Deloitte.

Fraunhofer IAIS

Superkraft Sprachmodell?

Wie generative KI einen Beitrag leisten kann, die Leistungsfähigkeit in der Verwaltung zu steigern

Einleitung 05

Anwendungsbeispiele von GenAI
in der öffentlichen Verwaltung 06

- Use Case 01 | Sprachmodell mit verwaltungsspezifischem Fachwissen 06
- Use Case 02 | Formularfreie Bedarfserfassung von Sozialleistungen 08
- Use Case 03 | Gesetzesumsetzung mit KI-Unterstützung 10

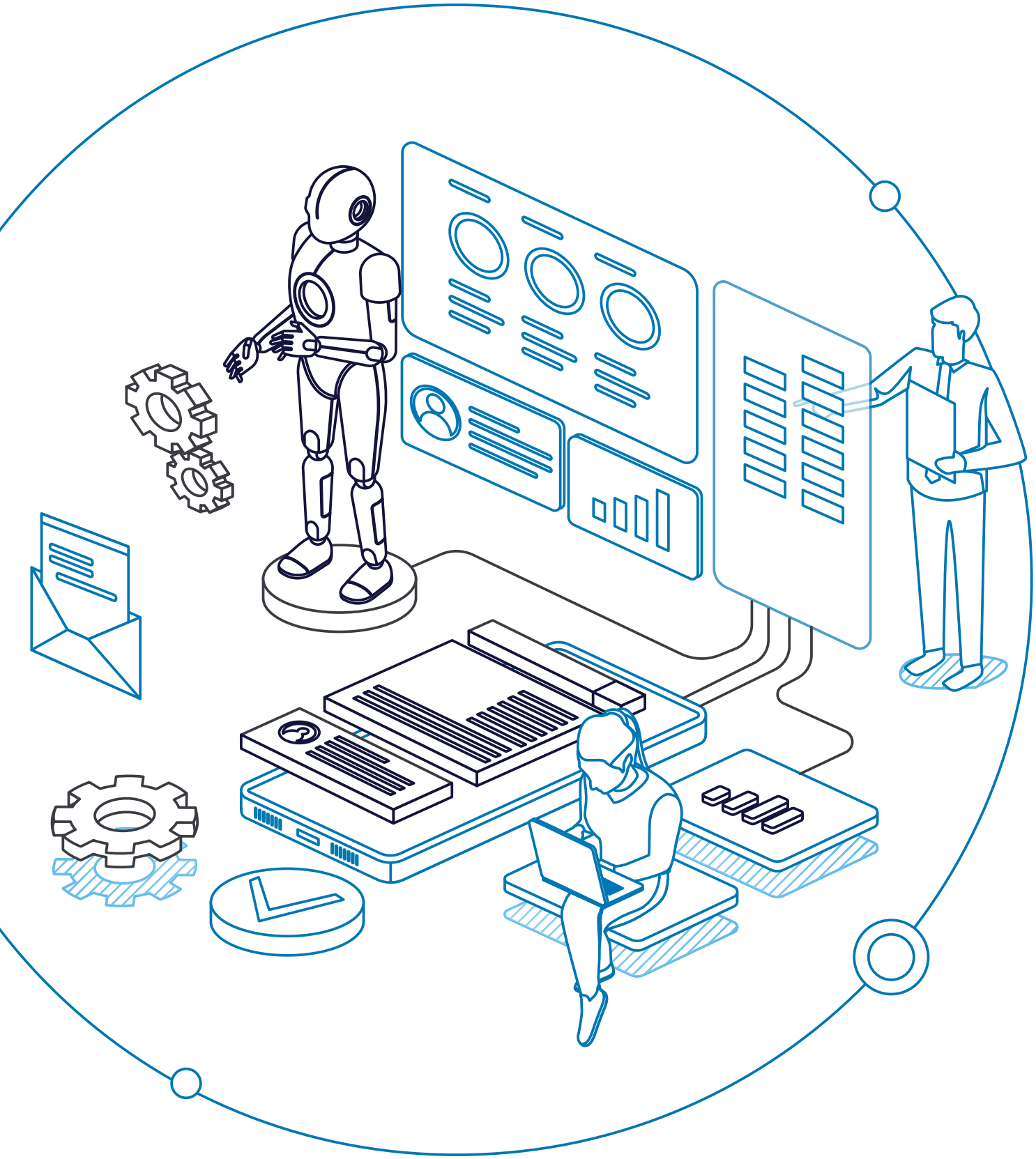
Implementierung und Rahmenbedingungen 13

GenAI in der Verwaltung – Zukunft gestalten 14

Anhang 16

- Funktionsweise von LLMs am Beispiel von GPT-Modellen 16
- Training und Funktionsweise von KNN, wie sie in GPT-Modellen vorkommen 18

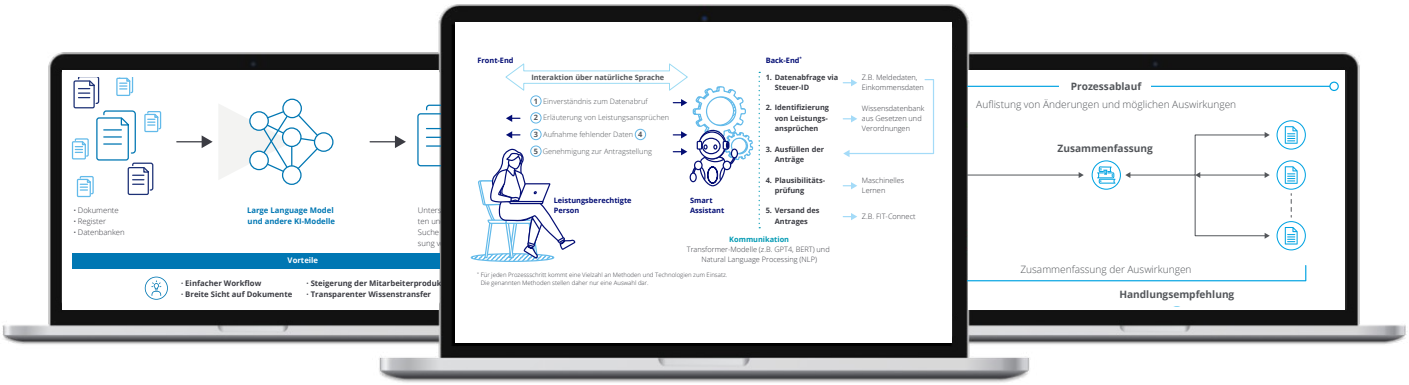
Ihre Ansprechpersonen 20



Einleitung

Die deutsche Verwaltungslandschaft steht angesichts der notwendigen Digitalisierung und Automatisierung von bisher manuellen Prozessen vor einer grundlegenden Transformation. Der Anstieg an Anträgen für Wohngeld, BAföG oder Einbürgerungsverfahren setzt Behörden zusätzlich unter Druck. Der entstehende Rückstau trägt zu einem sinkenden Vertrauen in die Leistungsfähigkeit der öffentlichen Verwaltung bei. Gleichzeitig muss sie die rückläufigen Mitarbeitendenzahlen infolge des demografischen Wandels kompensieren. Generative Künstliche Intelligenz (GenAI) und insbesondere große Sprachmodelle (Large Language Models, LLMs) spielen hier eine wichtige Rolle, um die Mitarbeitenden zukünftig in ihren Aufgaben zu unterstützen, zu entlasten und hierdurch Freiräume zu schaffen, um sich verstärkt der direkten Interaktion mit Bürgerinnen und Bürgern zu widmen. In diesem Briefing präsentieren Fraunhofer IAIS und Deloitte drei Anwendungsbeispiele großer Sprachmodelle, von welchen die öffentliche Verwaltung schon heute profitieren kann. Bei der Betrachtung zu etablierender Rahmenbedingungen muss zwischen den behördeninternen Voraussetzungen und der staatlichen Infrastruktur unterschieden werden. Unsere Publikation betrachtet die Voraussetzungen auf individueller Ebene der Behörden.

