Reinking, Ernst; Becker, Marco

Working Paper

Einsatzmöglichkeiten von KI in Unternehmen - Zeitnah erreichbare Personalisierung von Large Language Models (wie ChatGPT) in Zeiten der Industrie 5.0

IUCF Working Paper, No. 9/2023

Suggested Citation: Reinking, Ernst; Becker, Marco (2023) : Einsatzmöglichkeiten von KI in Unternehmen - Zeitnah erreichbare Personalisierung von Large Language Models (wie ChatGPT) in Zeiten der Industrie 5.0, IUCF Working Paper, No. 9/2023, ZBW - Leibniz Information Centre for Economics, Kiel, Hamburg

This Version is available at:
https://hdl.handle.net/10419/275737

Ernst Reinking, Marco Becker[1]

Einsatzmöglichkeiten von KI in Unternehmen

Zeitnah erreichbare Personalisierung von
Large Language Models (wie ChatGPT) in Zeiten der Industrie 5.0

Zusammenfassung

Mit der Einführung von ChatGPT als eines der bekanntesten Large Language Models wurde in der allgemeinen Wahrnehmung nicht nur ein neues Kapitel der Künstlichen Intelligenz aufgeschlagen - manche Autoren sprechen sogar von einer Ära der (Wirtschafts-)Informatik. Sie läutet auch die fünfte industrielle Revolution (Industrie 5.0) ein.

Ziel dieses Working Papers ist es, nicht nur den Widerspruch zwischen Hype und Realität im Kontext der Künstlichen Intelligenz zu versachlichen, sondern auch die Chancen und Perspektiven für die Analyse unstrukturierter, unternehmensinterner Daten aufzuzeigen. Dazu haben die Autoren auf Basis eigener Forschungsarbeiten mehrere Prototypen entwickelt, auf denen die Grundaussagen dieses Working Papers basieren.

Abstract

The introduction of ChatGPT as one of the best-known Large Language Models not only opened a new chapter in artificial intelligence in the general perception – some authors even speak of an era of (business) informatics. It also heralds the fifth industrial revolution (Industry 5.0).

The aim of this working paper is not only to objectify the contradiction between hype and reality in the context of artificial intelligence, but also to show the opportunities and perspectives for the analysis of unstructured, internal company data. To this end, the authors have developed several prototypes based on their own research work, on which the basic statements of this working paper are based.

[1] **Dipl.-Ing. Ernst Reinking** lehrt und forscht als Research Fellow an der NBS Northern Business School – University auf Applied Science in Hamburg zu Themenbereichen Wirtschaftsinformatik, digitale Ökonomie und Künstliche Intelligenz.
Prof. Dr. Marco Becker lehrt und forscht zu den Themen Controlling und Finanzmanagement an der NBS Northern Business School – University auf Applied Science in Hamburg und ist stellvertretender Leiter des Instituts für Unternehmensrechnung, Controlling und Finanzmanagement (IUCF). Darüber hinaus ist er Gründer und Partner der Marco Becker Management Consultants.

Ausgangspunkt: Aktueller Hype um Künstliche Intelligenz und ChatGBT

Der 28. November 2022 markiert den Beginn der fünften industriellen Revolution (Industrie 5.0) und läutet damit das Zeitalter der Künstlichen Intelligenz (KI) ein.[2] An diesem Tag veröffentlichte die Firma OpenAI ihre Anwendung ChatGPT, ein Chat-basiertes Large Language Model (LLM), d.h. ein komplexes Modell der künstlichen Intelligenz, das auf tiefen neuronalen Netzen basiert und entwickelt wurde, um komplexe Aufgaben im Bereich der Verarbeitung natürlicher Sprache zu lösen. Solche Modelle werden als "large" bezeichnet, weil sie aus einer beträchtlichen Anzahl von Parametern bestehen (GPT3 175 Mrd., GPT4 unveröffentlicht, geschätzt 1.000 Mrd.), was ihnen eine beeindruckende Fähigkeit verleiht, menschliche Sprachtexte zu erzeugen und zu verstehen.[3]

Die Anfänge der KI reichen bis in die 1950er Jahre zurück, aber nach einer sehr wechselvollen Geschichte erlebt sie erst in den letzten Jahren einen stetigen Bedeutungszuwachs, insbesondere durch die Verfügbarkeit sehr leistungsfähiger Hardware und begleitet von der Entwicklung neuer Algorithmen.[4] So ist auch GPT3 (auf dem ChatGPT basiert) ein evolutionärer Schritt. Allerdings wurde hier erstmals ein KI-Sprachmodell mit einer unvorstellbar großen Datenmenge und einem gigantischen Rechenaufwand über Jahre hinweg trainiert. Vom Ergebnis und der Leistungsfähigkeit waren selbst die Entwickler zunächst völlig überrascht. Denn das System basiert lediglich darauf, in einem n-dimensionalen Vektorraum gespeicherte Wörter mit Hilfe weniger Algorithmen nach den Prinzipien von Statistik und Wahrscheinlichkeit zu durchsuchen und Wort für Wort das nächste zur Frage passende Wort zu finden und als Antwort aneinanderzureihen. Dies geht so weit, dass das Modell, obwohl es sich um ein reines Textmodell handelt, auch einfache Fragen zu Programmiercode oder manchmal auch einfache mathematische Aufgaben richtig lösen kann.[5]

Die öffentliche Bereitstellung von ChatGPT und die entsprechende mediale Begleitung löste einen beispiellosen IT-Hype aus. Innerhalb von nur 5 Tagen überschritt die Zahl der Nutzer die Marke von 1 Million und ChatGPT mutierte in der öffentlichen Wahrnehmung zu einem allwissenden Werkzeug. Ein extremer Wettbewerb zwischen den großen IT-Unternehmen, aber auch einer sehr großen und schnell wachsenden Open-Source-Community, in der viele Wissenschaftler zu finden sind, begann. Inzwischen steht neben ChatGPT/GPT eine fast stetig wachsende Zahl weiterer LLMs zur Verfügung. Ein immenses mediales und auch politisches Echo begleitet diese Entwicklung mit meist realitätsfernen Informationen.[6]

Hauptfunktionen von Large Language Models

Im Laufe der Jahrzahnte haben sich sehr vielfältige Ausprägungen der Künstlichen Intelligenz mit unterschiedlichen Ausprägungen entwickelt. Das von der BITKOM entwickelte Periodensystem der Künstlichen Intelligenz gibt einen zusammenfassenden Überblick:[7]

[2] Vgl. Becker/Daube et al. (2023).
[3] Vgl. Heaven (2023).
[4] Vgl. Taukki (2022): S. 6 ff.
[5] Vgl. Heaven (2023) und Zhang/Li (2021).
[6] Vgl. Heaven (2023) und Zhang/Li (2021).
[7] Vgl. Bitkom e.V. (2023).

