SPD
Fraktion im Bundestag

Berlin, 26.06.2024

Leitlinien zur Nutzung von generativer KI in der parlamentarischen Arbeit

Im November 2022 hat OpenAI GPT-3.5 veröffentlicht und damit für einen echten Hype gesorgt. **ChatGPT** und andere Anwendungen generativer Künstlicher Intelligenz (KI) begegnen uns seitdem überall – in den Medien, auf den politischen Tagesordnungen und in den Diskussionen im Familien- und Freundeskreis.

Die vergangenen Monate haben gezeigt, wie sehr ChatGPT und Co. polarisieren. Einige begreifen sie als Allheilmittel, andere sehen in ihrem Aufstieg unseren Untergang. Wir plädieren für einen **ausgewogenen Umgang, indem wir Chancen ergreifen und Risiken erkennen**: Anwendungen generativer KI haben das Potenzial, unser Leben und Arbeiten zu erleichtern – denn sie können sowohl in Unternehmen als auch im öffentlichen Sektor und dem Deutschen Bundestag Informationen bündeln, Prozesse beschleunigen, die Produktivität steigern, Arbeitnehmer:innen entlasten und neue Denkanstöße geben. Gleichzeitig birgt die Nutzung von KI auch einige Risiken, mit denen wir bewusst umgehen müssen.

Daher begrüßt die SPD-Bundestagsfraktion ausdrücklich, dass die Mitgliedsstaaten der EU 2024 das weltweit erste umfassende Regelwerk für KI, den Artificial Intelligence Act (AIA), beschlossen haben. Für Anwendungen, die ein hohes Risiko für Gesundheit, Sicherheit und Grundrechte darstellen, sieht der AIA strenge Regeln vor. Anwendungen, von denen ein inakzeptables Risiko ausgeht, werden verboten.

Wir bekennen uns zum risikobasierten Ansatz des AIA und sind überzeugt, dass ein ausgewogener Regulierungsrahmen die Risiken von KI eindämmen und gleichzeitig die Chancen für Wirtschaft und Gesellschaft, aber auch Europas Souveränität heben kann.

Bis zum vollständigen Inkrafttreten der Verordnung und damit bis zum Eintritt aller rechtlichen Verpflichtungen dauert es allerdings noch etwa zwei beziehungsweise drei Jahre. Um diesen Zeitraum zu überbrücken, die Chancen von KI zu heben und die Risiken zu beherrschen, gibt es bereits Initiativen aus Politik, Wirtschaft und Zivilgesellschaft, die gemeinschaftliche Regeln im Umgang mit KI erarbeiten. Diesem Gedanken ergänzender Leitlinien sehen auch wir uns als SPD-Bundestagsfraktion verpflichtet.

Aussichtsreiche Anwendungsfelder von KI in der parlamentarischen Arbeit liegen beispielsweise in der Organisation, Textarbeit, Transkription oder Übersetzung. Als

Abgeordnete und Mitarbeiter:innen des Deutschen Bundestages und der Fraktion der SPD im Deutschen Bundestag sind wir jedoch zu einer **besonderen Sorgfalt,** insbesondere im Umgang mit den öffentlich verfügbaren Anwendungen – und langfristig auch mit internen KI-Lösungen – aufgerufen. Mithilfe dieses Papiers wollen wir deshalb für die Fallstricke der Systeme sensibilisieren und politische Leitlinien für unsere fraktionsinterne Arbeit vorlegen. Darüber hinaus gibt sich die SPD-Fraktion mit diesem Papier den Auftrag, unter Einbezug aller relevanten Akteur:innen **Rahmenbedingungen** für den Einsatz von KI innerhalb der Fraktion zu schaffen, auch in Form von **arbeitsrechtlichen Leitlinien.**

