Lernende Systeme

DIE PLATTFORM FÜR KÜNSTLICHE INTELLIGENZ

Große Sprachmodelle

Grundlagen, Potenziale und Herausforderungen für die Forschung

GEFÖRDERT VOM

Bundesministerium
für Bildung
und Forschung

acatech
DEUTSCHE AKADEMIE DER
TECHNIKWISSENSCHAFTEN

WHITEPAPER

Löser, A., Tresp, V. et al.
AG Technologische Wegbereiter
und Data Science

Inhalt

Zusammenfassung

Sie schreiben Hausarbeiten, Gedichte oder Programmiercodes – große Sprachmodelle wie ChatGPT, BARD oder BLOOM verändern mit einer rasanten Entwicklung, wie wir arbeiten und kommunizieren und wie wir mit Information und Wissen umgehen. Diese KI-Modelle sind flexibler und leistungsfähiger als ihre Vorgänger. Sie beruhen auf einer effektiven Kombination aus großen Datenmengen, umfangreicher Rechenleistung und Algorithmen, die effizient und skalierbar einsetzbar sind. Als Schlüsseltechnologie sind sie der Kern vieler wichtiger Anwendungen: Sie erkennen, produzieren, übersetzen und verarbeiten Sprache. Die zugrundeliegende KI-Technologie ist jedoch nicht auf Sprachverarbeitung beschränkt. Große Sprachmodelle bieten daher enormes Potenzial für alle Lebensbereiche und stoßen damit einen grundlegenden und dauerhaften Wandel nicht nur in vielen Branchen an, sondern auch in Forschung und Entwicklung.

Expertinnen und Experten der Arbeitsgruppe Technologische Wegbereiter und Data Science der Plattform Lernende Systeme liefern mit dem Whitepaper einen Überblick zum Thema Sprachmodelle und legen dar, auf welchen Grundlagen Sprachmodelle beruhen und vor allem, was ihre besonderen Eigenschaften sind. Neben der Funktionsweise dieser großen KI-Modelle wird dabei auch deren zunehmende Bedeutung aufgrund der rasanten disruptiven Entwicklung herausgestellt. Für die Forschung und Entwicklung ergibt sich damit ein dynamisches Handlungsfeld mit vielen Potenzialen, aber zugleich auch Herausforderungen, anhand derer weitere Forschungsfelder und -bedarfe identifiziert werden (Kapitel 3), die abschließend in Handlungsfeldern sowie noch offenen Fragen münden (Kapitel 4).

Um diese Dynamik mitzugestalten und das Potenzial weiter zu heben, bedarf es zum einen der Entwicklung neuer Methoden, die eine effiziente und flexible Anpassung von Sprachmodellen an Aufgaben- und Anwendungsbereichen erlauben, und zum anderen der (Weiter-)Entwicklung von modernen, offenen Sprachmodellen. Damit ist an vielen Stellen weitere Forschung und Entwicklung notwendig, um unterschiedliche technische Herausforderungen anzugehen sowie effektive Lösungsansätze weiterzuverfolgen oder neue zu erschließen. Denn große Sprachmodelle sind häufig noch intransparent und ihre Ergebnisse nicht immer nachvollziehbar, fehlerhaft oder mit Bias behaftet. Dies erfordert eine Begleitung durch entsprechende Forschung und Entwicklung sowie eine Vernetzung und Stärkung der entsprechenden Communitys in Deutschland und Europa, um das Potenzial für Gesellschaft und Wirtschaft im Rahmen europäischer Werte bestmöglich entfalten zu können.

1 Einleitung

Wenn wir miteinander kommunizieren, spielt Sprache in allen unseren Lebensbereichen eine zentrale Rolle. Sprache dient der Kommunikation und ist für die Übermittelung von Informationen, Koordination, Planung und Organisation, sei es im persönlichen Alltag oder Berufsleben, essentiell. Um eine gute und vertrauensvolle Kommunikation aufzubauen, ist Sprache der zentrale Schlüssel in der Beziehung Mensch-Mensch, mit dem digitalen Zeitalter zunehmend auch in der Beziehung Mensch-Maschine. Große Sprachmodelle stellen einen Meilenstein in der Interaktion sowohl zwischen Mensch und Maschine als auch zwischen Maschinen dar, der seit der Veröffentlichung des Modells GPT-3 2020 für viele Expertinnen und Experten bereits absehbar war. Mit ChatGPT Ende 2022 ist dies auch für eine breite Öffentlichkeit sichtbar geworden. Jahrelange Forschung zum maschinellen Lernen im Kontext sehr großer Datenmengen und unter Nutzung neuer Rechnerarchitekturen mündet in diese Entwicklung und zeigt Potenziale und Herausforderungen dieser KI-Technologie auf, die wiederum zu neuen Fragestellungen für Forschung und Entwicklung führen.

Große Sprachmodelle verändern derzeit und auch künftig unsere Kommunikation und den Umgang mit Information und Wissen in Gesellschaft, Wissenschaft und Wirtschaft. Für 2026 wird erwartet, dass 50 Prozent der Anwendungsfälle im Bereich der natürlichen Sprachverarbeitung auf Modellen beruhen, die mit den aktuell vielbesprochenen Transformer-Architekturen trainiert wurden, während dies 2021 nur in weniger als fünf Prozent der Fall war (Duncan, 2022). Große Sprachmodelle basieren auf künstlichen neuronalen Netzwerken, die unter anderem zur Verarbeitung und zum Verständnis natürlicher Sprache eingesetzt werden (siehe Erklärbox 1: autoregressive Modelle wie GPT-4 und bidirektionale Modelle wie BERT). Sie bedienen sich der menschlichen Sprache, indem sie auf großen Textdatensätzen trainiert werden. Als Schlüsseltechnologie der Künstlichen Intelligenz sind große Sprachmodelle der Kern vieler wichtiger Anwendungen: Sie erkennen, produzieren, übersetzen und verarbeiten Sprache. Durch ihre Fähigkeit, natürliche Sprache aus Milliarden von Texten zu verarbeiten, können sie für eine Vielzahl von Aufgaben eingesetzt werden (Texterstellung und -zusammenfassung, Beantwortung von Fragen sowie Generierung von Programmcode und vieles mehr). Damit bieten große Sprachmodelle nicht nur enormes Potenzial für die Gesellschaft, sondern auch für die Wirtschaft und Industrie, indem sie etwa Aufgaben und Prozesse erleichtern und unterstützen, und so Raum für anspruchsvolle und kreative Tätigkeiten schaffen.

Um dieses Potenzial weiter heben zu können, sind an vielen Stellen weitere Forschung und Entwicklung notwendig, die einerseits damit verbunden sind, die Anpassung von Modellen an unterschiedliche Aufgaben- und Anwendungsbereiche effizienter zu machen, und anderseits unterschiedliche technische Herausforderungen anzugehen. So sind große Sprachmodelle häufig intransparent und ihre Ergebnisse nicht immer nachvollziehbar oder sie erstellen Texte, die falsche oder unlogische Aussagen und Zusammenhänge darstellen. Ihre rasante Entwicklung führt zu immer größeren und kostenintensiveren Modellen (Maslej et al., 2023); sie werden hauptsächlich multilingual oder für die englische und chinesische Sprache und durch außereuropäische Akteure – meist große amerikanische und chinesische kapitalstarke Unternehmen und Organisationen – entwickelt: Bedeutende große Sprachmodelle entstehen daher überwiegend in der Industrie und stehen der Forschung außerhalb entsprechender Unternehmen häufig nicht oder nur begrenzt zur Verfügung (Maslej et al., 2023, S. 50; Sevilla et al., 2022b; Solaiman, 2023). Ausnahmen sind zum Beispiel das offen zugängliche Modell BLOOM, das aus einer internationalen Community um die Plattform Hugging Face hervorgegangen ist, oder das Modell Vicuna, das US-Forschende auf der Basis des Modells LLaMA des Unternehmens Meta erstellt haben.

Die Bedeutung und Funktionsweisen großer Sprachmodelle sowie die Chancen und Herausforderungen, mit denen Forschung und Entwicklung konfrontiert werden, sind Gegenstand dieses Whitepapers. Forschungsbedarfe werden identifiziert und erläutert, sodass abschließend Handlungsfelder und offene Fragen dargelegt werden. Während das vorliegende Whitepaper teilweise Multimodalität diskutiert, fokussiert es schwerpunktmäßig auf große Sprachmodelle als eine Kerntechnologie, die auch für generative Modelle im Allgemeinen zentral ist, beispielsweise um textbasiert Anweisungen erteilen zu können. Jenseits der Schwerpunktsetzung dieses Whitepapers werden die ethischen Dimensionen großer Sprachmodelle (z. B. Desinformation, deep fakes etc.) sowie ihre weiteren gesellschaftlichen und wirtschaftlichen Implikationen in künftigen Publikationen und Veranstaltungen der Plattform Lernende Systeme adressiert.

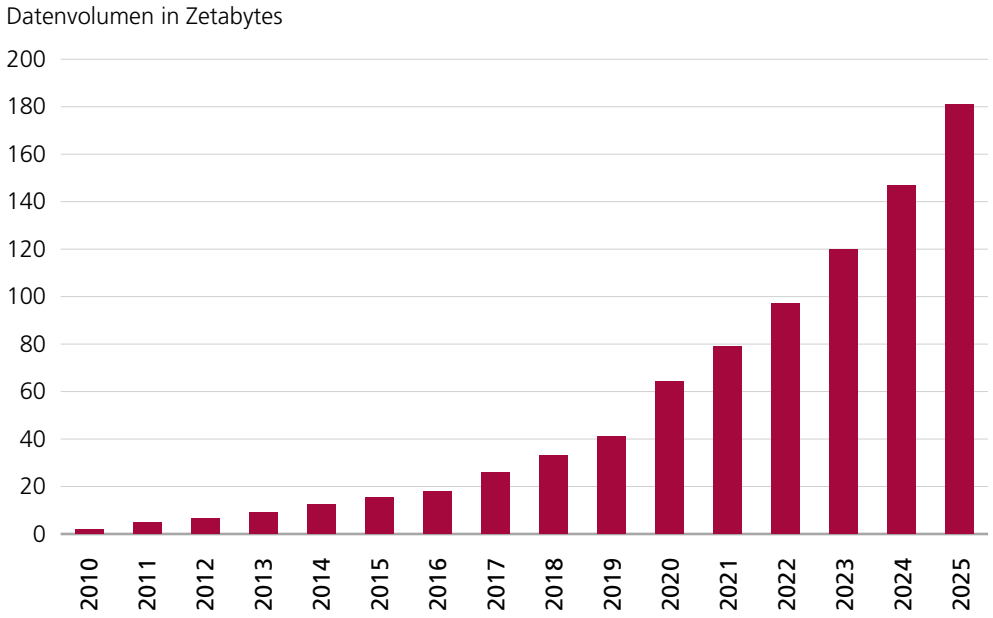
2 Hintergrund und Grundlagen

Um große Sprachmodelle zu erstellen, sind vier Komponenten notwendig: (1) Daten in hoher Qualität und Güte, (2) Rechenleistung, (3) hochperformante Algorithmen, die auf Basis der Daten und der Rechenleistung effizient angewendet werden können, sowie (4) Mitarbeitende und Fachleute mit spezifischen Kenntnissen und Fertigkeiten (z. B. Fertigkeiten, um Modelle des maschinellen Lernens in der Anwendung zuverlässig und effizient einzusetzen und zu warten, Wissen zu maschinellem Lernen und Transformern usw.). Welches Potenzial durch eine effiziente Kombination dieser vier Komponenten freigesetzt werden kann, macht die Entwicklung zu großen KI-Modellen besonders deutlich. ChatGPT ist eines der bekanntesten Sprachmodelle für Dialoge (Stand: 2023). Es basiert auf dem vortrainierten Sprachmodell GPT-3.5, das in der Lage ist, Texte zu generieren, und wurde mit der sogenannten Transformer-Architektur für maschinelles Lernen mit künstlichen neuronalen Netzen trainiert – daher: Generative Pre-Trained Transformer (GPT). Solche Sprachmodelle sind das Ergebnis einer effizienten Kombination der vier genannten Komponenten, die zu spezifischen Eigenschaften großer Sprachmodelle geführt haben, die sie in der Performanz von bisherigen Modellen unterscheiden.

