Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

26. April 2024 V1.0[1]

Merkblatt zur Sensibilisierung betreffend grossen KI-Sprachmodellen in der Bundesverwaltung

Was sind grosse KI-Sprachmodelle?

Grosse KI-Sprachmodelle («*Large Language Models*», LLMs) sind KI-Technologien, die den gebräuchlichsten KI-Systemen[2] im Bereich der natürlichen Sprache zu Grunde liegen. Sie sind die Basis vieler generativer KI-Werkzeuge.[3]
LLMs sind in der Lage, natürliche Sprache zu verarbeiten, zu verstehen, zu interpretieren und zu generieren, und können eine Vielzahl von Aufgaben ausführen, wie zum Beispiel Übersetzung, Textverständnis, Textgenerierung. Die modernsten LLMs sind mit grossen Datenmengen in unterschiedlichsten Formaten trainiert und können Texte produzieren, die oft nicht ohne Weiteres von menschengeschriebenen Texten zu unterscheiden sind. Es sind allerdings auch akustische Eingaben oder Bildeingaben denkbar, da diese inzwischen in vielen Fällen nahezu fehlerlos in Text konvertiert werden können. Auch akustische Sprachausgaben sind kaum mehr von menschlichen Stimmen unterscheidbar. Einige LLMs werden bereits zu sogenannten multimodalen Modellen, die z.B. neben Text auch Bilder und Videos verarbeiten und generieren können, erweitert.[4]
Beispiele für LLMs sind die GPT-Modellreihen (die in ChatGPT von OpenAI und Copilot von Microsoft verwendet werden), Gemini (verwendet in Googles Gemini, ehemals Bard), Metas LLaMA-Modelle, die Grok-Modellreihen von X und Anthropics Claude-Modelle.[5] LLMs erwerben die Fähigkeit zur Interpretation und Generierung von natürlicher Sprache für allgemeine Zwecke durch das «Erlernen» statistischer Beziehungen aus Textdokumenten während eines rechenintensiven iterativen Trainingsprozesses. Diese statistischen Modelle basieren auf Techniken und Methoden zur Verarbeitung natürlicher Sprache («*Natural Language Processing*», NLP[6]) und extrahieren so die Bedeutung und Zusammenhänge aus menschlicher Sprache.

Grundsätze zu grossen KI-Sprachmodellen (LLMs) in der Bundesverwaltung[7]

Als Nutzer oder Nutzerinnen in der Bundesverwaltung von generativen KI-Technologien (wie LLMs) sind Sie eingeladen, folgende Grundsätze für ein verantwortungsvolles Experimentieren zu beachten. Generative KI-Technologien (wie LLMs) können Sie bei Ihrer täglichen Verwaltungstätigkeit unterstützen. LLMs sind in der Lage, eine Vielzahl von textbasierten Aufgaben teil- oder vollautomatisiert zu übernehmen. Probieren Sie diese aus, lernen

---

1 Dieses Merkblatt wird regelmässig einer Überprüfung unterzogen, um neue Entwicklungen und ein besseres Verständnis der generativen KI-Technologien (wie grossen KI-Sprachmodellen) in der Bundesverwaltung zu berücksichtigen.

2 Ein «KI-System» ist ein maschinenbasiertes System, das für explizite oder implizite Ziele aus den empfangenen Inputs schlussfolgert, wie es Outputs wie Vorhersagen, Inhalte, Empfehlungen oder Entscheidungen erzeugen kann, welche die physische oder virtuelle Umgebung beeinflussen können. KI-Systeme können mit unterschiedlichem Ausmass an Autonomie ausgestattet werden. Eine «KI-Technologie» bezeichnet einzelne, in Computer implementierbare Funktionen für die Erreichung von KI (z.B. «maschinelles Lernen»). Ein KI-System bezeichnet somit eine strukturierte, kontextgebundene Kombination von KI-Technologien zwecks Erreichens von KI. *Quelle:* https://cnai.swiss/dienstleistungen/terminologie/

