Lernende Systeme

DIE PLATTFORM FÜR KÜNSTLICHE INTELLIGENZ

Große Sprachmodelle entwickeln und anwenden

Ansätze für ein souveränes Vorgehen

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

Das wirtschaftliche Potenzial großer Sprachmodelle ist enorm und vielversprechend. Ihre Anpassungsfähigkeit an unterschiedlichste branchen- und unternehmensspezifische Anforderungen verleiht ihnen eine hohe Wiederverwendbarkeit. Dies ermöglicht Anwendungsfälle, die ansonsten gar nicht möglich wären, aufgrund zu hoher Kosten oder fehlender Trainingsdaten. Für Unternehmen bieten sich damit viele, auch noch ungeahnte Potenziale. Gegenwärtig entstehen viele der fortgeschrittensten generativen Modelle in den USA und China. Diese sind jedoch oft nicht offen zugänglich und eine Ausrichtung an europäischen Werten und Vorgaben ist zudem nicht gesichert. Deutsche Unternehmen sind damit abhängig von der Qualität der Trainingsdaten und von den Modellen dieser Anbieter, wenn sie diese in Anspruch nehmen. Angesichts des zunehmenden Einflusses sowie der rasanten technologischen Entwicklung dieser KI-Modelle besteht ein Bedarf, Abhängigkeiten Europas in Bezug auf Technologie und Daten zu begegnen und Alternativen zu schaffen, um die Innovationskraft und Wettbewerbsfähigkeit in Deutschland und Europa voranzutreiben.

Expertinnen und Experten der Arbeitsgruppe Technologische Wegbereiter und Data Science der Plattform Lernende Systeme fokussieren mit dem Whitepaper die Anwenderperspektive großer Sprachmodelle und knüpfen damit an das Whitepaper „Große Sprachmodelle: Grundlagen, Potenziale und Herausforderungen für die Forschung“ (erschienen Mai 2023) an. Welche Potenziale, Herausforderungen sowie Lösungsansätze diese generativen Modelle insbesondere in der Anwendung liefern, wird im vorliegenden Papier an zwei konkreten Anwendungsfeldern – Geschäftsanwendung und Gesundheitswesen – gespiegelt. Um das wirtschaftliche Potenzial für deutsche wie europäische Unternehmen in einem global wachsenden Ökosystem hinsichtlich Wettbewerbsfähigkeit, Selbstbestimmtheit und Innovationskraft souverän auszuschöpfen, ist die Frage nach der Digitalen Souveränität entscheidend. Daher werden die wichtigsten technologischen und strukturellen Komponenten sowie die personelle Ressource der Talente als zentrale Kernpunkte hinsichtlich Digitaler Souveränität näher beleuchtet. Diese Grundvoraussetzungen, die für die Entwicklung und die Anwendung großer Sprachmodelle in Deutschland vorliegen, untermauern in den Gestaltungsoptionen den Stellenwert des Aufbaus eines umfangreichen, offenen wie kommerziell nutzbaren und im Sinne europäischer Werte und Regeln kuratierten Trainingsdatensatzes in deutscher Sprache. Ein solcher Datensatz kann wiederum Anstoß für die Entwicklung vieler unterschiedlicher Sprachmodelle in Forschung, Wirtschaft und Zivilgesellschaft sein und somit den Transfer in die Anwendung im Sinne der Digitalen Souveränität erleichtern.

SCHLAGWORTE

GENERATIVE KI – GROSSE SPRACHMODELLE – DIGITALE SOUVERÄNITÄT – OPEN SOURCE – KI-ÖKOSYSTEM

1 Einleitung

Um große Sprachmodelle und durch diese Modelle umgesetzte generative KI findet ein ausgeprägter Wettbewerb statt, der viele Innovationen und Anwendungsvarianten hervorbringt. So unterstützt generative KI bei Übersetzungen, bei der Erstellung von unterschiedlichen Textformaten und Programmcode oder illustrativen Bildern, aber auch beim Entdecken von Proteinstrukturen. Nach Schätzungen der Unternehmensberatung McKinsey (Chui et al., 2023) könnten die Auswirkungen der generativen KI auf die Produktivität der Weltwirtschaft einen Mehrwert in Billionenhöhe schaffen. Allerdings steht die Ära der generativen KI erst am Anfang und die Realisierung ihres Potenzials wird noch einige Zeit in Anspruch nehmen. Der globale Markt für generative KI allgemein lag 2022 bei 23,17 Milliarden Dollar und wird für 2030 auf einen Wert von 207 Milliarden Dollar geschätzt (Statista Market Insights, 2023). Das entspricht rund 30 Prozent des erwarteten globalen Marktwerts für Medizintechnik in 2027 (701,8 Milliarden) (Statista, 2022) oder rund 5 Prozent des deutschen Bruttoinlandsprodukts aus dem Jahr 2022.

Schon heute beginnen immer mehr Unternehmen, generative Modelle in ihre Produkte und Dienstleistungen zu integrieren, um das Potenzial dieser Technologie zu nutzen: So setzt Microsoft ChatGPT bereits für den Suchdienst Bing ein und Google nutzt das Modell PaLM2 für verschiedene Anwendungen (z. B. Docs, Slides und Sheets). Zudem kooperieren Konzerne wie ABB und Mercedes mit Microsoft, um generative KI in ihre Produkte zu integrieren, und Bosch arbeitet mit dem deutschen Start-up Aleph Alpha an einem eigenen KI-Modell, um das Entwicklungsteam zu entlasten.

Generative KI-Modelle wie Stable Diffusion, das aus deutscher Forschung hervorging, oder OpenAI's ChatGPT haben den KI-Wettbewerb seit 2020 im Bereich der Text-zu-Bild-Generatoren bzw. 2022 im Bereich der Chat-Bots vorangetrieben. Dabei überbieten sich vor allem die großen Technologiekonzerne aus den USA und China mit immer neuen Weiterentwicklungen sowie auch größeren Modellen (siehe Abbildung 5). In diesem Wettbewerb bringt sich zunehmend eine boomende Open-Source-Community-Szene ein (vgl. Modelle wie Bloom und das weitgehend offene LlaMA 2). Auch in Europa sind Entwicklungen zu beobachten, die diesem Trend und damit möglichen Abhängigkeiten von außereuropäischen großen Technologiekonzernen entgegenwirken: So befinden sich etwa mit Start-ups wie nyonic und Mistral AI in Deutschland und Frankreich Unternehmen in der Gründungsphase, die große Sprachmodelle aufbauen wollen, auch das private KI-Labor Silo.AI in Finnland beginnt ein solches Projekt und das 2019 gegründete deutsche Start-up Aleph Alpha hat mit SAP und Intel zwei große Unternehmen als Investoren gewinnen können. Darüber hinaus gibt es das offene Modell „Bloom", das von einer internationalen Community in Frankreich trainiert wurde; in Deutschland wird mit OpenGPT-X derzeit ein Sprachmodell auf Basis der GAIA-X-Infrastruktur aufgebaut. Diese Beispiele zeigen, dass auch von Europa aus Projekte und Unternehmungen vorangetrieben werden können, die an den globalen KI-Wettbewerb anschließen und dabei europäische Werte berücksichtigen.

Das im Mai 2023 veröffentlichte Whitepaper „Große Sprachmodelle: Grundlagen, Potenziale und Herausforderungen für die Forschung" der Plattform Lernende Systeme zeigt auf, dass Deutschland sich in der Forschung zu großen Sprachmodellen durchaus in einer guten Ausgangsposition befindet. Zudem stellt es vor allem das Potenzial von Eigenschaften großer Sprachmodelle wie deren Anpassbarkeit an Anwendungsdomänen und Aufgabenstellungen sowie deren Wiederverwendbarkeit und vielfältigen Einsatzmöglichkeiten heraus. Das vorliegende Papier greift daraus unter anderem offene Fragen auf. Zunächst wird aber die Anwendungsperspektive eingenommen, um Potenziale, Herausforderungen und auch konkrete Lösungsansätze an zwei Anwendungsfeldern darzustellen. In einem zweiten Schritt werden verschiedene Ebenen Digitaler Souveränität (DS) betrachtet, nämlich die Bedingungen für die souveräne Realisierung des ökonomi-

schen und gesellschaftlichen Potenzials großer Sprachmodelle im Sinne europäischer Werte im Rahmen eines fairen Wettbewerbs. 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. Die gesellschaftlichen Implikationen generativer KI werden in künftigen Publikationen und Veranstaltungen der Plattform Lernende Systeme thematisiert werden.

2 Anwendungsperspektiven – Potenziale und Herausforderungen

Große Sprachmodelle bergen ein enormes Potenzial für deutsche und europäische Unternehmen und Organisationen. Sie haben in vielen Fällen ältere Technologien in der Verarbeitung natürlicher Sprache (Natural Language Processing, kurz: NLP) abgelöst, sodass klassische NLP-Aufgaben wie die Extraktion von Wissen und Informationen oder von Beziehungen aus Texten optimiert werden können. Aber auch ganz neue Anwendungen, wie zum Beispiel die Suche nach Beschreibungen von Produkten oder Komponenten durch die Eingabe von Bildern, werden durch multi-modale Modelle ermöglicht.

Ein bemerkenswertes Potenzial liegt jedoch in der Wiederverwendbarkeit großer Sprachmodelle. Denn solche Modelle bieten die Möglichkeit der Anpassung an branchen- und unternehmensspezifische Anforderungen und Daten, sodass das kostenintensive Trainieren eines eigenen neuen großen, vortrainierten KI-Modells entfallen kann. Zudem können solche Anpassungen an spezifische Domänen und Aufgabenstellungen auch mit einem geringeren Umfang an unternehmenseigenen Daten durchgeführt werden, als dies ohne Rückgriff auf solche vortrainierten Modelle der Fall ist (siehe zu verschiedenen Möglichkeiten der Modellanpassung: Löser & Tresp, 2023, S. 14).

Dies macht Anwendungsfälle auch dort möglich, wo sich Unternehmen nur kleine oder gar keine eigenen Teams für maschinelles Lernen und Data Science leisten können und wo bisher möglicherweise nicht genügend Daten für Anwendungen des maschinellen Lernens zur Verfügung standen (siehe hierzu auch Abschnitt 3.2). Solche Daten können Verträge, E-Mail-Korrespondenz, Sammlungen wissenschaftlicher Arbeiten, Kundenrezensionen sein oder spezielle Textdatenbanken (z. B. Rechtstexte, Erfahrungsberichte, Produktbeschreibungen etc.). Viele Anwendungsfälle finden sich in der Verbesserung geschäftsunterstützender Bereiche (Supporting Functions) und im Service. Effizienzsteigerungen sind zum Beispiel durch einen verbesserten Informationszugang für Mitarbeitende oder durch die Automatisierung und Optimierung von repetitiven und potenziell fehleranfälligen Prozessen möglich.

In vielen Bereichen wird die Hebung des Potenzials nur dann möglich sein, wenn eine vertrauliche Datenverarbeitung gewährleistet ist. Zum einen geht es hierbei um die Wahrung des Datenschutzes von Nutzenden oder Patientinnen und Patienten, zum anderen aber auch darum, sensible Geschäftsdaten (z. B. Verträge, rechtliche Dokumente) zu schützen. Nur wenige Unternehmen in solchen Bereichen werden bereit sein, diese Daten zur Weiterverarbeitung mit großen Sprachmodellen an externe Anbietende herauszugeben, um beispielsweise Sprachmodelle für ihre Zwecke anzupassen. Daher sind Lösungswege nötig, um Herausforderungen dieser Art zu begegnen (siehe hierzu Abschnitt 3.5). Eine weitere Herausforderung für Unternehmen kann die mangelnde Rechts- und Planungssicherheit darstellen, wenn Unsicherheiten darüber bestehen, ob die Datengrundlage großer Sprachmodelle nach geltenden urheber-, lizenz- und datenschutzrechtlichen Bestimmungen erhoben und zusammengestellt wurde (siehe hierzu Abschnitt 3.2).

Beispiel für Anwendungsbereiche großer Sprachmodelle (nicht abschließend)

- IT-Branche: Unterstützung bei der Entwicklung und im Einsatz von Software (z. B. Generierung von Programmcode)
- Sales und persönliche Ansprachen
- Marketing/PR
- Medien- und Verlagsbranche (z. B. Unterstützung der Contentgenerierung)
- Dienstleistungssektor
- Wissensmanagement
- Gesundheitswesen: Chatbots können z. B. im Bereich Symptomerhebung verbessert werden

Nach Einschätzung der Unternehmensberatung McKinsey entfallen etwa 75 Prozent des Wertes, den generative KI-Anwendungsfälle liefern könnten, auf vier Bereiche: Kundengeschäft, Marketing und Vertrieb, Softwareentwicklung sowie Forschung und Entwicklung (Chui et al., 2023).

Beispiele für Aufgabenstellungen großer Sprachmodelle in der Anwendung (nicht abschließend)

- Überprüfung der Korrektheit von Informationen (z. B. in Kombination mit aktuellen und verifizierten Datenbanken). Generative Modelle allein können jedoch auch falsche Aussagen erstellen.
- Domänenspezifische Rechercheunterstützung (etwa Rechts- und Gesundheitswesen oder Forschung)
- Textanalyse (z. B. Sentiment- oder Stilanalyse)
- Extraktion von Wissen, Information, Beziehungen und Zusammenhängen aus Dokumenten bzw. Dokumentensammlungen sowie die Erkennung von spezifischen Inhalten in Dokumenten und aller Ausdrücke in Texten, die sich auf die gleiche Entität oder Sache beziehen (Co-Referenz-Auflösung)
- Abgleich, Aggregation und Verarbeitung von Dokumenten (z. B. Nachrichten, Geschäftsdokumente etc.)
- Information Retrieval inklusive der Verbesserung der Suche nach Informationen (z. B. semantische Suche, Bild-zu-Text-Suche etc.)
- Textgenerierung (z. B. Werbetexte, Zusammenfassung, Bildunterschriften etc.)
- Automatische Übersetzung
- Dialogsysteme, persönliche Assistenzsysteme, Chatbots
- Generierung von Programmcode und Unterstützung von Programmierenden (vgl. Github Copilot, Starcoder)
- Und vieles mehr

Beispiele für Aufgabenstellungen multi-modaler Modelle

- Generierung von Bildern, Videos aus Texteingaben (z. B. Stable Diffusion)
- Erstellung von 3D-Formen und -Modellen aus 2D-Zeichnungen und -Bildern.
- Erstellung von Bauteilen durch Eingabe geometrischer und physikalischer Spezifikationen (z. B. in der Raumfahrt, NASA)
- Steuerung von Robotern (z. B. die Modelle Palm-E, RoboCat)
- Suche von einer Modalität in die andere (z. B. Bild-zu-Text-Suche)
- Generierung von Bildunterschriften und -beschreibungen
- Abgleich und Bestätigung zwischen Modalitäten zur Verbesserung von Modellfähigkeiten (z. B. Bild- und Textverstehen)
- Und vieles mehr

NLP ist eine zentrale Achse in Daten-Ökonomien. In Service-orientierten Gesellschaften sind Textdaten allgegenwärtig. Angesichts des zunehmenden Talente- und Arbeitskräftemangels bedarf es im Bereich der Textverarbeitung weiterer Automatisierung. Die anzugehenden Probleme, oft beruhend auf komplexen semantischen Zusammenhängen, lassen sich mit klassischen NLP-Methoden kaum mehr abbilden. Um den nächsten Automatisierungsgrad für unsere Kunden erreichen zu können, benötigen wir Sprachmodelle auf dem neuesten Stand der Technik, die in der Lage sind, auf wenig Daten mit komplexen Sachverhalten umzugehen. Dafür ist es wichtig, vortrainierte Modelle zu nutzen, die möglichst viel „common sense", logisches Schlussfolgern („reasoning"), Semantik und Syntax abbilden – wie eben große, vortrainierte Sprachmodelle. (2022)[1]

Johannes Otterbach, ehemals Merantix, seit Juli 2023 nyonic

Große Sprachmodelle sind ein elementarer Baustein aktueller Modelle der natürlichen Sprachverarbeitung. Destillierte große Sprachmodelle haben bei uns die älteren Technologien der Named Entity Recognition und der Summerization abgelöst. Als Unternehmen versuchen wir vor allem KMU zu unterstützen. Unser Anspruch war es immer, Modelle zu haben, die ein gutes Preis-Leistungs-Verhältnis aufweisen. Gerade vor dem Hintergrund der Fortschritte von Modellen wie ChatGPT und GPT-4 untersuchen wir Methoden, um diese Technologien dem breiten Markt zugänglich zu machen. Beispielsweise indem wir durch Prompting ihre Zuverlässigkeit erhöhen und durch Modellkomprimierung ihre technischen Anforderungen reduzieren. (2022)[2]

Till Plumbaum, ehemals Neofonie GmbH, seit Februar 2023 Alexander Thamm GmbH

1 NLP steht für Natural Language Processing, also die Verarbeitung natürlicher Sprache.

2 Named Entity Recognition entspricht der Erkennung von benannten Entitäten, Summerization bezeichnet hier die automatisierte Textzusammenfassung.

KURZINFO

Ausblick – Planung und autonome KI-Assistenten

Sprachsysteme werden zukünftig auch **komplexe Planungsaufgaben übernehmen** und ganze Prozesse – statt nur einzelner Schritte – planen können: So werden Modelle in der Lage sein, Prozesse in einzelne Schritte, Aufgaben und Meilensteine zu zerlegen, um das Management von Einzelaufgaben, Werkzeugen, Arbeit und Kooperation zu unterstützen. Hier steht die Forschung allerdings erst am Anfang.

