Buxmann, Peter; Glauben, Adrian; Hendriks, Patrick

**Article — Published Version**

# Die Nutzung von ChatGPT in Unternehmen: Ein Fallbeispiel zur Neugestaltung von Serviceprozessen

HMD Praxis der Wirtschaftsinformatik

*Suggested Citation:* Buxmann, Peter; Glauben, Adrian; Hendriks, Patrick (2024) : Die Nutzung von ChatGPT in Unternehmen: Ein Fallbeispiel zur Neugestaltung von Serviceprozessen, HMD Praxis der Wirtschaftsinformatik, ISSN 2198-2775, Springer Fachmedien Wiesbaden, Wiesbaden, Vol. 61, Iss. 2, pp. 436-448,
https://doi.org/10.1365/s40702-024-01053-8

This Version is available at:
https://hdl.handle.net/10419/315877

HMD Praxis der Wirtschaftsinformatik (2024) 61:436–448
https://doi.org/10.1365/s40702-024-01053-8

SCHWERPUNKT

# Die Nutzung von ChatGPT in Unternehmen: Ein Fallbeispiel zur Neugestaltung von Serviceprozessen

**Peter Buxmann · Adrian Glauben · Patrick Hendriks**

Eingegangen: 6. Juli 2023 / Angenommen: 31. Januar 2024 / Online publiziert: 24. Februar 2024


**Zusammenfassung** Large Language Models (LLMs) revolutionieren die Art und Weise, wie Texte oder auch Software geschrieben werden. In diesem Artikel wollen wir insbesondere auf den Einsatz von ChatGPT in Unternehmen eingehen. Schwerpunkt ist ein Fallbeispiel zur Neugestaltung von Serviceprozessen, das gemeinsam mit einem mittelständischen Softwarehaus entwickelt wurde. Wir zeigen, wie LLMs Geschäftsprozesse transformieren können und welche wirtschaftlichen Effekte sich daraus ergeben.

**Schlüsselwörter** Generative KI · Große Sprachmodelle · Serviceprozesse · Wirtschaftlichkeit

## The Use of ChatGPT in Companies: A Case Study on the Redesign of Service Processes

**Abstract** Large Language Models (LLMs) are revolutionising the way texts and even software are written. In this article we will focus on the use of ChatGPT in companies. The emphasis is on a case study for redesigning service processes that was developed together with a medium-sized software company. We show how LLMs can transform business processes and what the economic impact is.

✉ Peter Buxmann · Adrian Glauben · Patrick Hendriks
Technische Universität Darmstadt, Darmstadt, Deutschland
E-Mail: peter.buxmann@tu-darmstadt.de

Adrian Glauben
E-Mail: adrian.glauben@tu-darmstadt.de

Patrick Hendriks
E-Mail: patrick.hendriks@tu-darmstadt.de

**Keywords** Generative AI · Large Language Models · Service Processes · Economic Viability

## 1 Einleitung

Die rasante Entwicklung der Künstlichen Intelligenz (KI) ist dabei, Gesellschaft und Wirtschaft grundlegend zu verändern. Ein Meilenstein in der Geschichte der KI war die Veröffentlichung von ChatGPT durch das US-amerikanische Unternehmen OpenAI im November 2022. Bei ChatGPT handelt es sich um ein Sprachmodell, das Texte erstellen, Software entwickeln und eine menschenähnliche Kommunikation führen kann. Das immense öffentliche Interesse an dieser Technologie spiegelt sich in den Nutzungszahlen wider. Nur zwei Monate nach der Markteinführung erreichte ChatGPT im Januar die Marke von 100 Mio. monatlich aktiven Nutzerinnen und Nutzern. Dies macht es zur am schnellsten wachsenden Verbraucheranwendung der Geschichte. Zum Vergleich: Die Social-Media-Plattform TikTok brauchte dazu neun Monate, Instagram sogar über zwei Jahre (Hu 2023).

Auch Investorinnen und Investoren sowie Technologieunternehmen haben das enorme Potenzial von ChatGPT erkannt. So sicherte sich Microsoft Anfang 2023 knapp 50 % der Anteile an OpenAI für rund 10 Mrd. US-Dollar (Bass 2023). Diese Investition dient der Strategie, ChatGPT in verschiedene Microsoft-Produkte wie Office, Bing und Azure zu integrieren, um die Produktqualität zu verbessern und die Wettbewerbsposition gegenüber Unternehmen wie Google zu stärken. Dadurch wird deutlich, dass die Einsatzmöglichkeiten von ChatGPT weit über den Unterhaltungssektor hinausgehen. Allerdings gibt es bislang kaum praxisorientierte Studien, die diskutieren oder aufzeigen, wie ChatGPT und Co. in betriebliche Prozesse integriert werden können und welche Unternehmensbereiche besonders stark davon beeinflusst werden. In einem Interview unterstreicht Sam Altman, CEO von OpenAI, dass aus seiner Sicht insbesondere der Kundenservice beachtliche Veränderungen erfahren wird (Fridman und Altman 2023). Vor diesem Hintergrund werden wir im Folgenden näher auf die Nutzung von ChatGPT zur Neugestaltung des Kundenservice eingehen. Dazu erläutern wir in Abschn. 2 zunächst die Grundlagen von Sprachmodellen. Darauf folgt eine detaillierte Beschreibung des Fallbeispiels in Abschn. 3. In Abschn. 4 schließen wir mit einer ökonomischen Bewertung und einem Ausblick auf die Zukunft.