Abb. 1 – Illustrative Darstellung der drei Use Cases



Use Case 01
Sprachmodell mit verwaltungsspezifischem Fachwissen

Use Case 02
Formularfreie Bedarfserfassung von Sozialleistungen

Use Case 03
Gesetzesaktualisierung mit KI-Unterstützung

Anwendungsbeispiele von GenAI in der öffentlichen Verwaltung

Use Case 01

Sprachmodell mit verwaltungsspezifischem Fachwissen

GenAI und LLMs haben insbesondere dank OpenAIs ChatGPT sowohl in der allgemeinen Öffentlichkeit als auch in der Geschäftswelt signifikante Aufmerksamkeit erhalten. Hingegen haben sprachbasierte Modelle mit Anwendungsbezug zur öffentlichen Verwaltung bislang nur begrenzte Verbreitung gefunden, da die Applikationen in diesem Kontext besondere Anpassungen im Hinblick auf die Datengrundlage erfordern. Dabei wäre ein Sprachmodell mit verwaltungsspezifischem Fachwissen, das für Wissensmanagement, Rechercheaufgaben und Dokumentenanalyse eingesetzt werden könnte, zweifellos auch innerhalb der öffentlichen Verwaltung von erheblichem Nutzen. Effektivitätssteigerungen könnten durch den Einsatz von großen Sprachmodellen insbesondere bei der Bekämpfung des Fachkräftemangels generiert werden. Vor diesem Hintergrund stellt sich die zentrale Frage, wie LLMs an die spezifischen Anforderungen im Umfeld der öffentlichen Verwaltung angepasst werden können.

Technische Lösung

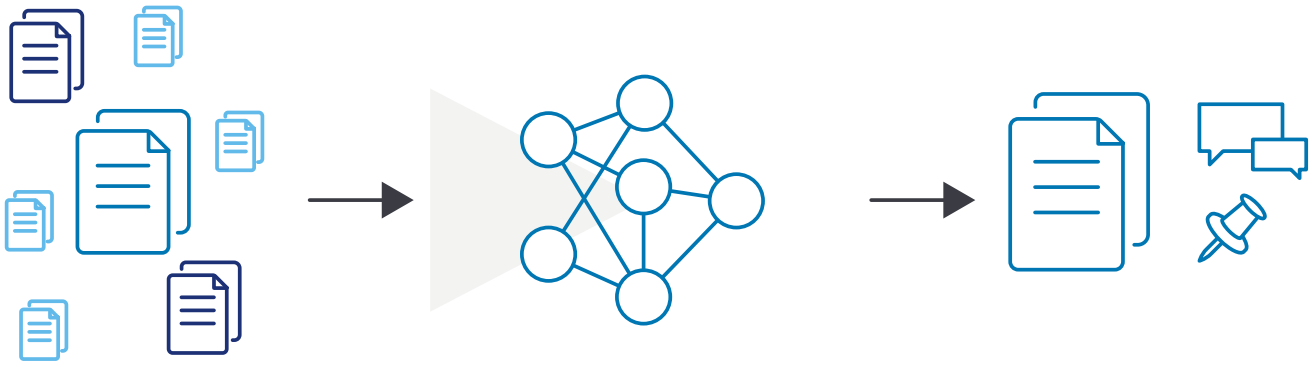
Ein großes Sprachmodell mit verwaltungsspezifischem Fachwissen bietet die technologische Grundlage für viele Anwendungsfälle. Für den internen Gebrauch in der Behörde lässt es sich zu einer chatbasierten Wissensdatenbank ausbauen. Statt des zeitaufwendigen Suchens von Dokumenten kann das Modell nach den gewünschten Informationen befragt werden. LLMs mit einem Aufbau wie bspw. ChatGPT oder BARD von Google beziehen sich im initialen Training des Modells häufig auf große, unkuratierte Datensätze. Um die Sprachmodelle für den Einsatz in der öffentlichen Verwaltung zu spezialisieren und sie mit Fachwissen zu versorgen, werden qualitativ hochwertige (Verwaltungs-)Daten identifiziert und aufbereitet. Anschließend erfolgt ein sogenanntes Finetuning des bestehenden Modells auf dieser Datengrundlage. Im weiteren Gebrauch bleibt das Modell durch die kontinuierliche Verwendung und Bereitstellung von relevanten Informationen aktuell. Eine Ergänzung zum Finetuning stellt die Einbettung eines bestehenden Modells in ein RAG Framework (Retrieval-Augmented Generation)[1] dar. Dabei werden relevante Dokumente wie Rechtstexte, Stellungnahmen und Drucksachen aus tagesaktuellen Datenbanken abgerufen und auf dieser Grundlage wird eine Antwort generiert. Eine ausführliche Darstellung zur Funktionsweise von LLMs findet sich im Anhang.

Die Stärke von LLMs liegt darin, dass es nun möglich ist, menschenähnliche Interaktionen mit einer KI durchzuführen. Ein Beispiel hierfür ist die Integration von LLMs in Sprachassistenten, die Bürger:innen beim Prozess von Anträgen durchgehend zur Seite stehen, unabhängig von Wochenenden und Öffnungszeiten der Behörden. Dabei können Anliegen in natürlicher Sprache geschildert sowie Plausibilitätsabfragen ergänzt werden, um so die Qualität der Anträge zu gewährleisten.

Mehrwerte für die Behörde

Die Einführung von Sprachmodellen mit verwaltungsspezifischem Wissen in Behörden kann eine erhebliche Steigerung der Effizienz mit sich bringen. Sachbearbeitende können so schnell und präzise auf benötigte Informationen zugreifen, wodurch sowohl Zeit eingespart als auch Fehler minimiert werden. Bei der täglichen Arbeit unterstützt das Sprachmodell die Sachbearbeitenden zum Beispiel bei langwierigen Recherchen. Dokumente können mithilfe von Sprachmodellen schnell zusammengefasst und ausgewertet werden.[2] Zudem steigert eine solche Lösung die Produktivität, indem sie relevante Informationen und Daten bereitstellt und Mitarbeitende unterstützt, schnell fundiertere Entscheidungen zu treffen.[3] Darüber hinaus dienen diese Sprachmodelle auch als ein wirksames Werkzeug für die Qualitätskontrolle, indem sie dazu genutzt werden können, Dokumente und Daten auf Konsistenz und Fehler hin zu überprüfen.

Abb. 2 – Use Case 01: Sprachmodell mit verwaltungsspezifischem Fachwissen



- Dokumente
- Register
- Datenbanken

Large Language Model und andere KI-Modelle

Unterstützung bei der intelligenten und kontextbezogenen Suche und der Zusammenfassung von Informationen

Vorteile

- **Einfacher Workflow**
- **Breite Sicht auf Dokumente**
- **Steigerung der Mitarbeiterproduktivität**
- **Transparenter Wissenstransfer**

[1] Ein RAG Framework bietet den Vorteil, dass das LLM insbesondere eine vordefinierte Dokumenten-Datenbank als Wissensgrundlage nutzt, die zudem schnell aktualisiert werden kann. Darüber hinaus bietet das RAG Framework eine gute Skalierbarkeit. Nachteile sind eine erhöhte Systemkomplexität durch die Kombination unterschiedlicher Prozesse und eine Leistungsbeschränkung bei den Ergebnissen, da die Antwortqualität mit den gewählten Dokumenten direkt verbunden ist. Im Vergleich bietet ein Finetuning Spezialisierungs- und Performancevorteile, da das Modell an die Charakteristika seiner Aufgaben angepasst wird. Nachteilig sind der Verlust der Generalisierbarkeit und die Datenabhängigkeit. Beides ist bei einer Implementierung in der öffentlichen Verwaltung unproblematisch, weil das Modell in einem einzigen Kontext genutzt wird und eine solide und valide Datengrundlage bereits vorhanden ist.

[2] https://www2.deloitte.com/us/en/pages/consulting/articles/gen-ai-use-cases.html#summarizing-legislative-documents%C2%A0(legislative-administration) abgerufen am 31.10.2023.

[3] https://www2.deloitte.com/us/en/pages/consulting/articles/gen-ai-use-cases.html#insights-for-all-(knowledge-management), abgerufen am 31.10.2023.