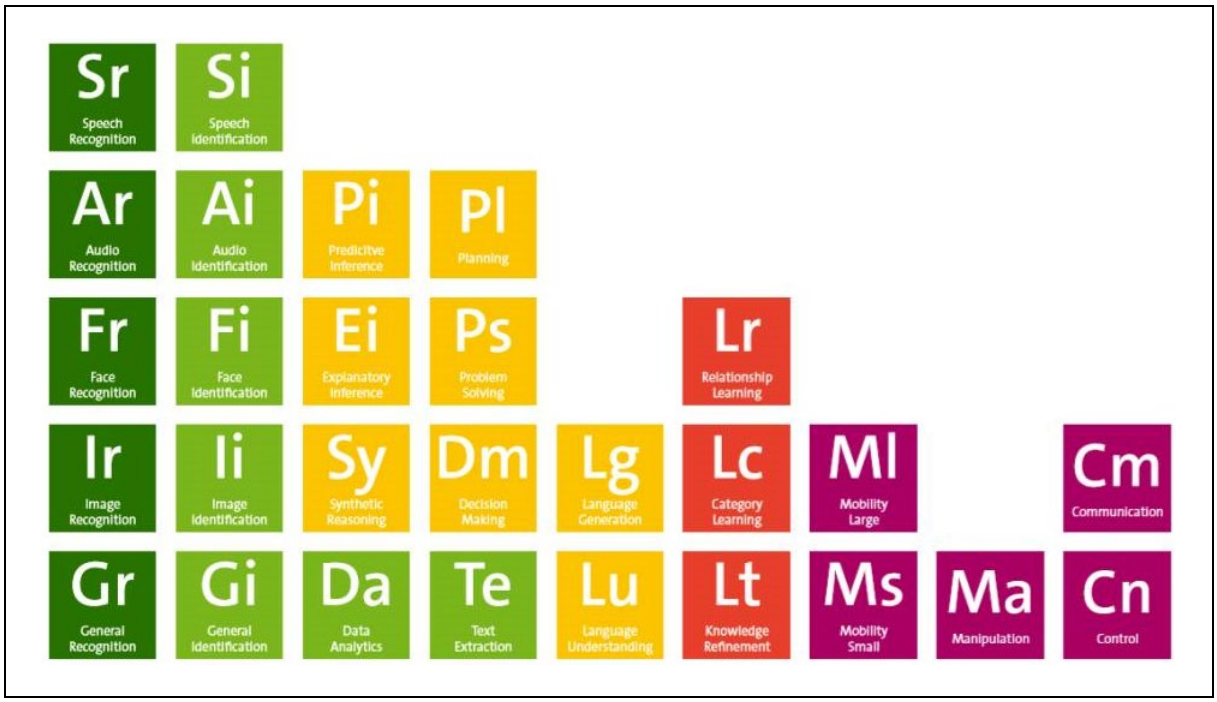


Abbildung 1: Periodensystem der Künstlichen Intelligenz[8]

Wie aus dem Periodensystem der Künstlichen Intelligenz ersichtlich ist, stellen LLM eine Kombination mehrerer Disziplinen (u.a. Te, Da, Te, Dm, Ps, Lu, Lg …) dar. Folgende Schlüsselfaktoren haben für LLMs dabei eine zentrale Bedeutung:[9]

- Textgenerierung:
  LLMs können Texte erzeugen, die dem menschlichen Schreibstil ähneln. Sie können in verschiedenen Kontexten eingesetzt werden, z.B. beim Verfassen von Nachrichtenartikeln, Marketingmaterial oder sogar literarischen Werken.

- Sprachverstehen:
  Diese Modelle können komplexe Anfragen interpretieren und darauf antworten. Sie sind in der Lage, die Bedeutung hinter Wörtern und Sätzen zu erfassen, was in Bereichen wie Kundendienst und virtuellen Assistenten nützlich ist.

- Sentiment-Analyse:
  LLMs können die Emotionen hinter einem Text erkennen und bewerten, ob er positiv, negativ oder neutral ist.

- Textübersetzung:
  Sie können Texte zwischen verschiedenen Sprachen übersetzen, ohne dass der Sinn verloren geht.

[8] Eigene Darstellung in Anlehnung an: Bitkom e.V. (2023).
[9] Vgl. DB Systel GmbH (2023) sowie eigene Überlegungen.

- Wissensextraktion:
  LLMs können Informationen aus großen Textmengen extrahieren und synthetisieren, was in Forschung und Analyse verwendet werden kann.

- Pretraining:
  Die großen LLMs sind in der Regel bereits mit einem sehr großen Allgemein-/ „Weltwissen“ trainiert, das Anwender nutzen können.

- Schnittstellen:
  LLMs verfügen über vielseitig nutzbare Programmschnittstellen über die andere Programme Fragen stellen und Antworten erhalten können. Dadurch können sie Teil anderer Programmumgebungen werden und so auch in Automatisierungen eingebunden werden.

LLMs bilden zentrale die Schnittstelle zwischen KI-Engine und Nutzer ab, wie dieses beispielsweise in ChatGPT erfolgt.[10]

Large Language Models in der Unternehmenswelt

LLMs haben das Potenzial, die Geschäftswelt grundlegend zu verändern, indem sie Prozesse automatisieren, das Kundenerlebnis verbessern und datengestützte Entscheidungen ermöglichen. LLM-Techniken und Funktionen der generativen KI werden auch in die heute verwendeten IT-Programme wie ERP, Office, Mailsysteme und andere Einzug halten. Unternehmen werden die Vorteile dieser fortschrittlichen Technologie erkennen, nutzen und als Wettbewerbsvorteil einsetzen.