Vor dem Hintergrund dieser Forschungsrichtung können Sprachmodelle auch zu einem **zentralen Bestandteil autonomer KI-Assistenten** werden. Erste Versuche in diese Richtung wurden bereits mit KI-Assistenten wie AutoGPT oder BabyAGI unternommen. Solche Assistenten könnten in Zukunft viele Geschäftsanwendungen ermöglichen, ohne dass eine umfangreiche Softwareentwicklung notwendig ist, indem sie…

… komplexere Aufgaben in Teilschritte aufteilen und automatisieren.

… verschiedene Arten von großen KI-Modellen (Sprache, Code, Bild etc.) integriert nutzen, um Aufgaben zu erfüllen, und dabei auch Komponenten einbinden, die nicht auf großen KI-Modellen basieren, wie z. B. Such- und Rechenmaschinen.

… die Ausgabe eines Modells überprüfen und neu schreiben und dabei auf die Ausgabe eines anderen Modells oder einer anderen Software zurückgreifen.

… etwas selbstständig ausprobieren, die Ergebnisse überprüfen und akzeptieren, wenn sie funktionieren, oder andernfalls einen anderen Versuch unternehmen, um die Aufgabe zu erledigen.

… laufende Eingaben kontinuierlich ausführen und verarbeiten (z. B. ein laufendes System über die Zeit steuern).

Autonome KI-Agenten stellen Forschung, Gesellschaft und Wirtschaft aber auch vor große Herausforderungen, da sie einerseits robuste und belastbare Reasoning-Fähigkeiten erfordern und andererseits umfangreiche Vorkehrungen und Einschränkungen notwendig sind, damit solche Agenten nicht missbräuchlich eingesetzt werden oder unbeabsichtigte Folgen haben. Diese Fragestellungen werden die Forschung weiter antreiben.

Quelle: Eigene Zusammenstellung basierend auf Löser & Tresp et al. (2023, S. 25), Schütze (2023) und Vogel (2023).

Zwei Anwendungsfelder werden im Folgenden detaillierter betrachtet: Geschäftsanwendungen sowie Anwendungen im Gesundheitswesen. Dabei wird auf das jeweilige Potenzial sowie auf die Herausforderungen eingegangen, aber es werden auch mögliche Lösungsansätze aufgezeigt. Es handelt sich dabei um Anwendungsfälle, in denen sensible Daten Teil der Verarbeitungsprozesse sein können, zum einen sensible Geschäftsdaten und internes Wissen und zum anderen datenschutzrechtlich besonders sensible medizinische Daten von Patientinnen und Patienten. In beiden Fällen sind Verarbeitungslösungen im Sinne Digitaler Souveränität (DS) wünschenswert und die Berücksichtigung europäischen und deutschen Rechts bedeutend, um Rechtssicherheit für Akteure zu sichern.

2.1 Geschäftsanwendungen

In Geschäftsanwendungen sind häufig komplexe und stark strukturierte Benutzeroberflächen zu entwickeln, die eine anwendungsspezifische Kommunikation zwischen Mensch und Maschine ermöglichen. Im Gegensatz dazu sind Sprache und Text die natürliche Form der Kommunikation zwischen Menschen und als sprach- und textbasierte Interfaces zunehmend auch zwischen Mensch und Maschine. Dementsprechend sind auch Sprachtechnologien und Sprachmodelle wichtige Bausteine für zukünftige KI-basierte Anwendungen in Unternehmen. Anhand der zwei Beispiele – Digitale Assistenten und Dokumentenverarbeitung – sollen zum einen das Potenzial von Sprachtechnologien aufgezeigt werden und zugleich einige Herausforderungen bei der Entwicklung von Sprachmodellen für Geschäftsanwendungen erläutert werden.

Digitale Assistenten

Während von Anwendenden komplexe und komplizierte Benutzerschnittstellen bei Geschäftsanwendungen in der Vergangenheit noch akzeptiert wurden (oder akzeptiert werden mussten), erwarten diese heutzutage, und ebenso Kundinnen und Kunden, in ihrem Arbeitsumfeld eine ähnlich komfortable Benutzererfahrung, wie sie es von ihren privat genutzten Apps und Plattformen gewohnt sind. Digitale Assistenten in Geschäftsanwendungen können hier einen wichtigen Beitrag zur Verbesserung der Benutzererfahrung leisten: beispielsweise durch Navigations- und Suchhilfen innerhalb einer Applikation („Ich möchte eine neue Stelle ausschreiben“) oder das Beantworten von Wissensanfragen („Welche offenen Stellen für Berufseinsteiger haben wir in unserer Abteilung aktuell?“). Die Eingabe kann flexibel in natürlicher Sprache erfolgen, was vor allem für neue Anwendende einen hohen Komfort darstellt und zudem einfacher zu erlernen ist als komplizierte Menüs oder spezielle Transaktionscodes. Im Gegensatz zu einfachen Chatbots, die in der Regel nur simple, transaktionale Aufgaben im Applikationskontext ausführen und FAQ-ähnliche Fragen beantworten können, werden digitale Assistenten über Applikationsgrenzen hinweg arbeiten, den Kontext der Nutzenden in Betracht ziehen und sich durch Personalisierung immer weiter auf diese einstellen können.

Große Sprachmodelle sind jetzt schon ein essenzieller Baustein für die NLU (Natural Language Understanding)-Komponenten eines digitalen Assistenten, zum Beispiel bei der Intentions- und Entitäten-Erkennung. Große Sprachmodelle dienen als Basismodelle, die mit einer geringen Anzahl von Beispielen an individuelle Kundenbedürfnisse wie kundeneigene Terminologien angepasst werden können. Dies verringert den Aufwand der Datensammlung sowie den Entwicklungsaufwand bei der Einführung von digitalen Assistenten auf Kundenseite.

Langfristig können immer leistungsfähigere Sprachmodelle direkte Interaktionsmöglichkeiten bieten, bei denen die Modelle nicht als Teil einer NLU eingesetzt werden, sondern direkt eine Antwort auf Eingaben von Nutzerinnen und Nutzern generieren. Dies wird durch die generativen Fähigkeiten von sehr großen Sprachmodellen ermöglicht, siehe beispielsweise die von OpenAI entwickelte Chat-Umgebung von ChatGPT.

Trotz der Fortschritte bestehen aber weiterhin Herausforderungen für eine erfolgreiche Anwendung von Sprachmodellen für digitale Assistenten im Unternehmenskontext. Ein Beispiel sind die verwendeten Daten. In der Forschung werden in der Regel große Datenmengen aus dem Internet (z. B. Common Crawl) verwendet. Diese Daten sind für kommerzielle Geschäftsanwendungen schwer nutzbar, einerseits aufgrund von möglichen Problemen mit Datenschutz und Lizenzen, andererseits, weil es eine Diskrepanz der Sprachdomänen gibt. Texte im Web sind nur bedingt mit Texten aus Geschäftsdokumenten vergleichbar. Eine weitere Herausforderung ist die Notwendigkeit, dass sich Modelle schnell an neue Gegebenheiten und Fakten anpassen müssen, insbesondere, wenn Nutzende mit relationalen Daten in einer Anwendungsdatenbank interagieren müssen. Wenn sich diese Daten verändern, muss der digitale Assistent darauf sofort reagieren und natürlichsprachliche Äußerungen müssen gegebenenfalls mit anderen Einträgen in der Datenbank verknüpft werden, zum Beispiel beim Umbenennen eines Produktes oder eines Geschäftspartners. Dies sollte für die Anwendenden möglich sein, ohne dass dazu ein aufwändiges Neutrainieren durch Expertinnen und Experten erforderlich ist. Wissensgraphen sind hier eine vielversprechende Technologie, um das dazu notwendige Wissen flexibel zu modellieren und um es mit natürlicher Sprache zu verknüpfen (siehe hierzu auch Kombinierte KI bzw. hybride KI bei Löser & Tresp, 2023, S. 20).

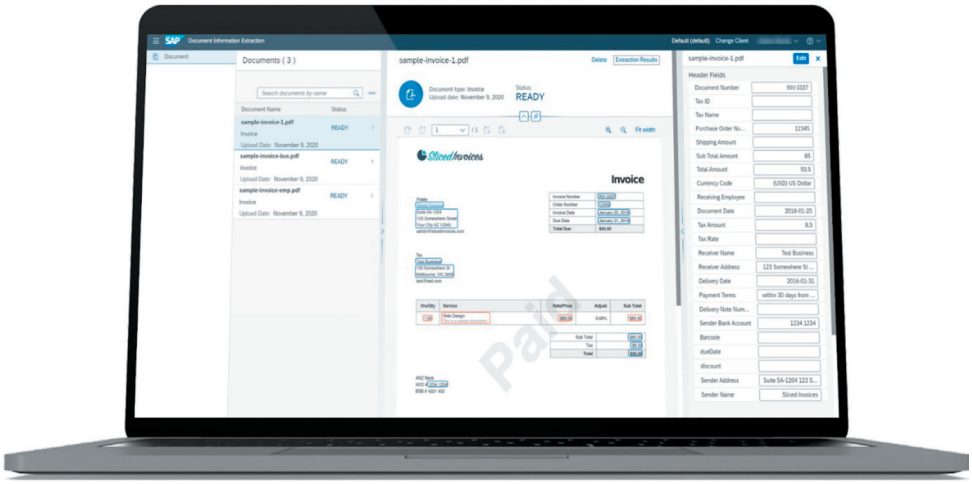
Dokumentenverarbeitung

Geschäftsdokumente sind Bestandteil fast aller Geschäftsprozesse. Dokumente werden zwischen Unternehmen (wie auch innerhalb einer Firma) oft im PDF-Format oder sogar noch auf Papier ausgetauscht. Mit der voranschreitenden Digitalisierung von Prozessen wird diese Form des Austausches seltener werden; doch bis dahin ist es noch ein weiter Weg, unter anderem, weil oft keine standardisierten Schnittstellen genutzt werden. Dokumente müssen auf der Empfängerseite in der Regel oft noch elektronisch erfasst werden, nicht selten durch manuelle Eingabe der Informationen in eine Geschäftsanwendung. Dies gilt vor allem für KMU, für die sich derartig aufwändige Systemintegrationen oft nicht rechnen.

Sprachtechnologie und Informationsextraktion haben das Potenzial, solche repetitiven Aufgaben zu erleichtern und zu automatisieren. Egal ob Rechnungen, Zahlungsavisen oder Bestellungen – Geschäftsdokumente enthalten reichhaltige Tabellenstrukturen, Sonderbegriffe, Referenznummern etc. Dies verdeutlicht die Komplexität der Problemstellung. Das Dokument muss nicht nur per OCR in Text umgewandelt werden, sondern die relevanten Informationen müssen im Dokument erkannt, extrahiert und mit Geschäftsdaten angereichert bzw. verknüpft werden (zum Beispiel die Verknüpfung einer erkannten Firma mit dem Geschäftspartner im System).

Das Unternehmen SAP hat eine Reihe eigener Deep Learning-Modelle speziell für die Dokumentenverarbeitung entwickelt: Dieses Chargrid-Modell verwendet eine Deep Learning-Architektur aus der Bildverarbeitung, repräsentiert das Dokument aber in einer neuartigen „Character Grid"-Repräsentation, die die Vorzüge von Text und Bild kombiniert. BERTgrid erweitert diesen Ansatz durch kontextbasierte Embedding-Repräsentationen auf Basis des großen Sprachmodells. Die Ansätze werden also kontinuierlich erweitert. Sowohl SAP als auch andere Unternehmen arbeiten an großen Sprachmodellen für visuell strukturierte Dokumente. Durch immer leistungsfähigere Sprachmodelle wird die Informationsextraktion flexibler. Lösungen können ohne große Datenmengen auf neue Dokumentenvarianten oder Dokumententypen angepasst und erweitert werden. Der geringere Aufwand und die höhere Genauigkeit ermöglichen auch die Nutzung in einer zunehmenden Anzahl von kritischen Prozessen.

Abbildung 1: **Interface eines Programms zur Dokumentenverarbeitung**



Quelle: Dokumentenverarbeitung mit Document Information Extraction von SAP, siehe Denk & Reisswig (2019); Katti et al. (2018); Klaiman & Lehne (2021).

2.2 Gesundheitswesen

Seit 2019 sind vortrainierte Sprachmodelle für den (bio-)medizinischen Bereich in englischer Sprache öffentlich verfügbar (Lee et al., 2020). In den vergangenen Jahren sind weitere Variationen hinzugekommen, die sich hauptsächlich in den genutzten Daten unterscheiden (Alsentzer et al., 2019; Huang et al., 2019). Üblicherweise wird auf große Datensätze zurückgegriffen, die medizinisches Fachwissen beinhalten, beispielsweise Publikationen aus der PubMed-Datenbank oder klinische Patientenbriefe und -daten aus der MIMIC-Datenbank (Medical Information Mart for Intensive Care). Im Vergleich zu domänenunabhängigen Modellen erreichen Sprachmodelle, die auf domänenspezifischer Sprache vortrainiert wurden, bessere Ergebnisse basierend auf medizinischen und klinischen Texten. Es hat sich zudem gezeigt, dass das alleinige Vortrainieren auf medizinischen Texten mitsamt eines spezialisierten Tokenizers oft zu Verbesserungen führt (Tinn et al., 2021).

Der größte Teil der bisherigen Anwendungen von Sprachmodellen im Gesundheitsbereich konzentriert sich auf englischsprachige medizinische und klinische Texte. Dies ist vor allem darauf zurückzuführen, dass große, öffentlich zugängliche medizinische Datensätze, die sich für das Vortrainieren von Sprachmodellen anbieten, überwiegend in englischer Sprache vorliegen. Daher wurde der multi-linguale medizinische Wissenstransfer vom Englischen in andere Sprachen untersucht (Papaioannou et al., 2022). Neben simplen Übersetzungen wurde die Wirksamkeit von sprachübergreifenden Modellen und sprachspezifischen Modell-Adaptern betrachtet. Die Analyse ergab, dass das Übersetzen ein geeignetes Mittel für den Wissenstransfer sein kann, sofern die Übersetzungen ein ausreichendes Niveau für medizinische Sprache erfüllen. Ist dies nicht der Fall, bieten sich sprachübergreifende Modelle in Kombination mit Sprach- und Aufgaben-Adaptern an.

Da Sprachmodelle für die Medizin überwiegend aus dem englischen Sprachraum stammen, wird deutlich, dass domänen-spezifische Sprachmodelle für die deutsche Sprache auch in der Medizin wünschenswert sind. Erste Schritte in diese Richtung wurden basierend auf dem BERT-Modell bereits unternommen (Bressem et al., 2023).

Konkrete Anwendungsbereiche solcher Sprachmodelle sind die Outcome-Vorhersage und die Entscheidungsunterstützung. Gerade im medizinischen Bereich sind möglichst genaue, robuste und erklärbare Resultate essenziell, was anhand der beiden folgenden Beispiele aufgezeigt wird.

Outcome-Vorhersage: Zahlreiche Ansätze untersuchen den Prozess der Differenzialdiagnose (siehe auch GAIA-X Case: BMWK, 2023a). Dieser Prozess ist Kern vieler ambulanter und stationärer Untersuchungen. Dabei erhält die Ärztin oder der Arzt eine Ersteinschätzung in der Anamnese der zu behandelnden Person, die textuelle Daten (klinische Anamnesebögen, engl. Clinical Admission Notes), Vitaldaten, Laborwerte etc. enthält. Das System schlägt den Ärztinnen und Ärzten mögliche Diagnostiken vor, zeigt Verteilungen von kritischen und auszuschließenden Diagnosen auf und empfiehlt Medikamente sowie auch Behandlungsmethoden. Sprachmodelle repräsentieren die Anamnesebögen, die oft Hinweise zu Anomalien enthalten, bzw. die Patientenhistorie. Aufgrund der großen Varianz von bis zu hunderttausend Fällen haben Sprachmodelle mehr Patientinnen und Patienten „gesehen" als selbst erfahrene Chefärztinnen oder Chefärzte in ihrer gesamten Berufskarriere. Dadurch könnten diese Systeme auf Basis von Sprachmodellen signifikant weniger erfahrene Ärztinnen und Ärzte auf potenzielle Anomalien und Risiken hinweisen, die abseits von Standards, wie Leitlinien, auftreten können.