Use Case 02

Formularfreie Bedarfserfassung von Sozialleistungen

Die Identifizierung von möglichen Leistungsansprüchen im Sozialleistungsbereich ist aufgrund der Vielzahl von Leistungen und zuständigen Behörden komplex. Aus der großen Anzahl an Menschen, die Sozialleistungen beantragen, resultiert ein hoher Beratungs- und Informationsbedarf, denn Beantragung und laufende Datenerfassung dieser Leistungen sind für viele Antragstellende oftmals kompliziert und zeitaufwendig. Selbst die Einführung von Online-Anträgen im Rahmen des Onlinezugangsgesetzes hat daran wenig geändert. Hierbei spielen unterschiedliche Faktoren eine Rolle, wie etwa die Vielfalt der verfügbaren Leistungen, die individuellen Umstände der Antragstellenden und technische Hürden bei der Antragstellung. Es stellt sich daher die Frage, wie LLMs Bürgerinnen und Bürgern dabei helfen können, ihre Leistungsansprüche zu ermitteln und schnell und formlos Leistungen zu beantragen.

Technische Lösung

In dem genannten Kontext stellen sprachbasierte Modelle in Form einer virtuellen Beratung und eines Antragsassistenten eine vielversprechende Lösung dar. Die Antragstellenden können in unstrukturierter natürlicher Sprache über Sprachein- und -ausgabe mit dem Antragsassistenten interagieren. Mithilfe des LLM extrahiert die Anwendung Informationen auf strukturierte Weise und leitet die Benutzenden an, weitere Informationen durch natürliche Sprache oder den Upload von Nachweisen anzugeben, um den Leistungsanspruch zu ermitteln und das korrekte Ausfüllen des Formulars durch den virtuellen Assistenten sicherzustellen. Letztlich bedarf es durch die Antragstellenden lediglich der Bestätigung, dass die aufgenommenen Angaben korrekt sind. Ein integrierter Konsistenzcheck überprüft die eingegebenen Informationen zusätzlich auf ihre Plausibilität und reduziert hiermit den Bearbeitungsaufwand in den Behörden.

Die Antragsdaten werden unmittelbar nach Versand an die zuständige Behörde gelöscht. Auf diese Weise kann sichergestellt werden, dass keine andere Entität als die zuständige Behörde Zugriff auf die personenbezogenen Daten erhält.

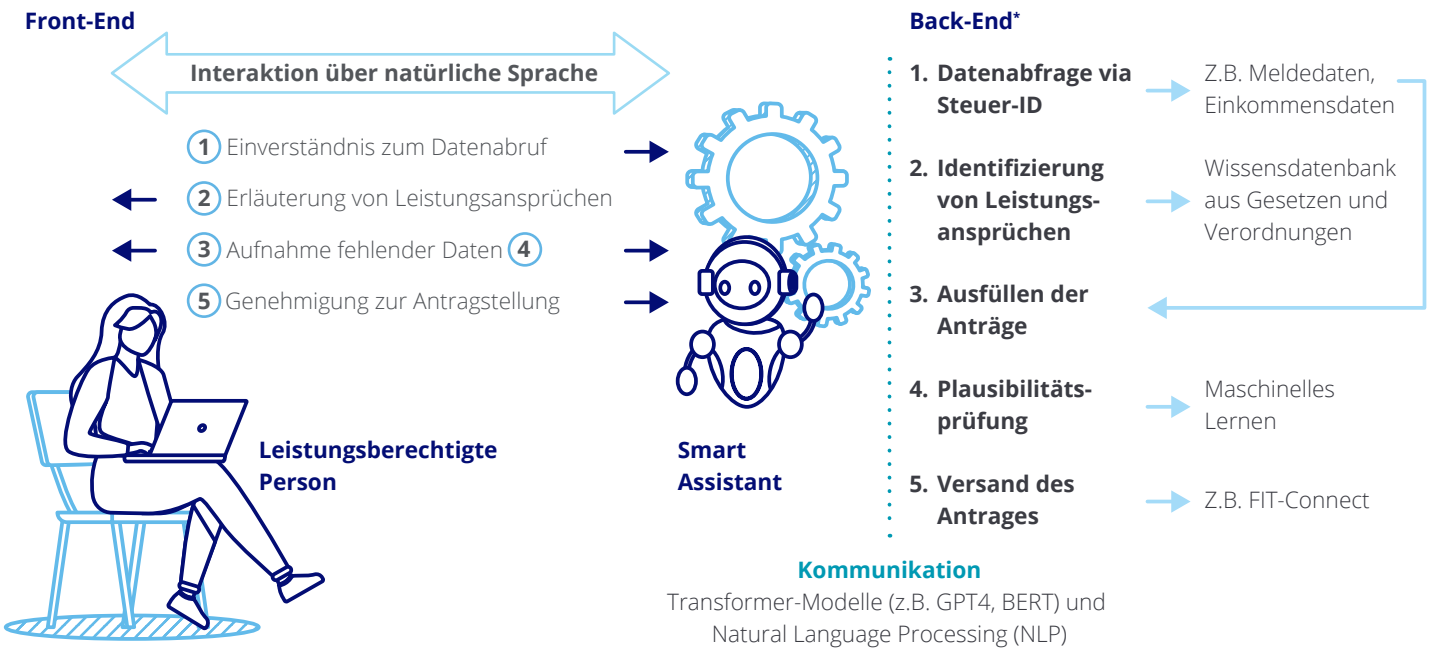
Umsetzbar ist dieser Anwendungsfall durch das Zusammenspiel von Methoden aus dem Bereich optische Zeichenerkennung (OCR) und KI. Hierbei werden zunächst gedruckte oder handgeschriebene Texte von Dokumenten oder Bildern erkannt und in maschinenlesbaren Text umgewandelt. Durch eine Texterkennung werden somit die relevanten Daten extrahiert und durch die KI validiert. Die Businesslogiken können über das Training der KI domainspezifisch gestaltet werden. Dieser Prozess kann voll automatisiert aufgesetzt werden, um die Sachbearbeitenden deutlich zu entlasten. Dabei ist klar, dass der Mensch immer die letzte Instanz in der Entscheidung bleibt.

Mehrwerte für Behörden und Antragstellende

Virtuelle Assistenten könnten dazu dienen, Bürgerinnen und Bürgern zu helfen, ihre Ansprüche auf Sozialleistungen besser zu verstehen und den Antragsprozess effizienter zu durchlaufen. Durch eine verständlichere und benutzerfreundlichere Erklärung der verschiedenen Leistungsarten können LLMs den Bürgerinnen und Bürgern ermöglichen, ihre individuellen Ansprüche genauer zu bestimmen. Durch die Unterstützung im Antragsprozess wird die Anzahl fehlerhafter Anträge reduziert, wodurch die Arbeitsbelastung in den Behörden sinkt und folglich eine schnellere Bearbeitungszeit von Anträgen erreicht wird. Durch die unmittelbare Möglichkeit, Fragen direkt während der Antragsstellung zu stellen, können etwaige Probleme zudem direkt gelöst werden. Hinzu kommt, dass die verkürzten Bearbeitungszeiten die Sozialleistungsempfangenden entlasten.

Der potenziell größte Mehrwert besteht in einer audiobasierten Interaktion mit dem virtuellen Assistenten. Statt des schriftlichen, manuellen Ausfüllens eines komplexen Formulars können Antragstellende fortan auf natürliche, sprachbasierte Art mit dem digitalen Assistenten kommunizieren, so ihre Sozialhilfe beantragen und währenddessen Rückfragen stellen. Wenngleich eine solche digitale Assistenz nicht den Bedarf von Beratungsangeboten vor Ort ersetzt, führt sie dennoch zur Reduzierung des persönlichen Beratungsbedarfs und folglich zur Entlastung der Mitarbeitenden in den Behörden und Beratungsstellen.

Abb. 3 – Use Case 02: formularfreie Bedarfserfassung von Sozialleistungen



* Für jeden Prozessschritt kommt eine Vielzahl an Methoden und Technologien zum Einsatz. Die genannten Methoden stellen daher nur eine Auswahl dar.