Beispielhaft sollen hier einige Anwendungsgebiete genannt werden:[11]

- Kundeninteraktion:
  LLMs können in Chatbots und Kundendienstsystemen eingesetzt werden. Mit ihrer Fähigkeit, natürliche Sprache zu verstehen und darauf zu reagieren, bieten sie personalisierte und effiziente Interaktionen. Die Automatisierung des Kundensupports führt zu reduzierten Kosten und gesteigerter Kundenzufriedenheit.

- Datenanalyse und -verarbeitung:
  Die Analyse von großen Datenmengen ist für viele Unternehmen von entscheidender Bedeutung. LLMs können unstrukturierte Daten analysieren, Trends erkennen und wertvolle Einblicke in Geschäftsprozesse liefern. Dies fördert datengetriebene Entscheidungsfindung und Strategieentwicklung.

- Personalisierung von Marketingkampagnen:
  Die Fähigkeit von LLMs, individuelle Präferenzen und Verhaltensmuster zu erkennen, ermöglicht es Marketingteams, personalisierte Kampagnen zu erstellen. Dies erhöht die Engagement-Rate und den ROI (Return on Investment) von Marketingstrategien.

[10] Vgl. DB Systel GmbH (2023) sowie eigene Überlegungen.
[11] Vgl. DB Systel GmbH (2023) und Kucharavy/Schillaci et al. (2023): S. 35 ff.

- Automatisierte Übersetzungen:
  Mit LLMs ist es möglich, Texte schnell und präzise in verschiedene Sprachen zu übersetzen. Dies fördert die globale Expansion von Unternehmen und erleichtert die Kommunikation mit internationalen Kunden und Partnern.

Large Language Models und ihr »Wissen«

„Wenn das Unternehmen wüsste, was es alles weiß…“[12]

Das Erzeugen von Wissen aus unternehmensinternen Daten ist einer der zentralen Erfolgsfaktoren der Industrie 5.0. Die Schwierigkeit liegt dabei im Datenformat, da ein Großteil der relevanten Daten (ca. 80 % bis 90 %) in unstrukturierter Form vorliegen. Sie gehen damit für herkömmliche Auswertungen verloren.[13]

Strukturierte vs. unstrukturierte Daten
Sie stellen zwei grundlegende Kategorien von Daten dar, die in der Datenverarbeitung und -analyse verwendet werden. Sie unterscheiden sich in vielerlei Hinsicht. Um die Unterschiede zu verdeutlichen, folgt eine Gegenüberstellung:

Strukturierte Daten[14]

- Definition: Strukturierte Daten sind Daten, die in einer festen Struktur oder einem Schema organisiert sind. Sie folgen einer spezifischen Ordnung und einem festen Format.
- Beispiele: Datenbanktabellen, Excel-Tabellen, CSV-Dateien.
- Format: Organisiert in Zeilen und Spalten, wobei jede Spalte einen bestimmten Datentyp (z.B. Ganzzahl, Zeichenfolge) repräsentiert.
- Analyse: Leicht zu analysieren und abzufragen, da sie in einem festen Schema organisiert sind.
- Speicherung: Oft in relationalen Datenbanken gespeichert.
- Größe: Im Allgemeinen kleiner als unstrukturierte Daten, da sie in einem festen Format gehalten werden.
- Anwendungen: ERP-Systeme, CRM-Systeme, Business Intelligence.

[12] Davenport/Prusak (1999).
[13] Vgl. MIT SLOAN SCHOOF OF MANAGEMENT (2023).
[14] Vgl. LexisNexis (2023).

Unstrukturierte Daten[15]

- Definition: Unstrukturierte Daten haben keine spezifische Struktur oder ein festgelegtes Schema. Sie können in verschiedenen Formaten und Typen vorliegen.
- Beispiele: E-Mails, Social-Media-Posts, Bilder, Videos, Textdokumente.
- Format: Kein festes Format. Kann in verschiedenen Formen und Typen vorkommen.
- Analyse: Schwieriger zu analysieren, da spezielle Tools und Techniken erforderlich sind, um die Daten zu verstehen (z.B. Textanalyse, Bilderkennung).
- Speicherung: Kann in verschiedenen Arten von Datenbanken oder Dateisystemen gespeichert werden, einschließlich NoSQL-Datenbanken.
- Größe: Oft umfangreicher als strukturierte Daten, da sie in vielen verschiedenen Formen und Formaten vorkommen können.
- Anwendungen: Big Data-Analysen, künstliche Intelligenz, Maschinelles Lernen, Social Media-Analyse.

Bei LLMs handelt es sich um Textmodelle. Sie sind ausschließlich für die Verarbeitung von unstrukturierten Textdaten geeignet. Bisher werden unstrukturierte Daten in Unternehmen in der Regel aus technischen Gründen nur in geringem Umfang verarbeitet bzw. ausgewertet. Ein LLM verspricht einen erheblichen Mehrwert und stellt eine lange gesuchte Lösung für die Analyse unstrukturierter Daten dar.[16]

Datenbasis des LLMs

Der Aufwand und damit die Kosten für die Erstellung eines LLM sind zu deutlich mehr als 90% den Daten zuzurechnen. Dies betrifft die Datenbeschaffung und -bereinigung ebenso wie die extrem rechenintensiven und langwierigen Lernprozesse des Modells. Für ChatGPT liegen hierzu keine belastbaren Kostenangaben vor, es ist jedoch allein für die Rechenzeiten der Anlernprozesse von einem höheren 2-stelligen Millionenbetrag auszugehen. Hinzu kommt die Aufbereitung der Daten, die über Jahre hinweg mit sehr viel menschlicher Arbeit erfolgt ist.[17]

Mit dem sehr wertvollen Wissen aus den beschriebenen Anlernprozessen steht ein LLM zur Nutzung zur Verfügung, das nur über ein – wenn auch sehr umfangreiches – Allgemeinwissen, ein „Weltwissen“ verfügt.[18] Sollen nun betriebliche Fragestellungen mit dem LLM beantwortet werden, so reicht dies allein nicht aus, sondern es ist zusätzlich Domänenwissen erforderlich:

[15] Vgl. LexisNexis (2023).
[16] Vgl. LexisNexis (2023) und Kucharavy/Schillaci et al. (2023): S. 4 ff.
[17] Vgl. DB Systel GmbH (2023).
[18] Vgl. DB Systel GmbH (2023).