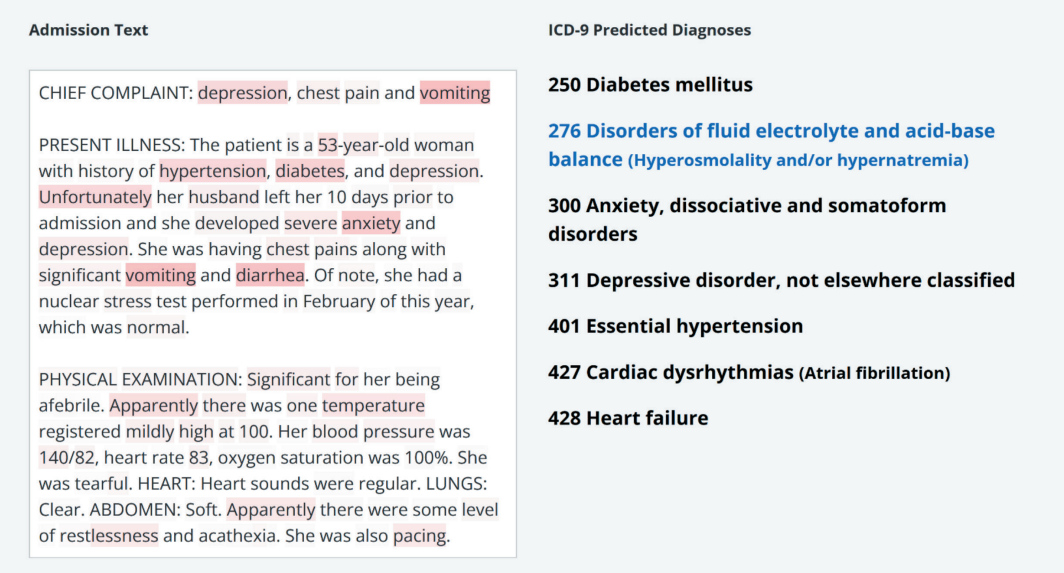
Eine Möglichkeit, Sprachmodelle in der Outcome-Vorhersage zu verbessern, stellt die Kombination solcher Modelle mit strukturiertem, bestehendem medizinischen Experten- und Hintergrundwissen dar (siehe hierzu auch Kombinierte KI bzw. hybride KI: Löser & Tresp, 2023, S. 20). So wurde beispielsweise ein Modell zusammen mit drei medizinischen Wissensgraphen genutzt. Die Relationen dieser Wissensgraphen wurden mit einem auf Ontologien angepassten Sprachmodellierungs-Verfahren gelernt und konnten somit in die Modellparameter integriert werden (siehe KIMERA-Modell; Winter et al., 2022). Dadurch erhöht sich die Robustheit der Modelle auch gegenüber ungewöhnlichen Vorhersagen. Auch tabellarische Daten (Miotto et al., 2016; Topol, 2019), Bilder (Esteva et al., 2021), Zeitreihendaten (Yang & Wu, 2021) oder Vorerkrankungen aus Begleitdiagnosen (Grundmann et al., 2021) können einbezogen und mit Sprachmodellen weiterverarbeitet werden. Hierdurch kann die Vorhersagekraft von Sprachmodellen verbessert werden.

Entscheidungsunterstützung: Vortrainierte Sprachmodelle eignen sich ebenfalls als Baustein für Fragen-Beantwortung (Question Answering) und für Chatbot-Aufgaben in medizinischen Bereichen (Arnold et al., 2020; Grundmann et al., 2021). So können sie als eine Grundlage für Systeme zur Entscheidungsunterstützung dienen. Da Sprachmodelle allerdings üblicherweise aus einer großen Menge an Parametern bestehen, lassen sich Vorhersagen nur schwer nachvollziehen und überprüfen. Dies ist vor allem im klinischen Betrieb – insbesondere in der Notfallmedizin – problematisch, da Ärztinnen und Ärzte Vorhersagen schnell verifizieren oder verwerfen müssen und somit auf Hinweise angewiesen sind, wie diese Vorhersagen zustande gekommen sind.

Aus diesem Grund müssen Entscheidungsunterstützungssysteme den Kriterien der Erklärbarkeit und Interpretierbarkeit Rechnung tragen können. Für diese Herausforderung wurden bereits verschiedene Lösungen vorgeschlagen. Beispielsweise kann dies durch die Integration von Sprachmodellen in prototypische Netzwerke gelingen (van Aken et al., 2022) oder durch Diagnose-spezifische Attention (Mullenbach et al., 2018), die den Fokus auf relevante Textpassagen legt, welche für eine Diagnose von besonderer Bedeutung sind.[3] Dadurch können Modell-Vorhersagen besser kommuniziert und schneller überprüft werden (siehe Abbildung 2 und 3 für Demonstratoren).

3 Hintergrund: Aus bestehenden Texten und Ergebnissen (Diagnosen) wird für spezifische Krankheitsbilder ein prototypisches, künstliches neuronales Netzwerk berechnet. Da die Datengrundlage bekannt ist, kann hier nachvollzogen werden, welche Buchstaben von Wörtern in den Texten und Diagnosen sich wie stark auf Gewichtungen im Netzwerk auswirken. Dieses Netzwerk dient dann als Vorlage für den Abgleich mit berechneten Repräsentationen von Informationen über neue Patienten. Finden sich Übereinstimmungen, geht aus dem Prototyp hervor, welche Buchstaben besonders bedeutend waren. Diese können dann wiederum angezeigt werden.

Abbildung 2: **Interface eines medizinischen Entscheidungs-unterstützungssystems**



Quelle: Demo für interpretierbare medizinische Entscheidungsunterstützungssysteme mittels Sprachmodellen und prototypischen Netzwerken. Das System erhält als Input eine kurze und oft unvollständige Beschreibung der Patientin/des Patienten, woraus es eine Diagnose-Vermutung als ICD-Code erstellt. Für jede Diagnose-Vermutung gibt es hervorgehobene Wörter im Text in unterschiedlicher Wichtigkeit (Grundermann et al., 2022; van Aken et al., 2021; van Aken et al., 2022).

Abbildung 3: **Abruf von Informationen aus großen Gesundheitsdatensätzen**

Quelle: Prototyp der Charité und Berliner Hochschule für Technik für Anfragen der Struktur <Disease, Thema> an Datenbanken mit medizinischen Dokumenten. Das System klassifiziert Sätze und Abschnitte, die eine Antwort enthalten, und unterstützt mehrere 10.000 Krankheiten und Themen. Das vortrainierte Sprachmodell ist essenziell, da pro Thema und Krankheit nur einige wenige Dokumente als Trainingsdaten vorliegen, sodass die Varianz der Wortsequenzen aus dem Sprachmodell benutzt wird (Arnold et al., 2019; Arnold et al., 2020).

3 Ebenen Digitaler Souveränität bei großen Sprachmodellen

Die verschiedenen Perspektiven der Anwendungen von Sprachmodellen zeigen das enorme Potenzial, welches in dieser KI-Technologie steckt. Sowohl die Geschäftsanwendungen als auch die Anwendungen im Gesundheitswesen verdeutlichen jedoch zugleich, dass es sich um potenziell sensible Daten handelt, was etwa Datenschutz und Rechtssicherheit betrifft. Damit diese Technologie im Sinne europäischer Werte und der Rechtssicherheit für Forschende, Entwickelnde und Anwendende (weiter-)entwickelt und in Deutschland und Europa in die Anwendung gebracht werden kann, ist auf die Digitale Souveränität zu achten. „Digitale Souveränität meint die Fähigkeit von Individuen, Unternehmen und Politik, frei zu entscheiden, wie und nach welchen Prioritäten die digitale Transformation gestaltet werden soll" (Kagermann et al., 2021), und damit auch die gegenwärtige Transformation durch KI. In loser Anlehnung an das Mehrebenen-Konzept für Digitale Souveränität der Deutschen Akademie der Technikwissenschaften (ebenda), das auf technologie- und daten-bezogenen Aspekten aufbaut, werden im Folgenden mehrere Ebenen betrachtet, die für große Sprachmodelle besonders bedeutend sind. Als weitere Ebene wurde zusätzlich der Aspekt der Talente hinzugefügt. So werden zum einen die Ebene der Recheninfrastruktur und damit auch die hierfür notwendigen Komponenten betrachtet und zum anderen die Ebenen des Zugangs zu Daten und Modellen sowie schließlich zu Fachkräften, um Aufgaben in Forschung, Entwicklung, Wartung, Vertrieb und Service für die Bereitstellung dieser Sprachmodelle auch am Standort Deutschland und/oder in Europa wahrnehmen zu können.

3.1 Europäisches Werte- und Rechtssystem

Sprachmodelle können verletzende oder falsche Ergebnisse generieren und zur Verbreitung von Desinformationen verwendet werden (siehe zu Bias auch Löser & Tresp, 2023, S. 25). Die Praktiken im Zusammenhang mit ihren Daten und ihrer Bereitstellung werfen viele rechtliche wie auch ethische Fragen auf (z. B. Datenschutz, Urheberrecht, ggf. Verbreitung von Falschnachrichten oder auch das Vortäuschen von Prüfungsleistungen etc.). Mit dem AI Act als Rechtsinstrument möchte die Europäische Union unter anderem einen rechtlichen Rahmen für KI bereitstellen, der einen risiko-basierten Ansatz verfolgt, um die Berücksichtigung europäischer Werte beim Aufbau und Einsatz von großen Sprachmodellen einzufordern. Eine Studie der Universität Stanford zeigt, dass viele prominente Sprachmodelle die in Europa diskutierten Evaluations-, Transparenz- und Dokumentationsanforderungen überwiegend nicht erfüllen (Bommasani et al., 2023). Auch für das weitgehend offene und kommerziell nutzbare Modell „LlaMA 2" kann Meta nicht garantieren, dass keine urheberrechtlich geschützten Arbeiten oder persönliche Daten verarbeitet wurden (Heikkilä, 2023).

Wie jedoch das Beispiel der Geschäftsanwendungen (siehe Abschnitt 2.1) deutlich betont, ist für Unternehmen eine rechtssichere Nutzung großer KI-Modelle bedeutend: Es sollte transparent sein, welche Daten für das Training solcher Modelle genutzt wurden, und zudem sollte europäisches Recht bei der Datenerfassung und -kuration eingehalten werden. Es ist allerdings darauf hinzuweisen, dass nicht nur der Datensatz im Sinne europäischer Werte und Regeln erstellt werden sollte, sondern auch Vorkehrungen, die getroffen werden, um das Modellverhalten im Sinne menschlicher Qualitätsvorstellungen zu optimieren. Dies gilt zum Beispiel für Bewertungsmodelle für Modellausgaben, wie sie aus dem verstärkenden Lernen auf Basis menschlicher Rückmeldungen (reinforcement learning from human feedback, RLHF) hervorgehen. Dabei handelt es sich um ein Verfahren, das auf der Basis menschlicher Bewertungen von Modellausgaben Feedback für einen

Algorithmus erstellt, der das generative Modell so anpasst, damit die Ausgaben näher an den menschlichen Präferenzen liegen (siehe hierzu Infografik 2 und Infografik 4 der Plattform Lernende Systeme).

Vor diesem Hintergrund liegt es nahe, monolinguale deutschsprachige wie auch multi-linguale europäische Sprachmodelle in Deutschland beziehungsweise Europa zu entwickeln. Auf diese Weise kann der ethische Rahmen bei der Auswahl der Trainingsdaten gewährleistet werden sowie eine Vermeidung von Bias und zugleich eine Wahrung der Privatsphäre erfolgen. Sprachmodelle „Made in Germany bzw. Europe" könnten sicherstellen, dass bei sensiblen Aufgaben, im Gesundheitswesen, im Katastrophenfall, in der Ausbildung und in den wichtigsten Kernbranchen Deutschlands (Justiz, Behörden, Forschung und Entwicklung, Sicherheit etc.) auf diese bedeutenden und einflussreichen KI-Modelle zugegriffen werden kann, ohne wirtschaftliche oder technologische Abhängigkeiten aufzubauen.

3.2 Daten – die Grundvoraussetzung

Der Zugang zu Datensätzen für das Training großer Sprachmodelle ist eine Grundvoraussetzung für die Erstellung qualitativ hochwertiger und leistungsfähiger Sprachmodelle für die deutsche Sprache und auch für multi-linguale Modelle. Sind Entwickelnde und Unternehmen auf Modelle angewiesen, die im Vergleich nur über einen geringen Anteil deutscher Textdaten im Trainingsdatensatz verfügen und zudem deutsche wie europäische Werte und Regeln bei der Datensatzerstellung nicht berücksichtigen, kann dies zu einem Nachteil im Wettbewerb um die besten Modelle und Anwendungen führen.

Es hat sich gezeigt, dass auch Sprachmodelle mit einer geringen Anzahl von Parametern eine ähnliche Leistungsfähigkeit wie sehr große Modelle erreichen können, wenn die Anzahl der Datensätze pro Parameter erhöht wird. So werden für Modelle derzeit Verhältnisse von Parametern zu Trainingsdaten von bis zu 1 zu 50 oder sogar 1 zu 100 verwendet (Stand 2023). Das größte Modell der populären und frei verfügbaren Modellreihe LLamA 2 wurde mit 2 Billionen Token[4] aus öffentlich zugänglichen Textdaten trainiert; das entspricht einem Parameter-Token-Verhältnis von 1 zu 28. Der Anteil der deutschen Sprache an diesem Datensatz beträgt jedoch nur 0,17 Prozent. Das bedeutet, dass ein multi-linguales Modell mangels deutscher Sprachbeispiele Antworten auf Basis der Datenlage anderer Sprachen im Datensatz, also insbesondere der englischen Sprache, ausgibt. Dies kann zu Ungenauigkeiten, Verzerrungen und Fehlern in den generierten deutschen Texten führen und damit zu einem Nachteil für Nutzende, Entwickelnde und Unternehmen in Deutschland.

Trainingsdatensätze sollten Transparenz, Nachvollziehbarkeit und Rechtssicherheit für Forschende, Entwicklerinnen wie Entwickler und Unternehmen gewährleisten. Wie bereits in Abschnitt 3.1 ausgeführt, ist dies bei prominenten Sprachmodellen in vielen Fällen nicht der Fall, da die Datensätze entweder nicht einsehbar sind oder nicht sichergestellt werden kann, ob urheberrechtlich geschütztes Material oder personenbezogene Daten verarbeitet wurden. Das Open-Source-Modell BLOOM, das auch die Trainingsdaten selbst zur Verfügung stellt, aber leider keine deutsche Sprache enthält, ist eines der wenigen Modelle, das viele der in Europa diskutierten Evaluations-, Transparenz- und Dokumentationskriterien erfüllt. Auch hier wurde ein besonderes Augenmerk auf die Sammlung von Textdaten und deren Kuratierung gelegt (siehe Infobox, Seite 29). Es zeigt sich, dass Unternehmen wie Forschende in Europa in vielerlei Hinsicht nicht auf einer rechtssicheren, nachvollziehbaren und bekannten Basis von Analysen und Regeln Sprachmodelle implementieren und Anwendungen entwickeln können.

4 Hinweis: Ein Token entspricht in etwa einem Wort, einem Satzzeichen oder einer anders zusammenhängenden Zeichenfolge.

In beiden Fällen, dem geringen Anteil deutschsprachiger Daten in bestehenden multi-lingualen Modellen sowie der mangelnden Transparenz und Rechtssicherheit der zugrundeliegenden Trainingsdatensätze, ergeben sich die Herausforderungen für Forschende und Unternehmen in Deutschland und Europa aus der relativen Abhängigkeit von prominenten Sprachmodellen, die nicht im europäischen Raum entwickelt wurden und daher eher andere Normen, Werte und Regeln berücksichtigen. Aus diesem Grund wäre ein umfangreicher, qualitativ hochwertiger und im Sinne europäischer Werte kuratierter Trainingsdatensatz eine große Chance sowohl für Forschende und Entwickelnde als auch für Unternehmen, insbesondere, wenn er nicht nur frei verfügbar, sondern auch kommerziell nutzbar wäre. Hinsichtlich des erforderlichen Umfangs ist eine Orientierung an aktuellen Projekten zur Reproduktion von Datensätzen bekannter Modelle wie LLamA für das Training von Open-Source-Modellen sinnvoll.

So hat das Projekt Redpajama einen entsprechenden Datensatz mit 1.200 Milliarden Token gesammelt und kuratiert, der überwiegend englischen Text umfasst. Wird ein mehrsprachiges Modell angestrebt, so werden gegebenenfalls nur einige 100 Milliarden Token in deutscher Sprache benötigt. Dies entspricht 10 bis 15 Terabytes deutschem Text, die auf wenigen Laptop-Festplatten Platz finden. Prinzipiell sind auch verschiedene größere Textdatensätze verfügbar (siehe Tabelle 1). Der Erstellung eines Datensatzes mit den oben genannten Charakteristika stehen jedoch häufig lizenzrechtliche Hürden entgegen oder frei zugängliche Textdaten müssen erst erschlossen, gesammelt und kuratiert werden, sodass es durchaus aufwändig sein kann, einen solchen Datensatz im gewünschten Umfang zu erstellen.

Exemplarisch zeigt sich dies am Deutschen Referenzkorpus (DeReKo), das mit 55 Milliarden Wörtern als weltweit größte Sammlung elektronischer Korpora geschriebener deutschsprachiger Texte für die linguistische Forschung ausgewiesen wird (IDS, 2023). Dies entspricht in etwa einem Zehntel der Textmenge, die für die angestrebten 10 bis 15 Terabyte Trainingsdaten benötigt wird. Das DeReKo ist aus urheber- und lizenzrechtlichen Gründen auf die wissenschaftliche, nichtkommerzielle Nutzung beschränkt, wobei die Nutzenden keinen Zugriff auf das Gesamtkorpus haben. 2017 wurden hierfür bereits Lizenzvereinbarungen mit über 200 Rechteinhabern abgeschlossen (Lüngen, 2017). Dies zeigt, vor welchen Herausforderungen der Aufbau eines großen Textdatensatzes für das Training großer Sprachmodelle steht, insbesondere, wenn eine möglichst breite Nutzbarkeit angestrebt wird, um das Potenzial großer Sprachmodelle in Deutschland bestmöglich ausschöpfen zu können.