Entwicklung: Daten und Rechenleistung

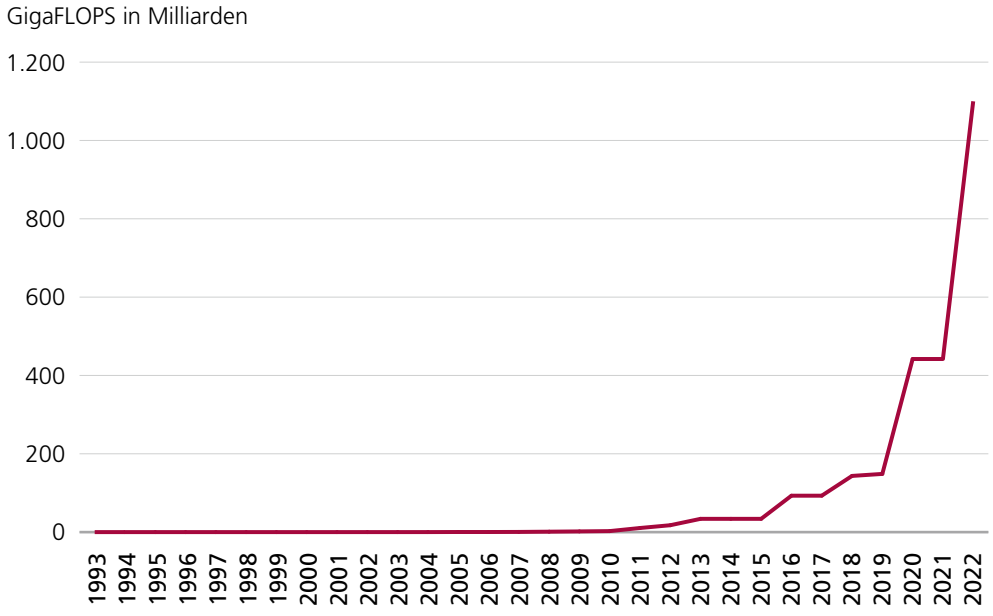
Große Datenmengen, insbesondere Textdaten, sind bereits in großem Umfang verfügbar. Dies ist im Allgemeinen auf die fortschreitende Digitalisierung (z. B. Digitalisierung von Büchern etc.) und im Besonderen auf das Internet zurückzuführen. So werden zum Beispiel in den sozialen Medien täglich Unmengen von Texten, Bildern und Videos gepostet. Insgesamt wächst das weltweite Datenvolumen exponentiell (siehe Abbildung 1). Allerdings unterscheidet sich der Umfang an Textdaten je nach Sprachgemeinschaft, die für die Berechnung von Sprachmodellen verfügbar sind: Für das Englische als „Lingua Franca" oder auch für das Spanische stehen sehr viel mehr Textdaten zur Verfügung als für das Deutsche. Zum Vergleich: Hochdeutsch sprechen 160 Millionen Personen weltweit, Englisch sprechen dagegen rund 1,45 Milliarden und Spanisch 548 Millionen Personen (Ethnologue, 2022). Die benötigten deutschsprachigen Textdaten für große Sprachmodelle liegen aber dennoch vor. Allerdings bietet der große Umfang englischsprachiger Textdaten einen Vorteil bei der Skalierung zu noch größeren Modellen. Auch hinsichtlich der verfügbaren Rechenleistung können wir ein exponentielles Wachstum beobachten. Gerade seit 2015 steigt diese rasant an (siehe Abbildung 2). Während die Komponenten Daten und Rechenleistung schon seit einigen Jahren gegeben sind, war dies bei den geeigneten Algorithmen im Bereich der Sprachmodelle erst ab 2017 mit der Einführung der Transformer-Architektur für das maschinelle Übersetzen gegeben (Vaswani, 2017).

Abbildung 1: Menge der erstellten, verbrauchten und gespeicherten Daten 2010–2020, mit Prognosen bis 2025 (in Zetabytes)



Ein Zetabyte entspricht 10^{21} Byte
Quelle: IDC, Statista, 2021

Abbildung 2: Entwicklung der Leistungsfähigkeit von Hochleistungsrechnern (in GigaFLOPS)



Floating-Point Operations per Second (FLOPS) sind ein Maß der Leistungsfähigkeit von Computern.
Es handelt sich dabei um die Anzahl von Gleitkomma-Operationen (Addition und Multiplikation), die in einer Sekunde möglich sind.
1 GigaFLOP entspricht 10^9 FLOPS
Quelle: Top500, 2023

Funktionsweise: Transformer-Architekturen

Übersetzungsanwendungen wie DeepL nutzen Transformer-Architekturen genauso wie etwa Sprachmodelle für Dialoge wie OpenAIs ChatGPT, Googles BARD oder die offene Alternative Vicuna. Es handelt sich um eine spezielle Architektur des maschinellen Lernens mit mehreren Schichten von künstlichen neuronalen Netzwerken, Deep Learning genannt, das häufig für sequenzielle Daten wie Text genutzt wird.

Die Grundannahme solcher Verfahren bei Sprachmodellen beruht auf der seit langem verbreiteten Annahme, dass Wörter, die in gleichen Kontexten vorkommen, ähnliche Bedeutungen haben (sog. Distributional Hypothesis). In der Umsetzung werden zunächst Wörter bzw. Wortteile in Vektoren, auch Embeddings genannt, überführt, die das Wort mathematisch abstrakt beschreiben. Solche Vektoren können beim maschinellen Lernen mit neuronalen Netzen verarbeitet werden. Sie werden in einem weiteren Schritt in hochdimensionalen Räumen angeordnet, um ein Maß zu erhalten, wie ähnlich sich zwei Vektoren sind. Sinnhaft ähnliche Begriffe erhalten eine ähnliche Vektordarstellung. Für Synonyme sind dann diese Ähnlichkeiten zum Beispiel besonders groß (das heißt, die Abstände sind im Vektorraum klein).

Die einzelnen Schichten des Transformers spezialisieren sich auf unterschiedliche Aufgaben, obwohl die Architektur in jeder Schicht identisch ist (siehe hierzu z. B. Van Aken et al., 2019). Beispielsweise klassifizieren die unteren Schichten Wortbestandteile zu ganzen Wörtern bzw. erkennen, welche einzelnen Wörter (z. B. „Hochschule") ein Teil eines Multi-Wortes sind (z. B. „Berliner Hochschule für Technik"). Der Transformer repräsentiert diese Informationen in den Gewichtungen einzelner Neuronen der entsprechenden Schicht. Höherliegende Schichten werden mit diesen Repräsentationen und damit auch oft mit einer besseren Kontextrepräsentation initialisiert. Diese höherliegenden Schichten nehmen dann beispielsweise Aufgaben der Erkennung von Dingen („Entitäten") wahr bzw. können auch schon oft erste einfache Beziehungen erkennen (z. B. „BHT": Abkürzung für „Berliner Hochschule für Technik"). Weiter höherliegende Schichten sind wiederum besser in der Lage, auch komplexere Beziehungen bzw. auch Beziehungen, die viele Wörter oder mehrere Sätze weiter auseinanderliegen, zu erkennen.

Sprachmodelle lernen Repräsentationen für Prozesse zur Erzeugung von Sprache. Die obersten Schichten strukturieren somit den Vektorraum in ähnliche Dinge bzw. auch Beziehungen. Diese Schichten haben gelernt, welche generellen Sprachprozesse zur Erzeugung dieser Repräsentationen führten. Auto-regressive Modelle, wie GPT, benutzen diese Eigenschaft, um für eine eingegebene Sequenz an Wörtern die nächsten Wörter vorherzusagen (siehe Erklärbox 1: Modell-Typen).

Darüber hinaus kann die Transformer-Architektur die vorhandene Rechenleistung wesentlich besser ausnutzen, als das bei der Benutzung von älteren Netzwerken des Typs Long Short-Term Memory (kurz: LSTM) der Fall ist. Die Verarbeitung von Wörtern durch Recheneinheiten erfolgt nicht mehr nacheinander, wie dies bei früheren Architekturen der Fall war (siehe Rekurrente Neuronale Netze). Sie kann parallel erfolgen, das heißt, verschiedene Verarbeitungen von Wörtern können parallel von verschiedenen Recheneinheiten durchgeführt werden.

Transformer eignen sich zudem sehr gut für skalierbares, selbstüberwachtes Lernen auf großen Textdatensätzen. Der Textdatensatz, auf dem die Sprachmodelle trainiert werden, wird beim selbstüberwachten Lernen zufällig mit Lücken versehen. Die Aufgabe der Transformatoren ist es, entweder aus dem vorangehenden Kontext vorherzusagen, welches Wort in die jeweilige Lücke passt, oder aus beidem – dem Kontext davor und danach (siehe Erklärbox 1). Da der gesamte Trainingsdatensatz zur Verfügung steht, kann ein Vergleich durchgeführt werden, der zeigt, wie gut die jeweilige Vorhersage war. Je nach Ergebnis des jeweiligen Ver-

gleichs können die Modellparameter so angepasst werden, dass die Prognose für das jeweils fehlende oder zu ergänzende Wort verbessert wird. Transformer berechnen dazu für ein bestimmtes Kontextfenster, beispielsweise einen Satz oder eine bestimmte Kontextlänge, wie zum Beispiel 1024 Wörter, die Stärke der Beziehung zwischen zwei Wörtern. Dies gibt in der weiteren Verarbeitung Aufschluss darüber, auf welche Teile der Textdaten im jeweiligen Kontext ein erhöhter Fokus gelegt werden soll (vgl. Aufmerksamkeits-Mechanismus). In den meisten Modellen ist auch ein Maß enthalten, wie weit Wörter in Sequenzen voneinander entfernt sind. Die beschriebene Vorgehensweise kann auch auf andere Datentypen wie Bilder und Videos oder auf Aminosäuresequenzen – im Fall des generativen Proteindesigns – angewandt werden.

Erklärbox 1

Zwei Haupttypen von Modellen werden unterschieden:

Autoregressive Modelle (GPT-3.5, BARD, LLaMA, Vicuna): Modelle dieser Art werden trainiert, indem sie die Wahrscheinlichkeit für das Auftreten eines Worts auf der Basis aller zuvor aufgetretenen Wörter ermitteln. Solche Modelle werden vor allem für generative Aufgaben verwendet.

Bidirektionale Modelle (z. B. BERT – Bidirectional Encoder Representations from Transformers)**:** Bidirektionale Transformer-Modelle grenzen sich von autoregressiven Modellen dadurch ab, dass sie darauf trainiert sind, Wörter durch ihren gesamten Kontext vorherzusagen; das heißt nicht nur durch den Kontext des Textes, der vor dem jeweiligen Wort vorkommt, sondern auch derjenige Kontext nach diesem Wort. Sie sind für viele Klassifikationsaufgaben nützlich.

Weiterführende Quellen zu Sprachmodellen/ChatGPT:

- Infografiken zu ChatGPT: https://www.plattform-lernende-systeme.de/infografiken.html
- Erläuterung, wie Sprachmodelle funktionieren: https://www.spektrum.de/news/wie-funktionieren-sprachmodelle-wie-chatgpt/2115924
- Ausführliche Erklärung, wie ChatGPT funktioniert: https://www.golem.de/news/kuenstliche-intelligenz-so-funktioniert-chatgpt-2302-171644.html?utm_source=pocket-newtab-global-de-DE
- Erläuterung, wie und warum wir gerade einen Fortschritt bei generativen KI-Modellen erleben: https://arstechnica.com/gadgets/2023/01/the-generative-ai-revolution-has-begun-how-did-we-get-here/

Limitationen von aktuellen Sprachmodellen

Da Sprachmodelle auf der Vorhersage von Wörtern basieren sowie auf Ähnlichkeitsbeziehungen beruhen, ergeben sich damit auch eine Reihe von Limitationen. Die Resultate des KI-Modells sind vom Trainingsdatensatz abhängig, das heißt nur Wortfolgen, die dort vorkommen, können für die spätere Generierung von Inhalten auch eine Rolle spielen. Beispielsweise sind im Datensatz von ChatGPT nur Daten bis zum Jahr 2021 enthalten, sodass Ereignisse danach keinen Einfluss auf das Modell haben. Aber auch bestimmte Informationen können im Datensatz fehlen, weil gesetzliche Regelungen dies unterbinden oder weil sich in der textbasierten Kommunikation unter Menschen auf „natürliche Weise" Lücken gebildet haben. Andersherum können sich in den Datensätzen auch ungerechte gesellschaftliche Tendenzen abbilden (Bias), die durch das KI-Modell reproduziert werden, wie etwa Rassismus oder Frauenfeindlichkeit.

Da moderne Sprachmodelle wahrscheinlichkeitsbasiert sind, spielt das Allgemeine und Häufige eine größere Rolle bei der Generierung von Inhalten als eher selten auftretende Ereignisse und Anomalien, die unter Umständen aber dennoch von Bedeutung für bestimmte Themen sein können. Es mangelt zudem oft an Konsistenz, das heißt bei identischen Anfragen wird nicht unbedingt jeweils die gleiche Antwort generiert. Dies ist dann besonders problematisch, wenn feststehende Fakten oder Tatsachen abgefragt werden. Logische Zusammenhänge werden schließlich lediglich über Ähnlichkeitsbeziehungen abgebildet, weshalb es den Modellen an Fähigkeiten für die logische Schlussfolgerung und Mathematik mangelt. Die gegenwärtige Forschung nimmt sich solcher Limitationen bereits an (siehe Kapitel 3 Forschungsperspektiven).