3 «Generative KI» ist ein weit gefasster Begriff, der sich auf KI-Systeme bezieht, die auf grosse Mengen von Daten aus der physischen und virtuellen Welt trainiert werden, um selbst Daten zu generieren (z.B. Texte, Bilder, Tonaufnahmen, Videos, Simulationen, Codes). Sie sind oft multimodal, z.B. mit Inputs und/oder Outputs in einer oder mehreren Modalitäten (z.B. Text, Bild, Video). *Quelle:* https://cnai.swiss/dienstleistungen/terminologie/

4 Diese sind auch als grosse multimodale KI-Modelle («*Large Multimodal Models*», LMMs) mit generativen Fähigkeiten bekannt (https://huyenchip.com/2023/10/10/multimodal.html).

5 Siehe «Merkblatt zur Verwendung von generativen KI-Werkzeugen in der Bundesverwaltung»; https://cnai.swiss/dienstleistungen-weitere-dienstleistungen-merkblaetter-zu-ki/

6 «*Natural Language Processing*» (NLP) ist ein Teilgebiet der KI, welches sich mit der Analyse, dem Verständnis und der Generierung von geschriebenen und gesprochenen Wörtern und Sätzen (natürlicher Sprache) beschäftigt. Die meisten NLP Techniken und Methoden basieren auf «maschinellem Lernen» und extrahieren so die Bedeutung und Zusammenhänge aus menschlicher Sprache. Anwendungsgebiete sind z.B. Texterkennung («*Text Recognition*»), Spracherkennung («*Speech Recognition*»), Bots, «*Chatbots*» und digitale Assistenten. *Quelle:* https://cnai.swiss/dienstleistungen/terminologie/

7 Inspiriert vom «*Generative AI Framework for UK Government*» (vom 18. Januar 2024); siehe https://www.gov.uk/government/publications/generative-ai-framework-for-hmg

Sie dazu! Mit etwas Kreativität tragen Sie so zu einer innovativen Verwaltung bei. Gehen Sie dabei aber vorsichtig und kritisch-reflexiv vor, und beachten Sie die geltenden Vorgaben.

1) Sie wissen, was generative KI ist, was LLMs sind und wo deren Grenzen liegen.
2) Sie setzen generative KI-Technologien (wie LLMs) rechtmässig, ethisch und verantwortungsbewusst ein. Sie kontrollieren, ob eine ausreichende Rechtsgrundlage für den geplanten Einsatz besteht. Die Anforderungen an die gesetzliche Grundlage unterscheiden sich, ob Sie generative KI-Technologien (wie LLMs) nur zur Unterstützung einer Übersetzung, zur stärkeren Unterstützung beim Verfassen von Texten oder für eine vollautomatische Entscheidung einsetzen.
3) Sie wissen, wie Sie den Einsatz generativer KI-Technologien (wie LLMs) sicher halten können.[8]
4) Sie haben eine sinnvolle menschliche Kontrolle in der richtigen Phase (von der Entwicklung bis zur Anwendung von generativen KI-Technologien wie LLMs).
5) Sie verwenden KI im Allgemeinen und LLMs im Speziellen transparent. Von LLMs erstellter Text kann den Anschein erwecken, von einem Menschen geschrieben worden zu sein. Als Bundesangestellte haben Sie sowohl in der internen als auch in der externen Kommunikation eine besondere Verantwortung. Die Kennzeichnung von durch KI erzeugtem Text wird empfohlen (z.B. «Dieser Text wurde mit Unterstützung von KI erstellt, redaktionell bearbeitet und geprüft.»). Die Kennzeichnung ist im Falle einer automatisierten Einzelentscheidung im Sinne von Art. 21 DSG[9] sogar obligatorisch.
6) Sie wenden diese Grundsätze zusammen mit den Leitlinien für den Umgang mit KI in der Bundesverwaltung und den Grundprinzipien des Bundes einer menschenzentrierten und vertrauenswürdigen Datenwissenschaft (und KI) an, und sorgen für die nötige (Qualitäts-)Sicherung.