Ein solcher Datensatz birgt aber auch große Chancen: Denn sobald er verfügbar ist, wird ein großes Potenzial für Unternehmen, Forschende sowie Entwicklerinnen und Entwickler auf verschiedenen Ebenen und in verschiedenen Bereichen freigesetzt. Denn dieser kann die Grundlage für die Erstellung vieler verschiedener Sprachmodelle sein, die Organisationen nutzen können, um Aufgaben durch oder unterstützend mit KI zu erledigen. Die Verfügbarkeit von Daten in Organisationen, die herangezogen werden können, um Sprachmodelle für verschiedene Aufgaben zu nutzen, sind allerdings ungleich verteilt: Es wird wenige Aufgaben in Organisationen geben, für die viele Daten verfügbar sind, und viele, für die nur wenige Daten verfügbar sind (Molino, 2023). Vor der Einführung großer Sprachmodelle waren meist nur solche KI-Lösungen realisierbar, für die Organisationen über sehr viele Daten für das Modelltraining verfügten. Jetzt können Organisationen durch Anpassung großer Sprachmodelle eigene domänenspezifische Modelle erstellen, die verschiedene Aufgaben ausführen können. Dafür benötigen sie weniger Daten als für das Training neuer (spezifischer) Modelle. Selbst wenn einer Organisation für bestimmte Aufgaben nur wenige Daten zur Verfügung stehen, können durch die Verknüpfung der Sprachmodelle mit Datenbankabfragen (vgl. z. B. Retrieval Augmented Generation) Lösungen gefunden werden. Schließlich sind in Organisationen für viele Aufgaben kaum Daten verfügbar, aber die Fähigkeit großer Sprachmodelle, auch aus wenigen Beispielen zu lernen und Aufgaben entsprechend umzusetzen, bietet auch hier Lösungsansätze. Da zu erwarten ist, dass sich die Datenlage in Organi-

sationen und die Fähigkeiten der Modelle in Zukunft weiter verbessern werden, können KI-Lösungen für weitere Aufgaben erschlossen werden. Auch für das Training dieser zukünftigen Modelle wäre ein Datensatz mit den oben beschriebenen Eigenschaften grundlegend.

Tabelle 1: **Übersicht zu Textdatensätzen in deutscher Sprache** (nicht abschließend)

Datensatz	Datenumfang (deutschsprachig)
OpenAssistant Conversations Dataset (OASST1)	Von Menschen erstellter und kommentierter Konversationskorpus im Assistentenstil: 5.279 Nachrichten Zum Vergleich: Spanisch: 43.061, Englisch: 71.956
Oscar Datensatz	Multilingualer Textdatensatz: ca. 73,8 Milliarden Wörter, ca. 20 Millionen Dokumente Zum Vergleich: Englisch: ca. 523,8 Milliarden Wörter, ca. 1,2 Millionen Dokumente
Colossal Oscar 1 Datensatz	Größte Veröffentlichung des OSCAR-Korpus auf der Grundlage von zehn verschiedenen monatlichen Snapshots von Common Crawl*
CulturaX	Multilingualer Textdatensatz basierend auf Common Crawl (27 Terabyte): 357 Milliarden Token (5,66 Prozent aller Token), 420 Millionen Dokumente Zum Vergleich: Englisch: 2847 Milliarden Token (45,13 Prozent aller Token), 3,24 Milliarden Dokumente
GermanQuAD	Von Hand annotierte Datensätze für die Beantwortung von Fragen: 13.722 Fragen und Antworten
Huge German Korpus	Sammlung deutschsprachiger Texte (Zeitungsartikel und Gesetzestexte): 204 Millionen Token
German colossal, cleaned Common Crawl corpus	Textdaten eines allgemein Webcrawls (bereinigt und von guter Qualität): HEAD: Besteht aus qualitativ hochwertigem Text (z. B. Zeitungen, Regierungswebseiten) = 181 GB MIDDLE: Mehr Umgangssprache wie Foreneinträge, Kommentare = 273 GB
DWDS-Korpora	Für die Recherche im DWDS: Über 57 Mrd. Token in historischen und gegenwartssprachlichen Textkorpora verfügbar
DeReKo – Deutscher Referenzkorpus	Nach eigenen Angaben weltweit größte Sammlung deutschsprachiger Korpora für linguistische Forschung: 55 Milliarden Wörter (Stand 08.03.2023)
Datensatz aus Parlamentsmaterialien	Debatten, Anfragen, Protokolle etc.: 131.835 Dokumente
German Wikipedia Data	Ca. 6,1 GB
Open Legal Data	57.193 Gesetzestexte, 251.037 Gerichtsentscheidungen
Clarin-d	Übersicht über weitere Korpora

Quelle: Eigene Zusammenstellung.
*Auf Common Crawl beruhende Datensätze beruhen auf Texten des gesamten Internets. Auch urheberrechtlich geschütztes Material ist darin enthalten. Die Verbreitung des Datensatzes findet unter Verweis auf das US-amerikanische Prinzip des Fair Use statt.

3.3 Komponente: Grafikprozessoren

Abhängigkeiten im Bereich der Halbleiterfertigung waren durch den Chip-Mangel Anfang der 2020er Jahre besonders spürbar, als Konsumenten bei Produkten wie Smartphones und Kühlschränken mit langen Wartezeiten rechnen mussten. Die Europäische Kommission schlug daher 2022 den EU Chips Act vor, über den im April 2023 eine vorläufige Einigung zwischen EU-Rat und Parlament erzielt wurde. Damit sollen öffentliche und private Anreize für Investitionen in den Ausbau von Produktionskapazitäten geschaffen werden. Ziel ist es, den europäischen Anteil am Weltmarkt von 10 Prozent auf 20 Prozent bis 2030 zu steigern. In Deutschland gibt es verschiedene Bestrebungen, durch Fördermaßnahmen Chip-Fabriken anzusiedeln bzw. auszubauen, wie etwa in Magdeburg (Intel) oder in Dresden (Infineon, TMSC). Abhängigkeiten gehen jedoch nicht nur in eine Richtung. Denn Chip-Hersteller selbst sind auf hochspezialisierte und komplexe Maschinen angewiesen, bei denen der niederländische Hersteller ASML weltweit führend ist. Dieser steht dabei in Kooperation mit deutschen Firmen wie Zeiss und Trumpf.

Chips sind allerdings nicht nur für Konsumgüter bedeutend, sondern vor allem auch für eine Recheninfrastruktur, die für die Entwicklung und Nutzung großer KI-Modelle genutzt wird. Für die Berechnung von KI-Modellen werden derzeit vor allem Grafikprozessoren als KI-Beschleuniger eingesetzt (Graphical Processing Units, kurz: GPU, siehe hierzu auch Abbildung 4). Angetrieben durch den Wettbewerb um die Entwicklung der besten Sprachmodelle und multi-modaler Modelle, steigt die Nachfrage nach Grafikprozessoren kräftig an. Dies zeigt sich nicht zuletzt am stark steigenden Aktienkurs des Chip-Herstellers Nvidia seit Anfang 2023. Es wird geschätzt, dass der globale Markt für KI-Chips 2033 ein Volumen von 256,7 Milliarden Dollar erreichen wird, was einer jährlichen Wachstumsrate von 24,4 Prozent entspräche (IDTechEx, 2023). Künftig könnten vermehrt Chips eingesetzt werden, die für sogenannte Transformer-Architekturen des maschinellen Lernens optimiert sind. Solche Transformer sind ein grundlegendes Prinzip aktueller, generativer KI-Modelle (siehe hierzu bei Löser & Tresp, 2023). Darüber hinaus wird an der Entwicklung von Chip-Typen geforscht, die auch große KI-Algorithmen auf kleinen Geräten prozessieren können (Whitten, 2022).

3.4 Recheninfrastruktur

Verschiedene Initiativen auf europäischer Ebene und in Deutschland tragen zur Digitalen Souveränität im Bereich der Recheninfrastruktur bei. Mit der Initiative „Gemeinsames Unternehmen für europäisches Hochleistungsrechnen“ (GU EuroHPC) wird seit 2018 in einer öffentlich-privaten Partnerschaft der Europäischen Union und der beteiligten Staaten mit Industrieverbänden ein europäisches Ökosystem für High Performance Computing vorangetrieben. Acht Hochleistungsrechner sind bereits in Betrieb, sechs davon sind unter anderem für den Bereich KI ausgewiesen (siehe Tabelle 2). Im Rahmen dieser Initiative wird Ende 2024 der erste europäische Exascale-Rechner „JUPITER“ am Forschungszentrum Jülich in Betrieb gehen (Exascale bedeutet: 1 Trillion Rechenoperationen pro Sekunde – eine „1“ mit 18 Nullen). Der Rechner LUMI in Finnland wird beispielweise genutzt, um ein großes Open-Source-Sprachmodell zu trainieren. Das verantwortliche Konsortium wird von dem privaten KI-Labor Silo.AI angeführt. Der Zugang zu den HPC-Rechnern für private Initiativen, etwa von Start-ups, soll künftig ausgeweitet werden (Lomas, 2023).

Tabelle 2: **Übersicht der Supercomputer des European High Performance Computing**

Hochleistungsrechner	Standort
JUPITER	Jülich, Deutschland (ab 2024)
LUMI	Kajaani, Finnland
LEONARDO	Bologna, Italien
MARENOSTRUM 5	Barcelona, Spanien
VEGA	Maribor, Slowenien
MELUXINA	Bissen, Luxemburg
KAROLINA	Ostrava, Tschechische Republik
DISCOVERER	Sofia, Bulgarien
DEUCALION	Guimarães, Portugal

Quelle: Eigene Zusammenstellung basierend auf EuroHPC (2023).

Von den 500 leistungsstärksten Computern der Welt sind 36 in Deutschland in Betrieb (zum Vergleich: USA: 150, China: 134, Frankreich: 24, UK: 14, siehe: Top500.org, 2023). Drei davon sind unter dem Dach des „Gauss Centre for Supercomputing“ (GCS) vereint, das über langjährige Erfahrung in der Zusammenarbeit mit der Wirtschaft verfügt (BMBF, 2021).

Hochleistungsrechner – Gauss Centre for Supercomputing:

- **HLRS:** Höchstleistungsrechenzentrum Stuttgart der Universität Stuttgart mit besonderer Ausrichtung auf die Zusammenarbeit mit der Industrie.

- **LRZ:** Das Leibniz-Rechenzentrum der Bayerischen Akademie der Wissenschaften in Garching bei München wurde 2022 mit einem speziell für KI-Methoden und neuronale Netze ausgelegten Chip ausgestattet (siehe CS-2-System), der bereits für Forschungsaufgaben eingesetzt wird, um den Fokus auf KI weiter zu vertiefen.

- **JSC:** Am Supercomputing Centre am Forschungszentrum Jülich ist der Rechner „JUWELS", ausgestattet mit einem Booster-Modul, in Betrieb, der bereits große KI-Modelle berechnet und auf dem auch KI-Beschleuniger getestet werden, z. B. im Rahmen des Projekts OpenGPT-X.

Ferner wird über vier KI-Servicezentren der Zugang zu KI-Recheninfrastruktur erleichtert (BMBF, 2022). So soll sowohl der Forschung als auch den Unternehmen, insbesondere den KMU, KI-Rechenleistung zur Verfügung gestellt werden. Die Servicezentren dienen als Angelpunkte für Innovationsökosysteme, um im Tandem mit der Forschung KI-Lösungen zu entwickeln.

Schließlich tragen auch aus privaten Investitionen hervorgegangene Rechenzentren zur Digitalen Souveränität bei. Das in Bayern ansässige Rechenzentrum alpha ONE gilt als leistungsstärkster kommerzieller Rechner in Europa und wird von der Firma Aleph Alpha vorangetrieben (Hahn, 2022). Solche kommerziellen Rechenzentren werden in Europa benötigt, um Sprachmodelle im kommerziellen Betrieb kontinuierlich betreiben und damit Alternativen zu außereuropäischen Angeboten nutzen zu können. Vor dem Hintergrund, dass 74 Prozent der Unternehmen externe Ressourcen vollumfänglich oder zumindest teilweise für ihre Kernapplikationen nutzen (IDC, 2022), wird deutlich, dass solche kommerziellen Rechenzentren in Deutschland und Europa besonders wertvoll sind, um Unternehmen Alternativen zu außereuropäischen Anbietern offerieren zu können, gerade auch bei Anwendungen großer Sprachmodelle. Allerdings bauen deutsche Unternehmen auch selbst Rechenkapazitäten aus: „25 Prozent bauen neue Rechenzentren, weitere 35 Prozent investieren in die umfassende Modernisierung ihrer Data Center" (ebenda). Solche eigenständigen Ressourcen können essenziell sein, damit künftig Unternehmen kleinere Sprachmodelle auch lokal betreiben können (siehe Option 2 in Abschnitt 3.5).

> *Wir leiden unter zu viel Regulierungen, vor allem bei der Nutzung von Rechner-Infrastruktur. Wir brauchen eine sinnvoll nutzbare, Datenschutz-konforme Cloud, der Unternehmen trauen und die als Ökosystem für einfach nutzbare Software-as-a-Service (NLP)-Lösungen von Unternehmen dient. (2022)*
>
> Timo Möller, deepset GmbH

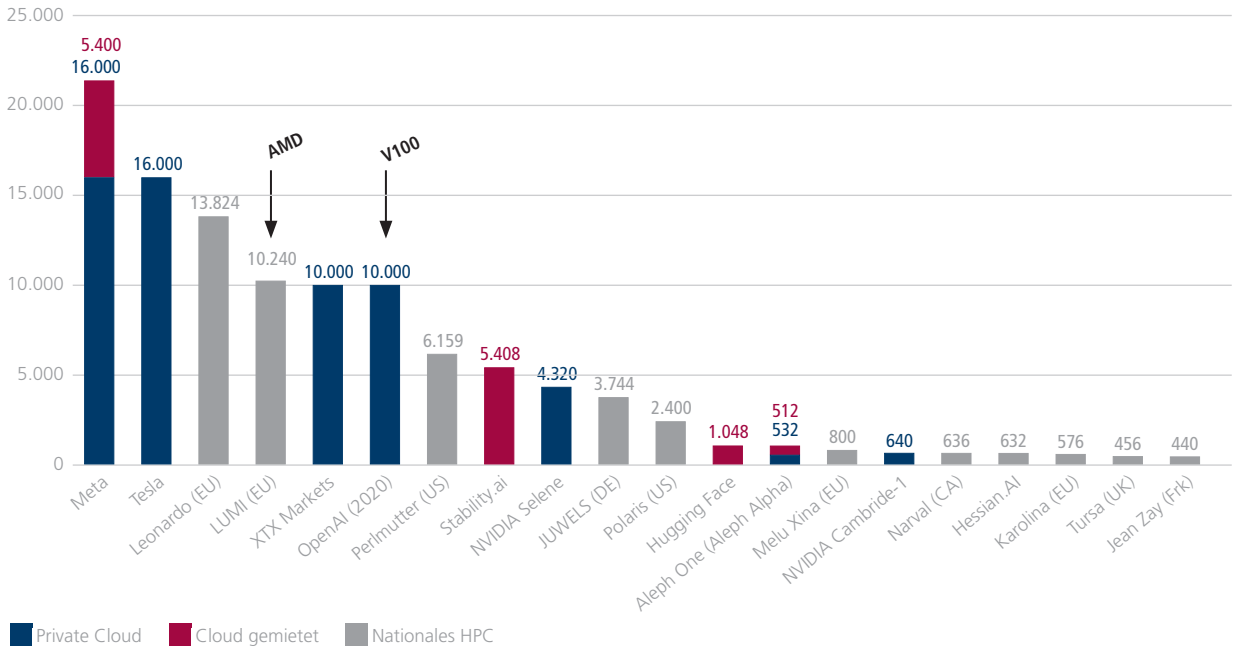
Da unter anderem GPUs für die schnelle Berechnung von KI-Modellen besonders relevant sind, lohnt sich ein Blick auf die Übersicht über die Anzahl der GPUs in Hochleistungsrechnern nach Unternehmen und Ländern (siehe Abbildung 4). Die Übersicht zeigt sechs Rechner, die in der Europäischen Union betrieben werden. Drei

davon stammen aus der EuroHPC-Initiative, darunter, an zweiter Stelle, der Rechner „Leonardo“ sowie die Rechner „MeLuxIna“ und „Karolina“. Im Mittelfeld ist der Rechner „JUWELS“ des Forschungszentrums Jülich vertreten und am flachen Ende der Verteilung befinden sich die deutsche Privatinitiative aleph ONE sowie ein nationaler Höchstleistungsrechner aus Frankreich.

Im Vergleich zu den privaten Unternehmen in der Übersicht ist die Anzahl der GPUs in nationalen Hochleistungsrechnern jedoch in vielen Fällen deutlich geringer. Hinzu kommt, dass die Rechner in Deutschland und Europa auch für andere Zwecke als nur für KI-Einsätze genutzt werden und somit ein Wettbewerb um die Rechenressourcen besteht, anders als beispielsweise bei der auf KI fokussierten Firma OpenAI. Insgesamt zeigt sich, dass Deutschland und Europa über eine Recheninfrastruktur verfügen, die für die Berechnung großer Sprachmodelle genutzt werden kann und im Fall von OpenGPT-X auch bereits genutzt wird. Die Initiative LEAM des KI-Bundesverbandes sieht in Deutschland allerdings eine Lücke bei den Rechenkapazitäten für große KI-Modelle und plädiert für den Aufbau einer ausschließlich auf die Anforderungen von KI ausgerichteten Recheninfrastruktur, die es so in Deutschland noch nicht gäbe (KI-Bundesverband, 2023).