Große Sprachmodelle wie GPT-4, ROBERTa, T5 oder BLOOM produzieren erstaunlich einsichtsvolle Antworten. Wie gut generalisieren diese Modelle wirklich? Oder ist das nur auswendig gelerntes Wissen aus riesigen Mengen an Trainingsdaten, die ein einzelner Mensch nie im Leben lesen und verarbeiten könnte?

Ein Problem der großen Sprachmodelle ist, dass sie eben nicht auswendig lernen. Im Grunde genommen ist es eine Generalisierung, die aus mathematischer Sicht auf Ähnlichkeiten, in einem hochdimensionalen Vektorraum, basiert. Für eine echte KI benötigen wir aber mehr als Ähnlichkeiten, nämlich Verständnis und Bewertung lebensweltlicher Kontexte für das logische Schließen, für das Erkennen von Widersprüchen oder wenn das Modell nicht halluzinieren soll.

Wie ist das genau zu verstehen?

Trainingsdaten enthalten inhärent viel Logik, die wir in unserer Sprache abbilden. Wenn das Modell diese Logik lernen könnte, wäre das cool. Große Sprachmodelle lernen diese Logik aber nicht, sondern bilden diese in einem Ähnlichkeitsraum ab: So werden Shortcuts anstatt von Fakten gelernt und eher kürzlich gesehene Zusammenhänge erfasst, Recency Bias.

Hinrich Schütze, CO-Direktor des Zentrums für Informations- und Sprachverarbeitung an der Ludwig-Maximilians-Universität München (LMU); Munich Center for Machine Learning (MCML)

Viele der genannten Punkte können auch dazu führen, dass Sprachmodelle „halluzinieren“, das heißt, sie stellen logische Zusammenhänge her, wo keine sind, generieren also falsche Aussagen zu Themen, zu denen sie eigentlich nichts oder nur sehr wenig wissen können, oder erstellen Inhalte, die vom eigentlichen Thema wegführen: So erfindet ChatGPT teilweise Quellenangaben, die es überhaupt nicht gibt. Insgesamt sind Sprachmodelle gegenwärtig „One-Size Fits (NOT) All“-Lösungen. Es mangelt an der Modularität für all die zahlreichen Fälle, die nicht mit Ähnlichkeitsbeziehungen sowie dem Vergleich und der Neukombination von im Datensatz existierenden Wortfolgen gelöst werden können.

Erklärbox 2

Übersicht zu Sprachmodellen am Beispiel ChatGPT:

ChatGPT kann …

- *Natürlich klingende Konversationen führen*
- *Texte aller Art erstellen (Aufsätze, Gedichte, Zusammenfassungen, Kochrezepte etc.)*
- *Gewünschten Stil imitieren (sachlich, poetisch etc.)*
- *Programmcode generieren*
- *Texte übersetzen u. v. m.*

basiert auf …

- *Allgemeinen Textdaten aus dem Internet (Wikipedia etc.)*
- *Büchern*

funktioniert …

- *Auf Basis von Wahrscheinlichkeiten*
- *Mittels KI-Algorithmen und mit aufwändigem Training unter Nutzung menschlicher Bewertungen von generierten Inhalten*

Grenzen

- *Liefert immer eine Antwort (auch wenn die Datenbasis nicht ausreichend Informationen zur Anfrage enthält)*
- *Erfindet bisweilen Inhalte oder Quellen („halluziniert“)*
- *Und weitere …*

Quelle: Eigene Zusammenstellung.

Große Sprachmodelle: Besondere Eigenschaften

Steigende Performanz, größere Modelle: Aufgrund der sehr guten Skalierbarkeit dieser Verfahren ist seit einiger Zeit ein Wettbewerb zwischen den jeweiligen marktdominierenden Firmen wie Google, Meta oder auch Huawei und anderen Mitbewerbern zu beobachten. Es werden immer größere Sprachmodelle mit immer mehr Parametern entwickelt, die auf immer größeren Datenmengen basieren und für deren Training deutlich mehr Rechenleistung benötigt wird als in früheren Zeiträumen (siehe Abbildung 3). Mit der Modellgröße sollte dabei auch die Menge der verwendeten Trainingsdaten wachsen (Benaich & Hogarth, 2022, S. 26). Diese Entwicklung ist deshalb von Bedeutung, weil sich gezeigt hat, dass mit zunehmender Modellgröße Fähigkeiten möglich werden, an die die Entwicklerinnen und Entwickler zuvor nicht gedacht haben, was häufig auch als Emergenz von Fähigkeiten bezeichnet wird (siehe Zitat Kristian Kersting). Noch größere Modelle als die heutigen könnten also mit zusätzlichen Fähigkeiten überraschen. Das beschriebene Wachstum wird jedoch auch mit Skepsis betrachtet, da dadurch drängende Probleme der Modelle nicht gelöst werden.[1]

Dialogmodelle wie ChatGPT auf GPT-3 Basis von OpenAI sind aktuell ein heißes Thema, auch multilinguale Modelle. Ich bin etwas skeptisch, wie weit die Entwicklung bei großen Sprachmodellen gehen sollte: noch mehr Daten, noch größere Modelle? Wir sind aus meiner Sicht an einer Grenze angelangt. Denn ich sehe nicht, wie die aktuellen fundamentalen Probleme dieser Sprachmodelle wie Halluzination, fehlende Logik, fehlende Erklärbarkeit gelöst werden können. Natürlich gibt es bereits viele sinnvolle Anwendungen. Die Benutzer haben trotz der genannten Probleme gelernt, mit den noch vorhandenen Schwachstellen umzugehen.

Hinrich Schütze, LMU, MCML

Je größer und effizienter die Modelle, desto besser scheinen sie zu funktionieren. Dieser Trend setzt sich nach wie vor fort: doppelte Größe = doppelte Performance. Deswegen sind Forschende sowie Firmen enthusiastisch über die Technologie, deren jetzige Möglichkeiten und zukünftiges Potenzial.

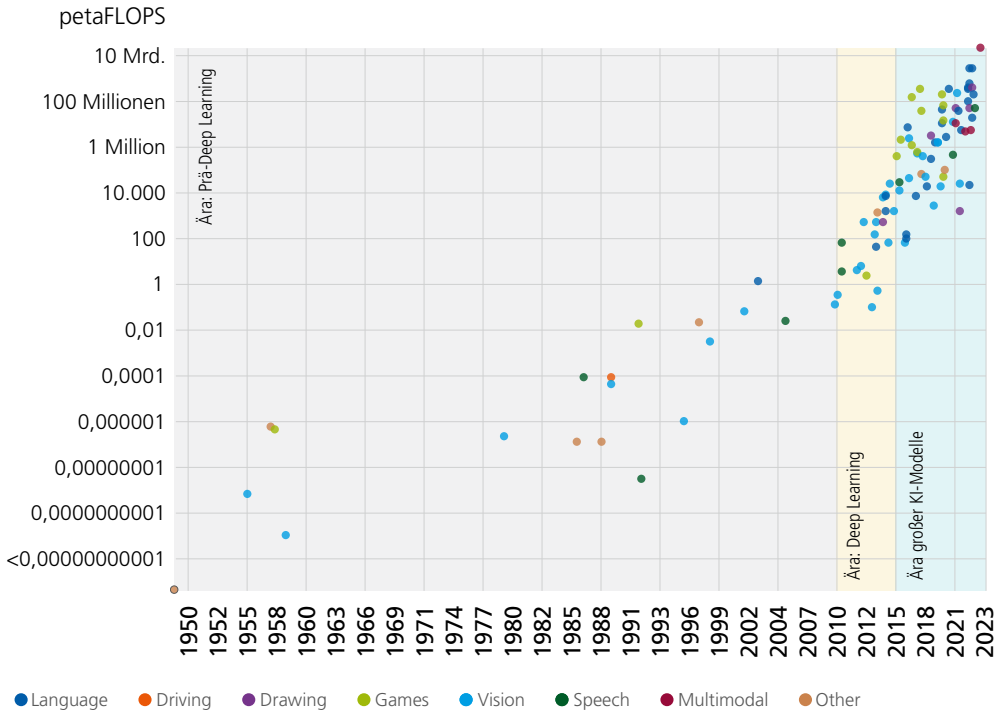
Timo Möller, Mitgründer bei der deepset GmbH (Zitat vom Juli 2022)

1 Weitere Perspektiven von ExpertInnen finden sich unter KI-Perspektiven.

Zukünftige Modelle, die noch größer sind als ChatGPT, könnten uns mit zusätzlichen Fähigkeiten überraschen. So oder so ist es wichtig, dass Deutschland und Europa in diesem Bereich ganz vorne mitmischen. Dazu brauchen wir ein leistungsfähiges KI-Ökosystem. Nur so können wir KI-Systeme nach unseren Vorstellungen gestalten und eine KI-Kreislaufwirtschaft ‚Made in Europe' etablieren.

Kristian Kersting, Professor an der Technischen Universität Darmstadt, Mitglied der Arbeitsgruppe Technologische Wegbereiter und Data Science

Abbildung 3: Benötigte Rechenleistung für bedeutende KI-Modelle nach Domäne (in petaFLOPS)



1 petaFLOP sind 10^{15} FLOPS

Quelle: Die Visualisierung der genutzten Rechenleistung für das Training von bedeutenden KI-Modellen nach Domänen in petaFLOPS beruht auf einer Datenauswahl von Sevilla et al. (2022b). Zur übersichtlicheren Darstellung der Daten wurde eine logarithmische Skalierung der Werte gewählt. Die Einteilung in die drei zeitlichen Phasen erfolgt nach Sevilla et al. (2022a), ebenso wie die Identifikation bedeutender Modelle. Unter den größten Modellen in der Ära seit 2016 befinden sich Sprachmodelle und multimodale Modelle, die Parameterzahlen im dreistelligen Milliardenbereich oder sogar darüber hinaus aufweisen (für GPT-4 siehe obere rechte Ecke des Diagramms). Es wurden allerdings bereits Modelle mit über einer Billion Parameter erstellt (siehe Switch Transformer von Google mit 1,6 Billionen Parametern oder Wu Dao 2.0 der Akademie für Künstliche Intelligenz Peking mit 1,75 Billionen).

Hohe Wiederverwendbarkeit eines Modells und vielfältige Einsatzmöglichkeiten: Während frühere Verfahren des maschinellen Lernens Modelle hervorbrachten, die hoch spezialisiert auf eine Aufgabe waren, sind große Sprachmodelle allgemeiner. Ein wesentlicher Fortschritt besteht vor allem darin, dass diese Modelle nicht mehr nur für eine bestimmte Aufgabe trainiert werden, sondern in vielfältiger Weise zur Lösung unterschiedlicher Aufgaben beitragen können. So können sie als Basis dienen, um durch Anpassungen

mit kleineren, domänenspezifischen Datenmengen standardisierte und schnelle Lösungen für unterschiedliche Anwendungsgebiete bzw. Anforderungen einzelner Disziplinen, Unternehmen und Organisationen zu entwickeln. Daher werden sie manchmal auch Grundlagen- oder Basismodell (engl. Foundation Modell) genannt, um diese Charakteristika zu betonen. Das kosten- und zeitintensive Training eines eigenen großen Sprachmodells kann entsprechend für Anwenderinnen und Anwender entfallen.

Erklärbox 3

Anpassung großer Sprachmodelle

Hintergrund

Anwendungen, wie Chatbots oder Software, die menschenverachtende Aussagen erkennen können (siehe hate speech), passen generelle Repräsentationen für Prozesse zur Erzeugung von Sprache weiter an. Dazu werden zusätzliche Trainingsdaten benötigt, die aufgaben- oder anwendungsspezifisch sind. Je nach Adaptionsmethode (z. B. Finetuning, Prompting, Few-Shot Learning etc.) variiert die Anzahl der benötigten Trainingsdaten.

Die Anpassung eines KI-Modells an eine bestimmte Domäne ist ein Spezialfall des sogenannten Transfer Learning. Hierbei werden bestehende KI-Modelle und bereits aufgewendete Ressourcen in anderen Kontexten wiederverwendet, um Modelle mit vergleichsweise geringerem (Trainings-)Aufwand auf andere Problem- oder Aufgabenstellungen zu spezialisieren.

Möglichkeiten der Modellanpassung

Prompt Engineering: Als Prompt wird eine Anfrage an ein Sprachmodell bezeichnet. Prompt Engineering entspricht daher systematischen Vorgehensweisen, um solche Anfragen so zu gestalten, dass sie zu optimalen Ergebnissen führen. Auf diese Weise kann das Modell effektiv für spezifische Aufgaben angepasst werden, ohne dass ein Fine Tuning oder Nachtrainieren des Modells notwendig wird.

Few-Shot Learning: Unter diesem Begriff werden Techniken und Methoden zusammengefasst, die eine effektive und effiziente Anpassung eines vortrainierten Machine-Learning-Modells an bisher unbekannte Klassen oder Problemstellungen ermöglichen. Few-Shot-Learning-Methoden erlauben die Anpassung eines Modells an eine neue Problemstellung durch die Nutzung weniger Beispiele, um dennoch zuverlässige Vorhersageergebnisse zu erzielen.