## 2 Large Language Models – Ein Überblick

Obwohl die Öffentlichkeit fast ausschließlich über ChatGPT diskutiert, ist es bei weitem nicht das einzige LLM auf dem Markt. Neben OpenAI entwickeln Tech-Giganten wie Google und Meta, Startups wie ElutherAI und Universitäten wie Stanford University oder UC Berkeley eigene LLMs mit vergleichbaren Fähigkeiten (siehe Abb. 1). Den Grundstein für diese Entwicklung legte Google mit der Entwicklung und Veröffentlichung der *Transformer-Architektur* (Vaswani et al. 2017)

– eine spezielle Art eines Neuronalen Netzwerks – welche die Basis für nahezu alle großen Sprachmodelle bildet. Abb. 1 gibt einen Überblick.

## 2.1 Funktionsweise und Fähigkeiten

Abgesehen von der Modellarchitektur teilen LLMs grundlegende Gemeinsamkeiten in ihrer Funktionsweise. Im Wesentlichen handelt es sich um Systeme, die die Wahrscheinlichkeiten von Wortsequenzen in Texten berechnen. Die Wortsequen-

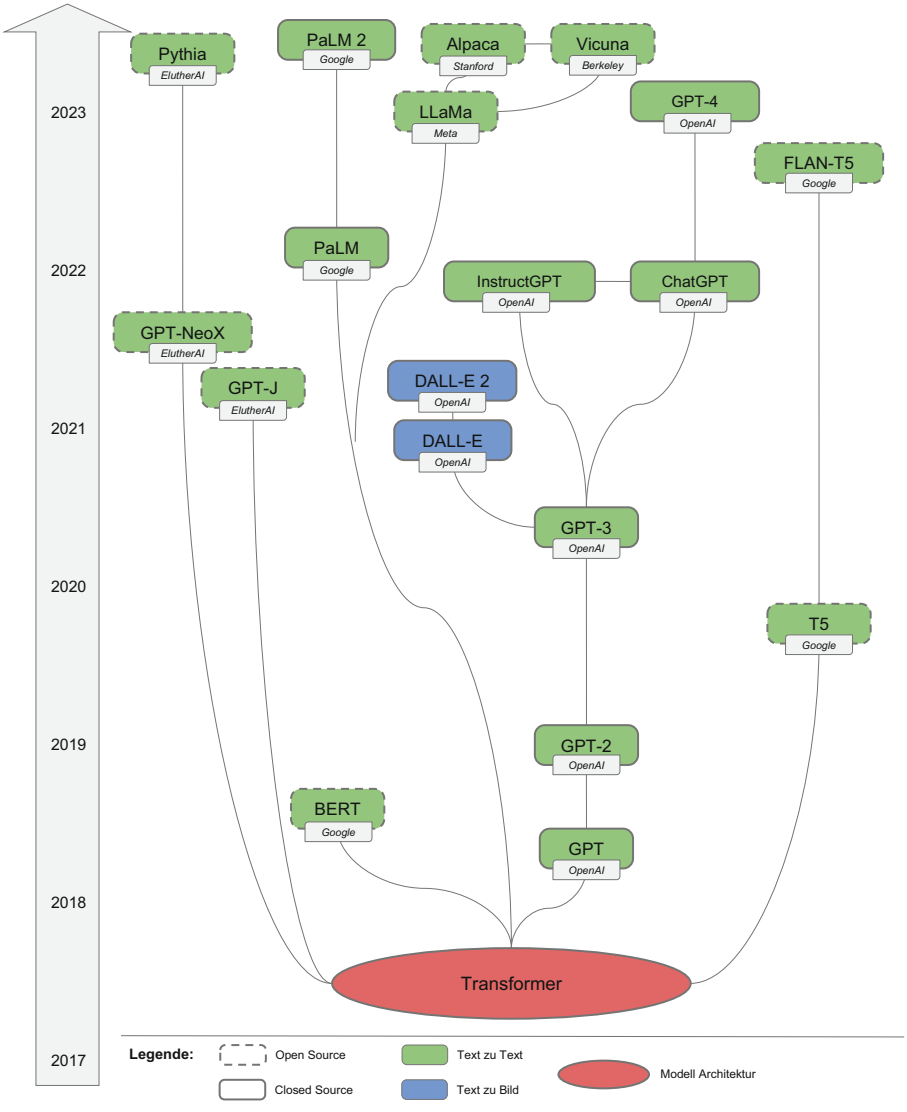


**Abb. 1** Entwicklung verschiedener Sprachmodelle seit 2017. (Eigene Abbildung nach Garske (2023))

zen können dabei sowohl kurze Sätze als auch längere Textabschnitte umfassen (Li 2022). Solche Systeme können sowohl zur Klassifikation (vgl. Xu et al. 2020, Choudrie et al. 2021) als auch zur Generierung (vgl. Bubeck et al. 2023; Luo et al. 2021) von natürlicher Sprache eingesetzt werden. Bei der Klassifikation gibt das Sprachmodell eine Wahrscheinlichkeit für jede der zuvor festgelegten Klassen aus (vgl. Sun et al. 2019). Ein anschauliches Beispiel für die Klassifizierung ist die automatische Unterscheidung von E-Mails in die Klassen „Spam“ und „kein Spam“. Bei der Generierung von natürlicher Sprache sagen Sprachmodelle das nächste Wort einer Sequenz vorher, indem sie für jedes mögliche Wort der vorgegebenen Sprache eine Wahrscheinlichkeit berechnen (vgl. Vaswani et al. 2017). Während ältere Ansätze auf rein statistischen Verfahren basieren (vgl. Nestor und Gonzalez-Abascal 1997), legen neuere Ansätze neuronale Netzwerke zugrunde (vgl. Bengio et al. 2003; Graves 2014). Aufbauend auf neuronalen Netzwerken entwickelten Vaswani et al. (2017) die bereits erwähnte Transformer-Architektur. Diese stellt dabei einen Paradigmenwechsel dar: Sie verwendet sogenannte *Attention-Mechanismen*, um zu bestimmen, welche Teile eines Textes in einem gegebenen Moment relevant sind und ermöglicht