Dass es auch in Europa möglich ist, große offene Sprachmodelle zu bauen, hat die Initiative Big Science gezeigt, die mit dem Modell „Bloom“ ein ähnlich großes Modell wie GPT-3 von OpenAI erstellt hat und dafür 117 Tage Rechenzeit auf dem Rechner „Jean Zay“ in Frankreich benötigte (vgl. Infobox, Seite 29). Der Rechenaufwand, die Anzahl der Modellparameter sowie die zu verarbeitenden Daten steigen jedoch bei großen KI-Modellen in den letzten Jahren stark an (Epoch.AI, 2023). Dieser Trend dürfte sich in Zukunft insbesondere aufgrund der multi-modalen Modelle fortsetzen.

Abbildung 4: **Anzahl von A100 Grafikprozessoren in Hochleistungsrechnern verschiedener Firmen und Länder**



Quelle: Eigene Zusammenstellung basierend auf Benaich & Hogarth (09.2023), Wiggers (2022) sowie LUMI (2021). Cloud (gemietet) = von Hyperscalern gemietete Kapazität; Private Cloud = im Besitz/Betrieb des Unternehmens; National HPC = im Besitz/Betrieb der öffentlichen Hand. Alle Angaben beziehen sich auf A100 GPUs, außer diejenigen zu OpenAI, die sich auf V100 GPUs beziehen, wie auch die GPUs von LUMI auf AMD-Prozessoren. Google hat im Mai 2023 zudem einen auf KI spezialisierten Supercomputer mit 26.000 H100 GPUs angekündigt (Miller 2023).

3.5 Cloud-basierte und lokal ausführbare Modelle

Auf Modellebene berührt der Zugang zu diesen Modellen für Forschung wie Unternehmen die Digitale Souveränität. Dieser Zugang eröffnet sich den verschiedenen Akteuren vor allem über drei Optionen (siehe hierzu auch Goel, 2023):

- Option 1: Trainieren eines eigenen Modells
- Option 2: Nutzung von APIs großer KI-Unternehmen (OpenAI, Cohere etc.)
- Option 3: Anpassung von Open-Source-Modellen aus frei zugänglichen Sammlungen wie Hugging Face (siehe zu verschiedenen Möglichkeiten der Modellanpassung Löser & Tresp, 2023, S. 14)

Option 1 ist zwar im Sinne Digitaler Souveränität eine gangbare Lösung; sie wird jedoch nur wenigen Akteuren offenstehen, die über genügend Know-how, Fachkräfte und finanzielle Ressourcen verfügen. Option 2 kann grundsätzlich dazu führen, dass weite Teile der KI-Forschung und -Anwendung von großen Sprachmodellen im Sinne von Basismodellen (engl. Foundation Models) vereinnahmt werden könnten, das heißt, dass künftig vermehrt auf Lösungen gesetzt wird, die ausgehend von einigen wenigen sehr großen und fortgeschrittenen Modellen – im Sinne eines „Foundation Model plus X" – entwickelt werden. Da es sich bei den Anbietern der Basismodelle meist um große, außereuropäische Unternehmen und Einrichtungen handelt[5], ist nicht garantiert, dass bei der Entwicklung und Umsetzung solcher Modelle europäisches Recht und europäische Werte berücksichtigt werden (siehe Abschnitt 3.1). Kostenpflichtige API-basierte Zugänge zu großen Sprachmodellen könnten zudem die Eintrittsschwelle für deutsche Unternehmen erhöhen, sodass es sich gegebenenfalls nicht mehr lohnt, bestimmte innovative Ideen zu verfolgen. Die Gründe hierfür können neben der Kostenschwelle auch darin liegen, dass die Zugänge zu Sprachmodellen weitgehend an Großkunden ausgerichtet würden oder Deutsch als weniger weit verbreitete Sprache unter Umständen weniger prioritär behandelt würde, wenn neue Funktionen bei Sprachmodellen eingeführt werden. Neben APIs bieten Unternehmen wie OpenAI auch Modellanpassungen für Kundinnen und Kunden an, wofür dem Unternehmen allerdings Domänen oder aufgabenspezifische Daten zur Verfügung gestellt werden müssen. Auch hier stehen europäische Unternehmen vor der Frage, ob sie eigene Daten teilen möchten oder können. Initiativen bzw. europäische Start-ups wie Aleph Alpha, Mistral.AI oder nyonic und Silo.AI könnten hierbei sicherlich eine Alternative darstellen, die einige der genannten Herausforderungen adressieren und so zur Digitalen Souveränität Europas beitragen. Auch auf angepassten Open-Source-Modellen beruhende und lokal betriebene Modelle können hier einen Beitrag leisten.

Open-Source-Software kann grundsätzlich dabei helfen, Abhängigkeiten abzubauen, und ist daher unterstützenswert. Seit 2016 geht mit zunehmender Tendenz die Mehrheit an bedeutenden Sprachmodellen und multi-modalen Modellen aus der Industrie hervor und nicht etwa aus Hochschulen, Kooperationen oder Forschungskollektiven (Epoch.AI, 2023). Entsprechend stellt sich auch die Verteilung von offen zugänglichen Sprach- und multi-modalen Modellen zu nicht offen zugänglichen Modellen dar, auch wenn große Unternehmen sich zum Teil ebenfalls im Bereich Open Source engagieren (siehe Abbildung 5). Einer Angebotskonzentration bei den Anbietern großer Sprachmodelle wirkt seit 2021 eine Bewegung hin zu Open-Source-Communities und dezentralen Forschungsgemeinschaften entgegen. Modelle aus der Open-Source-Community weisen jedoch häufig auch kleinere Parameterzahlen auf (siehe Abbildung 5 sowie Benaich & Hogarth, 2022, S. 34 ff. & 84; Solaiman, 2023). Wie schon im vorangegangenen Abschnitt 3.2 betont, bedeutet dies

5 Laut AI Index stammten 54 Prozent der Forschenden im Jahr 2022, die an der Veröffentlichung neuer großer Sprachmodelle und multi-modaler Modelle beteiligt waren, aus US-Institutionen und 21,88 Prozent aus dem Vereinigten Königreich. 2021 hatte China noch einen Anteil von 27,65 Prozent verglichen mit 8 Prozent im Jahr 2022 (siehe Zhang et al., 2023, S. 39). Ein chinesischer Bericht sieht China im Jahr 2021 bei der Zahl veröffentlichter Modelle mit den USA gleichauf und 2022 lediglich um neun Modelle hinter den USA (Reuters, 2023). Es zeigt sich eine Diskrepanz für das Jahr 2022 im Verhältnis zum AI Index.

jedoch nicht, dass solche kleineren Modelle nur eingeschränkt leistungsfähig sind. Werden diese mit größeren Mengen an qualitativ hochwertigen Daten pro Parameter trainiert, zeigen diese Modelle ebenfalls umfangreiche Fähigkeiten, gerade auch dann, wenn sie mit domänen-spezifischen Daten für bestimmte Aufgaben angepasst werden. Es ist festzustellen, dass einige der bedeutenden Sprachmodelle hinsichtlich der Parameterzahl eher rückläufig sind (siehe LLaMA, PALM 2) und sich der Wettbewerb zu immer größeren Modellen etwas abgekühlt hat (siehe Abbildung 5).

> *Retraining war für uns bisher nicht relevant, da es teuer ist, große Sprachmodelle von Grund auf zu trainieren, und darüber hinaus eine Menge Daten erfordert. Plattformen wie Hugging Face bieten außerdem eine große Auswahl an vortrainierten Sprachmodellen in vielen Sprachen und verschiedenen Domänen an, die komplett kostenlos sind – warum diese also nicht nutzen? Bei einem kleineren Datensatz mit Labels ist es in der Regel am sinnvollsten, ein vortrainiertes Modell zu wählen, das von der Domäne so gut wie möglich passt, und es dann auf der Basis des Datensatzes zu feinzutunen. (2022)*[6]
>
> Johannes Otterbach, ehemals Merantix, seit Juli 2023 nyonic

Diese Entwicklung bildet die Grundlage für Option 3, also die Anpassung von Modellen, die Open Source zur Verfügung stehen und an spezifische Domänen und Aufgaben sowie an die rechtlichen Bedürfnisse in Deutschland und Europa angepasst werden können. Diese Variante gilt zugleich als relativ kostengünstig, bietet Transparenz sowie den Vorteil, dass in Kooperation mit der Community und der Forschung an der (Weiter-)Entwicklung von offenen Modellen sowie an Lösungen für Herausforderungen gearbeitet werden kann. Ein Beispiel stellt die Firma Hugging Face dar, die neben privaten Aktivitäten vor allem eine offene Community-Plattform für den Austausch von Modellen und Daten aufgebaut hat. Mit Unterstützung französischer Forschungseinrichtungen und vielen Freiwilligen hat Hugging Face zudem das offene, multi-linguale Modell Bloom trainiert. Rund um den „Leak" des LlaMA-Modells von Meta wurden zudem diverse offene Modellvarianten entwickelt. In Deutschland wird beispielsweise mit OpenGPT-X basierend auf der GAIA-X-Infrastruktur ebenfalls ein offenes Modell entwickelt. Auch wenn die Open-Source-Entwicklung für die Digitale Souveränität grundsätzlich förderlich ist, kann diese auch zu eher abträglichen Dynamiken führen, wenn diese Entwicklung durch große, außereuropäische Unternehmen getrieben wird. Denn in diesem Fall kann das Unternehmen die „Entwicklergemeinde durch solche Angebote näher an sich und ihr Ökosystem binden und [erhält damit bereits] früh Einblicke in Trends und Anwendungsgebiete" (Kagermann et al., 2021). So kann das jeweilige Unternehmen Entwicklungen und Lösungswege der Community aufgreifen und dann closed source weiterentwickeln, um eigene Produkte zu verbessern und somit die Marktstellung zu festigen.

Vor dem Hintergrund der skizzierten Entwicklung zeichnen sich drei Tendenzen als wahrscheinlich ab:

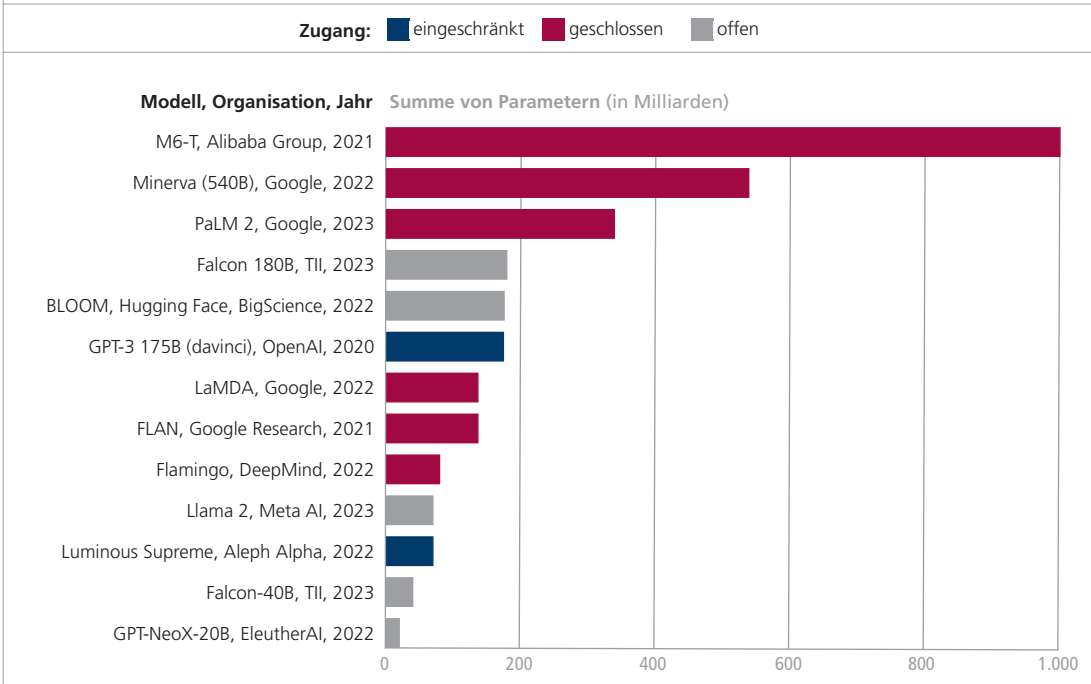
1. Große Technologieunternehmen (OpenAI, Microsoft etc.) werden weiterhin große Basismodelle in Maßstäben veröffentlichen, die viele andere Akteure nicht erreichen können. Viele Unternehmen werden sich bei diesen Modellen einmieten, die Kostenschwelle in Kauf nehmen und unverfängliche Daten an den Anbieter übermitteln. Europäische Start-ups könnten hierbei jedoch eine Alternative darstellen. Vor allem auch deshalb, weil es mittlerweile möglich ist, große Basismodelle zu moderaten Kosten aufzubauen.

6 Retraining bedeutet, dass ein bestehendes Modell von Grund auf neu trainiert wird, etwa, weil mit neuen Trainingsdaten die Leistungsfähigkeit des Modells in einer veränderten Modellumgebung erhalten oder verbessert werden soll.

2. Daneben wird es weiterhin zahlreiche Open-Source-Modelle geben, die gegebenenfalls Spezialaufgaben (wie StarCoder in der Programmierunterstützung) wahrnehmen, für bestimmte Sprachen, wie Deutsch, optimiert sind und an rechtliche, unternehmerische und andere Bedürfnisse in Deutschland und Europa angepasst sind. Diese Modelle könnten dann sogar klein genug sein, um lokale große Sprachmodelle zu ermöglichen (local LLM); das heißt, Unternehmen könnten diese dann hinter ihrer Firewall betreiben. Da nur wenige Unternehmen bereit sein werden, sensible Programmcodes oder Daten, wie etwa rechtliche Dokumente und Verträge oder medizinische Dokumente, an externe Unternehmen zu senden oder gegebenenfalls sogar rechtliche Bestimmungen diesem entgegenstehen, ist diese Option eine realistische Alternative.

3. Die beiden wirtschaftlichen Tendenzen werden auch die Forschung anregen sowie die KI-Ökosysteme inklusive der Start-ups. Vor allem die Open-Source-Entwicklung und das Thema Modellanpassung im Sinne der Option 2 wird auch ein bedeutendes Thema für die Forschung und Entwicklung darstellen, zum Beispiel hinsichtlich kostengünstiger und einfach anwendbarer Anpassungswerkzeuge (siehe hierzu auch Löser & Tresp, 2023, S. 28).

Abbildung 5: **Bedeutende Sprachmodelle sowie multi-modale Modelle nach Parameterzahl und Grad des Zugangs**



Quelle: Für das Diagramm wurden die 15 größten Modelle ausgewählt, die bei Epoch.AI (2023) als bedeutende Modelle gelistet sind und für die Informationen zur Zugänglichkeit verfügbar waren. Mit „Bloom" wurde ein offenes Modell aufgenommen, das durch eine internationale Community vorangetrieben und mit Rechenkapazität von Frankreich unterstützt wurde. Mit dem Modell „Luminous Supreme" wurde ein kommerzielles Modell der Firma Aleph Alpha aus Deutschland hinzugefügt. Für eine graduell abgestufte Analyse von KI-Modellen zwischen geschlossenen und offenen Ansätzen der Zugänglichkeit siehe Solaiman (2023). Für Übersichten zur Modellzugänglichkeit siehe etwa CRFM (2023) oder das Open LLM Leadershipboard (Hugging Face, 2023). Neben dem Modell „M6-T" existieren weitere Modelle in der Größenordnung von Trillionen Parametern aus den USA und China (siehe z. B. „Switch Transformer", „Wu Dao 2.0"), die allerdings bei Sevilla et al. (2022) nicht die Kriterien für die Einbeziehung in die Analyse erfüllten. GPT-4 ist nicht offen zugänglich; da über die Parameterzahl [Stand: September 2023] keine gesicherten Informationen zur Verfügung stehen, wurde es in die Übersicht nicht aufgenommen.

> *Stellen Sie sich eine Zukunft vor, in der vielleicht in 20 Jahren oder vielleicht noch länger jede einzelne unserer Interaktionen mit der digitalen Welt durch ein KI-System vermittelt wird [...] dann werden diese Systeme zum Aufbewahrungsort allen menschlichen Wissens werden, und es ist sehr wichtig, dass zumindest die Basis dafür Open Source ist. [...] Wenn alle unsere Informationen über alle Bürgerinnen und Bürger im Grunde durch diese KI-Systeme gefiltert werden, muss die Art und Weise, wie diese Systeme trainiert werden, crowd sourced sein, ähnlich wie bei Wikipedia, um kulturelle Informationen und Wissen aus der ganzen Welt zu sammeln, nicht nur aus der Weltsicht in Palo Alto oder anderenorts. Deshalb bin ich ein großer Befürworter von Open-Source-Basismodellen für KI [...], weil sie dadurch sicherer und leistungsfähiger werden. Sie entwickeln sich schneller weiter. Sie sind kulturell vielfältiger, wenn mehr Menschen sie trainieren können, und es entsteht ein ganzes Ökosystem von Start-ups und Forschungsprojekten, die darauf aufbauen können.*
>
> Yann LeCun, Chief AI Scientist, Meta AI Research (München, 2023) [Übersetzung d. Red.]