Use Case 03

Gesetzesumsetzung mit KI-Unterstützung

Die Verabschiedung neuer Gesetze hat vielfältige Auswirkungen auf die Verwaltung. Insbesondere die für den Vollzug von Leistungen vorgesehenen Behörden sind davon stark betroffen. In der Folge kommt es zu zeitaufwendigen und fehleranfälligen Interpretationen, manuellen Anpassungen von Folgedokumenten wie Durchführungsanweisungen sowie hohen Schulungsaufwänden. Es stellt sich daher die Frage, wie Sprachmodelle genutzt werden können, um die Auswirkungen von Gesetzen im Vollzug und die damit verbundenen Handlungsbedarfe in unterschiedlichen Bereichen einer Behörde kenntlich zu machen.

Technische Lösung

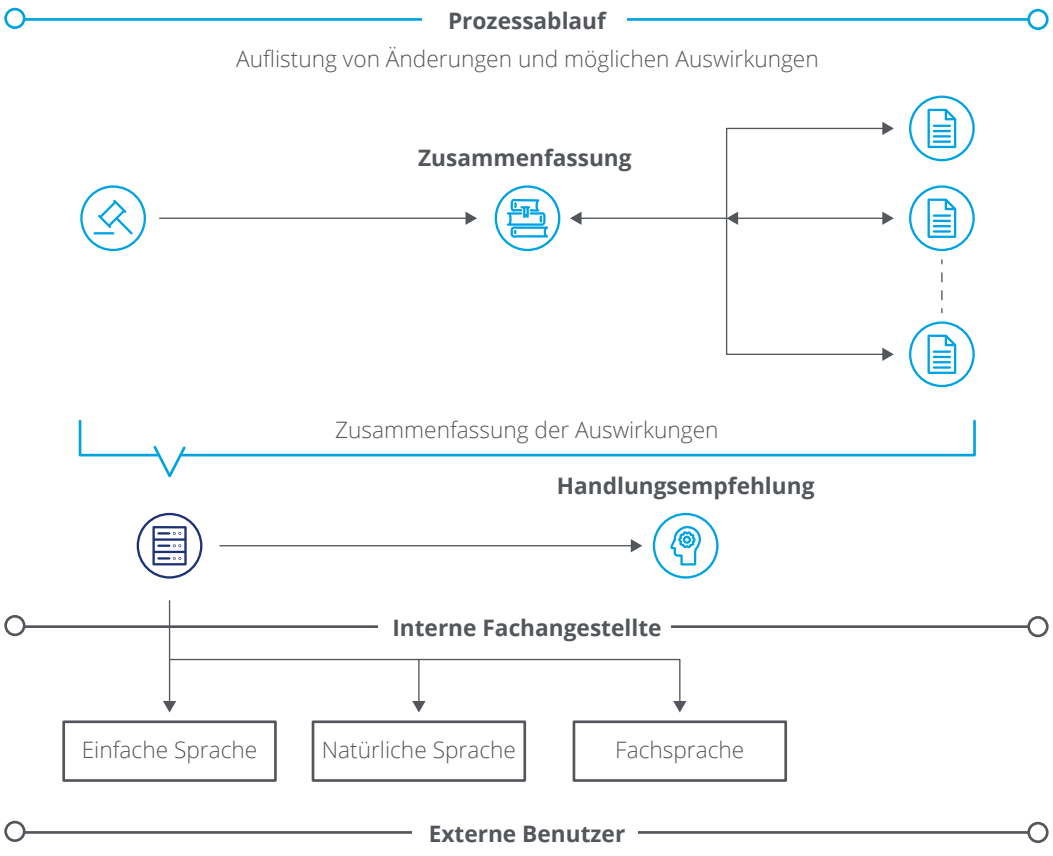
Quelldokumente wie Gesetzestexte werden durch ein Sprachmodell analysiert, um die Informationen aufzubereiten und zu organisieren. Aus einem Pool aus Zieldokumenten (z.B. Durchführungsanweisungen, Broschüren) können die anzupassenden Dokumente automatisch identifiziert oder manuell durch die Nutzenden ausgewählt werden. Angepasste oder neue Inhalte für Zieldokumente werden als Vorschläge markiert und durch das Sprachmodell generiert. Darüber hinaus ermöglicht der Einsatz dieser Lösungen in diesem Zusammenhang die nutzer:innenspezifische Zusammenfassung von Dokumenten. Auch hier gilt eine Prüfung durch Menschen als unabdingbar, da semantische Sprachmodelle nur die Ergebnisse ausgeben, die am wahrscheinlichsten sind.

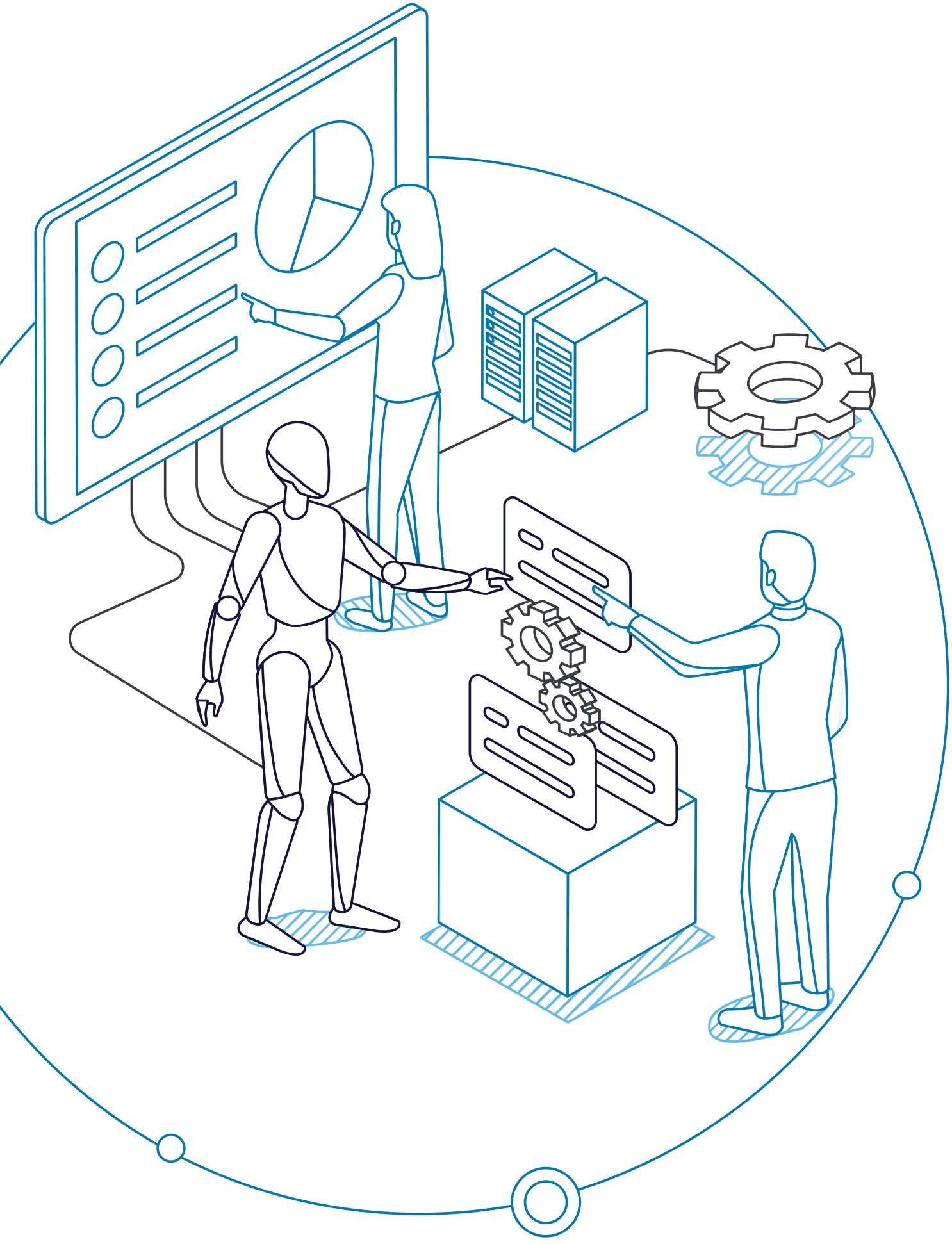
Die technische Umsetzung erfolgt in mehreren Schritten: Zunächst werden aus verschiedenen Gesetzestexten Zusammenfassungen mithilfe von LLMs erstellt und maschinell in einer Datenbank systematisch abgespeichert. Die KI ist in der Lage, ein Verständnis zum Inhalt der Zusammenfassungen zu entwickeln und Beziehungen durch Korrelationen zwischen den Inhalten aufzuzeigen. Daraus lassen sich zum einen Zusammenfassungen in verschiedenen Formen erstellen (z.B. externe Artikel) oder zum anderen durch das Zusammenspiel mit einem Recommendersystem Handlungsempfehlungen aussprechen. Je nach Komplexität des Recommendersystems kann dieses entweder regelbasiert aufgesetzt oder über ein domainspezifisches Training der KI angeeignet werden.