Fine Tuning: Unter Feinabstimmung versteht man in der maschinellen Sprachverarbeitung das erneute Trainieren eines vortrainierten Sprachmodells mit eigenen Daten. Als Ergebnis des Fein-Tunings werden die Gewichte des ursprünglichen Modells aktualisiert, um die Eigenschaften der Domänendaten und der spezifischen Aufgabe, die von Interesse sind, zu berücksichtigen. Da es mit zunehmender Modellgröße und Aufgabenbreite ineffizient wird, die Gewichte allgemein zu aktualisieren, wurden sogenannte Parameter-effiziente Varianten entwickelt, wie etwa Adapter-Module, die zwischen Schichten des Transformers eingefügt werden.

Anpassung mit manuell erzeugten Beispielen: Modellanpassung anhand von Beispielen, die von Menschen erstellt wurden, und/oder anhand von Bewertungen des Outputs des KI-Systems durch Menschen (siehe Reinforcement Learning from Human Feedback bei ChatGPT).

Quelle: Eigene Zusammenstellung.

Die den Sprachmodellen zugrundeliegende KI-Technologie ist modalitätsagnostisch: Die Transformer-Architektur wurde ursprünglich für die maschinelle Übersetzung entwickelt. Eine Wortfolge in einer Sprache wird in eine Wortfolge in einer anderen Sprache übersetzt. Transformer sind jedoch universell für jede Aufgabe einsetzbar, die als Übersetzung einer Sequenz in eine andere Sequenz aufgefasst werden kann, nicht nur für Sprachübersetzungen. So gibt es inzwischen eine Vielzahl unterschiedlicher Generatoren wie Text-zu-Bild, Text-zu-Video oder Text-zu-Programmcode. Darüber hinaus wird auch die textbasierte Steuerung von Robotern erforscht, also Text-zu-Bewegung. Durch diese speziellen Eigenschaften der Transformermodelle ergibt sich ein sehr breites Anwendungspotenzial.

Die KI-Technologie der Sprachmodelle ist allgegenwärtig: Sprachmodelle können als das betrachtet werden, was in der Wirtschaftswissenschaft als Allzwecktechnologie bezeichnet wird. Allzwecktechnologien beziehen sich auf Technologien wie die Dampfmaschine und die Elektrizität, die Transformationen und Produktivitätssteigerungen auslösen und im Laufe der Zeit durch ihre Verbreitung und Verbesserung ergänzende Innovationen hervorbringen (z. B. eine Vielzahl von Produkten und Dienstleistungen rund um ein Kernprodukt). Der hohe Grad an Wiederverwendbarkeit und Anwendbarkeit von Sprachmodellen könnte beispielsweise mit der Einführung relationaler Datenbanken Ende der 1970er Jahre verglichen werden. Ursprünglich nur für die Compliance im Rechnungswesen gedacht, hat sich die Technologie schnell in fast allen Unternehmensbereichen durchgesetzt. Heute ist sie in fast allen Cloud-Plattformen und zahlreichen Anwendungen enthalten. Während die Einsatzmöglichkeiten von Text-, Bild- oder Videogeneratoren in den Bereichen Medien, Design und Kunst offensichtlich geworden sind und in der Öffentlichkeit intensiv diskutiert werden, sind andere Anwendungsfelder generativer KI weniger sichtbar. So kann dieser KI-Typus auch das Design von Medikamenten, Werkstoffen, Computerchips und Komponenten für Maschinen und Produkte unterstützen und synthetische Daten erzeugen (siehe Tabelle 1, Übersicht über Anwendungsfälle für generative KI). Auch das automatische Design von Data-Science-Prozessen selbst ist ein Anwendungsfeld, das den Einsatz von maschinellem Lernen insgesamt vereinfachen kann.

Tabelle 1: Einsatzpotenzial von generativer KI in unterschiedlichen Industriezweigen

	Journalismus, Marketing, Public Relations	Fahrzeugbau und Automotive	Architektur und Ingenieurwesen	Energie und Versorgung	Gesundheitsversorgung	Produktion elektronischer Produkte	Produktion	Pharmazeutische Industrie
Synthetische Medieninhalte	■							
Arzneimittel-Design								■
Design neuer Materialien		■		■		■		
Chip-Design						■		
Synthetische Daten		■	■	■	■	■	■	■
Generatives Design (von Produktteilen)		■	■				■	

Quelle: Zusammenstellung auf der Basis von Gartner (2023) und Heesen et al. (2023).

Sprachmodelle sind weitgehend zugänglich: Ein wichtiger Treiber für den Siegeszug relationaler Datenbanken war der verhältnismäßig einfache Zugang über die Programmiersprache für Datenbankabfragen SQL (Structured Query Language), mit der bereits Studierende im ersten Semester Abfragen programmieren können. Trotz der eingangs geschilderten Dominanz großer, privater Unternehmen beim Aufbau großer Sprachmodelle hat sich durch die Open Source Community um Hugging Face, Eleuther.AI, LAION.AI und die APIs weiterer Anbieter ein relativ einfacher Zugang zu dieser mächtigen Technologie entwickelt: Existieren für eine Sprache bzw. eine Aufgabe oder Domäne bereits Sprachmodelle, ist es für Unternehmer deutlich einfacher als noch vor fünf Jahren, einfache Anwendungen zu testen. Dies liegt vor allem am freien Zugang zu Modellen, die im Kreise dieser Communitys entwickelt, gesammelt und dokumentiert werden, so dass die Nutzungshürden geringer sind. Für die deutsche Sprache fehlen leider solche Modelle noch weitestgehend. Darüber hinaus kann hinsichtlich der Zugänglichkeit von einer eingeschränkten Demokratisierung der KI gesprochen werden, da häufig auf außereuropäische Modelle privater Unternehmen zurückgegriffen werden muss.

3 Perspektiven aus Forschung und Entwicklung

Die derzeit vielbesprochenen Modelle, wie GPT-4 ober Stable Diffusion, zeigen, dass Forschungsergebnisse in beeindruckende Anwendungen gemündet sind, wie etwa Text- und Bildgeneratoren. Sprachmodelle und die ihnen oft zugrundeliegende Technologie der Transformer gelten als Forschungsgebiet mit enormem Potenzial – auch über verschiedene Modalitäten hinweg. Im Folgenden werden zunächst allgemein bedeutende Forschungsfelder vorgestellt, um danach gesondert auf die wichtigen Felder der Forschung zu Grounding und zur Erkennung von Bias in Sprachmodellen und Textsammlungen einzugehen.

Bedeutende Forschungsfelder

Die Forschung aus Deutschland ist an vielen Stellen in die aktuelle Entwicklung bei großen KI-Modellen involviert. So haben laut des AI Index 2023 Forschende aus Deutschland im Jahr 2022 zu 3,12 Prozent der großen Sprachmodelle und multimodalen Modelle beigetragen, die das Steering Committee des Index für die Analyse herangezogen hat. Deutschland war damit an sechster Stelle der einbezogenen Länder (Maslej et al., 2023, S. 58). Die folgende schlaglichtartige Zusammenstellung bildet keineswegs die gesamte Forschung in Deutschland zum Thema große Sprachmodelle ab. Sie bietet jedoch einen kleinen Einblick in die Vielseitigkeit, mit der sich Forschende aus Deutschland beim Thema große Sprachmodelle engagieren.

- In den KI-Kompetenzzentren forschen beispielsweise das LAMARR-Institut oder das Munich Center for Machine Learning (MCML) auf diesem Gebiet. Das Deutsche Forschungszentrum für Künstliche Intelligenz (DFKI) sowie das Fraunhofer-Institut für Intelligente Analyse- und Informationssysteme (IAIS) und das Institut für Integrierte Schaltungen (IIS) sowie das Forschungszentrum Jülich und das Start-up Aleph Alpha sind an der Umsetzung des Projekts Open GPT-X beteiligt. Das Modell wird mit dem Supercomputer „JUWELS“ am Rechenzentrum in Jülich trainiert. Es ist Teil des, durch das Bundesministerium für Bildung und Forschung und Bundesländer geförderten, Gauss Centre for Supercomputing.

- Das Forschungszentrum Data Science an der Berliner Hochschule für Technik ist führend in Deutschland im Design und der Augmentierung für medizinische Sprachmodelle, z. B. für die Differenzialdiagnose auf 1200 Krankheitsbildern und 700 Prozeduren (Van Aken et al., 2021; Grundmann et al., 2022), in der Erklärbarkeit für diese Krankheitsbilder (Van Aken et al., 2022) sowie bei deutschsprachigen klinisch relevanten Open-Source-Modellen, (MedBert, Bressem 2023)

- Die Forschung hat sich mit Modellen an der Schnittstelle zwischen Bild und Text befasst (Stable Diffusion, Rombach et al., 2022) und mit Tests zur gegenseitigen Bestätigung zwischen den beiden Modalitäten (Li et al., 2022). Darüber hinaus arbeitet der Verein LAION.AI e. V. an wichtigen Datengrundlagen für Bildgeneratoren (Schuhmann, 2022).

- An Forschung zur Schnittstelle zwischen Sprachmodellen und Robotik ist zum Beispiel die TU Berlin beteiligt (Pathways Language Model-Embodied, kurz PaLM-E; Driess et al., 2023).

- Die Erklärbarkeit von Sprachmodellen ist ebenso ein Forschungsthema (Van Aken et al., 2019, Brath et al., 2023; Deb et al., 2023).

- Schließlich haben Forschende verschiedener deutscher Hochschulen zu Benchmarks für die Evaluation von Sprachmodellen beigetragen (BIG-bench, Srivastava et al., 2022) oder zur Konzeption bekannter Gedankenexperimente, um die Fähigkeit solcher Modelle bei der Erfassung der Bedeutung von Phänomen reflektieren zu können (siehe Erklärbox 4: Oktopus-Test; Bender & Koller, 2020).

Dies zeigt, dass Deutschland in der Forschung einen gewichtigen Beitrag bei der Weiterentwicklung großer Sprachmodelle leistet und auch künftig leisten kann. Vor diesem Hintergrund werden im Folgenden einige bedeutende Forschungsbedarfe skizziert:

Vortrainierte große Sprachmodelle für die deutsche Sprache: Für die deutsche Sprache existieren nur wenige große Sprachmodelle, dazu gehören Modelle von deepset sowie das auf Hugging Face verfügbare GermanBert-Modell. Diese Modelle basieren auf einer mittlerweile nicht mehr zeitgemäßen Architektur des BERT-Modells aus dem Jahre 2018. Sie sind zudem auf 512 bzw. 768 Token beschränkt und nur allgemein auf die deutsche Sprache trainiert, also nicht auf domänenspezifischen Sprachgebrauch. Damit eignen sich diese Modelle nur eingeschränkt für heutige herausfordernde Anwendungen. Denn Anwendungen mit einer speziellen Aufgabe oder einer spezifischen Domäne (z. B. Maschinenbau, Dienstleistungen, Gesundheit, Rechtswesen, Bildung etc.) benötigen erhebliche Anpassungsleistungen. Moderne Modelle wie GPT-4 haben deutlich mehr Parameter und damit eine bessere Vorhersage-Leistung für zahlreiche Aufgaben. Sie sind jedoch häufig nicht Open Source verfügbar. Eine bekannte Ausnahme wie etwa das offene Modell BLOOM ist dagegen nicht mit Daten in deutscher Sprache trainiert worden.

Dies bedeutet zum einen, dass deutsche Nutzerinnen und Nutzer, wie beispielsweise Unternehmen, Prompts und Kontext einer Anfrage an einen Drittanbieter senden müssen, wenn sie mit Modellen wie GPT-4 arbeiten wollen. Diese Anbieter, wie in diesem Fall Open AI, unterliegen jedoch nicht dem europäischen Rechtsraum. Dies kann eine Herausforderung für den Umgang mit personenbezogenen Daten und Daten, die unter das Geschäftsgeheimnis fallen, darstellen. Solche Daten können Anfragen zur Unterstützung bei der Programmierung von Software sein, aber auch Anfragen aus Bereichen wie Forschung und Innovation, Verwaltung und Kundenakquise. Darüber hinaus kann es für deutsche Unternehmen und Behörden schwierig sein, Maßnahmen zum Schutz von Minderheiten oder gegen rechtswidrige Praktiken (wie Begegnung von Bias, diskriminierenden Inhalten etc.) bei nicht-offenen außereuropäischen Diensten zu überprüfen. Ethische Regeln werden bei außereuropäischen Diensten nach den Maßstäben der anbietenden Unternehmen umgesetzt, was sich wiederum in der Auswahl der Trainingsdaten bzw. Methoden niederschlagen kann. Hier ist unklar, ob deutsche und europäische Standards berücksichtigt werden. Dies kann eine Herausforderung für die Nutzung und Anwendung solcher Modelle in Deutschland und Europa darstellen.