3.6 Talente

Aus technischer Sicht sind für den Aufbau von Sprachmodellen die Zutaten Daten, Rechenleistung und geeignete Algorithmen sehr wichtig. Jedoch kann das Potenzial von großen KI-Modellen ohne eine Community mit entsprechendem Wissen und Know-how nicht gehoben werden. Sprachmodelle, die aus einer Community heraus entstanden sind, wie Bloom, aber auch Organisationen wie OpenAI, die geschickt essenzielle Talente gebunden haben, zeigen, wie bedeutend es für den erfolgreichen Aufbau von großen Sprachmodellen ist, eine kritische Masse an Talenten aufzubauen, zu vernetzen und zu koordinieren. Zwei Communities sind für Deutschland und Europa besonders relevant, um das Potenzial von Sprachmodellen realisieren zu können:

1. Technische Community mit einem Hintergrund in der Verarbeitung natürlicher Sprache und im maschinellen Lernen, da Communities zu diesen beiden Bereichen sowie zu Data Science noch zu oft getrennt sind.

2. Community aus Unternehmen, Beratungsfirmen und Hochschulen, die auf die Anpassung von großen Sprachmodellen fokussiert ist.

Für beide Community-Typen werden entsprechende KI-Talente benötigt. Daher wird im Folgenden die allgemeine Lage zu KI-Talenten in Deutschland betrachtet und im Anschluss auf die Situation bei den Talenten für die Entwicklung von Sprachmodellen sowie auf notwendige Kenntnisse und Fähigkeiten eingegangen.

Allgemeine Situation zu KI-Talenten

Nach einer Analyse des Risikokapitalgebers Sequoia verfügt Europa über einen Talentpool von 200.000 Ingenieurinnen und Ingenieuren, die Erfahrung mit KI haben, und einen Kernpool von 43.000 Personen, die sich

stärker zu KI engagieren (Sequoia, 2023a; 2023b). Pro Kopf ist, der Studie zufolge, der Pool an Talenten in Europa größer als in den USA und dreimal so groß wie in China. Die Universität Edinburgh trägt in der Ausbildung dieser Talente am meisten bei, aber auch die Technische Universität München wird, neben den Universitäten Amsterdam und Madrid, als eine der bedeutendsten Ausbildungsstätten genannt. Der größte Anteil am allgemeinen KI-Talentpool in Europa findet sich mit Abstand in London (12,29 Prozent), Deutschland ist mit Berlin (2,65 Prozent) an vierter Stelle gleich nach Paris und Zürich. Mit München (Platz 7) und Stuttgart (Platz 14) ist Deutschland zudem mit zwei weiteren Städten im Ranking vertreten. Auch in anderen Studien finden sich positive Indizien für die allgemeine Lage zu KI-Talenten. Die OECD-Studie zeigt anhand von LinkedIn-Daten, dass Arbeitnehmende in Deutschland 1,7-mal häufiger KI-relevante Fähigkeiten angeben als der Durchschnitt der G20-Staaten in den Jahren 2015 bis 2022 (OECD, 2023b). Nur in den USA und in Indien ist dies noch ausgeprägter. Im Ranking des Medienunternehmens Tortoise befindet sich Deutschland auf Platz 3 in der Kategorie Talente – ebenfalls hinter Indien und den USA (Cesareo & White, 2023). Die Kategorie basiert auf 15 Indikatoren, unter anderem auf der Anzahl der KI- und Data Science-Fachkräfte und der Absolvierenden von MINT-Fächern und Computerwissenschaften, aber beispielsweise auch auf der Anzahl an Antworten auf KI-relevante Fragestellungen auf der Internetplattform Stackoverflow. Insgesamt ergibt sich so im Allgemeinen ein vergleichsweise positives Bild für Europa und Deutschland beim Faktor KI-Talente und damit eine Chance, die im KI-Wettbewerb verstärkt ausgeschöpft werden kann. Dennoch zeigt sich in der Praxis der KI-Start-ups, dass viele Stellen unbesetzt bleiben (35 Prozent im Frühjahr 2023 oder im Laufe des Jahres 2022, siehe BMWK, 2023b).

Bei KI-Toptalenten zeichnet eine auf LinkedIn-Daten basierende Studie der Stiftung Neue Verantwortung jedoch auch ein anderes Bild (Maham et al., 2022). Der Studie zufolge sind 63 Prozent der Top-Promovierenden nach ihrer Promotion in Deutschland beschäftigt. Davon sind gut 30 Prozent in der Forschung geblieben und fast 70 Prozent haben eine Stelle in der Wirtschaft angetreten. Die Top fünf der Arbeitgeber promovierter KI-Talente in Deutschland sind Amazon, Bosch, Siemens und die Technischen Universitäten in München sowie Berlin. 80 Prozent der Top-Doktorandinnen und -Doktoranden, die Deutschland nach ihrer Promotion in Richtung USA, Großbritannien oder Schweiz verlassen haben, arbeiten in Unternehmen. Zu den Top fünf der Arbeitgeber in den USA gehören Meta, Amazon, Apple, Google und die Stanford University. Die Studie kommt zu dem Schluss, dass „deutsche Universitäten und Forschungsinstitutionen […] ein wichtiger Teil des KI-Talentpools [sind], in dem die großen Tech-Firmen fischen“ (Maham et al., 2022). Obwohl Deutschland junge Forscherinnen und Forscher aus Asien und Osteuropa anzieht und ausbildet, gehen jedoch viele dieser vielversprechenden Talente in die weltweit führenden KI-Standorte.

Talente für die Entwicklung von Sprachmodellen

Die Arbeit an Sprachmodellen in Unternehmen in Deutschland ist komplex: Von den Mitarbeitenden wird einerseits ein Domänen- und Kundenverständnis verlangt, zum Beispiel beim Erkennen von Problemen und Geschäftsmodellen beim Kunden, und andererseits auch die Übersetzungsleistung in Verfahren des maschinellen Lernens sowie eine Abschätzung der Machbarkeit und Kosten. Schließlich wird eine möglichst kostengünstige Umsetzung erwartet sowie geringe Wartungskosten.

Diese Tätigkeiten erfordern fast immer einen Masterabschluss oder oft sogar eine einschlägige Promotion in den Gebieten Natürliche Sprachverarbeitung, Data Engineering, Verteilte Systeme und maschinelles Lernen. Diese Gebiete sind wie viele Disziplinen der Informatik, aber auch anderer MINT-Fachgebiete, sehr komplex und anspruchsvoll: Doktorandinnen und Doktoranden müssen ihre Leistung auf internationalen Tagungen darstellen; die Ausbildung beträgt im Schnitt vier bis fünf Jahre.

Tabelle 3: **Bedeutende Kenntnisse und Fähigkeiten zu Entwicklung und Umsetzung großer Sprachmodelle**

Basis	• Allgemeine Grundkenntnisse in der Verarbeitung natürlicher Sprache (Natural Language Processing, kurz NLP) • Erfahrung in der Programmiersprache Python und entsprechenden Programmbibliotheken für NLP und maschinelles Lernen (Maschine Learning, kurz ML) etc. • Kenntnisse zu gängigen Bibliotheken für Open-Source-Modelle und Trainingsdaten (z. B. Hugging Face usw.) • ML-Grundkenntnisse, speziell für große Sprachmodelle, insbesondere zu Transformern • Kenntnisse in Software-Engineering, Backend-Engineering bzw. Service-basierten Infrastrukturen bei Nutzung externer Dienstleister zur Anpassung großer Sprachmodelle • Kenntnisse gängiger Evaluierungsmethoden und Benchmarks (z. B. HELM, MTEB, BEIR, GLUE, SUPERGLUE etc.) und der qualitativen und quantitativen Fehleranalyse • Kenntnisse in der Anpassung von großen Sprachmodellen, insbesondere im Transfer Learning • Kenntnisse in der Anreicherung großer Sprachmodelle etwa durch komplementäre multi-modale, ontologische (symbolische) oder multi-linguale Daten
Zusatz	• Im Rahmen der Anpassung von auto-regressiven, großen Sprachmodellen wie ChatGPT: Kenntnisse im Prompt Engineering, z. B. via Transfer Learning von existierenden GPT-Modellen, sowie Grundkenntnisse des Instruct-GPT-Ansatzes • Grundkenntnisse zu DevOps und MLOps, wenn kleinere Modelle (100 Millionen bis einige Milliarden Parameter) lokal trainiert werden sollen • Vertiefte Kenntnisse von Frameworks für das verteilte Berechnen des Stochastic Gradient Descent sowie vertiefte DevOps- und ML MLOps-Kenntnisse für verteilte Systeme, wenn größere Modelle (100 Milliarden Parameter) lokal trainiert werden sollen • Erfahrung mit Visual-Language-Modellen sowie multi-modalen Modellen im Allgemeinen • Im Unternehmenskontext: Übersetzungsleistung von der jeweiligen Domäne auf Aufgaben der Sprachmodelle bzw. Anpassungen, Abschätzungsleistung der Kosten sowie „Sprechen" der Produkt- und Technologiesprache zur Vermittlung in interdisziplinären Teams

Quelle: Eigene Zusammenstellung. Machine Learning Operations (MLOps) bezieht die Development Operations (DevOps) auf das maschinelle Lernen, das heißt, es geht hierbei darum, die Bereiche des maschinellen Lernens, der Softwareentwicklung und des laufenden Betriebs zusammenzubringen. Insbesondere wird eine effiziente, zuverlässige und qualitativ hochwertige Gestaltung der Entwicklung, Bereitstellung, Verwaltung und Überwachung von KI-Modellen angestrebt.

Aufbauend auf der Übersicht in Tabelle 4 über Hochschulgruppen, Forschungseinrichtungen und -institute existieren annäherungsweise 30 bis 40 Gruppen an der Schnittstelle zwischen natürlicher Sprachverarbeitung und maschinellem Lernen. Das bedeutet, dass circa 60 bis 80 Promovierende pro Jahr mit Schwerpunkt auf Sprachmodelle und ingenieurwissenschaftlicher Ausbildung für den deutschen Markt bereitstehen: Davon werden nicht alle in der Wirtschaft arbeiten wollen, ein kleiner Teil wird an den Hochschulen verbleiben. Daher ist die Anzahl der verbleibenden Fachkräfte für den Bedarf der deutschen Wirtschaft deutlich zu klein. Allerdings zählen auch weitere Absolvierende aus den Bereichen des maschinellen Lernens zu diesem Talentpool. Auch Absolvierende von MINT-Fächern können durch Weiterqualifikation zu spezifischen Kenntnissen und Fähigkeiten von Sprachmodellen (siehe Tabelle 3) Teil dieses Talentpools werden.

> *Wissen und Talent sind dabei unserer Ansicht nach die raren Ressourcen. Wir schaffen unsere eigene Community bei inovex durch eine Vielzahl an Masterarbeiten, dadurch finden wir Talente. Dieser Community-Aufbau ist ein nicht-technisches Kernprodukt: die gegenseitige Befruchtung von Unternehmen und Community. (2022)*
>
> Hans-Peter Zorn, inovex GmbH

Ähnlich wie bei der allgemeinen Situation um Top-Talente im Bereich KI in Deutschland ist auch im spezifischen Fachgebiet an der Schnittstelle zwischen natürlicher Sprachverarbeitung und maschinellem Lernen davon auszugehen, dass viele Top-Talente abwandern. Zum einen dürfte dies darin begründet sein, dass nach wie vor noch zu wenige Unternehmen überhaupt erkennen, mit welchen Maßnahmen diese Fachkräfte gebunden werden können. Zum anderen, dass die spannendsten Entwicklungen im Bereich großer Sprachmodelle beispielsweise in US-amerikanischen Unternehmen und Organisationen vorangetrieben werden und Top-Talente dort an den interessantesten Anwendungen mit Zugang zu umfangreicher Ausstattung arbeiten können. Das führt dazu, dass wichtige Dienstleistungen und Sprachmodelle für deutsche Kunden von diesen Fachkräften bei Unternehmen im Ausland entwickelt werden, wie bei Deep Mind, Hugging Face oder Cohere.

Tabelle 4 bietet eine nicht abschließende Übersicht zu führenden Hochschulen und Einrichtungen in Deutschland im Bereich Sprachmodelle bzw. an der Schnittstelle zwischen der Verarbeitung natürlicher Sprache und maschinellem Lernen mit starker Informatik-Ausrichtung und eigenen Lehrstühlen bzw. größeren Gruppen.

Tabelle 4: **Hochschulen und Einrichtungen in Deutschland im Themenbereich Sprachmodelle** (nicht abschließend)

Hochschule	Zentrum / Abteilung / Gruppe / Schwerpunkt
Ludwig-Maximilians-Universität München	Center for Information and Language Processing
Universität des Saarlands	Department of Language Science and Technology
Ruprecht-Karls-Universität Heidelberg	Institute for Computational Linguistics
Universität Tübingen	Department of General and Computational Linguistics
Universität Mannheim	The Data and Web Science Group
Technische Universität Hamburg	Language Technology Group
Technische Universität Darmstadt	Ubiquitous Knowledge Processing (UKP) Lab
Berliner Hochschule für Technik	Data Science and Text-based Information Systems (DATEXIS); Data Science +X Research Center
Universität Paderborn	Data Science Group
Bauhaus-Universität Weimar	Webis Group
Universität Hannover	NLP Group
Universität Stuttgart	Institute for Natural Language Processing
Friedrich-Schiller-Universität Jena	Datenbanken und Informationssysteme

Quelle: Eigene Zusammenstellung.

Institut / Forschungszentrum	Zentrum / Abteilung / Gruppe / Schwerpunkt
Fraunhofer IAIS	Document Analytics / Natural Language Understanding
Hasso-Plattner-Institut	Information Systems Group
Max-Planck-Institut Saarbrücken	Databases and Information Systems
Forschungszentrum Jülich	Scalable Learning and Multi-Purpose AI (SLAMPAI) Lab
Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI)	Sprachtechnologie
LAMARR-Institut	Bereich: Natural Language Processing
ScaDS.AI (Center for Scalable Data Analytics and Artificial Intelligence)	Bereich: Understanding Language
Munich Center for Machine Learning (MCML)	Unter anderem Bereich B1 Computer Vision und B2 Natural Language Processing (u. a. Multi-modale Modelle, Generative KI, Visual Question Answering etc.)
Hessian.AI	Natural Language Processing, multi-modale Modelle

Quelle: Eigene Zusammenstellung.

KURZINFO

Community-getriebener Aufbau von offenen Modellen – der Fall „Bloom"

Der Fall „Bloom" zeigt anschaulich, wie offene KI-Modelle aus einer Community heraus zusammen mit einem Unternehmen wie Hugging Face vorangetrieben werden können, die auch überwiegend Kriterien erfüllen, wie sie in Europa diskutiert werden (siehe Bommasani et al., 2023). Dieser macht deutlich, wie verschiedene Rahmenbedingungen zusammenwirken: Community-Building und -Organisation, Datensammlung und -kuratierung, Recheninfrastruktur, KI-Know-how und KI-Talente greifen ineinander, um die Entwicklung allgemein und den Transfer von der Forschung zu anwendbaren Sprachmodellen voranzutreiben. Diese Organisations- und Koordinationsleistung kann auch durch andere Instanzen erbracht werden, wie etwa durch öffentlich geförderte Konsortialprojekte (siehe OpenGPT-X) oder auch durch Non-Profit-Organisationen und wissenschaftliche Einrichtungen (siehe KI-Kompetenzzentren).

Allgemeines

Bloom ist ein multi-linguales Sprachmodell, hervorgegangen aus dem Big Science Projekt. Es hat mit 176 Milliarden Parametern eine ähnliche Größe wie GPT 3 der Firma OpenAI. Eine Community aus Freiwilligen hat insgesamt ein Jahr lang an der Umsetzung des Projekts gearbeitet.

Datensatz

- Datensatz mit 341 Milliarden Token (Wörtern). Das Verhältnis von Parametern zu Token ist mit 1:2 etwas kleiner als es für Modelle dieser Größe üblich ist.
- 46 natürliche Sprachen und Dialekte, 13 Programmiersprachen, aber ohne Deutsch.
- Datenbasis: Bücher, wissenschaftliche Publikationen, Radiotransskripte, Podcasts, Webseiten.
- Fast zwei Drittel des Datensatzes wurden von Hand aus 500 Quellen ausgewählt. Bei der Auswahl wurde auf Datenschutz geachtet und nach Qualität gefiltert, um z. B. Bias zu reduzieren.

Training

- 117 Tage Trainingszeit
- Rechenzeit gespendet von französischen Einrichtungen
- 48 Knoten, je 8 NVIDIA A100 80GB GPUs (Insgesamt 384 GPUs)
- Theoretische Spitzenleistung waren 312 TeraFLOP/s, erreicht wurden 156 TeraFLOP/s
- Trainingskosten: 2,29 Millionen Dollar (zum Vergleich: Megatron-Turing NGL mit 530 Milliarden Parametern entsprach 11,35 Millionen Dollar)

Im Vergleich

- Forschende schätzen, dass ein Modell mit 175 Milliarden Parametern auf einem Datensatz mit 300 Milliarden Token auf 1024 A100 GPUs in 34 Tagen berechnet werden kann, wenn 140 TeraFLOP/s per GPU erreicht werden.
- Ein Modell mit einer Billion Parametern könnte mit einem Datensatz von 450 Billionen Token auf 3.072 A100 GPUs in 84 Tagen berechnet werden, wenn 163 TeraFLOP/s erreicht werden.