so ein tiefgreifenderes Verständnis des Kontextes – sprich der umgebenden Wörter und ihrer Bedeutungszusammenhänge. Dies führt dazu, dass LLMs nicht nur Wortsequenzen auf Basis statistischer Häufigkeit generieren, sondern auch die semantischen Beziehungen zwischen den Wörtern erfassen und interpretieren können, wodurch Texte entstehen, die kohärenter und inhaltlich nuancierter sind (Bouschery et al. 2023).

Wie der Begriff „Large Language Models“ bereits andeutet, zeichnen sich diese Modelle durch ihre enorme Größe aus, die sich aus drei maßgeblichen Faktoren ergibt (Kaplan et al. 2020):

1. Die Anzahl der Parameter, d. h. die Anzahl der veränderbaren Werte eines neuronalen Netzes, die durch den Trainingsprozess angepasst werden können;
2. die Größe des Datensatzes, d. h. die Anzahl der Datenpunkte im Trainingsdatensatz;
3. die Trainingszeit, d. h. die für das Training aufgewendete Rechenzeit.

Überraschenderweise führt die Vergrößerung von LLMs nicht nur zu einer linearen Verbesserung ihrer bestehenden Fähigkeiten, sondern bringt zudem unerwartete neue Fähigkeiten hervor (Wei et al. 2022). Das heißt, obwohl sie nicht explizit dazu trainiert wurden, entwickeln sie zunächst Kompetenzen in Textklassifikation, dann in mehrstufiger Argumentation und schließlich in semantischem Textverständnis (Wei et al. 2022). Die fortschrittlichsten LLMs – wie z. B. GPT-4 – zeigen bereits Fähigkeiten, die Bubeck et al. (2023) als erste Anzeichen von einer allgemeinen künstlichen Intelligenz bezeichnen. Dazu zählen mentale Fähigkeiten wie Kreativität, Argumentation und Deduktion, aber auch ihre Expertise in Bereichen wie Literatur, Medizin und Programmierung sowie die Fähigkeit, Werkzeuge zu verwenden und sich selbst zu erklären (Wei et al. 2022).

Wohlgemerkt sind all diese Fähigkeiten in einem einzigen Modell enthalten, ohne dass dieses Modell für eine der Aufgaben spezialisiert werden muss. Erreicht wird diese *Generalität* durch ein sogenanntes *self-supervised pre-training*. Dabei

lernt das Modell eigenständig – d. h. ohne menschliche Vorgaben – die zugrunde liegenden Strukturen und Muster innerhalb eines riesigen Datensatzes (Li 2022). Übliche Trainingsdatensätze sind hunderte bis tausende Gigabyte groß und enthalten unter anderem Bücher, Code, Webseiten, wissenschaftliche Artikel und Daten aus sozialen Netzwerken (Zhao et al. 2023).