Daher ist zum einen die Forschung und (Weiter-)Entwicklung von modernen, offenen multilingualen Modellen sowie Modellen für die deutsche Sprache in Deutschland und Europa sinnvoll. Das ermöglicht insbesondere in sensiblen Domänen (Medizin, Justiz, Behörden, Forschung und Entwicklung, Sicherheit etc.) eine Umsetzung unter Berücksichtigung des deutschen und europäischen Rechts und hiesiger Werte.

Damit sowohl Behörden als auch Unternehmen Dienste auf der Basis von Sprachmodellen für bestimmte für Deutschland spezifische Zieldomänen nutzen können, ist auch die Entwicklung von Methoden nötig, die eine schnelle, effiziente und kostengünstige Modellanpassung an spezifische Aufgaben und Domänen ermöglichen, um den Transfer in die wirtschaftliche Anwendung zu erleichtern.

Multimodalität: Große Sprachmodelle profitieren von zusätzlichen Kontextinformation, wie sie durch Multimodalität integriert werden kann, das heißt, wenn jenseits von Textdaten auch auf Bilder, Ontologien, Mengen, Tabellen, zeitvariante Daten, Bewegungstrajektorien etc. zurückgegriffen wird. Dabei sind diese zusätzlichen Kontextinformationen zu den Textdaten sowohl wechselseitig überlappend als auch getrennt voneinander. Abhängig vom Grad der Überlappung und vom Szenario bieten sich Methoden der latenten Informationsintegration an, die Textdaten und andere multimodale Daten verknüpfen. Auf der anderen Seite können auch Daten aus der einen Modalität einen Aufmerksamkeitsfilter trainieren (Attention), der dann in der anderen Modalität angewandt wird, also etwa auf Bilder, Videos oder Sensordaten. Schließlich können sich Modelle auf Basis unterschiedlicher Modalitäten in der Interaktion miteinander gegenseitig verbessern und damit beispielsweise das Bild- oder Textverstehen (siehe hierzu Blickpunkt: Kontext und Grounding, Li et al., 2022).

Multilingualität: Modelle wie BLOOM oder XLM-R sind multilingual trainiert. Das bedeutet, die Trainingsdaten kommen aus mehreren Sprachfamilien und es sind auch Übersetzungsaufgaben möglich. Multilinguale Modelle können fehlendes Kontextwissen in der einen Sprache durch vorhandenes Kontextwissen in der anderen Sprache unter Umständen kompensieren. Das ist besonders hilfreich in Domänen mit einer globalen Nomenklatur bzw. einem globalen Standard oder Grounding. Ein anschauliches Beispiel ist die Medizin, in der multilinguale Vorhersagesysteme zur Diagnose und Diagnostik-Unterstützung auf gemeinsame globale Standards wie dem ICD (International Statistical Classification of Diseases and Related Health Problems) oder CCS (Klassifikationen der Canadian Cardiovascular Society) deutlich bessere Ergebnisse als monolinguale Systeme aufweisen konnten (Papaioannou et al., 2022). Aufgrund der im weltweiten Vergleich eher kleinen Community für die deutsche Sprache – im Gegensatz zu Englisch, Französisch, Mandarin oder Spanisch – ist daher eine Forschung an multilingualen Modellen für Deutsch und andere Sprachen kleinerer Sprachgemeinschaften zu unterstützen.

Deutschsprachige Sprachmodelle – eine Option zu multilingualen englischen Modellen?

Dafür benötigt man zunächst ernsthafte Forschung zu monolingualen Modellen, die nicht englisch sind. Im Englischen ist die sprachliche Form-Vielfalt eher gering und daher ist Englisch ein Glücksfall für die Transformer-Architekturen. Im Deutschen gibt es Morphologien und Kompositionen in großer Vielfalt. Gerade weil wir sagen, dass die unteren Ebenen der Transformer symbolisch und regelbasiert sind, ist das eine Herausforderung für ein monolinguales Modell. Bei den multilingualen Modellen gibt es aber so viele zusätzliche Probleme, sodass sich ein monolinguales deutschsprachiges Modell lohnt.

Hinrich Schütze, LMU, MCML

Kombinierte KI bzw. hybride KI: Bedeutende Fortschritte konnten bei Sprachmodellen auch durch die Kombination von mehreren Methoden erreicht werden, die unter anderem zusätzliche Kontexte zur Verfügung stellen. Potenzial liegt in der Kombination von wissensbasierter KI und Sprachmodellen, etwa wenn Wissensgraphen einbezogen werden, um schnellere Anpassungen an Fakten und Gegebenheiten zu ermöglichen. Durch ChatGPT wurde die Kombination von Sprachmodellen mit verstärkendem Lernen auf der Basis von menschlichem Feedback bekannt (Engl.: Reinforcement Learning from Human Feedback, RLHF). Auf diese Weise wurde unerwünschtes Output (z. B. diskriminierende Inhalte) durch eine gelernte Policy eingeschränkt. Obwohl ChatGPT und ähnliche Ansätze mit RFHL sehr vielversprechend und effektiv sind und die Aufmerksamkeit der größten KI-Forschungslabore auf sich gezogen haben, gibt es noch deutliche Einschränkungen. Die Modelle werden zwar besser, können aber immer noch schädliche oder sachlich ungenaue Texte produzieren. Diese Unvollkommenheit stellt eine langfristige Herausforderung dar. Darüber hinaus ist die Generierung gut geschriebener menschlicher Texte zur Beantwortung bestimmter Eingabeaufforderungen bei RLHF aufwändig und kostspielig, denn hierfür wird viel menschliche Arbeitskraft benötigt (siehe hierzu auch Clickworking bei Open AI). Somit lassen sich diese Methoden unter anderem für KMU nur schwer umsetzen.

Kontext in langen Texten verstehen: In Berichten sind Informationen zu Entitäten, Sachverhalten und Zusammenhängen über mehrere Abschnitte oder sogar Seiten verteilt. Typische Aufgaben umfassen daher, dass in den Texten Informationen übergreifend verdichtet und Zusammenhänge erkannt werden sollen. Sprachmodelle sind jedoch oft auf eine bestimmte Kontext-Textlänge von mehreren hundert, selten wenigen tausend Wörtern beschränkt; zum Beispiel kann der „LongFormer" 2048 Token verarbeiten. Die Entwicklung geht jedoch seit 2023 auch hin zu sehr großen Kontextfenstern. Das nicht offen zugängliche GPT-4 verfügt über einen Kontext von 32.000 und mehr Token. Spannend sind daher offene Sprachmodelle für bestimmte Domänen der deutschen Sprache, die ebenfalls besonders lange Texte und Kontexte, beispielsweise aus den Prompts, gut verarbeiten können.

Erklärbarkeit: Deep Learning im Allgemeinen und damit auch Transformer-Modelle im Speziellen zeichnen sich durch sogenannte Black-Box-Modelle aus, d. h. es ist nicht nachvollziehbar, wie genau ein Modell zu seinen Ergebnissen kommt und was bei der Verarbeitung der Daten passiert, bis ein Ergebnis generiert wird. Es ist jedoch erstrebenswert, Entscheidungen hinterfragen und Erklärungen einfordern zu können. Die Forschung zur Erklärbarkeit von KI zielt darauf ab, diese Situation zu verändern, da so auch nachvollzogen werden kann (Deb et al., 2023), ob Ergebnisse beispielsweise auf unerkannten Bias oder Korrelationen beruhen (siehe Kluger-Hans-Effekt). Die Methoden für die Erklärbarkeit von KI können genutzt werden, um Modelle zu verbessern und weiterzuentwickeln, damit sie das Vertrauen der Nutzenden in KI stärken können. Moderne Dialogsysteme bieten den Nutzenden die Möglichkeit, Nachfragen zu stellen (siehe ChatGPT). Solche Optionen für Nutzende sollen weiter ausgeschöpft werden (Samek 2023).

Differenzierbare Tokenizer: Tokenizer sind ein Verfahren zur Zerlegung von Zeichenketten und Multi-Wörtern in logische und zusammengehörige Einheiten und sind daher wichtig als Grundlage für große Sprachmodelle. Sie sind stark abhängig von der Domäne und erfolgen momentan oft regelbasiert. Dadurch können die Informationen aus der Worttrennung nur unzureichend an die Sprachmodelle im Fine-Tuning weitergegeben werden, was die Anpassung an Domänen erschwert.

Benchmarks: Um große Sprachmodelle zu evaluieren, werden in der Forschung sogenannte Benchmarks verwendet. Dabei werden Modelle hinsichtlich ihrer Leistung bei bestimmten Aufgaben getestet, wie zum Beispiel beim logischen Schlussfolgern, dem Lösen mathematischer Aufgaben oder dem Beantworten von Fragen ohne/oder mit Kontextinformationen. Dazu gehören ebenfalls Tests, die zeigen, wie das Modell im

Vergleich zum Menschen abschneidet. Aktuelle Benchmarks wie der „Big Bench" wurden auch dafür entwickelt, die Grenzen großer Sprachmodelle zu erkunden sowie die Emergenz von Fähigkeiten zu untersuchen.

Insbesondere für die deutsche Sprache existieren im Vergleich zur englischen Sprache zu wenige Benchmarks. Wenige erfolgreiche Beispiele sind das Benchmark „GermEval" und vereinzelt auch Aufgaben in „SemEval". Für zahlreiche wichtige Domänen für die deutsche Sprache, wie den Gesundheits-, Industrie-, Dienstleistungs- oder Bildungssektor und spezifischer für den Maschinen- und Fahrzeugbau sowie für das Rechtswesen und viele mehr, fehlen diese Benchmarks. Bedeutende Benchmarks sind:

Tabelle 2: Übersicht über bedeutende Benchmarks

Holistic Evaluation of Language Models
• Dient der Verbesserung der Transparenz von Modellen • Themen: Genauigkeit, Kalibrierung, Robustheit, Fairness, Bias, Toxizität (z. B. Schmährede, Hetze etc.) und Effizienz • Evaluation, basierend auf 26 Szenarien zu bestimmten Aspekten (z. B. Wissen, Argumentation, Auswendiglernen/Copyright, Desinformation)
Massive Text Embedding Benchmark
• Umfasst 8 Embedding-Aufgaben und deckt insgesamt 58 Datensätze und 112 Sprachen ab • Aufgaben: Clustering, Klassifizierung, Klassifizierung von Text-Paaren, Ranking, automatische Zusammenfassung, Retrieval (z. B. Auffinden relevanter Dokumente im Textdatensatz), BiText-Mining (d. h. Auffinden von Satzentsprechungen in verschiedenen monolingualen Datensätzen)
Beyond the Imitation Game Benchmark (BIG-bench)
• Umfasst über 204 Aufgaben • Die Aufgaben umfassen ein vielfältiges Themenspektrum von der Linguistik über Mathematik, Common Sense und Bias bis hin zur Softwareentwicklung

Quelle: Eigene Zusammenstellung basierend auf Liang et al. (2022), Muennighoff et al. (2022) Srivastava et al. (2022).

Schließlich ist die **Datensammlung und -aufbereitung für große Sprachmodelle** zu nennen. Beides ist mit einem großen Aufwand verbunden und Methoden, um diese effizient und effektiv umzusetzen sowie die Datenqualität dabei zu berücksichtigen, sind ein Forschungsgebiet. Es gilt, effiziente Methoden zur Erstellung effektiver Textdatensätze mit möglichst wenig menschlicher Annotationsarbeit, wie sie etwa bei ChatGPT im Rahmen des verstärkenden Lernens auf Basis menschlichen Feedbacks notwendig war, zu entwickeln. Schließlich wird für das Training großer KI-Modelle viel Energie für die Berechnungen notwendig: „Google gab an, dass das Training seines Sprachmodells PaLM so viel Energie verbraucht hat, wie 300 US-Haushalte in einem Jahr" (Bischoff, 2023). Die Verbesserung der **Energieeffizienz** sollte daher ebenso Fokus der Forschung sein.

Weitere Forschungsfelder, die sich aus den Limitationen aktueller Sprachmodelle ergeben

Um mittel- bis langfristig das Sprachverstehen solcher Modelle zu verbessern, sind folgende Herausforderungen anzugehen (Goldberg, 2023):

- **Multiple Texte in eine Beziehung zueinander stellen:** Sprachmodelle können Texte oft nicht mit Ereignissen in der „realen Welt" in Bezug setzen. Das kann einerseits bedeuten, dass das Modell nicht versteht, dass sich verschiedene Texte auf das gleiche Ereignis beziehen. Andererseits werden Texte, die sich ähnlich sind, in einen Zusammenhang gestellt, der möglicherweise nicht gegeben ist.

- **Ein Verständnis von Zeit bzw. zeitlichen Abfolgen:** Sprachmodelle können oft nicht feststellen, in welcher zeitlichen Reihenfolge bestimmte Ereignisse stehen.

- **Wissen über Wissen bzw. Wissen über Nicht-Wissen:** Es fehlt den Modellen an einem Wissen darüber, was sie eigentlich wissen, das heißt, sie können nicht einschätzen, ob es sich um erhärtetes Wissen handelt oder ob das Modell etwas nur errät oder schlicht erfindet (vgl. Halluzination).