Projekt-Koordination und -Organisation

Es handelt sich um ein internationales Projekt, das hauptsächlich durch den Aufbau und die Koordination einer Gemeinschaft von mehr als 1.000 Freiwilligen vorangetrieben wurde, die jedoch nicht in Vollzeit arbeiteten (Multi-Stakeholder-Ansatz mit Ethikern, Philosophen, Juristen, Ingenieuren aus Start-ups und großen Unternehmen).

- Die Ursprünge von Big Science gehen auf Gespräche zwischen dem wissenschaftlichen Leiter der Firma Hugging Face und Vertretern französischer Einrichtungen zurück (GENCI – Grand équipement national de calcul intensif und P IDRIS – Institut du développement et des ressources en informatique scientifique).
- Es wurden Lenkungsausschüsse gebildet, die die Mitglieder von Big Science – aus über 60 Ländern und 250 Institutionen – wissenschaftlich und allgemein berieten, gemeinsame Aufgaben entwickelten sowie Workshops, Hackathons und öffentliche Veranstaltungen organisierten.
- Verschiedene Arbeitsgruppen wurden beauftragt, sich mit Herausforderungen wie Datenmanagement, Beweisen von Theoremen in der Mathematik und Archivierungsstrategien zu befassen.

Implikationen für Deutschland und Europa

Wie das Projekt zeigt, ist es möglich, in einem relativ überschaubaren Zeitraum ein großes Sprachmodell aufzubauen, das auch in Europa diskutierte Kriterien berücksichtigt. Wird ein solches Projekt angestrebt, sollten mehrere Phasen eingeplant werden:

- Der Aufbau eines deutschsprachigen Open-Source-Datensatzes von 10–15 Terabyte, der im Sinne europäischer Werte und Regeln kuratiert wird und idealerweise auch kommerziell nutzbar ist.
- Die Berechnung des Modells inkl. einer Moderationsinstanz zur Begrenzung von Bias und anderen Qualitätsmängeln, deren Daten und Modell ebenfalls Open Source sein sollten und die sich an europäischen Werten und Regeln orientieren sollte, sowie das Testen und die Evaluierung des Modells.

- Über die Modellerstellung hinaus könnten Folgeprojekte angestrebt werden. Aufbauend auf dem Open-Source-Modell können Unternehmen damit beginnen, das Modell an ihre Domänen und Aufgabenstellungen anzupassen und Anwendungen zu entwickeln, die das KI-Wissen in die Anwendung bringen.

Die Kosten für ein solches Projekt belaufen sich auf ca. 15 bis 20 Millionen Euro (Stand: September 2023). Dabei ist von einem kleineren Personenkreis als im Big Science-Projekt auszugehen, der in den Bereichen Data Science, Data Engineering, DevOps, verteilte Systeme und im Produktmanagement entsprechend hoch qualifiziert ist. Für die Erstellung des Modells sind circa 20 bis 30 Vollzeitmitarbeitende mit einem Volumen von circa 60 bis 80 Personenjahren anzusetzen. Die Kosten für die Berechnung eines deutschen Sprachmodells auf dem Niveau von ChatGPT 3.5 belaufen sich auf etwa 2 bis 3 Millionen Euro, wobei von circa 2 bis 3 Millionen Rechenstunden auf einem GPU-Cluster ausgegangen wird. Dies würde eine Rechenkapazität von 1000 GPUs (z. B. Nvidia Server) erfordern, die, wie in Abschnitt 3.4 gezeigt, in Europa verfügbar sind.

Quellen: Eigene Zusammenstellung basierend auf Narayanan et al. (2021), Scao et al. (2022), Wiggers (2022) und Zhang et al. (2023).

4 Fazit und Gestaltungsoptionen

In Anbetracht des enormen Potenzials, das große Sprachmodelle heute schon in ihrer Anwendung(-svielfalt) aufweisen und künftige durch ihre Weiterentwicklung noch bieten werden (siehe Abschnitt 2), ist es zum einen bedeutend, sich darüber Klarheit zu verschaffen, was Digitale Souveränität gerade bei dieser zentralen Technologie ausmacht und was zu solchen Ebenen Digitaler Souveränität – technologischer, talent-bezogener wie gesellschaftlich-normativer Art – beiträgt (siehe Abschnitt 3). Im Folgenden werden nach einer Zusammenfassung von Kernpunkten zu den Ebenen Digitaler Souveränität einige Gestaltungsoptionen dargelegt.

Tabelle 5: **Ebenen Digitaler Souveränität bei großen Sprachmodellen – Zusammenfassung**

Ebenen Digitaler Souveränität (DS)	Erläuterungen
Europäisches Werte- und Rechtssystem	• Der AI Act kann ein Instrument werden, um europäische Werte bei großen Sprachmodellen durchzusetzen. **(+)** • Aktuell werden bei vielen bekannten Modellen Kriterien nicht eingehalten (bzw. es bleiben Unsicherheiten). **(-)**
Daten	• Mehrere größere, deutschsprachige Textkorpora bereits vorhanden. **(+)** • Umfangreicher (10 bis 15 Terabyte), breit verfügbarer und im Sinne europäischer Werte und Regularien kuratierter, deutschsprachiger Textdatensatz fehlt. **(-)** • Ein vergleichsweise kleiner Anteil deutscher Textmengen an bestehenden, bekannten Modellen kann die Qualität von Modellausgaben in deutscher Sprache beeinträchtigen. **(-)** • Urheber- und lizenzrechtliche Herausforderungen bei Erstellung umfangreicher, breit nutzbarer Korpora für Modelltraining. **(-)**
(Grafik-) Prozessoren	• Die EU hält nur 10 Prozent Marktanteil am Chip-Markt. **(-)** • Abhängigkeit der Chip-Hersteller von komplexen Produktionsmaschinen, für die europäische Produzenten marktführend sind. **(+)** • Mittels Chips Act und Förderung der Ansiedlung von Produktionsstätten wird Abhängigkeiten generell entgegengewirkt. **(+)** • Es bleibt bei einer allgemeinen Abhängigkeit, gerade auch bei den besten und leistungsfähigsten GPUs (siehe NVIDIA). **(-)**
Recheninfrastruktur	• GU EuroHPC sowie GSC-Initiativen tragen zur DS bei, vor allem, aber nicht nur, in der Forschung. **(+)** • Vereinzelte private Initiativen tragen zu DS bei. **(+)** • 25 Prozent der Unternehmen wollen in eigene Ressourcen investieren oder ausbauen. **(+)** • Aber: Mitwachsen der Infrastruktur mit Anforderungen nötig sowie mehr europäische kommerzielle Lösungen. **(-)** • Aber: 74 Prozent der Unternehmen sind von externen Ressourcen abhängig und damit oft auch von außereuropäischen Cloud-Anbietern (IDC, 2022). **(-)**
Modelle	• Open Source mit lokalen Sprachmodellen (local LLM) kann zu DS beitragen.**(+)** • Europäische Start-ups sind eine Chance für mehr DS. **(+)** • Aber die meisten Modelle entstehen in USA und China. Große, neue und aufwendig zu trainierende Modelle werden weiterhin von großen Tech-Firmen und Einrichtungen entwickelt werden. **(-)**
Talente	• Im Vergleich mit anderen Ländern besteht in Deutschland eine gute allgemeine Situation. **(+)** • 35 Prozent von Stellen bei KI-Start-ups werden nicht besetzt (BMWK, 2023b). **(-)** • Abwanderung von Top-Talenten. **(-)** • Talente an der Schnittstelle von Verarbeitung natürlicher Sprache und maschinellem Lernen nicht ausreichend vorhanden. **(-)**

Quelle: Eigene Zusammenstellung.

Textdaten sammeln, kuratieren und als Open Source verfügbar stellen: Grundvoraussetzung für das Training großer Sprachmodelle – im Sinne deutscher und europäischer Werte und Regeln – sind entsprechend umfangreiche und kuratierte Trainingsdatensätze. Ein möglichst breit nutzbarer, idealerweise auch kommerziell verwendbarer, deutschsprachiger Datensatz im Umfang von 10 bis 15 Terabyte wäre nicht nur bereits heute für alle Seiten, seien es Forschende, Entwickelnde oder Unternehmen, von Vorteil, sondern auch für das Training zukünftiger, weiterentwickelter Modelle. Darüber hinaus würde die breite Nutzbarkeit die weitere Entwicklung des KI-Ökosystems fördern. Wenn ein solcher Datensatz im Sinne des künftigen AI Acts erstellt werden würde, könnte er auch Rechts- und Investitionssicherheit für Unternehmen und weitere Akteure bieten und damit ebenfalls das KI-Ökosystem stärken. Ein solches Datenprojekt kann über Community Manager oder Projektmitarbeitende umgesetzt werden, die auf die Unterstützung von Data-Engineering-Spezialistinnen und -Spezialisten und Lizenzanwältinnen und -anwälten zurückgreifen können.

Komponenten gewährleisten: Die ausreichende Verfügbarkeit von KI-Beschleunigern ist ein Faktor, der für den Aufbau einer leistungsfähigen Recheninfrastruktur zum Trainieren großer KI-Modelle im Allgemeinen und Sprachmodellen im Besonderen grundlegend ist. Künftige Maßnahmen sollten auch darauf hinwirken, die Entwicklung von KI-Beschleunigern in Europa voranzutreiben und so die Grundlage für eine KI-adäquate, schnelle und ressourcenschonende Recheninfrastruktur der Zukunft zu gewährleisten.

Recheninfrastruktur anforderungsadäquat ausbauen: Auch künftig dürfte sich der Trend zunehmender nötiger Rechenleistung für KI-Modelle fortsetzen, sodass die Recheninfrastruktur in Deutschland und Europa mit den steigenden Anforderungen mitwachsen sollte, um Unternehmen und vor allem der Forschung eine schnelle und bezahlbare Berechnung großer KI-Modelle zur Verfügung stellen zu können. Bestehende Kapazitäten in Europa sollten daher optimal genutzt werden und die Leistungsfähigkeit der Recheninfrastruktur in Deutschland sollte kontinuierlich ausgebaut werden. Daher möchte das Bundesministerium für Bildung und Forschung die Recheninfrastruktur ausbauen und den Zugang für KI-Forschende, KMU und Start-ups verbessern (BMBF, 2023), und auf EU-Ebene sollen die Zugangsregeln zu Superrechnern für KMU und Start-ups ausgeweitet werden. Beides ist ein Schritt in die richtige Richtung (Lomas, 2023). Für die zukünftige strategische Planung eines solchen „Mitwachsens" von KI-Rechenkapazitäten fehlen jedoch wichtige Indikatoren, wie die OECD feststellt: So sollte zum Beispiel zwischen allgemeinen und KI-bezogenen Rechenkapazitäten unterschieden werden können, um genauer Lageüberblicke entwickeln zu können (OECD, 2023).

Modelle gemäß europäischen Werten entwickeln: Im Sinne der Digitalen Souveränität werden Sprachmodelle für die deutsche Sprache und nach unserem Wertesystem benötigt. Neben einem Fokus auf offenen, kommerziell nutzbaren großen Sprachmodellen für die deutsche Sprache (und auch multi-lingual) kann ein Fokus auf proprietären Modellen europäischer Unternehmen und Start-ups liegen. Solche Modelle sollten im Sinne des europäischen und deutschen Rechts- und Wertesystems aufgebaut werden, um Unternehmen, Behörden und Vereinen einen sicheren Ausgangspunkt zu bieten. Offene Modelle bieten einerseits die Möglichkeit, eigenständig Anpassungen an diesen Modellen nach eigenen Bedürfnissen mit eigenen Daten vornehmen zu können und andererseits so local LLM aufbauen zu können. Von einem solchen Open-Source-Modell würden nicht nur Unternehmen profitieren (insbesondere auch kleinere und weniger technologienahe), sondern auch Forschende. Im besten Fall würde ein solches Modell eine ähnliche Dynamik in der Weiterentwicklung, Anpassung und Wiederverwendung durch die Open-Source-Community entfalten, wie dies beim Modell „LlaMA" von Meta der Fall war. Eine Dynamik, die wiederum allen Teilen der Community zugutekommen kann. Der Aufbau eines solchen Open-Source-Modells könnte öffentlich oder privat gefördert werden und in Kooperation mit einer starken Innovations-Community aus kleineren und größeren Unternehmen erfolgen. Entscheidend für die Qualität eines solchen Modells sind vor allem umfangreiche und hochqualitative Trainingsdaten, auf deren Auswahl und Kuratierung große Sorgfalt zu legen ist.

Community-Building und -Entwicklung vorantreiben: Eine starke Innovations-Community ist zentral, um das Ineinandergreifen von Datensammlung und -kuratierung, Recheninfrastruktur sowie KI-Know-how so zu gestalten, dass Aufbau und Weiterentwicklung großer Sprachmodelle nach den Bedürfnissen deutscher und europäischer Unternehmen vorangetrieben wird. Insbesondere bei dem – im Sinne einer verantwortungsbewussten Gestaltung von KI – wichtigen Anspruch nach Transparenz und Nachvollziehbarkeit der Modelle verspricht eine starke, interdisziplinäre Community große Vorteile: Datengrundlagen lassen sich schnell überprüfen, Bias erkennen. Um Kooperationen zwischen kleinen und größeren Unternehmen verschiedener Branchen sowie Talentpools in Deutschland wie Europa verstärkt auszuschöpfen und somit die Weiterentwicklung und Anwendung von großen Sprachmodellen in der Wirtschaft zu stärken, ist es notwendig, Community-Building vor allem in zwei Richtungen voranzutreiben:

1. Technische Communities aus den Bereichen der Verarbeitung natürlicher Sprache, dem maschinellen Lernen und der Data Science noch besser zu vernetzen und zu koordinieren.

2. Unternehmen aller Größen und Branchen, Beratungsfirmen und Hochschulen und weitere wissenschaftliche Einrichtungen sowie Vereine, die auf die Anpassung von großen Sprachmodellen auf die Bedürfnisse verschiedener Branchen und Disziplinen ausgerichtet sind, als Community zu vernetzen und zu entwickeln.

Als vernetzende und koordinierende Instanz für derartige Innovations-Communities kommen unterschiedliche Akteure in Frage (siehe Infobox, S. 29). Dies können zum Beispiel (Unternehmens-)Verbünde sein, Vereine sowie staatliche, wissenschaftliche oder wissenschaftsbasiert beratende Einrichtungen.

Talente für die Communities in Forschung und Industrie fördern und aufbauen: Damit der Talentpool – vor allem an den oben genannten Schnittstellen – ausgebaut werden kann, bedarf es einer Stärkung und Förderung von Praktika und Programmen in Forschungsprojekten von wissenschaftlichen Einrichtungen, in der Industrie sowie in Kooperationsprojekten, um den Nachwuchs für die Community rund um den Aufbau und die Anpassung von Sprachmodellen zu fördern. Dies kann auf vier Ebenen geschehen:

- **Curricula:** Grundsätzlich ist eine Orientierung an Handreichungen zu Studien- und Weiterbildungsangeboten zu empfehlen, wie zum Beispiel zu Masterstudiengängen für Data Science der Gesellschaft für Informatik in Zusammenarbeit mit der Plattform Lernende Systeme (Gesellschaft für Informatik, 2021). In dieser Publikation wird im Curriculum auch bereits auf generative Modelle verwiesen. Multimodalität kommt dagegen bisher nicht als Stichwort vor, kann jedoch an verschiedenen Stellen des Curriculums aufgegriffen werden, sei es bei Einheiten zum Deep Learning oder zur Datenverarbeitung.

- **Master-Absolvierende:** Im Rahmen von Projekten können Master-Absolvierende die GPU-Cluster sowie die Infrastruktur nutzen und von den Vollzeitangestellten bzw. Promovierenden wichtige Grundlagen erlernen, die derzeit kaum im Studium des maschinellen Lernens in der Tiefe vermittelt werden. Damit ließen sich Lücken in der Ausbildung an der Schnittstelle zwischen maschinellem Lernen und natürlicher Sprachverarbeitung sowie insbesondere bei großen Sprachmodellen weiter schließen.

- **Promovierende:** Während der Promotion bietet sich häufig die Gelegenheit zur Zusammenarbeit mit Unternehmen oder es entsteht das Interesse, eine solche Kooperation mit einem Unternehmen zu vertiefen, zum Beispiel im Rahmen eines Forschungspraktikums. Dabei sollte es den Promovierenden sowohl möglich sein, an Publikationen mitzuarbeiten, die zu ihrer

Promotion zählen könnten, als auch zugleich praxisnahe Erfahrung im Schreiben von Software auf Produktionsniveau zu sammeln. Sinnvoll wäre es hierbei, Anreize zu setzen, die einen solchen Exkurs in die Praxis im vierten oder fünften Jahr der Promotion erleichtern.