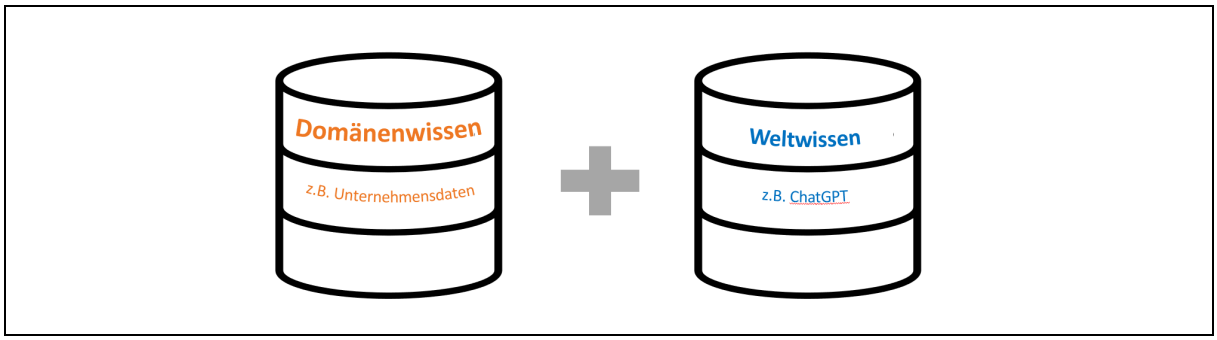


Abbildung 2: Domänenwissen und Weltwissen[19]

Möglichkeiten zur Verarbeitung von Unternehmensdaten in einem pretrained LLM

Grundsätzlich bestehen zwei Möglichkeiten Domänenwissen zu ergänzen, Fine-Tuning des LLMs und Content Injection.

Fine-Tuning

Beim Fine-Tuning wird das vortrainierte LLM-Modell in einer zusätzlichen Trainingsphase mit weiteren spezifischen Datensätzen weiter trainiert. Das Verfahren ist somit eine Fortsetzung des ursprünglichen Trainings und kann nur von dem durchgeführt werden, der auch über das Modell in Gänze verfügt.[20]

Im Falle von GPT bietet OpenAI diese Möglichkeit den Kunden gegen Entgelt an. Der Kunde muss seine Daten in einem vorgegebenen Format anliefern, OpenAI führt dann das aufwändige Training durch und hält anschließend eine eigene GPT-Version für den Kunden bereit. Dieses Verfahren ist mit sehr hohen einmaligen und laufenden Kosten verbunden. Es kommt, wenn überhaupt, nur für sehr langlebige Datenergänzungen in Frage, da das Modell selbst permanent verändert wird.[21]

Erste Erfahrungen zeigen aber auch, dass dieses Verfahren nur unzureichend funktioniert. Dies liegt daran, dass das Modell bereits auf einer großen Menge allgemeiner Textdaten trainiert wurde und die domänenspezifischen Daten in der Regel schon von der Menge her nicht ausreichen, um das bereits Gelernte zu überschreiben.

Content-Injection

Dieses Verfahren nutzt die Tatsache, dass die an das LLM gestellte Frage eine für normale Fragen recht komfortable Länge haben darf.[22]

Die Länge wird in sogenannten Token definiert. Die Tokenisierung ist der Prozess, bei dem ein Text in kleinere Teile oder "Token" zerlegt wird, damit er von einem Modell verarbeitet werden kann. Token sind unterschiedlich lang und haben eine empirische Durchschnittslänge von 3,5 Buchstaben. Im weiteren Verlauf der Verarbeitung werden Token im Übrigen in Zahlenwerte konvertiert.[23]

[19] Eigene Darstellung.
[20] Vgl. OpenAI (2023c) und Grosse/Hallgarten et al. (2023) und Wang/Li et al. (2019): S. 3.
[21] Vgl. OpenAI (2023c) und Grosse/Hallgarten et al. (2023).
[22] Vgl. Ferus (2022).
[23] Vgl. Ferus (2022).

Die folgende Tabelle enthält die unterschiedlichen Eingabelängen ausgewählter Modelle:

GPT-Modell	Technischer Name	Max. Token	~ DIN A4 Seiten*
GPT-4 (32k)	gpt-4-32k-0613	32.768	6,3
GPT-4 (8k)	gpt-4-0613	8.192	1,5
ChatGPT (16k)	gpt-3.5-turbo-16k-0613	16.384	3,0
ChatGPT (Standard – 4k)	gpt-3.5-turbo-0613	4.096	0,8

*) Eine DIN A4 Normseite entspricht 1.500 Zeichen inkl. Leerzeichen[24]

Abbildung 3: Maximale Dialoglänge in ausgewählten LLMs[25]

Bei der Content-Injection wird die Frage z.B. mit vorgeschalteten Programmen aufgebaut aus

- einem fest definierten Text
- mitgeliefertem Inhalt eigener Dokumente
- der eigentlichen Frage[26]

> *Beantworte die Frage am Ende des Textes anhand der folgenden Informationen. Wenn Du die Antwort nicht weißt, sage einfach, dass Du es nicht weißt, versuche nicht, eine Antwort zu erfinden.*

> *3.3. Umfang der Arbeiten*
>
> *Für Hausarbeiten wird ein Umfang von 15 Seiten und für Bachelorarbeiten eine Länge von 40 bis 60 Seiten als Regelfall definiert. Für eine Masterarbeit sind zwischen 60 und 80 Seiten einzuplanen. Der Umfang der Praktikumsarbeit beträgt 20 bis 30 Seiten. Wich-…….*

> Frage: Wie lang sollte eine Hausarbeit im Fach Wirtschaftsinformatik sein?