Diskriminierung

In der Nutzung von Anwendungen generativer KI fällt immer wieder auf, dass die sogenannten Large Language Models (LLM) Stereotype oder diskriminierende Narrative reproduzieren. Der Ursprung dafür liegt unter anderem in den **Trainingsdatensätzen** der Modelle: Diese setzen sich aus einer Vielzahl unterschiedlicher Texte von qualitativ hochwertigen wissenschaftlichen Publikationen bis hin zu Texten geringerer Qualität wie Forenbeiträgen zusammen. LLM bauen ihre Antworten zunächst auf **Wahrscheinlichkeiten von Wortfolgen** aus ihren Trainingsdatensätzen auf. Im Umkehrschluss heißt das: Je mehr Texte von niedriger Qualität im Datensatz vorhanden sind, desto eher haben sie Gewicht für die Richtung, die das Modell in seiner Antwort einschlägt. Dazu, wie Trainingsdatensätze der populären LLM im Detail zusammengesetzt sind, halten sich die Anbieter bedeckt. Bei GPT-3 ist bekannt, dass rund **82 Prozent** der Daten aus verschiedenen Kompilationen von **Internetinhalten** stammen, 16 Prozent aus Büchern und drei Prozent aus Wikipedia[1]. Allein die schiere Verfügbarkeit von Texten wie Forenbeiträgen gegenüber einer kleineren, weil aufwendiger zu erstellenden Zahl wertigerer Texte, lassen eine ähnliche Verteilung auch bei anderen Modellen vermuten. Verschwiegen zeigen sich die Anbieter außerdem in Hinblick auf ihre **Trainingsprozesse**, die ebenfalls auf den Output der Modelle Einfluss nehmen: Weder die **Gewichtung** noch etwaige **Leitplanken**, die verhindern, dass die Modelle sozial unerwünschte oder potenziell gefährliche Antworten geben, werden kommuniziert.

Beim Output gilt es für Nutzer:innen deshalb, **genau hinzusehen** und den Blick für stereotypisierende oder diskriminierende Inhalte zu schärfen, insbesondere solange die Anbieter ihre Trainingsdaten und -prozesse nicht transparent machen. Klar sein muss außerdem, dass selbst bereinigte Modelle nur einen Teil der Wirklichkeit abbilden können: Denn die Menschen und Lebensrealitäten, die bereits in den Trainingsdaten nicht ausreichend repräsentiert sind, werden bei einem auf Wahrscheinlichkeiten basierenden Modell auch im Output nur wenig Platz haben.

Falschinformation

LLM neigen außerdem zu sogenannten **Halluzinationen**: Sie erfinden Informationen

[1] Research Article „Language Models are Few-Shot Learners" von Tom Brown et al. im Rahmen der Neural Information Processing Conference, Dezember 2020: https://dl.acm.org/doi/abs/10.5555/3495724.3495883

oder geben solche Antworten aus, die in Widerspruch zu ihren Trainingsdaten stehen. Darüber, ob dieses strukturelle Problem behoben werden kann, wird derzeit gestritten. Falschinformationen können aber auch daher rühren, dass die Trainingsdaten der Modelle faktisch falsch sind (man denke an nicht vertrauenswürdige Quellen) oder die Gegenwart nicht abdecken.

Die Antworten der Systeme sind also eloquent – aber **keinesfalls gebrauchsfertig**. Besonderer Nutzen ist dann zu erwarten, wenn der zu erstellende Text oder die zu bewältigende Aufgabe keine Faktizität voraussetzt oder wenn zu dem erfragten Themenbereich bereits vertieftes Wissen besteht, das eine kritische Auseinandersetzung erleichtert.

Desinformation

Anwendungen generativer KI arbeiten schnell und kostengünstig. Das macht sie attraktiv für die, die Desinformation streuen wollen. Versuche der Aufstachelung oder gezielten Verunsicherung, wie wir sie bereits auf den Social-Media-Kanälen rechtspopulistischer und rechtsextremer Parteien in Form KI-generierter Bilder gesehen haben, fallen durch die Funktionsweise der Plattformen auf fruchtbaren Boden. Das ist auch mit Blick auf politische Wahlen zutiefst beunruhigend.

Als Vertreter:innen der demokratischen Parteien haben wir die Aufgabe, gegenzuhalten: Einerseits mit **Fakten und Gegendarstellungen**, damit Nutzer:innen sozialer Netzwerke Informationen genauer einordnen können. Andererseits mit einer **Kennzeichnung** von Inhalten, die in Zusammenarbeit mit KI entstanden sind. Das schafft nicht nur Transparenz – die Kennzeichnung bietet außerdem Gelegenheit, über die Risiken der Systeme aufzuklären, Sensibilität gegenüber möglicher Desinformation zu fördern und so Bürger:innen zu einem mündigen, informierten Umgang mit den Systemen zu befähigen.

Die benannten Risiken erkennt auch der AIA an und stellt richtigerweise Regeln in Bezug auf die Kennzeichnung generierter Inhalte auf. Allerdings müssen wir anerkennen, dass Akteur:innen mit böswilligen Absichten sich von solchen Regeln nicht aufhalten lassen werden. Dazu kommt: Formen der Kennzeichnung wie Wasserzeichen oder entsprechende Hinweise in den Metadaten lassen sich nach bisherigem Stand der Entwicklung noch umgehen oder manipulieren. Auf eine allein technische Lösung können und sollten wir deshalb nicht setzen. Essentiell ist deshalb der flankierende **Aufbau und die Stärkung** der **KI- und Medienkompetenz** in der Bevölkerung. Beides werden wir vorantreiben.