- **Domänenexpertinnen und -experten:** Auch geschulte fachfremde Personen mit Domänen- und Prozesswissen sollten in die Lage versetzt werden, wichtige Zusammenhänge und Geschäftsmodelle zu vermitteln. Ein gutes Beispiel ist das DFG-Projekt zur Ausbildung von Digital Clinical Scientists an der Charité; angehendem fachärztlichem Personal wird eine Reduktion der klinischen Verpflichtungen zugestanden, damit dieses stattdessen in dieser Zeit zusammen mit einer Hochschule sowie gemeinsam mit Expertinnen und Experten Sprachmodelle und KI-Modelle für den klinischen Alltag mitentwickeln kann. Analoge Programme könnten auch für andere Disziplinen und Domänen gewinnbringend sein, indem Schnittstellenakteure ausgebildet werden, die in der Community zwischen Anwendungsdomänen und KI-Expertinnen und -Experten vermitteln können. Weiterbildungsprogramme durch Unternehmen in Kooperation mit der Forschung können weiterhin dazu dienen, Domänenexpertinnen und -experten zu befähigen, eine solche Schnittstellenrolle einzunehmen.

5 Offene Fragen

Die gesellschaftlichen und ethischen Fragestellungen wie Implikationen wurden in diesem Whitepaper ausschließlich aus der Perspektive Digitaler Souveränität betrachtet.

Eine ausführlichere Behandlung dieser Themen parallel zu den sich abzeichnenden und bereits vorhandenen Anwendungsfällen ist jedoch bedeutend und wird in künftigen Publikationen und Veranstaltungen der Plattform Lernende Systeme adressiert.

- Welche konkreten gesellschaftlichen Auswirkungen generativer KI-Modelle stellen wir bereits fest und welche sind zu erwarten?

- Welche Handlungsansätze für einen wertorientierten Einsatz generativer KI können vor dem Hintergrund dieser Auswirkungen entwickelt werden?

Literatur

Alsentzer, E. et al. (2019): Publicly Available Clinical BERT Embeddings. In Proceedings of the 2nd Clinical Natural Language Processing Workshop (S. 72–78). Minneapolis, Minnesota, USA: Association for Computational Linguistics. https://doi.org/10.48550/arXiv.1904.03323

Arnold, S. et al. (2019): SECTOR: A Neural Model for Coherent Topic Segmentation and Classification. Computational Linguistics, 7, 169–184.

Arnold, S. et al. (2020): Learning Contextualized Document Representations for Healthcare Answer Retrieval. (A. f. Machinery, Hrsg.) roceedings of The Web Conference 2020, S. 1332–1343. https://doi.org/10.1145/3366423.3380208

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

Benaich, N. & Hogarth, I. (2022b): State of AI Report Compute Index. Online unter: https://www.stateof.ai/compute

Bundesministerium für Bildung und Forschung (2021): Hoch- und Höchstleistungsrechnen für das digitale Zeitalter. Forschung und Investitionen zum High-Performance-Computing. Online unter:https://www.bmbf.de/SharedDocs/Publikationen/de/bmbf/5/31669_Hoch_und_Hoechstleistungsrechnen_fuer_das_digitale_Zeitalter.pdf?__blob=publicationFile&v=7 (abgerufen am 24.10.2023)

Bundesministerium für Bildung und Forschung (2022): Förderung von vier KI-Servicezentren gestartet. Online unter: https://www.bmbf.de/bmbf/shareddocs/kurzmeldungen/de/2022/11/foerderung-von-4-ki-zentren-gestartet.html (abgerufen am 23.01.2023)

Bundesministerium für Bildung und Forschung (2023): BMBF-Aktionsplan Künstliche Intelligenz. Neue Herausforderungen chancenorientiert angehen. Online unter: https://www.bmbf.de/SharedDocs/Publikationen/de/bmbf/5/837380_Aktionsplan_Kuenstliche_Intelligenz.pdf?__blob=publicationFile&v=5 (abgerufen am 13.11.2023)

Bundesministerium für Wirtschaft und Klimaschutz (2023a): Differentialdiagnose. Online unter: https://www.bmwk.de/Redaktion/DE/Artikel/Digitale-Welt/GAIA-X-Use-Cases/differentialdiagnose.html (abgerufen am 31.07.2023)

Bundesministerium für Wirtschaft und Klimaschutz (Hrsg.) (2023b): Das Ökosystem für KI-Startups in Deutschland. Online unter: https://www.de.digital/DIGITAL/Redaktion/DE/Digitalisierungsindex/Publikationen/publikation-download-ki-start-ups-2023.pdf?__blob=publicationFile&v=4 (abgerufen am 24.08.2023)

Bommasani, R., Klyman, K., Zhang, D. & Liang, P. (2023): Do Foundation Model Providers Comply with the Draft EU AI Act? Center for research on Foundation models, Human-Centered Artificial Intelligence. Stanford: Stanford University. Online unter: https://crfm.stanford.edu/2023/06/15/eu-ai-act.html (abgerufen am 31.07.2023)

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

Cesareo, S. & White, J. (2023): The Global AI Index. Online unter: https://www.tortoisemedia.com/intelligence/global-ai/#rankings (abgerufen am 03.08.2023)

Chui, M. et al. (June 2023): The economic potential of generative AI: The next productivity frontier. (McKinsey Special Report). Online unter: https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/the-economic-potential-of-generative-ai-the-next-productivity-frontier (abgerufen am 31.07.2023)

CRFM (2023) HELM: Online unter: https://crfm.stanford.edu/helm/latest/?models=1

Denk, T. I. & Reisswig, C. (2019): BERTgrid: Contextualized Embedding for 2D Document Representation and Understanding. https://doi.org/10.48550/arXiv.1909.04948

Epoch.AI (2023): Parameter, Compute and Data Trends in Machine Learning. Online unter: https://docs.google.com/spreadsheets/d/1AAIebjNsnJj_uKALHbXNfn3_YsT6sHXtCU0q7OIPuc4/edit#gid=1917852922 (abgerufen am 28.03.2023)

Esteva, A. et al. (2021): Deep learning-enabled medical computer vision. NPJ Digital Medicine, 4(1). https://doi.org/10.1038/s41746-020-00376-2

EuroHPC (2023): Our supercomputers. Online unter: https://eurohpc-ju.europa.eu/supercomputers/our-supercomputers_en (abgerufen am 25.09.2023)

Gesellschaft für Informatik (2021): Masterstudiengänge „Data Science" – Auf Basis eines Bachelors in (Wirtschafts-) Informatik oder Mathematik. Online unter: https://gi.de/fileadmin/GI/Hauptseite/Service/Publikationen/Empfehlungen/Empfehlungen_Masterstdiengaenge_DataScience_2021.pdf (abgerufen am 25.09.2023)

Goel, A. (2023): Unraveling GPU Inference Costs for Fine-tuned Open-source Models V/S Closed Platforms. Online unter: https://mlops.community/unraveling-gpu-inference-costs-for-fine-tuned-open-source-models-v-s-closed-platforms/

Grundermann, P., Oberhauser, T. & Gers, F. (2022): Attention Networks for Augmenting Clinical Text with Support Sets for Diagnosis Prediction. COLING, 4765–4775.

Grundmann, P., Arnold, S. & Löser, A. (2021): Self-supervised Answer Retrieval on Clinical Notes. https://doi.org/10.48550/arXiv.2108.00775

Hahn, S. (17. August 2022): Europas schnellstes kommerzielles KI-Rechenzentrum feierlich in Berlin eröffnet. Online unter: https://www.heise.de/news/Europas-schnellstes-kommerzielles-KI-Rechenzentrum-feierlich-in-Berlin-eroeffnet-7267438.html (abgerufen am 23.01.2023)

Heikkilä, M. (2023): Meta's latest AI model is free for all. Online unter: https://www.technologyreview.com/2023/07/18/1076479/metas-latest-ai-model-is-free-for-all/ (abgerufen am 03.08.2023)

Huang, K., Altosaar, J. & Ranganath, R. (2019): ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission. In I. a. Proceedings of ACM Conference on Health (Hrsg.), Proceedings of ACM Conference on Health, Inference, and Learning, CHIL 2020.

Hugging Face (2023): Open LLM Leaderboard. Online unter: https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard (abgerufen am 25.09.2023)

IDC (2022): IDC-Studie: 60 Prozent der deutschen Unternehmen investieren in zukunftsfähige Data Center. International Data Corporation (IDC). Online unter: https://www.idc.com/getdoc.jsp?containerId=prEUR149762022

IDS (2023): Ausbau und Pflege der Korpora geschriebener Gegenwartssprache. Online unter: https://www.ids-mannheim.de/digspra/kl/projekte/korpora/ (abgerufen am 25.09.2023)

IDTechEx (2023): AI Chips 2023–2033. Online unter: https://www.idtechex.com/en/research-report/ai-chips-2023-2033/937

Kagermann, H., Streibich, K.-H. & Suder, K. (2021): Digitale Souveränität – Status quo und Handlungsfelder (acatech IMPULS). München. Online unter: https://www.acatech.de/publikation/digitale-souveraenitaet-status-quo-und-handlungsfelder/

Katti, A. R. et al. (2018): Chargrid: Towards Understanding 2D Documents. In A. f. Linguistics (Hrsg.), Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (S. 4459–4469). Brussels, Belgium: Association for Computational Linguistics. http://dx.doi.org/10.18653/v1/D18-1476

KI-Bundesverband (2023): Große KI-Modelle für Deutschland. Online unter: https://leam.ai/wp-content/uploads/2023/01/LEAM-MBS_KIBV_webversion_mitAnhang_V2_2023.pdf (abgerufen am 16.03.2023)

Klaiman, S. & Lehne, M. (2021): DocReader: Bounding-Box Free Training of a Document Information Extraction Model. ArXiv, abs/2105.04313. https://doi.org/10.48550/arXiv.2105.04313

LeCun, Y. (2023): From Machine Learning to Autonomous Intelligence (Vortrag vom 29.09.2023, München). Online unter: https://www.youtube.com/watch?v=pd0JmT6rYcI

Lee, J. et al. (2020): BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics, 36(4), S. 1234–1240. https://doi.org/10.1093/bioinformatics/btz682

Lomas, N. (2023): EU to let 'responsible' AI startups train models on its supercomputers. Online unter: https://techcrunch.com/2023/09/13/eu-supercomputers-for-ai/ (abgerufen am 25.09.2023)

LUMI (2021): LUMI's full system architecture revealed. Online unter: https://www.lumi-supercomputer.eu/lumis-full-system-architecture-revealed/ (abgerufen am 25.09.2023)

Lüngen, H. (2017): „DeReKo – Das Deutsche Referenzkorpus: Schriftkorpora der deutschen Gegenwartssprache am Institut für Deutsche Sprache in Mannheim", Zeitschrift für germanistische Linguistik, vol. 45, no. 1, S. 161–170. https://doi.org/10.1515/zgl-2017-0008

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

Maham, P. et al. (2022): Deutschland als KI-Standort: Destination oder Drehscheibe? Berlin: Stiftung Neue Verantwortung (SNV).

Miotto, R. et al. (17. Mai 2016): Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records. Scientific reports, (1)(6), S. 1–10. https://doi.org/10.1038/srep26094

Molino, P. (2023): Will there be one large general model to rule them all or will there be thousands of specialized ones? LinkedIn. Online unter: https://www.linkedin.com/posts/pieromolino_llm-ai-machinelearning-activity-7099458821449814016-BOXd?trk=public_profile_like_view (abgerufen am 26.10.2023)

Mullenbach, J. et al. (Juni 2018): Explainable Prediction of Medical Codes from Clinical Text. (A. f. Linguistics, Hrsg.) Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), S. 1101–1111. http://dx.doi.org/10.18653/v1/N18-1100

Narayanan, D. S. et al. (2021): Efficient Large-Scale Language Model Training on GPU Clusters. Online unter: https://cs.stanford.edu/~matei/papers/2021/sc_megatron_lm.pdf (abgerufen am 25.03.2023)

OECD (2023a): A blueprint for building national compute capacity for artificial intelligence. Paris: OECD Publishing. https://doi.org/10.1787/876367e3-en

OECD (2023b): AI talent concentration by country. Online unter: https://oecd.ai/en/data?selectedArea=ai-jobs-and-skills&selectedVisualization=ai-talent-concentration-by-country (abgerufen am 16.03.2023)

Papaioannou, J.-M. et al. (2022): Cross-Lingual Knowledge Transfer for Clinical Phenotyping. In E. L. Association, Proceedings of the Thirteenth Language Resources and Evaluation Conference (S. 900 – 909). Marseille, France.

Reuters (2023): Chinese organisations launched 79 AI large language models since 2020, report says. Online unter: https://www.reuters.com/technology/chinese-organisations-launched-79-ai-large-language-models-since-2020-report-2023-05-30/ (abgrufen am 03.08.2023)

Statista Market Insights (2023): Generative AI – Worldwide. Online unter: https://www.statista.com/outlook/tmo/artificial-intelligence/generative-ai/worldwide#market-size (abgerufen am 25.09.2023)

Scao, T. F. et al. (2022): BLOOM: A 176B-Parameter Open-Access Multilingual Language Model. Online unter: https://arxiv.org/pdf/2211.05100.pdf (abgerufen am 25.03.2023)

Schütze, Hinrich (2023): Keynote. Munich LLM Conference 2023. Online unter: https://www.youtube.com/watch?v=qD50IN6DLS8 (abgerufen am 25.03.2023)

Sequoia (2023a): Atlas Sequoia's interactive guide to Europe's technical talent. Online unter: https://atlas.sequoiacap.com/?_skill=ai

Sequoia (2023b): A Talented Home for AI. Sequoia. Online unter: https://atlas.sequoiacap.com/a-talented-home-for-ai/

Solaiman, I. (2023): The Gradient of Generative AI Release. Online unter: https://arxiv.org/pdf/2302.04844.pdf (abgerufen am 24.03.2023)

Statista (2022): Medical Technology – Worldwide. Online unter: https://www.statista.com/outlook/hmo/medical-technology/worldwide (abgerufen am 30.01.2023)

Tinn, R. et al. (2021): Fine-tuning large neural language models for biomedical natural language processing. https://doi.org/10.1016/j.patter.2023.100729

Top500.org. (16. März 2023): Top500. Online unter: https://www.top500.org/statistics/sublist/ (abgerufen am 03.08.2023)

Topol, E. J. (07. Januar 2019): High-performance medicine: the convergence of human and artificial intelligence. Nature Medicine, 25, S. 44 – 56. https://doi.org/10.1038/s41591-018-0300-7

van Aken, B. et al. (2021): Clinical Outcome Prediction from Admission Notes using Self-Supervised Knowledge Integration. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics (Bd. Main Volume, S. 881 – 893). http://dx.doi.org/10.18653/v1/2021.eacl-main.75

van Aken, B. et al. (10 2022): This Patient Looks Like That Patient: Prototypical Networks for Interpretable Diagnosis Prediction from Clinical Text. AACL/IJCNL 2022.

Vogel, M. (2023): ChatGPT, Next Level: Meet 10 Autonomous AI Agents: Auto-GPT, BabyAGI, AgentGPT, Microsoft Jarvis, ChaosGPT & friends. Online unter: https://medium.com/the-generator/chatgpts-next-level-is-agent-ai-auto-gpt-babyagi-agentgpt-microsoft-jarvis-friends-d354aa18f21 (abgerufen am 24.08.2023)

Whitten, A. (2022): New Chip Expands the Possibilities for AI. Quantamagazine. Online unter: https://www.quantamagazine.org/a-brain-inspired-chip-can-run-ai-with-far-less-energy-20221110/

Wiggers, K. (16. November 2022): techcrunch.com. Online unter: https://techcrunch.com/2022/11/16/microsoft-and-nvidia-team-up-to-build-new-azure-hosted-ai-supercomputer/ (abgerufen am 16.03.2023)

Wiggers, K. (12. Juli 2022): techcrunch.com. Online unter: https://techcrunch.com/2022/07/12/a-year-in-the-making-bigsciences-ai-language-model-is-finally-available/ (abgerufen am 25.03.2023)

Winter, B. et al. (2022): KIMERA: Injecting Domain Knowledge into Vacant Transformer Heads. Proceedings of the Thirteenth Language Resources and Evaluation Conference, (S. 363 – 373). Marseille, France.

Yang, B. & Wu, L. (2021): How to leverage the multimodal EHR data for better medical prediction? Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021 (S. 4029 – 4038): Virtual Event / Punta Cana, Dominican Republic, 7-11 November: Association for Computational Linguistics.

The AI Index 2023: Annual Report. Stanford University, AI Index Steering Committee, Institute for Human-Centered AI, CA. Online unter: https://aiindex.stanford.edu/report/#individual-chapters

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,
Munich Center for Machine Learning (MCML)
Dr. Johannes Hoffart, SAP

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

Befragte Experten
Timo Möller, deepset GmbH
Johannes Otterbach (ehemals Merantix, seit Juli 2023 nyonic)
Till Plumbaum (ehemals Neofonie GmbH, seit Februar 2023 Alexander Thamm GmbH)
Hans-Peter Zorn, inovex 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

Gestaltung und Produktion
PRpetuum GmbH, München

Stand
Dezember 2023

Bildnachweis
nuttapong punna/iStock/Titel

Empfohlene Zitierweise
Löser, A., Tresp, V. et al. (2023): Große Sprachmodelle entwickeln und anwenden. Ansätze für ein souveränes Vorgehen. Whitepaper aus der Plattform Lernende Systeme, München.
DOI: https://doi.org/10.48669/pls_2023-6

Bei Fragen oder Anmerkungen zu dieser Publikation kontaktieren Sie bitte Dr. 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.