Mehrwerte für Behörden

LLMs ermöglichen automatisierte und präzise Textinterpretation und -verarbeitung, was zu einer erheblichen Zeitersparnis führt. Durch die schnelle Verarbeitung großer Mengen von Gesetzestexten können Folgedokumente wie Durchführungsanweisungen und Gesetzeskommentare effizienter und zeitnah aktualisiert werden. Eine technische Lösung zur Gesetzesaktualisierung mithilfe von KI trägt somit zur Effizienzsteigerung, Präzision und Aktualität von Behördenprozessen bei. Damit erhalten Mitarbeitende die notwendige zeitliche Entlastung, was zu reibungsloseren Abläufen, Minimierung von Wissenslücken und verbesserter Servicequalität beiträgt.

Abb. 4 –Use Case 03: Gesetzesaktualisierung mit KI-Unterstützung





Implementierung und Rahmenbedingungen

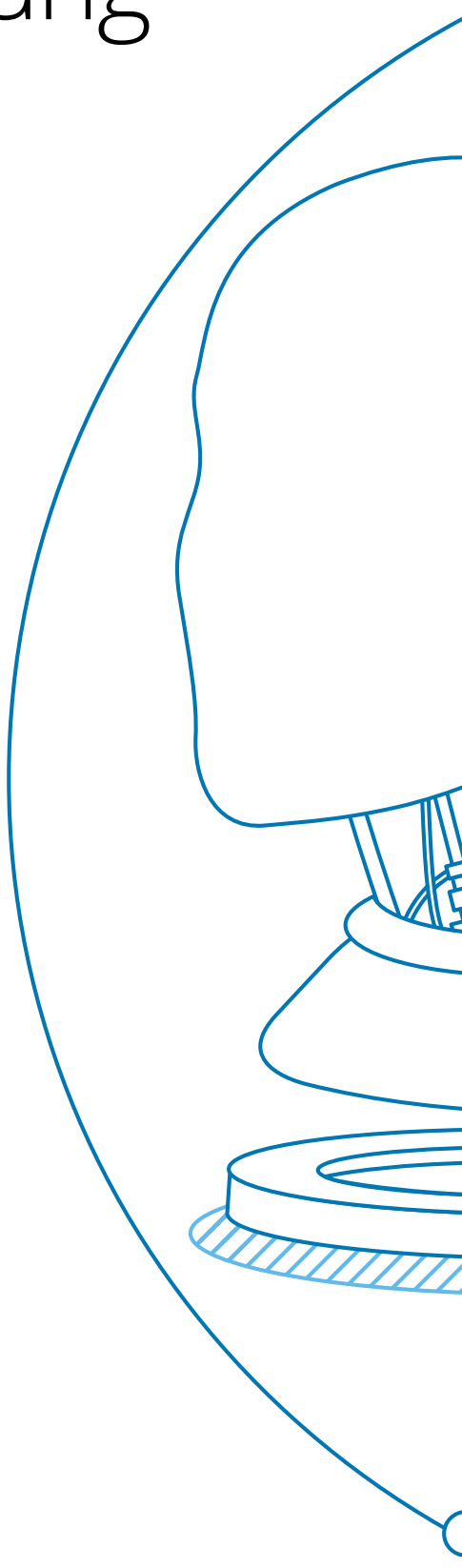
Bei der Pilotierung von GenAI Use Cases sollten Behörden auf bereits bestehende, vorkuratierte Sprachmodelle aufbauen (vgl. Use Case „Sprachmodell mit Fachwissen“). Im nächsten Schritt bedarf es des Finetunings für den Anwendungsfall in der Behörde, bei dem das Modell mit einem spezifischen, relevanten Datensatz trainiert wird, wie mit Gesetzestexten und Dokumenten, die sich bei Veränderungen anpassen. Je nach Anwendungsfall und den behördlichen Anforderungen kann es notwendig sein, zusätzliche Schritte einzufügen. Hierzu zählen unter anderem die Implementierung von Datenschutzmaßnahmen, die Anpassung des Modells an bestimmte technische Anforderungen oder dessen Integration in bestehende IT-Systeme und Datenpipelines sowie die organisatorischen Arbeitsabläufe und eine Schulung der Beschäftigten. Nach erfolgreicher Testung und Validierung des Modells bedarf es im nächsten Schritt der Schulung von Mitarbeitenden, bevor nach erfolgreicher Pilotierung die Skalierung und ggf. Weiterentwicklung beginnen können.

Abb. 5 – Entwicklungsprozess eines LLM

* Wird in der Regel nicht durch die Verwaltung durchgeführt.

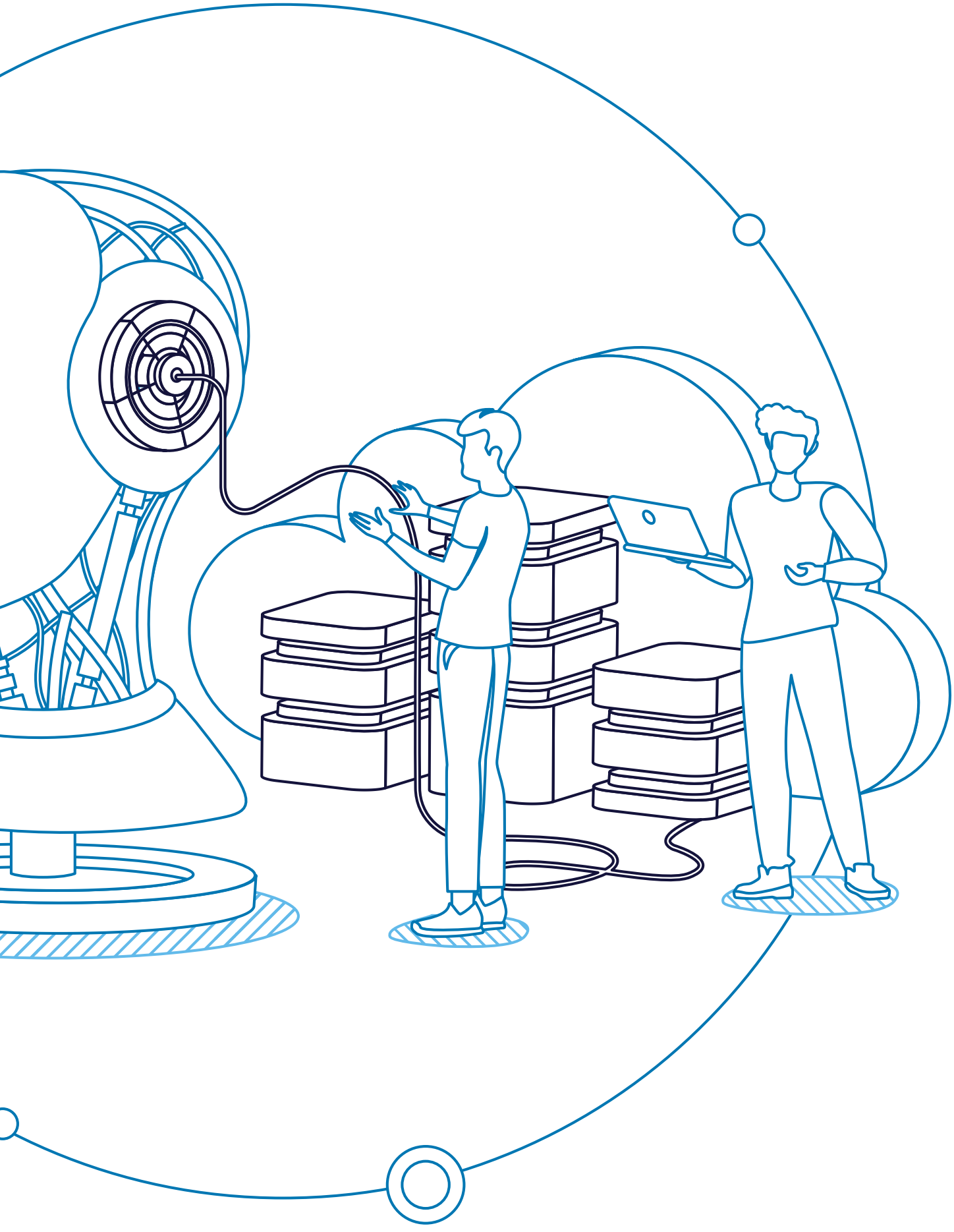
GenAI in der Verwaltung – Zukunft gestalten