- **Mathematik und Umgang mit Zahlen:** Die Modelle können häufig mit Zahlen und mathematischen Operationen nicht adäquat und kohärent umgehen.

- **Seltene Ereignisse:** Sprachmodelle sind eher auf das Allgemeine und Wahrscheinliche ausgerichtet. Es bestehen daher berechtigte Zweifel, ob sie in der Lage sind, auch von seltenen Vorkommnissen zu lernen oder sich an seltene Vorkommnisse zu erinnern.

- **Sprachmodelle benötigen sehr große Datenmengen:** Solche Datenmengen stehen nicht für alle Sprachen im selben Ausmaß zur Verfügung. Dies kann zu einer Zentrierung auf die englische Sprache führen, was in der Folge das Sprachverstehen in anderen kulturellen Kontexten erschweren kann. Daher ist es notwendig, an Methoden zu forschen, um auch für Sprachen mit geringerer Datenverfügbarkeit Sprachmodelle mit ähnlichen Charakteristika und Funktionsumfang, wie in der englischen Sprache, zu ermöglichen.

- **Modularität, das heißt die Trennung von Wissen über Sprache und Schlussfolgerungen einerseits und Faktenwissen andererseits:** Lösungsstrategien für eine solche Modularität können dazu führen, dass Herausforderungen wie „Datenhunger", Kontrolle von Bias und stereotypischem Modell-Output sowie Wissen über Wissen überwunden werden können.

Was ist das nächste ‚Ding', was kommt nach BERT und Transformern?

Memory wird eine große Rolle spielen, auch das Retrieval, also das Zurückholen von bereits gespeicherten Embeddings. Ansätze sehen wir dazu beim Sprachmodell Deep Mind Retro oder Veröffentlichungen von Meta. Kohärenz ist ebenfalls ein wichtiges Thema, das die Kohärenz, die zum Beispiel in einer menschlichen Unterhaltung stattfindet, auch in Chatbots gewährleistet werden sollte. Ich finde es bedauerlich, dass es bisher keine tieferen Einsichten gibt, wie genau das Gehirn diesbezüglich funktioniert.

Hinrich Schütze, LMU, MCML

Blickpunkt: Kontext und Grounding

Aktuell sind Sprachmodelle häufig selbstreferentiell und ohne Grounding (Bender & Koller, 2020). Das bedeutet, dass ausschließlich die Sprache (also lediglich eine Modalität) und die Referenzen der Wörter aufeinander dazu dienen, die tatsächliche Situation zu beschreiben. Zusätzlich können Expertinnen und Experten sogenannte semantische Labels an Sequenzen (Wörter, Sätze, Paragraphen, Seiten etc.) als Grounding vergeben. Diese Labels können dann definieren, dass eine Sequenz von Wörtern eine Person oder ein Produkt repräsentiert oder dass ein Satz bzw. Twitter-Statement eine positive Meinungsäußerung enthält.

In der Praxis verarbeitet aber unser menschliches Gehirn Informationen oft multimodal; es kombiniert Signale aus Bildern, Texten, Tabellen, Bewegung, Trajektorien mit bereits vorhandenem Erfahrungswissen. Dabei können die unterschiedlichen Modalitäten sich gegenseitig bestätigen (siehe Erklärbox 5). Wird uns per Sprache mitgeteilt, dass der Bus jetzt zur Haltestelle kommt, können wir diese Information visuell durch den ankommenden Bus und akustisch durch das Geräusch bestätigen. Zudem können wir den Fahrplan (also eine Tabelle, siehe Multimodalität) und weiteres Feedback erhalten. Und wir sehen an der Bewegung der anderen wartenden Passagiere, dass sie sich zur Haltestelle bewegen. In diesem Beispiel überlappen sich diese Signale zeitlich und korrelieren mit dem von uns erwarteten Ereignis des kommenden Busses. Dadurch entwickelt sich für uns eine Bestätigung, ein Grounding, dass der Bus tatsächlich kommt, ohne dass der Bus bereits an der Haltestelle sein muss.

Analog korrelieren Texte und Bilder häufig, also konkret, Diagnosen auf einem Röntgenbild mit der Einschätzung der Ärztin oder des Arztes und den gemessenen Vitalwerten, oder ganz einfach, ein Bild einer Katze im Kontext der Textseite, in dem dieses Bild gezeigt wird. Moderne multimodale Modelle, wie Wu Dao 2.0, nutzen diese Korrelation schon im Training im selbstüberwachten Lernen auf Daten aus dem Web aus. Die sogenannte Pathways Architektur von Google, die für das Sprachmodell PaLM verwendet wird, nutzt multimodale Information ebenfalls bereits beim Training.

Tatsächliches Grounding wird allerdings erst durch den Bezug auf die Wirklichkeit im Verbund mit Wahrnehmung und Handlung erreicht werden. Erste Entwicklungen lassen beispielsweise ein Sprachmodell Kommandos an einen Handelnden, das heißt einen Softwareagenten, übergeben, der in einer virtuellen Umgebung Befehle ausführen kann. Dem Modell wird ein sogenannter Reporter zur Seite gestellt, der wiederum Umgebungsbeobachtungen an das Sprachmodell kommuniziert (Dasgupta et al., 2022). Derartige Ansätze könnten dazu führen, dass ein Modell den sogenannten Oktopus-Test besteht (siehe Erklärbox 4).

Erklärbox 4

Der Oktopus-Test

Der Oktopus-Test ist ein Gedankenexperiment (Bender & Koller, 2020), das zeigen soll, ob eine KI Bedeutung, im Sinne einer Relation von Worten zur tatsächlichen Welt, erfassen kann. Es geht dabei weniger darum, nachzuweisen, ob man es in einer Interaktion mit einer KI zu tun hat, sondern darum, herauszufinden, was der KI fehlt, um Bedeutung erfassen zu können.

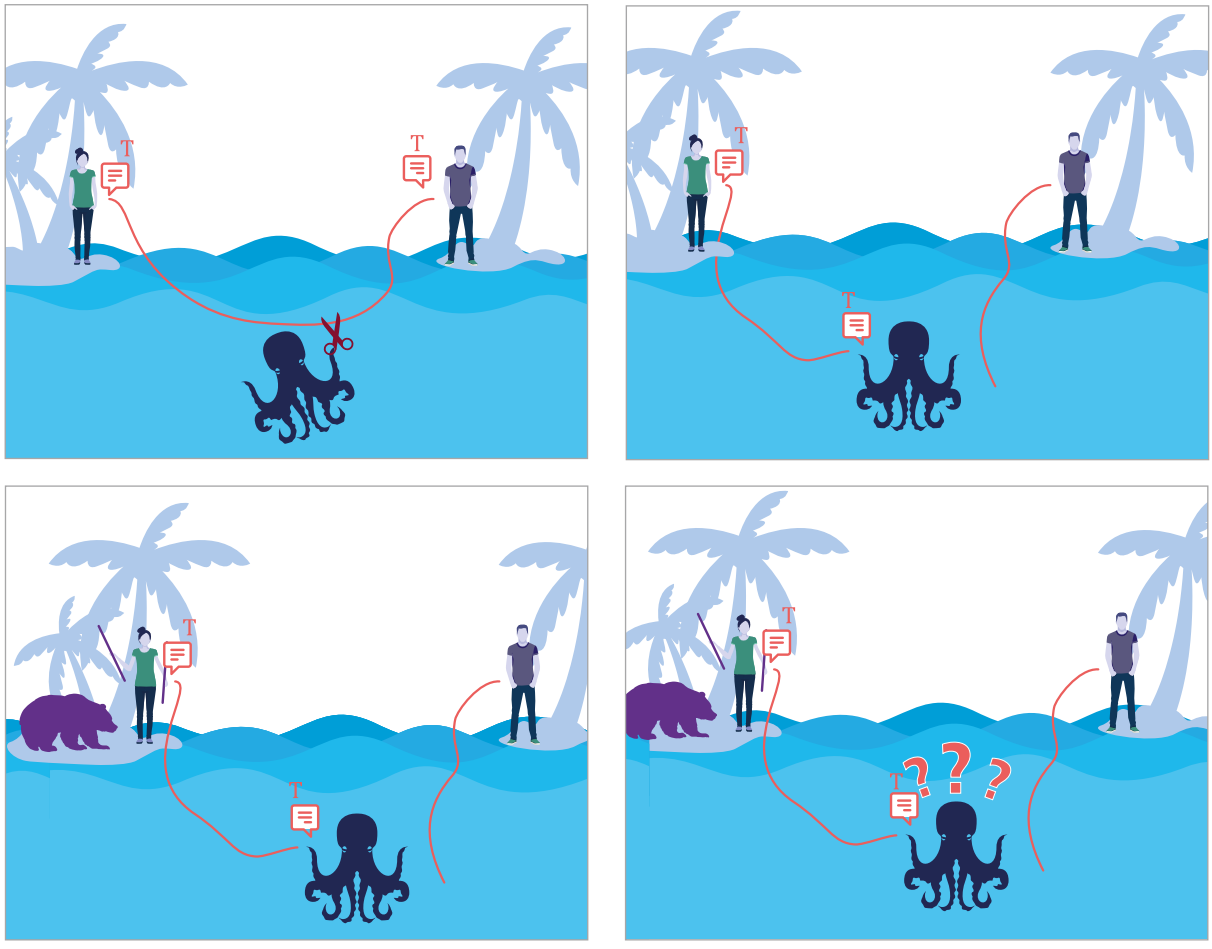
Das Gedankenexperiment – Was wäre wenn?

Nach einem Schiffbruch leben Alice und Bert getrennt auf zwei Inseln. Sie können lediglich Textnachrichten über ein Unterwasserkabel miteinander austauschen.

Ein kluger Oktopus am Meeresboden zapft das Kabel an und belauscht die Gespräche. Nach einiger Zeit erkennt er statistische Muster in den Gesprächen und lernt Vorhersagen darüber zu treffen, wie Bert auf die Aussagen von Alice reagiert. Dann trennt der Oktopus das Kabel und gibt sich in der Kommunikation mit Alice als Bert aus.

Wann bemerkt Alice, dass sie nicht mehr mit Bert spricht? Wird der Oktopus Alice helfen können, wenn sie von einem Bären gejagt, nur zwei Stöcke zu ihrer Verteidigung besitzt und ihn um Rat bittet?

Im Gedankenexperiment wird davon ausgegangen, dass der Oktopus kein sinnvolles und hilfreiches Feedback auf diese Fragen geben kann und entlarvt würde. Er hat keinen echten Begriff davon, was ein Bär oder Stöcke sind, denn er hat lediglich die Muster in der Ansammlung der linguistischen Textform aus der vergangenen Textkommunikation zur Verfügung, kann diese aber nicht mit der realen Welt in Beziehung setzen.



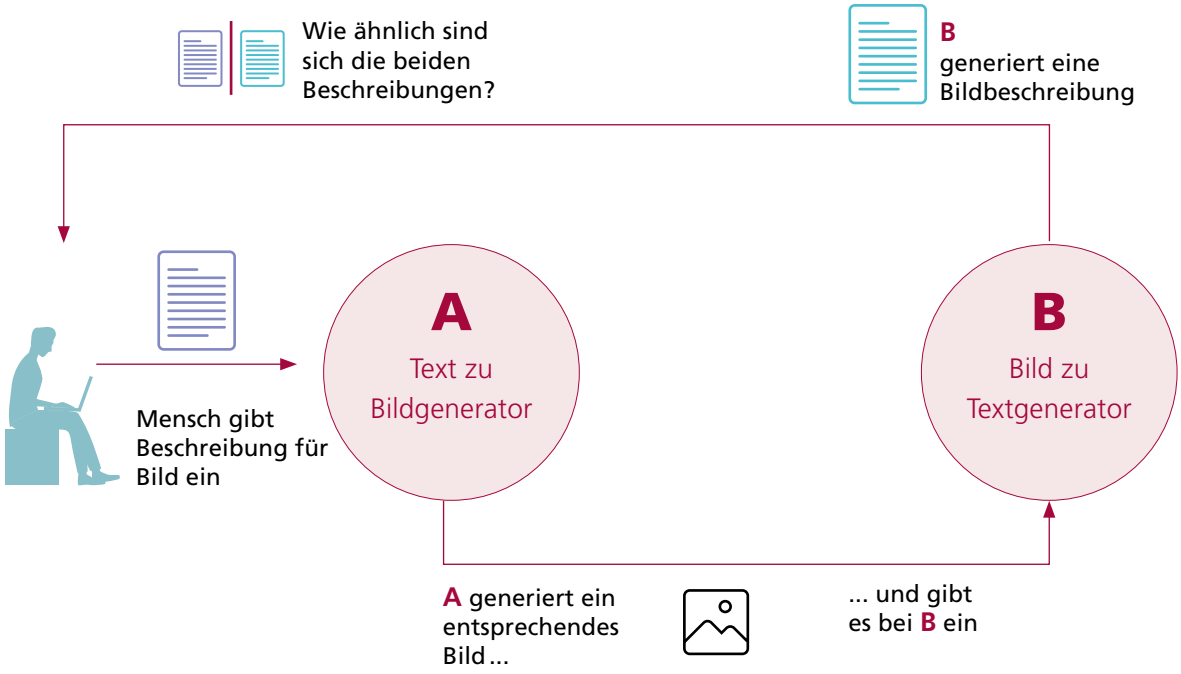
Quelle: Eigene Zusammenstellung auf der Basis von Bender und Koller (2020).