Grenzen von grossen KI-Sprachmodellen (LLMs)

LLMs erstellen ihre Ausgaben/Ergebnisse aufgrund der statistischen Wahrscheinlichkeiten für das nächste Wort, nicht unter Berücksichtigung des Wahrheitsgehalts. Dies kann zu Faktizitäten und Halluzinationen führen, also Ausgaben/Ergebnisse, die nicht der Realität entsprechen. Dies lässt sich zwar einschränken, indem dem Modell für die Antwortfindung ein Kontext (z.B. eine Menge an Textdokumenten) zur Verfügung gestellt wird. Allerdings kann es auch hier zu Halluzinationen kommen, wenn die Antworten Inhalte ausserhalb dieses Kontexts beinhalten. Weitere Grenzen von LLMs im Rahmen der ordnungsgemässen Nutzung beinhalten beispielsweise:[10]

- unerwünschte Ausgaben/Ergebnisse, wörtliches Erinnern und Verzerrungen im Modell (Bias) resultierend aus der unausgewogenen Zusammenstellung und Inhalten der Trainingsdaten (z.B. persönliche oder urheberrechtlich geschützte Daten sowie Texte mit fragwürdigen, falschen oder diskriminierenden Inhalten);
- fehlende Aktualität, da LLMs Text generieren auf Basis der verarbeiteten Trainingsdaten, welche sich zwangsweise auf Inhalte beschränken, die zum Zeitpunkt des Trainings des jeweiligen Modells bereits existierten;
- fehlende Reproduzierbarkeit, da selbst wenn ein LLM wiederholt eine gleichbleibende Eingabe erhält, die/das jeweils erzeugte Ausgabe/Ergebnis sowohl sprachlich als auch inhaltlich unterschiedlich sein kann;
- fehlende Sicherheit von generiertem Code (falls ein LLM auf Code trainiert wurde, kann es auch solchen generieren, der potenziell Sicherheitslücken in der Ausführung enthält);

---

[8] Siehe zu deren Einsatz innerhalb generativer KI-Werkzeuge das «Merkblatt zur Verwendung von generativen KI-Werkzeugen in der Bundesverwaltung»; https://cnai.swiss/dienstleistungen-weitere-dienstleistungen-merkblaetter-zu-ki/

[9] https://www.fedlex.admin.ch/eli/oc/2022/491/de#art_21

[10] Als Basis diente das Kapitel «2.3 Risiken von LLMs» des Berichts (Version 1.1 vom 27. März 2024) des deutschen Bundesamts für Sicherheit in der Informationstechnik über die Chancen und Risiken der LLMs; siehe https://www.bsi.bund.de/SharedDocs/Downloads/DE/BSI/KI/Generative_KI-Modelle.pdf

- fehlerhafte Reaktion auf spezifische Eingaben, da bereits kleine Abweichungen in den Eingaben zu grossen Unterschieden in den erzeugten Ausgaben/Ergebnissen führen können. Weichen Eingaben an ein LLM z.B. von den Texten ab, die zum Training verwendet wurden, kann das Modell diese häufig nicht mehr korrekt verarbeiten und generiert fehlerhafte Ausgaben/Ergebnisse;
- LLMs generieren in der Regel z.B. sprachlich fehlerfreien und inhaltlich überzeugenden Text und sind zudem in vielfältigen Themenbereichen aussagefähig. Dadurch kann der Eindruck eines menschenähnlichen Leistungsvermögens und damit ein zu grosses Vertrauen in die Aussagen sowie die Leistungsfähigkeit der Modelle entstehen («*automation bias*»);
- Wiederverwendung der eingegebenen Daten in vielen Fällen als Trainingsdaten (auch objektiv falsche Eingaben von Nutzern und Nutzerinnen werden so genutzt);
- Anfälligkeit für die Interpretation von Text als Anweisung, da LLMs alle Eingaben auf die gleiche Weise interpretieren und nicht zwischen Anweisungen und sonstigen Texten unterscheiden;
- fehlende Vertraulichkeit der eingegebenen Daten, da LLMs häufig als Service über das Internet mittels geeigneter Schnittstellen angeboten werden (z.B. unter Verwendung eines Webbrowsers). Die Übermittlung von personenbezogenen Daten an solche Services über das Internet kann eine Weitergabe von Daten an Dritte darstellen. Es gelten die Regeln des Datenschutzgesetzes;
- die Entwicklung und der Betrieb von LLMs durch Anbieter auf deren Infrastruktur kann mit einer grossen Abhängigkeit einhergehen (z.B. fehlende Datenhoheit, fehlende Kontrollierbarkeit des Modells);
- LLMs generieren zum Teil auch Quellenangaben, die nicht immer stimmen oder gar nicht existieren, und halten beharrlich und überzeugend daran fest.