Angesichts der großen Herausforderungen, vor denen die öffentliche Verwaltung in den nächsten Jahren steht, sollte sie den Potenzialen von GenAI und Large Language Models mit Offenheit begegnen. In diesem Briefing wurden drei Vorschläge präsentiert, wie sich Behörden bereits heute die Potenziale von GenAI zunutze machen können. Die Chancen, die diese Technologie mit sich bringt, können einen Beitrag leisten, den Rückstand der Verwaltungsmodernisierung in Deutschland aufzuholen und das Vertrauen der Bürgerinnen und Bürger in die Leistungsfähigkeit des Staates wiederherzustellen. Der Einstieg in diese Technologie gelingt über einzelne Anwendungsfälle. Die Skalierbarkeit für eine echte Transformation innerhalb der Verwaltung kommt nur durch eine erfolgreiche Verzahnung von Implementierung und Betrieb auf einer leistungsstarken Infrastruktur zustande. – Mit geeigneten Ansätzen zur Skalierbarkeit von KI-Anwendungen befasst sich das nächste GenAI-Briefing.



„Wissensarbeit und Prozessumsetzung prägen die Verwaltung. GenAI bietet die Chance, beides neu zu gestalten und so die Leistungsfähigkeit des Staates zu sichern."

Felix Dinnessen, Partner, Government & Public Services



Anhang

Funktionsweise von LLMs am Beispiel von GPT-Modellen

Während der Fokus bisheriger KI-Modelle auf einzelnen Anwendungen lag, sind in den vergangenen Jahren Modelle aufgekommen, die weitaus komplexeren Output genieren können. Vom Erstellen umfassender Texte (z.B. ChatGPT) über Bild- (Dall-E-2, Midjourney etc.) und Videogeneratoren (z.B. Pictory, Synthesia) bis hin zur Kreierung neuer Musik[4]: KI ist längst in der Lage komplexere Problemstellungen zu lösen, wenn eine entsprechende Datengrundlage vorhanden ist. Dabei gibt es unterschiedliche Methoden und Techniken, um ein KI-Modell für eine Problemlösung zu entwickeln.

Die Basis vieler großer Sprachmodelle sind Methoden des Machine Learning (ML). Im Gegensatz zu einer Simulation, bei der ein fertiges Modell bekannt ist, ist beim ML das finale KI-Modell zunächst unklar. Abbildung 6 zeigt den Prozess des ML als Adaption einer Darstellung von Von Rüden et al. (2020, 551)[5]: Man beginnt mit einem Datensatz und trainiert diesen auf eine vorläufige Modell-Architektur (Hypothesis Set), man bekommt einen ersten Algorithmus und optimiert diesen bis zum finalen KI-Modell (Final Hypothesis). Abbildung 7 zeigt das klassische Vorgehen bei einer Programmierung, ebenfalls nach Von Rüden et al. (2020, 552)[6]. Diese beruht auf der Grundlage einer Simulation. Man wählt ein bekanntes Modell, versorgt dieses mit entsprechenden Parametern und erhält ein simuliertes Ergebnis. Die beiden Abbildungen verdeutlichen den signifikanten Unterschied zwischen ML und klassischem Programmieren. Beim ML beginnt man auf einer Basis aus Daten und erhält am Ende ein KI-Modell, welches wiederum Daten verarbeiten kann. Dieser Entwicklungsprozess des ML veranschaulicht die Notwendigkeit qualitativer Daten für ein hochwertiges Modell. Die Daten sind der Ursprung, aus dem es entsteht.