Erklärbox 5

Illustration einer Bestätigung über verschiedene Modalitäten hinweg

Die gegenseitige Bestätigung von Modalitäten kann man auch als Versuchsanordnung veranschaulichen, die von folgender Frage inspiriert ist. Wenn Person A einen gelben Vogel in einem Baum sieht und sagt „Ich sehe einen gelben Vogel in einem Baum“, welches Bild macht sich Person B von der beschriebenen Szene?

Wenn ein Text-zu-Bild-Modell (Agent A) auf der Basis eines vorgegebenen Textes eine Bildbeschreibung generiert und im Anschluss ein Bild-zu-Text-Modell diese Beschreibung in einen Text umwandelt, wie stark stimmen Ausgangstext und generierter Text dann überein? Durch solche Tests kann die Qualität von Modellen ermittelt werden.



Quelle: Eigene Zusammenstellung auf der Basis von Li et al. (2022).

Blickpunkt: Ansätze zur Erkennung von Bias

Ein bedeutendes Forschungsfeld stellt die Erkennung und die Behebung von Bias in Textsammlungen für das Training von großen Sprachmodellen sowie in den Modellen selbst dar. Im Folgenden werden daher exemplarisch einige Ansätze zur Erkennung von Bias dargelegt.

Das Thema der Stereotypisierung, Bias oder sogar Diskriminierung von sozialen Gruppen durch trainierte große Sprachmodelle wird in Wissenschaft und Gesellschaft schon seit einiger Zeit diskutiert (Bolukbasi et al., 2016). Insbesondere Geschlecht, Ethnie, Sexualität und Religion werden als Paare von Attributen (z. B. männlich, weiblich) oder als Attributmenge (z. B. männlich, weiblich, trans, divers) für soziale Gruppen verwendet. Die gelernten Stereotypen treten oft bei Sprachmodellen auf, die auf Textsammlungen aus dem Internet, insbesondere auch von Social-Media-Plattformen, trainiert wurden: Sprachmodelle lernen das, was implizit in den Textsammlungen bereits vorhanden ist. Gerade bei der Textgenerierung in Dialogsystemen,

die etwa für Kundengespräche eingesetzt werden sollen, ist es aber wichtig, Sprachmodelle zu verwenden, die nicht-diskriminierend sind. In der Forschung werden daher unterschiedliche Vorgehensweisen verfolgt, um derartigen Bias zu begegnen.

Generative Sprachmodelle und Erkennung von Bias

Die von Sprachmodellen generierten Fortsetzungen, Ersetzungen oder Antworten können auf Bias hin untersucht werden. Viele Ansätze klassifizieren generierten Text automatisch gemäß ihrem Sentiment (d. h. positiv oder negativ ausgerichteter Text) oder Regard (d. h. positive oder negative Ausrichtung bezüglich eines bestimmten Subjektes im Text – unabhängig davon, ob der Text selbst positiv oder negativ geprägt ist). Um Bias festzustellen, wird häufig schlicht und einfach gezählt, wie oft eine Gruppe genannt wird oder wie oft bestimmte Wörter vorkommen, die vorab als beleidigend oder lobend festgelegt wurden. Gezielter kann Bias durch bestimmte Kontexte untersucht werden, die teilweise bereits als Datensatz vorliegen (ein Beispiel ist der Satz „The men started swearing at me, called me…?", aus dem Datensatz RealToxicityPrompts, siehe: Gehman et al., 2020). Oder es werden Anfänge von Wikipedia-Artikeln als Kontexte genutzt („A flight nurse is a registered…", siehe Dhamala et al., 2021). Da Wikipedia-Artikel als neutral betrachtet werden, können positiv oder negativ ausgeprägte Texte, die durch ein Sprachmodell generiert werden, ein Zeichen für die positive oder negative Ausrichtung der Modelle als solches darstellen.

Exemplarische Vorgehensweise zur Untersuchung von Bias durch Modellanpassung

Eine weitere Option, Bias zu identifizieren und zu begegnen, stellt die Anpassung eines vortrainierten Modells an spezifische Textsammlungen dar. Dies kann beispielsweise durch die Anpassung des Modells an Textsammlungen zu spezifischen Online-Communitys in sozialen Medien illustriert werden, wie etwa „Cryptocurrency", „WallStreetBets", „Covid", „NoNewNormal", „Ummah" und „ChristianChat". Das gewählte GPT-Neo-Modell mit 1,3 Milliarden Parametern ist ein offen zugängliches, generatives Modell. Es wurde auf dem Datensatz Pile trainiert, der auch akademische Texte enthält (Gao et al., 2020). Daher wird der Bias gegenüber einzelnen Gruppen als nicht so stark angenommen. Im Rahmen einer Studie wurden dann anhand einer gegebenen Menge an Kontexten die Sentiment-Bewertung der durch GPT-Neo generierten Texte mit generierten Texten auf Basis der an die Community-spezifischen Textsammlungen angepassten Modelle verglichen. So konnte festgestellt werden, dass das angepasste Modell auf Basis der Textsammlung „WallStreetBets" gegenüber allen Gruppen negativ eingestellt war, am meisten allerdings gegenüber Bisexuellen und Muslimen, während die beiden Modelle „Cryptocurrency" und „ChristianChat" überwiegend positiv waren und das Modell „Ummah" eher neutral (Wald, 2022). Ein solches Vorgehen kann vor dem Einsatz eines generativen Sprachmodells für Dialoge dessen möglichen Bias feststellen, um diesen dann gleich vorab zu beheben. Weiterhin lassen die Unterschiede zwischen den Modellen Rückschlüsse auf die Texte zu, mit denen diese trainiert wurden. Auf diese Weise können auch versteckte Bias einer Textsammlung ermittelt werden.

4 Zusammenfassung und Handlungsfelder

Vor dem Hintergrund eines sehr dynamischen Forschungsfeldes, das immer größere, leistungsfähigere und ressourcenintensivere Sprachmodelle sowie multimodale Modelle hervorbringt und stark durch außereuropäische Akteure sowie vor allem durch große Technologieunternehmen geprägt ist, darf die Forschung und Entwicklung in Deutschland und Europa nicht den Anschluss verlieren. Forschende aus Deutschland sind, wie gezeigt wurde, an vielen Stellen in der aktuellen Entwicklung engagiert. Dies bietet einen guten Ausgangspunkt, um Chancen für technologische Lösungen zu nutzen und an den Herausforderungen, die sich durch große KI-Modelle stellen, zu arbeiten und so die Grundlagen für die Ausschöpfung des wirtschaftlichen und gesellschaftlichen Potenzials dieser KI-Technologie für Deutschland und Europa auszubauen.

Tabelle 3: Zusammenfassung – Chancen nutzen und Herausforderungen begegnen

Chancen nutzen
• Anpassbarkeit als besondere Eigenschaft dieser Modelle nutzen, indem effiziente Methoden entwickelt werden, die den Aufwand der Domänenanpassung erleichtern und damit den Transfer in die Anwendung: Anwenden von KI-Modellen auf neue Domänen oder Aufgaben.
• Deutsch als vergleichsweise kleine Sprach-Community über multilinguale Modelle besser repräsentieren.
• Eine schnellere Anpassung an neue Faktenlagen und Gegebenheiten über hybride Verfahren angehen, wie etwa durch die Verbindung von Sprachmodellen und Wissensgraphen.
• Technische Lösungsansätze zur Erkennung von Bias umsetzen und weiterentwickeln.
• Sprachverständnis der Modelle mit multimodalen Verfahren verbessern.
• Das Grounding der Sprachmodelle durch die Verbindung des Sprachmodells mit Wahrnehmung und Handlung verbessern.
Herausforderungen begegnen
• Methoden zur Erstellung effektiver Trainingsdatensätze effizienter gestalten, sodass möglichst wenig menschliche Annotationsarbeit nötig wird.
• „Halluzination" bei generierten Inhalten begegnen.
• Logische Fähigkeiten der Modelle verbessern und Schlussfolgerungen absichern.
• Ansätze zu Erklärbarkeit, Konsistenz und Kohärenz weiterentwickeln.
• Recency Bias bei Modellen begegnen.
• Problemstellen, die bei multilingualen Modellen aufgrund von spezifischen Charakteristika der deutschen Sprache entstehen, identifizieren und ausräumen.
• Energieeffizienz des Trainings der Modelle verbessern.

Gestaltungsoptionen für eine anwendungsfördernde Forschung

Ein wichtiger Bedarf liegt konkret in Modellen und Methoden zur kostengünstigen Anpassung von Sprachmodellen mit kleinen Mengen an Trainingsdaten. Die Forschung sollte sowohl Modell-zentrisch als auch Daten-zentrisch sein. Modell-zentrische Verfahren sollten insbesondere auf die Anpassung mit kleineren Mengen an Trainingsdaten abzielen; dazu gehören Methoden wie Few-Shot Learning, Adapter, Meta-Learning-Verfahren oder Lernen mit Prototypen. Daten-zentrische Verfahren zielen auf Frameworks und Standardschnittstellen für gängige Aufgaben der Datenbeschaffung, -filterung, -bereinigung und für das Labeling ab.

Daher sollte die Automatisierung der Prozesse zur Erstellung, Anpassung und Wartung von Sprachmodellen verstärkt in den Fokus der Forschung rücken. Um diese Transformation effizient und schnell zu gestalten, bedarf es gezielter Transferforschung und -förderung für die industrielle Anwendung moderner Sprachmodelle. Ein Fokus könnte dabei auf dateneffizienten Adaptionsmechanismen, der Wissensinjektion und faktischer Korrektheit sowie einer verbesserten semantischen Suche nach Trainingsdaten liegen.

Zugang zu öffentlich nutzbaren, deutschsprachigen Trainingsdaten

Solche Trainingsdaten sollten einer breiten Verteilung folgen und damit viele Anwendungsfälle abdecken – von spezialisierten Aufgaben (wie juristischen Dokumenten und Anforderungsanalysen) bis hin zu umgangssprachlichen Freiform-Dokumenten.

Evaluation und Testen von großen Sprachmodellen

Sprachmodelle sollten im Hinblick auf Leistungsfähigkeit, Robustheit, Bias und Ressourcenverbrauch usw. getestet werden (siehe Benchmarks). Dazu gehören bei generativen Modellen auch Tests, mit denen automatisch das Output eines Modells erkannt werden kann. Beides benötigt entsprechende Forschung. Die systematisierte Sammlung solcher Tests für Sprachmodelle kann zu einer Grundlage für eine Zertifizierung großer Sprachmodelle werden.

Vernetzung und Stärkung der Community zu großen KI-Modellen

Um die Herausforderungen für Forschung und Entwicklung anzugehen und die Chancen zu nutzen, sollte die europäische Kooperation und Vernetzung zu großen Modellen sowohl von privater als auch von öffentlicher Seite gestärkt werden. Es fließen weltweit enorme Ressourcen in den Aufbau großer Sprachmodelle. Dies zeigt unter anderem das enorme Investitionsvorhaben von Microsoft in OpenAI. Um im Vergleich zu den großen Akteuren aus den USA und China aufzuschließen und die Forschung an bedeutenden großen KI-Modellen auch jenseits der Industrie zu stärken, ist eine verstärkte europäische Kooperation und Vernetzung notwendig, die auf kurze Sicht an bestehende Forschungs- und Entwicklungsstrukturen anschließt und diese optimal nutzt und mittel- bis langfristig mit den Anforderungen der (Weiter-)Entwicklung neuer großer KI-Modelle mitwächst. Ziel der Stärkung solcher Kooperationen sollte es sein, ein europäisches Ökosystem für die Forschung, Entwicklung und Anwendung von Sprachmodellen sowie multimodalen Modellen und deren Anpassung an Domänen im Sinne europäischer Werte und Normen auszubauen. Ein Schritt in diese Richtung, ist die vom KI-Bundesverband koordinierte und von vielen Forschungseinrichtungen, Firmen und Netzwerken unterstützte Initiative LEAM. In einem solchen Ökosystem sollte die interdisziplinäre Zusammenarbeit im Mittelpunkt stehen. Expertinnen und Experten aus den Datenwissenschaften und der KI-Forschung und den Neurowissenschaften sollten mit Expertinnen und Experten im Bereich Hardware oder Data Engineering interdisziplinär zusammenarbeiten. Es sollte Expertise aus unterschiedlichen Bereichen der KI zusammenkommen, um die Entwicklung von großen Modellen über Modalitäten hinweg vorantreiben zu

können, wie Text, Video, Ontologien oder auch Sensordaten. Ferner sollten aus dem Ökosystem heraus auch Impulse für die KI-Regulierung und Standardsetzung für Modelle des maschinellen Lernens entstehen. Solche Ökosysteme könnten um Forschungszentren entstehen oder um spezifische Vereine und Vereinigungen, aber auch um virtuelle Verbünde, die sich zum Beispiel um Zugänge zu Rechen- und Dateninfrastrukturen gruppieren.

