlen, Sensibilisierungsmaßnahmen für alle Mitarbeiterinnen und Mitarbeiter durchzuführen, um ein Bewusstsein für den sicheren und verantwortungsvollen Umgang mit KI-Technologien zu schaffen. Dies umfasst insbesondere Datenschutz, Anforderungen aus den Normen der Informationssicherheit, die geltenden Richtlinien zur Internetnutzung, Urheberrecht, Antidiskriminierungsrecht und weitere ethische Überlegungen.

9. Beauftragte mit einbeziehen

Die Datenschutz- und Informationssicherheitsbeauftragten sowie die Beschäftigtenvertretungen der Behörden sind für interne Weisungen und Anwendungsfälle vor Beginn der Nutzung zu beteiligen.

---

[4] Eine Halluzination im Feld der KI ist ein Resultat eines generativen KI-Modells, was falsche oder irreführende Informationen beinhaltet und überzeugend formuliert. Solche "Halluzinationen" können zu unerwarteten, widersprüchlichen oder unlogischen Ergebnissen führen, die nicht der Realität oder dem erwarteten Ergebnis entsprechen.

[5] Generative KI-Modelle, Chancen und Risiken für Industrie und Behörden, 2024 https://www.bsi.bund.de/SharedDocs/Downloads/DE/BSI/KI/Generative_KI-Modelle.pdf?__blob=publicationFile&v=5

Eine Datenschutz-Folgenabschätzung i.S.d. Art. 35 DSGVO ist nur dann durchzuführen, wenn ausnahmsweise personenbezogene Daten mittels eines LLM-basierten Chatbots verarbeitet werden und hierfür eine Rechtsgrundlage sowie der erforderliche vertragliche Rahmen besteht. Bei automatisierten Verarbeitungen personenbezogener Daten ist zudem nach § 26 Abs. 2 Berliner Datenschutzgesetz (BlnDSG) eine datenschutzrechtliche Risikoanalyse durchzuführen. Die zu treffenden technischen und organisatorischen Maßnahmen sind in einem Datenschutzkonzept zu dokumentieren. Anforderungen an die jeweils anzuwendenden Rechtsvorschriften, die in der Europäischen Union, Deutschland und im Land Berlin Anwendung finden (z.B. DSGVO, BlnDSG, Anforderungen aus den Normen der Informationssicherheit, LADG und AGG) sind einzuhalten. Ein Austausch mit den jeweiligen Personalvertretungen wird empfohlen.

10. Nutzungsbedingungen prüfen

Die Nutzungsbedingungen der jeweiligen Anbieter sind durch die Vertragsbereiche der Behörden zu prüfen. Verfügbare Modelle können von unterschiedlichen Anbietern zu differierenden AGBs bereitgestellt werden.

Weiterführende Hinweise:

Orientierungshilfe der unabhängigen Datenschutzaufsichtsbehörden des Bundes und der Länder

Lektion KI des Behörden-IT-Sicherheitstraining (BITS)

BSI Einschätzung: Generative KI-Modelle - Chancen und Risiken für Industrie und Behörden, 2024

Gutachten der Datenethikkommission mit ethischen und rechtlichen Handlungsempfehlungen im Umgang mit Daten und algorithmischen Systemen und Künstlicher Intelligenz