Generative KI-Technologien (wie LLMs) bieten vielfältige Chancen und Anwendungsmöglichkeiten und entwickeln sich aktuell mit hoher Dynamik weiter. Damit einhergehend treten neue Sicherheitsbedenken rund um die Entwicklung, den Betrieb und die Nutzung dieser Modelle auf. Ein sicherer Umgang mit diesen KI-Technologien setzt die Durchführung einer systematischen Risikoanalyse voraus.

In diesem Sinne sollen auch alle Nutzer oder Nutzerinnen in der Bundesverwaltung vor der Integration von generativen KI-Technologien (wie LLMs) in die eigenen alltäglichen Arbeitsabläufe eine individuelle Risikoanalyse durchführen. Dabei können die vorausgegangen Grundsätze einen Orientierungsrahmen für ein verantwortungsvolles Experimentieren liefern.

Bei Fragen:

- Zu KI in der Bundesverwaltung: Arbeitsgruppe KI und Knotenpunkte im Kompetenznetzwerk für KI (CNAI)
- Zu Informationssicherheit und Datenschutz: die ISBOs und DSBOs Ihrer Verwaltungseinheit
- Für konkrete Dienstleistungen im Bereich Datenwissenschaft und KI: DSCC

Weitere Hinweise:

Leitlinien für den Umgang mit KI in der Bundesverwaltung

Die sieben Leitlinien für den Umgang mit KI in der Bundesverwaltung gelten weiterhin: Den Menschen in den Mittelpunkt stellen, Rahmenbedingungen für Entwicklung und Anwendung von KI gewährleisten, Transparenz, Nachvollziehbarkeit und Erklärbarkeit einfordern, Verantwortlichkeit klar definieren, Sicherheit gewährleisten, aktive Mitgestaltung der Gouvernanz von KI vorantreiben und dabei alle relevanten nationalen und internationalen Akteuren einbeziehen.

Verhaltenskodex des Bundes für menschenzentrierte und vertrauenswürdige Datenwissenschaft (und KI)

Durch den Verhaltenskodex werden die Verwaltungseinheiten des Bundes im Sinne einer Orientierungshilfe mit praktischer Erläuterung zum einen für die Grundprinzipien einer menschenzentrierten und vertrauenswürdigen Datenwissenschaft (und KI) sensibilisiert, und zum anderen zu deren Umsetzung im Arbeitsalltag befähigt. Die Grundprinzipien lauten: Daten- und Informationsschutz, Informationssicherheit, Datensicherheit, Datengouvernanz, Nichtdiskriminierung, Erklärbarkeit, Nachvollziehbarkeit, Transparenz, Reproduzierbarkeit, Neutralität, Objektivität und ethischer Umgang mit Daten und Ergebnissen.

Das Merkblatt ist in der Arbeitsgruppe «KI in der Bundesverwaltung» im CNAI unter Mitwirkung von Vertreterinnen und Vertretern aus allen Departementen und der BK entstanden.