Fine-Tuning vs. Content-Injection

Fine-Tuning kommt nur für dauerhafte Modelländerungen in Frage, der Aufwand ist sehr hoch, die Wirkung bleibt gering, da die Datenmenge für das Fine-Tuning in der Regel sehr viel kleiner ist als für das Pretraining.[27]

Content-Injection kann sehr dynamisch von Query zu Query geändert werden, die Wirkung ist gut, aber die Menge des Contents pro Query ist sehr begrenzt.

[24] Vgl. VG WORT (2023): S. 7.
[25] Eigene Darstellung auf Basis von: OpenAI (2023a) und OpenAI (2023b).
[26] Vgl. Ferus (2022).
[27] Vgl. OpenAI (2023c).

Erweiterter Lösungsansatz für Content-Injection

Nachdem sich die Content-Injection-Methode als die derzeit mit Abstand sinnvollste Möglichkeit zur Integration eigener Daten in ein LLM darstellt, geht es bei allen Überlegungen darum, mit der begrenzten Menge an Content pro Query umzugehen.

Ein sehr vielversprechender Ansatz besteht darin, in einem vorgeschalteten eigenen KI-Modell jeweils die für die Beantwortung der Frage relevanten eigenen Dokumententeile zu ermitteln und nur diese Fragmente im Rahmen der zulässigen Tokenlänge als Content an das LLM zu liefern.

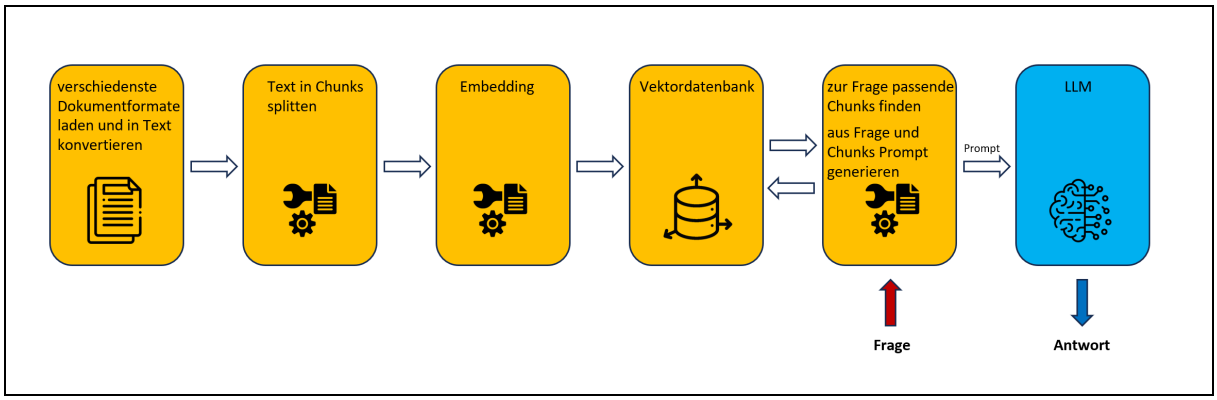


Abbildung 4: Beispielhafter Aufbau eines vorgeschalteten Modells[28]

Die zentralen Schritte des vorgeschalteten Modells werden im Folgenden kurz charakterisiert:

1. Dokumente laden und konvertieren:
   In einem ersten Schritt müssen die unterschiedlichen Dokumentformate mit ihren unstrukturierten Daten eingelesen werden. Sie reichen von .pdf über Textverarbeitungsformate (.doc, .docx, .odt, .txt, .rtf, .md, ...), Mailformate (.html, .eml, ...), Webtexten bis hin zu Formaten, die mit OCR-Texterkennung verarbeitet werden müssen. Sie werden in reines ASCII umgewandelt und von allen Steuerzeichen befreit.

2. Texte splitten:
   Der Text eines eingelesenen Dokuments wird in Teile, sogenannte Chunks, zerlegt, die je nach Einstellung einige hundert bis einige tausend Zeichen umfassen. Die Aufteilung erfolgt so, dass sich die Chunks um eine ebenfalls vorgegebene Anzahl von Zeichen überlappen. Damit soll erreicht werden, dass bei der technisch notwendigen Aufteilung nicht die inhaltliche Kontinuität bzw. der Sinn verloren geht oder verfälscht wird.[29]

3. Embedding:
   Beim Embedding geht es allgemein darum, die Texte der Chunks mit anderen Texten oder Chunks vergleichbar zu machen. Dies geschieht durch die Umwandlung der Texte in jeweils eindeutige numerische Vektoren. Dazu werden sogenannte Embedding-Modelle verwendet,

[28] Eigene Darstellung.
[29] Vgl. Lee (2023).

es existiert z.B. ein von OpenAI veröffentlichtes Embedding-Modell, das die Vektoren für einen 1536-dimensionalen Raum generiert.[30]

4. Vektordatenbank:
   Eine Vektordatenbank ist ein spezieller Typ von Datenbank, der dazu verwendet wird, Vektordaten zu speichern, zu organisieren und abzufragen. Eine häufige Anwendung von Vektordatenbanken ist die Ähnlichkeitssuche. Man kann eine Abfrage mit einem gegebenen Vektor starten und die Vektoren in der Datenbank finden, die diesem am ähnlichsten sind. Dies wird durch verschiedene Ähnlichkeits- oder Distanzmetriken wie Kosinus-Ähnlichkeit oder euklidische Distanz erreicht.