Abb. 6 – Phasen einer datengetriebenen Modellierung

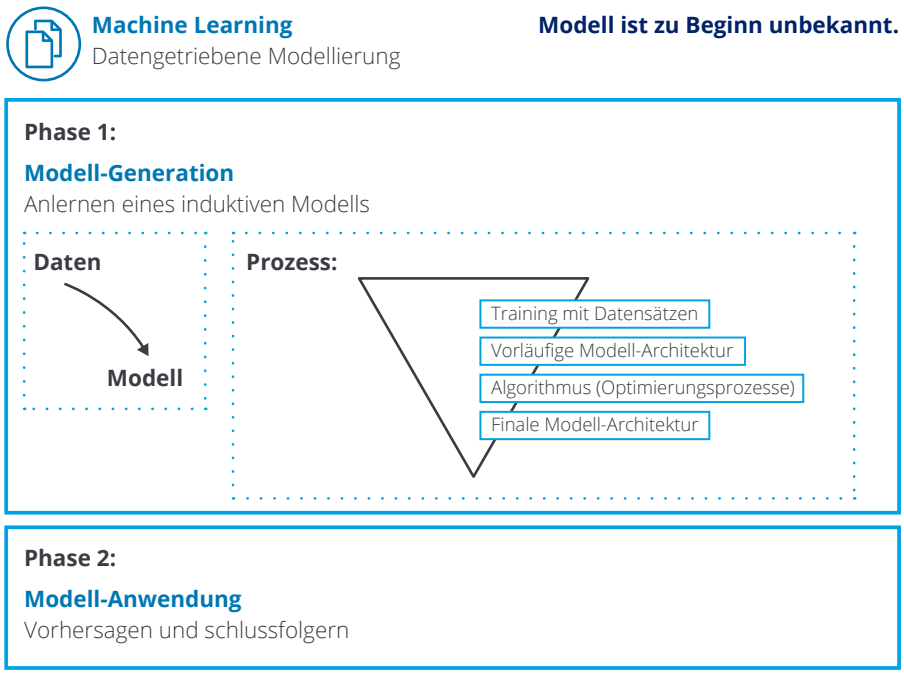
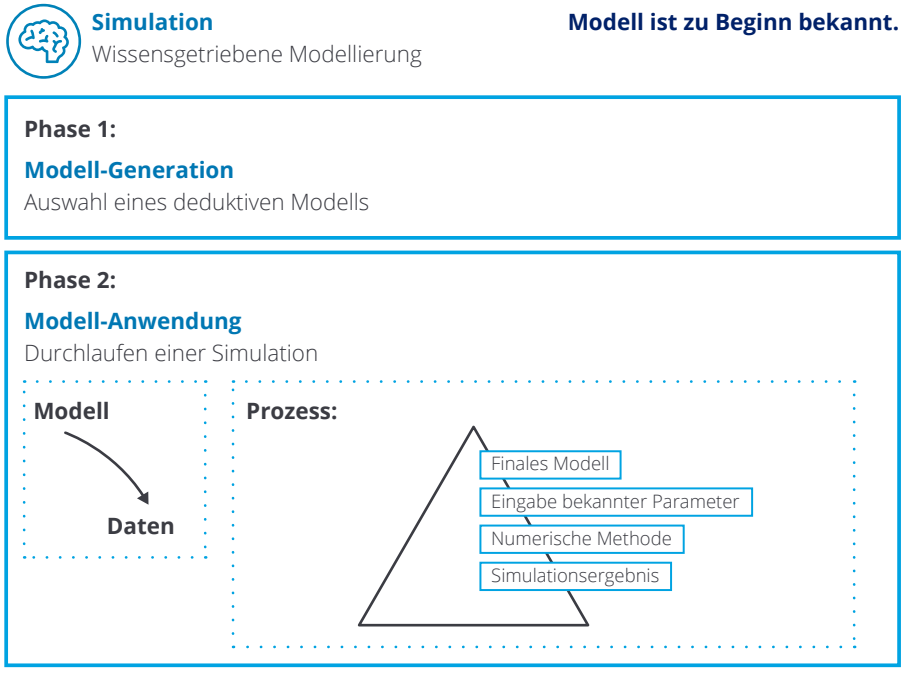
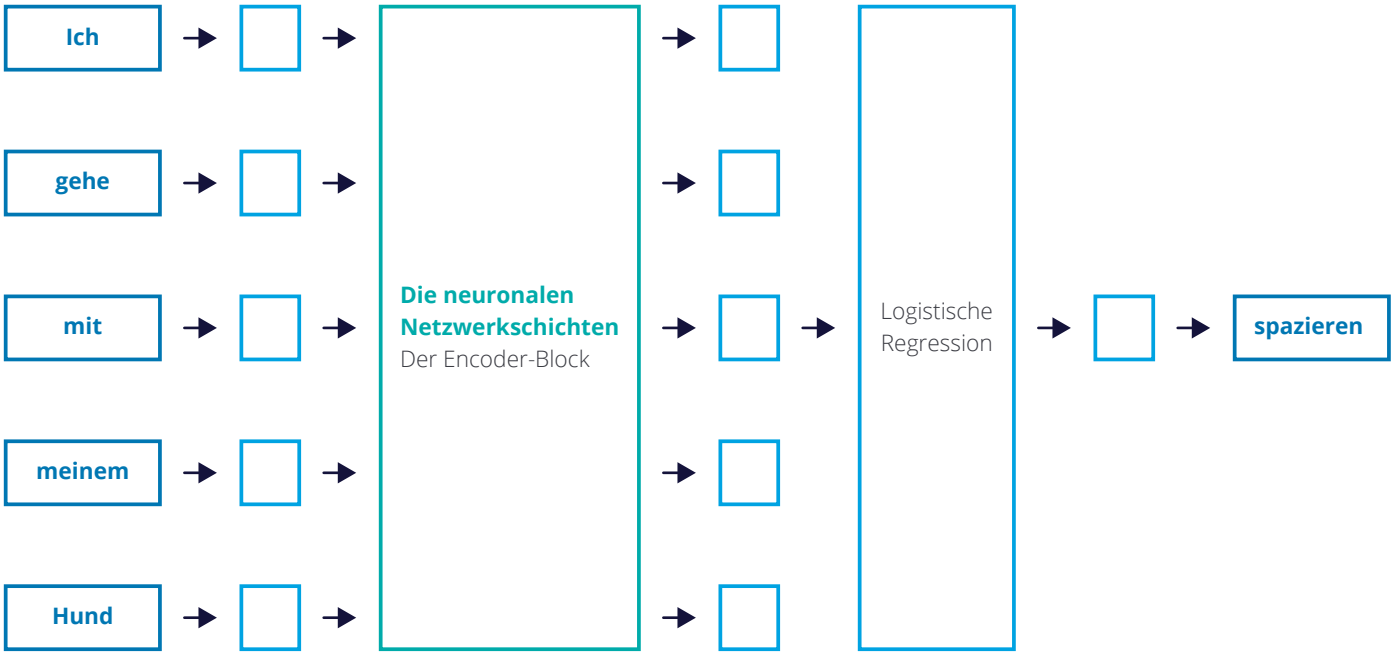


Abb. 7 – Phasen einer wissensgetriebenen Modellierung



[4] https://www.faz.net/aktuell/wirtschaft/digitec/ki-generiert-fake-song-von-drake-und-the-weeknd-18829126.html, abgerufen am 25.11.2023.
[5] Von Rueden, Laura, et al.: "Combining machine learning and simulation to a hybrid modelling approach: Current and future directions." Advances in Intelligent Data Analysis XVIII: 18th International Symposium on Intelligent Data Analysis, IDA 2020, Konstanz, Germany, April 27–29, 2020, Proceedings 18. Springer International Publishing, 2020, S. 551–552.
[6] Ebd.

Abb. 8 – Tokenisierung



Es gibt unterschiedliche Methoden im ML, wie man von den Daten zum finalen KI-Modell gelangen kann, und unterschiedliche Strukturen, die eine Modell-Architektur aufweisen kann. Die Methoden variieren nach den Daten, z.B. deren Umfang, Struktur und Formaten. Die spätere Struktur orientiert sich in der Regel an der finalen Aufgabe des Modells. Bei LLMs sind bspw. künstliche neuronale Netzwerke (KNNs), in unterschiedlichen Formen eine weit verbreitete Struktur, so zum Beispiel bei BERT[7] und weiteren GPT-Modellen[8] wie ChatGPT und OpenGPT-X. Ein KNN in einem KI-Modell ist dabei der Funktionsweise eines realen neuronalen Netzes aus der Biologie nachempfunden und stellt eine komplexe Verbindung unzähliger Parameter dar. Eine Informationseingabe in ein solches KI-Modell wird in ihre Bestandteile aufgeschlüsselt, durchläuft die unterschiedlichen Schichten des KNN und ermöglicht so die Ausgabe von neu generierten Inhalten.

[7] BERT (Bidirectional Encoder Representations from Transformers) ist ein Modell für die Verarbeitung natürlicher Sprache und in Fachkreisen ein beliebtes Beispiel für die Verwendung der Transformer-Architektur. Es wurde, ebenfalls wie die zuvor genannte Konversations-KI Bard, von Google entwickelt. Doch im Gegensatz zu Bard, welches eine vollständige KI-Anwendung darstellt, wird BERT in der Regel wie ein Bauteil verwendet, das erst mit anderen Elementen vereint eine fertige Anwendung ergibt.

[8] Paaß, Gerhard; Giesselbach, Sven: "Foundation Models for Natural Language Processing Pre-trained Language Models Integrating Media." arXiv preprint arXiv:2302.08575 (2023).