Datenschutzverletzung

Aus datenschutzrechtlicher Perspektive gibt es bei LLM gleich mehrere Bedenken. Durch das breitflächige Datensammeln werden auch personenbezogene und private Daten in die Trainingsprozesse von Sprachmodellen aufgenommen und können so in KI-generierten Inhalten verbreitet werden. Zudem nutzen viele Anbieter von

Anwendungen generativer KI die Eingaben der Nutzer:innen (sogenannte Prompts) für weitergehendes Training ihrer Systeme. So fließen auch etwa **sensible Informationen** an die Anbieter zurück und bilden potenziell den Ausgangspunkt für eine künftige Antwort des Systems. Hacker:innen ist es immer wieder gelungen, den Anwendungen diese Informationen zu entlocken. Eine wirksame Maßnahme gegen den Abfluss sensibler Informationen liegt deshalb darin, in öffentlich verfügbaren Anwendungen **keine persönlichen oder vertraulichen Informationen in den Prompts** zu teilen.

Urheberrechtsverletzung

Die zugrundeliegenden Trainingsdatensätze von LLM treffen auch beim Urheberschutz auf rechtliche Probleme. Beim Auslesen des Internets (sogenanntes Scraping) greifen generative KI-Modelle auch auf verfügbare, aber urheberrechtlich geschützte Materialen (z. B. Texte, Bilder, Videos, Tonaufzeichnungen) zurück und reproduzieren in ihren generierten Inhalten somit Teile dieser Materialien wieder. Über Art, Umfang und Herkunft der Daten wie auch über diese Prozesse wird von den Plattformen bislang keine Transparenz hergestellt. Diese Untergrabung des Urheberrechts ist existenzbedrohend für die Kreativen, deren Werke ohne Zustimmung oder Entlohnung genutzt werden.

Training und Betrieb generativer KI-Systeme war und ist weiterhin nur möglich, indem geschützte Inhalte genutzt werden. Daher müssen Transparenz über die Nutzung sichergestellt und Urheber:innen und ausübende Künstler:innen an der Wertschöpfung angemessen beteiligt werden, die durch und mit generativer KI erfolgt.

Prekäre Arbeitsbedingungen

Das breite Abschöpfen von Textmaterial aus dem Internet zu Trainingszwecken hat zur Folge, dass auch **schädliche Inhalte** wie Darstellungen von Gewalt in die Daten gelangen. Um sie dahingehend zu bereinigen und die späteren Nutzer:innen zu schützen, werden bestimmte Filter eingesetzt. Die Filter wiederum werden mithilfe von Textausschnitten trainiert, die zuvor von Menschen als schädlich markiert worden sind. Diesen Prozess hat OpenAI laut einer investigativen Recherche des TIME-Magazins in den **Globalen Süden** verlagert: In Kenia wurden sogenannte Clickworker:innen für einen Stundenlohn von nicht mehr als zwei US-Dollar potenziell traumatisierenden Darstellungen von Suizid bis Tierquälerei ausgesetzt[2].

Aus sozialdemokratischer Perspektive ist das ein unhaltbarer Zustand. Unsere Aufgabe ist deshalb, die Big-Tech-Unternehmen an ihre Verantwortung zu erinnern und uns energisch für **gute Arbeitsbedingungen** und **faire Bezahlung** für alle Akteur:innen entlang der KI-Wertschöpfungskette einzusetzen.

[2] Recherche des TIME-Magazins, 18. Januar 2023: https://time.com/6247678/openai-chatgpt-kenya-workers/

Ressourcenverbrauch

LLM werden mit riesigen Datensätzen trainiert, die gegenwärtig mit jeder neuen Version und jedem neuen Launch immer schwindelerregendere Dimensionen annehmen. Allein für GPT-3 kalkuliert eine Studie einer US-amerikanischen Universität einen Stromverbrauch von **1.300 Megawattstunden** und einen Wasserverbrauch von **700.000 Litern**[3]. Das entspricht dem Jahresstrombedarf von etwa 650 Ein-Personen-Haushalten und dem täglichen Trinkwasserverbrauch von etwa 5600 Menschen in Deutschland. Dazu kommen die **Kühlung** der Rechenzentren sowie der **Betrieb** der Systeme, die weitere Ressourcen verbrauchen.

Auch hier braucht es ein stärkeres **Bewusstsein** für die Zusammenhänge zwischen der Entwicklung und dem Betrieb der Anwendungen und den negativen Umweltfolgen. Klar ist: Jeder Prompt kostet.

[3] Studie der University of California in Riverdale, 2023: https://arxiv.org/pdf/2304.03271.pdf