5. Relevante Chunks finden und Prompt generieren:
   In diesem Schritt wird die Anfrage des Benutzers entgegengenommen und nach vorheriger Aufbereitung eine Ähnlichkeitssuche gegen die in der Vektordatenbank gespeicherten Chunks gestartet. Das Ergebnis, der oder die zur Anfrage passenden Chunks, werden wiederum in Text umgewandelt und als Content zusammen mit der Anfrage wie oben beschrieben zu einem Prompt für das LLM zusammengefasst.[31]

Bewertung

Auf Basis der eigenen Forschung wurde mehrere Prototypen entwickelt werden. Mit Hilfe dieser selbst entwickelten IT-Systeme haben die Autoren nachgewiesen, dass der Umfang der an das LLM zu übermittelnden Inhalte ohne Informationsverlust minimiert werden kann. Damit kann ein LLM sinnvoll mit eigenen – auch unternehmensinternen – Daten genutzt werden.[32]

Dieser Ansatz bietet noch sehr viele Möglichkeiten zur Optimierung und Weiterentwicklung sowie die Adaption an spezifische Gegebenheiten und Aufgaben. Die Weiterentwicklung dieses Ansatzes steht im Fokus der aktuellen Forschung. So gibt es inzwischen erste Vektordatenbankmodelle in einer Größe und Leistungsfähigkeit, die es erlauben, alle unstrukturierten Daten eines Unternehmens grundsätzlich vorzuhalten und permanent auf einem aktuellen Stand zu halten, und nicht nur bei Bedarf zu erheben.

Erweiterungen

<u>Chains</u>
Eine aus der Programmierung bekannte Technik zur Strukturierung von Abläufen und Algorithmen ist die Chain, ein Verfahren, bei dem Tasks oder Operationen in einer festgelegten Reihenfolge ausgeführt werden. Diese Reihenfolge kann sowohl streng sequentiell (Sequential Chain) als auch konditional, d.h.

[30] Vgl. Lee (2023).
[31] Vgl. Lee (2023).
[32] Bei der Verarbeitung von unternehmensinternen Daten sind die allgemeinen Vorschriften zum Datenschutz (insbesondere DSGVO und BDSG) zu beachten. Darüber hinaus sind geeignete Schutzmaßnahmen zur Intelligence Protection (IP) sowie zur Einhaltung der Compliance-Vorschriften zu etablieren. Insbesondere die IP-Problematik ließe sich durch die lokale Installation eines LLMs lösen.

abhängig von Daten („if..then..else“) sein (Router Chain). Diese Verfahren lassen sich leicht in das oben beschriebene Modell integrieren.

Sequential Chain:
Als Beispiel betrachten wir folgenden Ablauf:

- Anfrage an das LLM zu einem bestimmten Thema
- Anfrage an das LLM, die Antwort auf das Thema in eine andere Sprache zu übersetzen
- Speichern der Übersetzung auf der Festplatte

In eine Chain werden alle Schritte automatisch in der richtigen Reihenfolge und konsistent ausgeführt. Es können auch andere Ressourcen als das LLM einbezogen werden, sofern das Programm selbst Zugriff darauf hat.

Um ein weiteres Beispiel zu geben: Es ist auch möglich, mehrere Schritte in einem Folgeschritt zusammenzufassen:

- Anfrage an das LLM zu Thema 1
- Anfrage an das LLM zu Thema 2
- Anfrage an das LLM zu Thema 3
- Anfrage an das LLM: „Antworten zu 1 bis 3 zusammenfassen“.

Router Chain:
Eine Router Chain ist ähnlich wie eine sequentielle Kette, aber anstatt, dass jede Operation nacheinander ausgeführt wird, bestimmt eine Routing-Entscheidung, welche Operation als nächstes ausgeführt wird. Auch hier ein Beispiel:

- Schritt 1: Abfrage an LLM zu einem Thema als Sentimentanalyse (Ergebnis positiv, neutral oder negativ)
- Schritt 2: (wird nur ausführt, wenn Antwort zu Schritt1 = positiv
- Schritt 3: (wird nur ausgeführt, wenn Antwort zu Schritt1 nicht positiv)

Agenten

- Ein Agent in der Künstlichen Intelligenz (KI) ist eine autonome Einheit, die Aktionen in einer Umgebung ausführt, um ein bestimmtes **Ziel** zu erreichen.
- Im Kontext von Sequential Chains oder Router Chains kann ein Agent die Rolle einer Einheit übernehmen, die durch die Kette navigiert und Aktionen ausführt. Beispielsweise könnte ein Agent in einer Router Chain die Ausgabe einer vorhergehenden Operation empfangen, eine Entscheidung darüber treffen, welche Operation als nächste ausgeführt werden soll, und dann die entsprechenden Daten an diese Operation senden.
- Agenten können unterschiedlich komplex sein. Einfache Agenten folgen festgelegten Regeln oder Algorithmen, um Entscheidungen zu treffen. Zielorientiert arbeitende Agenden, die auf Machine Learning oder Deep Learning basieren, sind wesentlich komplexer und in der Lage, aus Erfahrungen zu lernen und ihre Strategien im Laufe der Zeit anzupassen. Agenten können auch miteinander interagieren und in einer Umgebung mit mehreren Agenten voneinander lernen.
- Das oben beschriebene Modell kann ohne großen Aufwand durch einfache Agenten ergänzt werden. Die Herausforderung besteht darin, einem solchen autonom agierenden Agenten

ausreichende Grenzen zu setzen, so dass es nicht zu unvorhersehbarem Verhalten kommen kann.