Training und Funktionsweise von KNN, wie sie in GPT-Modellen vorkommen

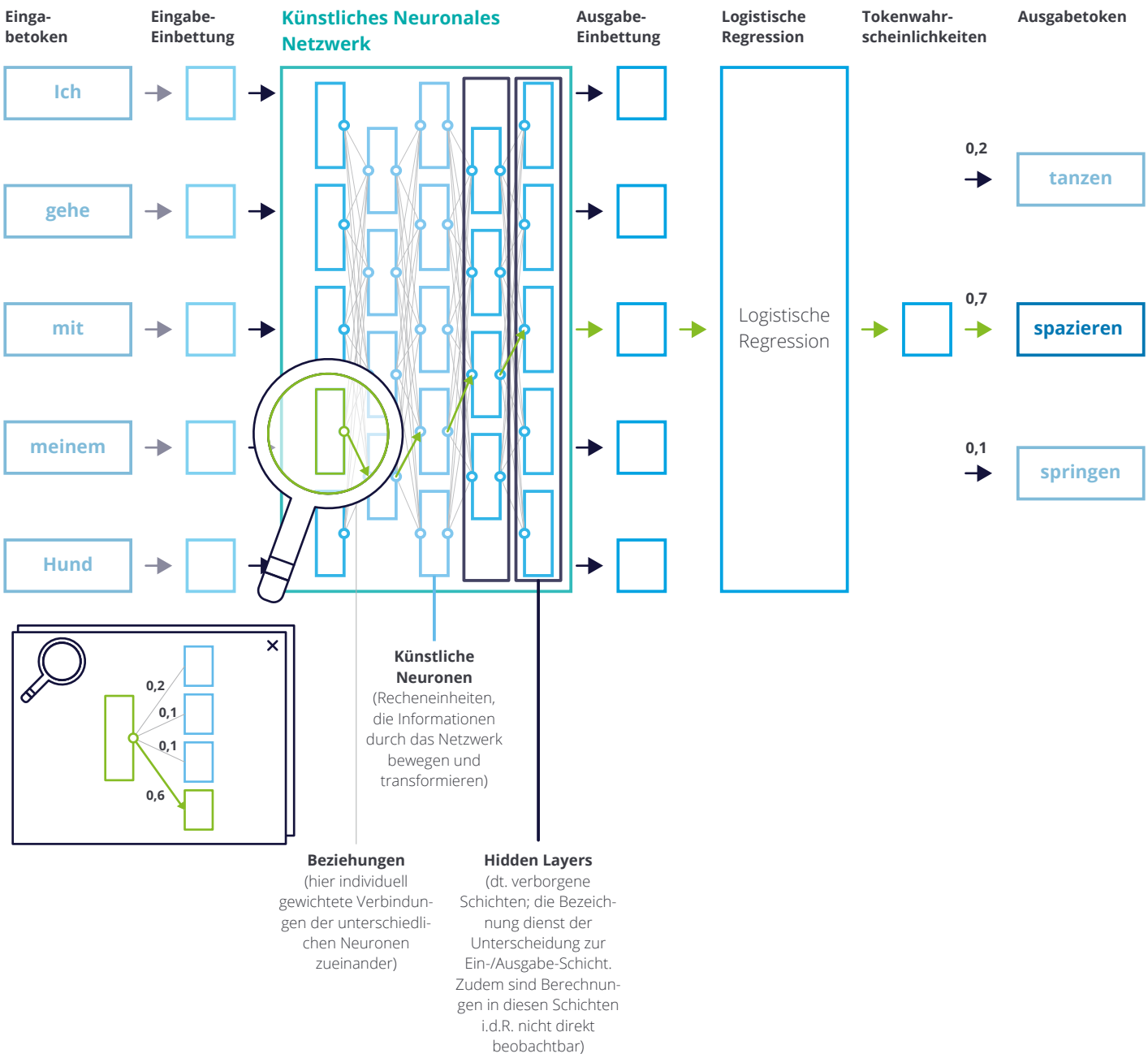
Damit ein LLM die Muster einer Sprache erlernt und ein NN entwickelt, wie bspw. GPT-Modelle das tun, trainiert man ein vorläufiges Modell mit einer großen Menge an Textdaten. Das Training funktioniert in diesem Beispiel mittels Tokenisierung (Abb. 8). Dabei werden die eingegebenen Daten, hier in Form von Texten, in kleinere Einheiten aufgespalten, welche dann als „Token" bezeichnet werden. Oftmals stellt ein Wort oder Satzzeichen einen Token dar. Im Training lernt das Modell die Beziehung zwischen den einzelnen Token. Diese Beziehungen stellen sich in Form eines KNN dar.

Abb. 8, als Adaption nach Gerhard Paaß und Sven Giesselbach (2023, 38)[9], veranschaulicht die Tokenisierung einer Eingabe in ein GPT-Modell. Für das Modell stellt sich die Eingabe als Token dar. In der Eingabe-Einbettung werden die Token in Eingabevektoren umgewandelt. Erst in diesem Format sind sie für das KNN verarbeitbar. Es folgt eine Berechnung der kontextuellen Beziehungen der eingegebenen Token. Im Rahmen dieser Berechnung kommt es zu einer Einordnung in die unterschiedlichen Schichten des KNN. Das letzte Token findet sich dabei in der höchsten Schicht wieder. In Kombination mit den gesamten eingeordneten Eingabetoken lässt sich durch ihre bekannten Beziehungen, die zuvor gelernten Muster, die Wahrscheinlichkeit für das folgende Ausgabetoken ermitteln. (Paaß und Giesselbach 2023, 38).

Während Abbildung 8 den Prozess der Tokenisierung darstellt, zeigt Abbildung 9 eine vereinfachte Übersicht der Struktur eines KNN, mit seinen Knotenpunkten und deren individuellen Beziehungen zueinander. In dem verwendeten Beispiel stellt der Satzanfang „Ich gehe mit meinem Hund" die Eingabe von Daten in Form von Token dar. Im nächsten Schritt findet eine Transformation der Token in Eingabevektoren statt, erst in diesem Format kann das KNN die Informationen verarbeiten. Nun durchlaufen die Informationen das KNN. Abbildung 9 veranschaulicht den Aufbau des Netzwerks. Es besteht aus künstlichen Neuronen, auch Knotenpunkte genannt. Diese sind die grundlegenden Recheneinheiten des KNN, welche sich wiederum aus einem gewichteten Eingabevektor (umgewandelter Token, Tokensequenz), einer Aktivierungsfunktion (Anwendung der gewichteten Summe der Eingabe) und einem Ausgabewert zusammensetzen. Ein Token ist ab diesem Punkt nur ein Wert in einer Funktion, der in Kombination mit anderen Informationen zu einem Ergebnis führt, welches dann weiterverarbeitet werden kann. Die Neuronen sind durch gewichtete Beziehungen miteinander verbunden. Diese gewichteten Beziehungen entstehen im Modelltraining, jede einzelne Verbindung, von Neuron zu Neuron, hat eine individuelle Wahrscheinlichkeit für eine mögliche Kombination miteinander. In dem verwendeten Beispiel ergeben die gewichteten Beziehungen der Neuronen, dass bei der Eingabe von Ich, gehe, mit, meinem und Hund in dieser Konstellation zueinander die Wahrscheinlichkeit für eine folgende Kombination mit dem Wort „spazieren" die höchste ist. Nachdem die Eingabe-Informationen das KNN durchlaufen haben, wird die so entstandene numerische Ausgabe in der Ausgabe-Einbettung in das Tokenformat zurücktransformiert. Bei dem dargestellten GPT-Modell erfolgt nun eine logistische Regression. Diese ermöglicht das gleichzeitige Erfassen komplexer Merkmale und das Generieren einer Vorhersage. Das in den Abbildungen 8 und 9 dargestellte KNN ist ein Encoder-Block, das bedeutet, es hat die Aufgabe, die Bedeutung und den Kontext der Eingabe zu erfassen. Viele LLMs bestehen aus einem Netz von Encoder- und Decoder-Blöcken, also mehreren KNNs mit unterschiedlichen Aufgaben. Ein Decoder-Block würde, auf Grundlage der Informationen von einem oder mehreren Encoder-Blöcken, eine Ausgabe generieren. Bei Paaß und Giesselbach (2023, 38) folgt auf die logistische Regression die Bestimmung der Tokenwahrscheinlichkeit mit anschließender Ausgabe. Dieser Vorgang dient der Verarbeitung und Umwandlung der Informationen des KNN, ähnlich wie in einem Decoder-Block. Am Ende des Verarbeitungsprozesses steht die Ausgabe des Wortes „spazieren". Die Tokenisierung wird während des Trainings des Modells, wenn die einzelnen Beziehungen entstehen und gewichtet werden, wie auch bei der späteren Anwendung verwendet. Auf diese Weise kann Sprache verstanden, vorhergesagt und generiert werden.

[9] Ebd

Abb. 9 – Aufbau Künstliches Neuronales Netzwerk



Ihre Ansprechpersonen

Felix Dinnessen
Partner Government & Public Services
Tel: +49 221 8772 3721
fdinnessen@deloitte.de

Dr. Björn Bringmann
Managing Director
Lead AI Institute
Tel: +49 89 29036 6131
bbringmann@deloitte.de

Dr. David Dang
Enterprise AI Leader
Tel: +49 89 29036 6385
dadang@deloitte.de

Prof. Dr. Rafet Sifa
Abteilungsleiter Media Engineering
Geschäftsfeldleiter Cognitive Business Optimization
Tel: +49 2241 142405
rafet.sifa@iais.fraunhofer.de

Sandra Halscheidt
KI-Business-Developerin
Cognitive Business Optimization
Tel: +49 170 2179804
sandra.halscheidt@iais.fraunhofer.de

Unter Mitwirkung von:
Jana Lilian Birr (Fraunhofer IAIS), Tabea Weiss, Julius Sicken, Jonathan Kreilaus und Marc Hermanns

Deloitte.

Stand 03/2024