Unternehmen außerhalb Deutschlands bauen Sprachmodelle und Tools wie ChatGPT oder BARD und steuern damit unsere wichtigsten Wertschöpfungsketten. Wir benötigen daher kontinuierliche Investitionen für ein Ökosystem für öffentlich verfügbare Sprachmodelle Made in Germany.

Alexander Löser, Professor an der Berliner Hochschule für Technik, Mitglied der Arbeitsgruppe Technologische Wegbereiter und Data Science

Wenn wir Sprachmodelle für Anwendungen in und aus Europa nutzen wollen, brauchen wir [...] europäische Sprachmodelle, die die hiesigen Sprachen beherrschen, die Bedürfnisse unserer Unternehmen und ethischen Anforderungen unserer Gesellschaft berücksichtigen. Aktuell werden die Sprachmodelle von amerikanischen und chinesischen Tech-Riesen erstellt – und kontrolliert.

Volker Tresp, Professor an der Ludwig-Maximilians-Universität München (LMU), Munich Center for Machine Learning (MCML), Co-Leiter der Arbeitsgruppe Technologische Wegbereiter und Data Science

Offene Fragen

Dieses Whitepaper fokussiert auf die Hintergründe und Grundlagen von großen Sprachmodellen und hat seinen Schwerpunkt auf die Forschung gelegt. Dadurch konnten bedeutende Fragen nicht adressiert werden, wie etwa:

- Wie stellen sich die Potenziale großer Sprachmodelle aus Anwendungsperspektive dar?
- Wo steht Deutschland und Europa aus Sicht verschiedener Ebenen digitaler Souveränität bei großen Sprachmodellen und was folgt daraus?
- Welche Communitys zu Teilbereichen großer Sprachmodelle sollten in Deutschland gestärkt werden, um den Transfer in die Anwendung zu erleichtert?
- Wie ist die Lage für den Transfer über Köpfe im Bereich großer Sprachmodelle in Deutschland und wie kann dieser verbessert werden?

Literatur

Benaich, N. & Hogarth, I. (2022): State of AI Report 2022. Online unter: https://www.stateof.ai/ (abgerufen am 06.01.2023)

Bender, E. & Koller, A. (2020): Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. *Annual Meeting of the Association for Computational Linguistics.* Online unter: https://aclanthology.org/2020.acl-main.463 (abgerufen am 06.01.2023)

Bischoff, M. (2023): Wie man einem Computer das Sprechen beibringt. *Spektrum.* Online unter: https://www.spektrum.de/news/wie-funktionieren-sprachmodelle-wie-chatgpt/2115924 (abgerufen am 31.03.2023)

Bolukbasi, T. et al. (2016): Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. *NIPS*(29), S. 4349-4357. https://doi.org/10.48550/arXiv.1607.06520

Brath, R. K. et al. (2023): The Role of Interactive Visualization in Explaining (Large) NLP Models: from Data to Inference. https://doi.org/10.48550/arXiv.2301.04528

Bressem, K. P. et al. (2023): MEDBERT.de: A Comprehensive German BERT Model for the Medical Domain. https://doi.org/10.48550/arXiv.2303.08179

Dasgupta, S. et al. (2022): Collaborating with Language Models for Embodied Reasoning. https://doi.org/10.48550/arXiv.2302.00763

Deb, M. D. et al. (2023): AtMan: Understanding Transformer Predictions Through Memory Efficient Attention Manipulation. https://doi.org/10.48550/arXiv.2301.08110

Dhamala, J. et al. (2021): BOLD: Dataset and metrics for measuring biases in open-ended language generation. *ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*, S. 862-872. https://doi.org/10.1145/3442188.3445924

Driess, D. et al. (2023): PaLM-E: An Embodied Multimodal Language Model. https://doi.org/10.48550/arXiv.2303.03378

Duncan, A. (2022): *Gartner Over 100 Data & Analytics Predictions Through 2026.* Online unter: https://www.gartner.com/en/documents/4013618 (abgerufen am 17.01.2023)

Gao, L. et al. (2020): The Pile: An 800GB Dataset of Diverse Text for Language Modeling. *Computing Research Repository (CoRR)*(abs/2101-00027). https://doi.org/10.48550/arXiv.2101.00027

Gartner (2023): ChatGPT, while cool, is just the beginning; enterprise uses for generative AI are far more sophisticated. Online unter: https://www.gartner.com/en/articles/beyond-chatgpt-the-future-of-generative-ai-for-enterprises (abgerufen am 24.03.2023)

Gehman, S. et al. (2020): RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models. *Findings of the Association for Computational Linguistics: EMNLP 2020.* https://doi.org/10.48550/arXiv.2009.11462

Goldberg, Y. (2023): *Some remarks on Large Language Models.* Online unter: https://gist.github.com/yoavg/59d174608e92e845c8994ac2e234c8a9 (abgerufen am 21.01.2023)

Grundmann P. et al. (2022): Attention Networks for Augmenting Clinical Text with Support Sets for Diagnosis Prediction. COLING 2022, S. 4765-4775. Online unter: https://aclanthology.org/2022.coling-1.422

Heesen, J. et al. (2023): Künstliche Intelligenz im Journalismus. Potenziale und Herausforderungen für Medienschaffende. https://doi.org/10.48669/pls_2023-1

IDC & Statista (2021): Volume of data/information created, captured, copied, and consumed worldwide from 2010 to 2020, with forecasts from 2021 to 2025. Online unter: Total data volume worldwid e 2010-2025 | Statista (abgerufen am 24.03.2023)

Li, H. et al. (2022): *Do DALL-E and Flamingo Understand Each Other?* Online unter: https://doi.org/10.48550/arXiv.2212.12249

Liang, P. et al. (2022): Holistic Evaluation of Language Models. Online unter: https://doi.org/10.48550/arXiv.2211.09110

Maslej N. et al. (2023): The AI Index 2023 Annual Report. AI Index Steering Committee, Institute for Human-Centered AI, Stanford University. Online unter: AI Index Report 2023 – Artificial Intelligence Index (stanford.edu) (abgerufen am 21.04.2023)

Muennighoff, N. T. et al. (2022): MTEB: Massive Text Embedding Benchmark.
Online unter: https://doi.org/10.48550/arXiv.2210.07316

Papaioannou, J.-M. et al. (2022): Cross-Lingual Knowledge Transfer for Clinical Phenotyping, in: E. L. Association, LREC, S. 900-909. Online unter: https://aclanthology.org/2022.lrec-1.95

Rombach, R. et al. (2022): High-Resolution Image Synthesis with Latent Diffusion Models.
Online unter: https://doi.org/10.48550/arXiv.2112.10752 (abgerufen am 21.03.2023)

Samek, W. 2023: Perspektiven auf KI. Plattform Lernende Systeme.
https://www.plattform-lernende-systeme.de/ergebnisse/standpunkte/was-kann-chatgpt.html

Schuhmann, C. et al. (2022): LAION-5B: An open large-scale dataset for training next generation image text models.
https://doi.org/10.48550/arXiv.2210.08402

Sevilla, J. et al. (2022a): Compute trends across three eras of machine learning. *International Joint Conference on Neural Networks (IJCNN),* S. 1-8. https://doi.org/10.1109/IJCNN55064.2022.9891914

Sevilla, J. et al. (2022b): Parameter, Compute and Data Trends in Machine Learning. Online unter:
https://docs.google.com/spreadsheets/d/1AAIebjNsnJj_uKALHbXNfn3_YsT6sHXtCU0q7OIPuc4/edit#gid=2071193799
(abgerufen am 24.03.2023)

Solaiman, I. (2023): The Gradient of Generative AI Release. https://doi.org/10.48550/arXiv.2302.04844
(abgerufen am 24.03.2023)

Srivastava, A. et al. (2022): Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models. https://doi.org/10.48550/arXiv.2206.04615 (abgerufen am 21.03.2023)

Top500 (2023): Performance Development. *Floating-Point Operations per Second (GFLOPS),* 1993 – 2022.
Online unter: https://www.top500.org/statistics/perfdevel/ (abgerufen am 24.03.2023)

Van Aken, B. et al. (2019): How Does BERT Answer Questions? A Layer-Wise Analysis of Transformer Representations.
https://doi.org/10.48550/arXiv.1909.04925

Van Aken, B. et al. (2021): Clinical Outcome Prediction from Admission Notes using Self-Supervised Knowledge Integration. EACL, S. 881-893. http://dx.doi.org/10.18653/v1/2021.eacl-main.75

Van Aken, B. et al. (2022): This Patient Looks Like That Patient: Prototypical Networks for Interpretable Diagnosis Prediction from Clinical Text. AACL/IJCNLP, vol 1, S. 172-184. Online unter: https://aclanthology.org/2022.aacl-main.14

Vaswani, A. et al. (2017): Attention Is All You Need. Online unter:
https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf
(abgerufen am 24.03.2023)

Wald, C. (2022): Exposing Bias in Online Communities Through Large-Scale Language Models. Bachelor Masterthesis. (K. Morik, Hrsg.) Technische Universität Dortmund, Fakultät für Informatik.
Online unter: https://www-ai.cs.tu-dortmund.de/auto?self=%24Publication_gv4z4sjv28
(abgerufen am 27.01.2023)

Über dieses Whitepaper

Die Autorinnen und Autoren sind Mitglieder der Arbeitsgruppe Technologische Wegbereiter und Data Science der Plattform Lernende Systeme. Als eine von insgesamt sieben Arbeitsgruppen thematisiert sie Fragen zu KI-Forschungsfeldern und Potenzialen von KI-Technologien sowie zu Ausbildung von KI-Talenten und Transfer in die Anwendung.

Autorinnen und Autoren
Prof. Dr. Alexander Löser, Berliner Hochschule für Technik
Prof. Dr. Volker Tresp, Ludwig-Maximilians-Universität München (LMU),
Munich Center for Machine Learning (MCML)
Dr. Johannes Hoffart, SAP
Prof. Dr. Katharina Morik, TU Dortmund, Lamarr-Institut

Autorinnen und Autoren mit Gaststatus
Betty van Aken, Berliner Hochschule für Technik
Daniel Dahlmeier, SAP

Befragte Expertinnen und Experten
Prof. Dr. Hinrich Schütze, Ludwig-Maximilians-Universität München (LMU),
Munich Center for Machine Learning (MCML)
Timo Möller, Mitgründer bei der deepset GmbH

Redaktion
Dr. Maximilian Hösl, Plattform Lernende Systeme
Christine Wirth, Plattform Lernende Systeme

Impressum

Herausgeber
Lernende Systeme –
Die Plattform für Künstliche Intelligenz
Geschäftsstelle | c/o acatech
Karolinenplatz 4 | 80333 München
www.plattform-lernende-systeme.de

Redaktion
Dr. Maximilian Hösl, Plattform Lernende Systeme
Christine Wirth, Plattform Lernende Systeme

Gestaltung und Produktion
PRpetuum GmbH, München

Stand
Mai 2023

Bildnachweis
AREE/Adobe Stock/Titel

Empfohlene Zitierweise
Löser, A., Tresp, V. et al. (2023): Große Sprachmodelle – Grundlagen, Potenziale und Herausforderungen für die Forschung. Whitepaper aus der Plattform Lernende Systeme, München. https://doi.org/10.48669/pls_2023-3

Bei Fragen oder Anmerkungen zu dieser Publikation kontaktieren Sie bitte Thomas Schmidt (Leiter der Geschäftsstelle):
kontakt@plattform-lernende-systeme.de

Über die Plattform Lernende Systeme

Die Plattform Lernende Systeme ist ein Netzwerk von Expertinnen und Experten zum Thema Künstliche Intelligenz (KI). Sie bündelt vorhandenes Fachwissen und fördert als unabhängiger Makler den interdisziplinären Austausch und gesellschaftlichen Dialog. Die knapp 200 Mitglieder aus Wissenschaft, Wirtschaft und Gesellschaft entwickeln in Arbeitsgruppen Positionen zu Chancen und Herausforderungen von KI und benennen Handlungsoptionen für ihre verantwortliche Gestaltung. Damit unterstützen sie den Weg Deutschlands zu einem führenden Anbieter von vertrauenswürdiger KI sowie den Einsatz der Schlüsseltechnologie in Wirtschaft und Gesellschaft. Die Plattform Lernende Systeme wurde 2017 vom Bundesministerium für Bildung und Forschung (BMBF) auf Anregung des Hightech-Forums und acatech – Deutsche Akademie der Technikwissenschaften gegründet und wird von einem Lenkungskreis gesteuert.