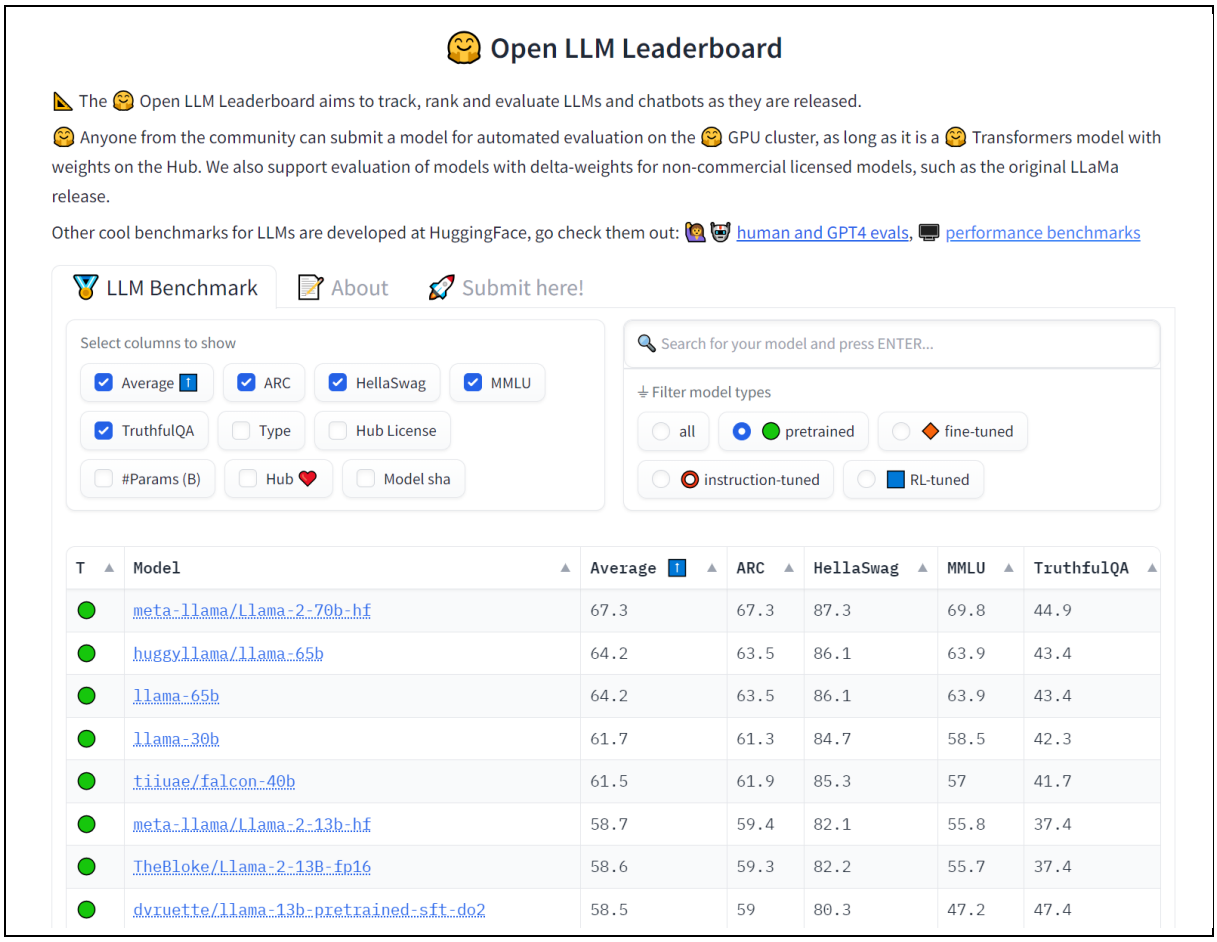
- Agenten, die ihre Strategien auf Basis von Rückmeldungen eines LLMs treffen, sind grundsätzlich fehlerbehaftet, da ihre Ergebnisse auf Wahrscheinlichkeiten basieren. Der Einsatz dieser Agenden kann derzeit nur unter menschlicher Kontrolle verantwortet werden.
- Agenten sind ein aktives Forschungsgebiet. Es gibt laufende Bemühungen, Methoden und Techniken zu entwickeln, um den genannten Herausforderungen zu begegnen. Zusätzlich werden Vorschriften und Standards diskutiert, um KI-Agenten sicher, fair und verantwortungsvoll einzusetzen.

Entwicklungsstand von Large Language Models

In der Vergangenheit wurden wesentliche Entwicklungen in wissenschaftlichen Fachbeiträgen veröffentlicht. ChatGPT bildet hier die erste Ausnahme. Dennoch haben alle großen IT-Unternehmen in dem sich nun entwickelnden Wettbewerb sehr ähnliche Startbedingungen. So finden LLM-Entwicklungen in den meisten namhaften Software-Unternehmen statt. Unter den kommerziellen Anbietern sind heute in erster Linie OpenAI (mit Microsoft als Lizenznehmer) und Google mit jeweils mehreren Modellen, Meta mit mehreren kleineren Modellen auch als Open Source sowie Produkte von Start-ups, die jeweils von ehemaligen OpenAI- und Google-Entwicklern gegründet wurden (z.B. Anthropic), zu nennen.

Gleichzeitig entwickelt sich eine leistungsfähige Open-Source-Community, an der u. a. auch renommierte Forschungseinrichtungen beteiligt sind. Als erstes Unternehmen hat Meta sein KI-Modell der KI-Community zur Verfügung gestellt. Mittlerweise haben sich bereits zahlreiche Varianten entwickelt. Das gerade veröffentlichte, derzeit leistungsstärkste Open-Source-LLM Falcon-40B kommt mit seinen Leistungswerten inzwischen sehr nahe an die kommerzielle Spitzengruppe heran. Die Community-Website huggingface.co veröffentlicht ein Leaderboard der Open-Source-LLMs. Hier waren zum Zeitpunkt der Abfrage 101 vortrainierte Modelle gelistet und herunterladbar.[33]

[33] HUGGINGFACE (2023).



T	Model	Average	ARC	HellaSwag	MMLU	TruthfulQA
	meta-llama/Llama-2-70b-hf	67.3	67.3	87.3	69.8	44.9
	huggyllama/llama-65b	64.2	63.5	86.1	63.9	43.4
	llama-65b	64.2	63.5	86.1	63.9	43.4
	llama-30b	61.7	61.3	84.7	58.5	42.3
	tiiuae/falcon-40b	61.5	61.9	85.3	57	41.7
	meta-llama/Llama-2-13b-hf	58.7	59.4	82.1	55.8	37.4
	TheBloke/Llama-2-13B-fp16	58.6	59.3	82.2	55.7	37.4
	dvruette/llama-13b-pretrained-sft-do2	58.5	59	80.3	47.2	47.4

Abbildung 5: Open LLM Leaderboard[34]

Die Quantisierung von Modellen ist ein sehr aktives Forschungsgebiet. In der Informatik werden Fließkommazahlen einfacher Genauigkeit in einem Datenfeld fester Länge von 32 Bit gespeichert. Für ein einfaches Modell mit 1 Milliarde Parametern benötigt einen (Haupt-)Speicherbedarf von 4 GB. Wird das Zahlenformat auf 8bit reduziert, so verringert sich der Speicherbedarf auf 1 GB. Gleichzeitig verbessern sich Rechenzeit und Anlernzeit in ähnlichem Maße. Somit lassen sich LLMs auch auf verbreiteten Servern betreiben.

Durch die Veränderung des Zahlenformats verringert sich die Genauigkeit der dargestellten Zahl. In der Anwendung von LLMs wirkt sich dies aber deutlich weniger auf die Performance aus, als erwartet. In der Liste der Open-Source-Modelle finden sich mehrere gleiche Modelle mit jeweils unterschiedlichen Datenformaten.

[34] HUGGINGFACE (2023).

Literaturverzeichnis

Becker, Marco; Daube, Carl Heinz; Ernst, Reinking (2023): *Industrie 5.0*, IUCF Working Paper Nr. 4/2023, ZBW - Leibniz Information Centre for Economics, Kiel, Hamburg.

Bitkom e.V. (2023): *Periodensystem der Künstlichen Intelligenz*, Online-Quelle: https://periodensystem-ki.de/Mit-Legosteinen-die-Kuenstliche-Intelligenz-bauen, letzter Zugriff am: 14.08.2023.

Davenport, Thomas; Prusak, Laurence (1999): *Wenn Ihr Unternehmen wüßte, was es alles weiß... Das Praxisbuch zum Wissensmanagement*, Landsberg/Lech, Moderne Industrie.

DB Systel GmbH (2023): *Large Language Models - Was hat ChatGPT dazu zu sagen?, in: Digital Foresight Januar 2023*, Online-Quelle: https://www.dbsystel.de/resource/blob/10372378/bc7c882c57e69eeb2cee72812a139be4/Download-Digital-Trend-Impuls-ChatGPT-data.pdf, letzter Zugriff am: 14.08.2023.

Ferus, Jacob (2022): *Improving ChatGBT With Prompt Injection*,Online-Quelle: https://levelup.gitconnected.com/improving-chatgpt-with-prompt-injection-b0c0c27b7df7, letzter Zugriff am: 14.08.2023.

Grosse, Tobias; Hallgarten, Philipp; Zwaller, Lukas; Hoellig, Christoph; Asgharzadeh, Pouyan (2023): *ChatGPT & enterprise knowledge: "How can I create a chatbot for my business unit?"*, Online-Quelle: https://medium.com/next-level-german-engineering/chatgpt-enterprise-knowledge-how-can-i-create-a-chatbot-for-my-business-unit-4380f7b3d4c0, letzter Zugriff am: 14.08.2023.

Heaven, Will Douglas (2023): *Generative KI: Die Geschichte hinter ChatGPT - Der Durchbruch von OpenAI kam gefühlt über Nacht – aber dahinter steht jahrzehntelange Forschung, in: MIT Technologie Review*, Online-Quelle: https://www.heise.de/hintergrund/Generative-KI-Die-Geschichte-hinter-ChatGPT-8243968.html, letzter Zugriff am: 14.08.2023.

HUGGINGFACE (2023): *Open LLM Leaderboard*, Online-Quelle: https://huggingface.co/spaces/gsaivinay /open_llm_leaderboard, letzter Zugriff am: 14.08.2023.

Kucharavy, Andrei; Schillaci, Zachary; Maréchal, Loic; Wünsch, Maxime; Dolamic, Ljiljana; Sabonnadiere, Remi; Davic, Dimitri Percia; Mermoud, Alain; Lenders, Vincent (2023): *Fundamentals of Generative Large Language Models*, Online-Quelle: https://arxiv.org/pdf/2303.12132.pdf, letzter Zugriff am: 14.08.2023.

Lee, Ernesto (2023): *Chunking Strategies for LLM Applications*, Online-Quelle: https://drlee.io/chunking-strategies-for-llm-applications-7a37d56e2b15, letzter Zugriff am: 14.08.2023.

LexisNexis (2023): *Datenanalyse: Wertvolle Informationen aus Daten - Mit Datenanalysen gewinnen Sie Erkenntnisse aus Daten. Hier erfahren Sie alles rund um die Vorteile, Methoden und Anwendungsgebiete der Datenanalyse.*, Online-Quelle: https://www.lexisnexis.de/begriffserklaerungen/data-integration/methoden-der-datenanalyse, letzter Zugriff am: 14.08.2023.

MIT SLOAN SCHOOF OF MANAGEMENT (2023): *Tapping the power of untructured data*, Online-Quelle: https://mitsloan.mit.edu/ideas-made-to-matter/tapping-power-unstructured-data, letzter Zugriff am: 14.08.2023.

OpenAI (2023a): *Models - GPT-3-5*, Online-Quelle: https://platform.openai.com/docs/models/gpt-3-5, letzter Zugriff am: 14.08.2023.

OpenAI (2023b): *Models - GPT-4*, Online-Quelle: https://platform.openai.com/docs/models/gpt-4, letzter Zugriff am: 14.08.2023.

OpenAI (2023c): *Fine-tuning - Learn how to customize a model for your application*, Online-Quelle: https://platform.openai.com/docs/guides/fine-tuning, letzter Zugriff am: 14.08.2023.

Taukki, Tom (2022): *Grundlagen der Künstilchen Intelligenz - Eine nichttechnische Einführung*, Delaware, Springer Nature.

VG WORT (2023): *Merkblatt zur VG WORT für Urheber und Verlage*, Online-Quelle: https://www.vgwort.de/fileadmin/vg-wort/pdf/dokumente/Merkblaetter/Auszahlungen/Merkblatt_Allgemeines_VG_Wort_2022.pdf, letzter Zugriff am: 14.08.2023.

Wang, Chenguang; Li, Mu; Smola, Alexander J. (2019): *LanguageModelswithTransformers*, Online-Quelle: https://arxiv.org/pdf/1904.09408.pdf, letzter Zugriff am: 14.08.2023.

Zhang, Min; Li, Juntao (2021): *A commentary of GPT-3, in: MIT Technology Review 2021*, Online-Quelle: https://www.sciencedirect.com/science/article/pii/S2667325821002193, letzter Zugriff am: 14.08.2023.