Auf diese Weise trainierte Modelle werden auch als *Foundation Models* bezeichnet (vgl. Bommasani et al. 2021). Obwohl solche Modelle bereits viele Probleme ohne weitere Spezialisierung lösen können, kann es für bestimmte Aufgaben notwendig sein, ein weiteres Training durchzuführen. Für diesen als *fine-tuning* bezeichneten Prozess wird das mittels self-supervised pre-training vortrainierte Foundation Model mit Hilfe eines typischerweise kleineren und von Menschen gelabelten Datensatzes an ein bestimmtes Problem angepasst (Li 2022).

### 2.2 Vektordatenbanken

Gerade für die Integration in Unternehmensprozesse kann es notwendig sein, das LLM mit den proprietären Daten des Unternehmens anzureichern und ihm regelmäßig neue Informationen bereitzustellen. Neben dem fine-tuning werden hierfür in der Praxis vor allem *Vektordatenbanken* genutzt (bspw. Pinecone[1] und Milvus[2]; vgl. Lu et al. 2023; Bryan 2023). Im Gegensatz zum fine-tuning werden mit Vektordatenbanken keine neuen Informationen in das LLM eintrainiert, sondern lediglich bei Bedarf bereitgestellt. Dazu enthalten Vektordatenbanken neben den gespeicherten Dokumenten (z. B. PDFs) jeweils einen Vektor, der die Bedeutung und den Inhalt des Dokuments in einem abstrakten mathematischen Vektorraum repräsentiert. Zur Berechnung dieser Vektoren werden vortrainierte KI-Modelle verwendet, sogenannte *Embedding Models*, die sowohl Open Source als auch von namhaften Herstellern wie Google oder OpenAI verfügbar sind.

### 2.3 Ausgewählte Probleme

Mit der zunehmenden Intelligenz von LLMs und deren verstärkten Integration in Unternehmensprozesse gewinnt die Auseinandersetzung mit den damit verbundenen Problemen an Bedeutung. Eine Herausforderung sticht dabei besonders hervor: LLMs können ohne Vorwarnung Fehler erzeugen. Dazu zählen mathematische, logische und konzeptionelle Fehler, aber auch schlichtweg inkorrekte Aussagen. Diese Art von Fehlern wird als *Halluzinationen* bezeichnet, da sie dazu neigen, vernünftig zu erscheinen und logischen Schlussfolgerungen zu entsprechen. Häufig sind sie zusätzlich zwischen korrekten Informationen versteckt und werden überzeugend sowie selbstbewusst präsentiert (Bubeck et al. 2023).

Neben Halluzinationen ist *Bias* eines der derzeit schwerwiegendsten Probleme von LLMs. Ein Großteil der Trainingsdatensätze stammt aus dem Internet (Zhao et al. 2023) und enthält daher erhebliche Vorurteile, die von den darauf trainierten Sprachmodellen wiedergegeben oder sogar verstärkt werden können (Bubeck et al.

[1] https://www.pinecone.io/.

[2] https://milvus.io/.

2023). Dazu zählen neben stereotypen Assoziationen und negativen Empfindungen gegenüber bestimmten Gruppen auch gesellschaftspolitische Vorurteile (Bender et al. 2021).

## 3 Neugestaltung von Serviceprozessen mit ChatGPT

Im Folgenden soll gezeigt werden, wie Sprachmodelle genutzt werden können, um Prozesse in Unternehmen zu transformieren. Im Fokus liegt dabei ein Fallbeispiel, welches die Transformation eines Serviceprozess mittels ChatGPT aufzeigt. Die Bedeutung eines hervorragenden Kundenservice ist in der heutigen Geschäftswelt unbestritten und eng mit den informationstechnologischen Fähigkeiten eines Unternehmens verwoben (Ray et al. 2005). Sam Altman – Gründer und CEO von OpenAI – weist darauf hin, dass Sprachmodelle, mit ihren Fähigkeiten, unstrukturierte Daten zu verarbeiten und Dialoge zu führen, besonders geeignet sind, um Serviceprozesse zu transformieren (Fridman und Altman 2023). Insbesondere deswegen da Serviceprozesse neben der Interaktion mit Kundinnen oder Kunden mittels natürlicher Sprache häufig auch Informationen aus Dokumentenarchiven einbeziehen (vgl. Subramani et al. 2021). Einer der Kernbereiche von Serviceprozessen bezieht sich auf die Bearbeitung von Kundenanfragen durch Supportmitarbeiterinnen und -mitarbeiter (Kumar und Telang 2012). Unser Fallbeispiel basiert auf der Zusammenarbeit mit einem mittelständischen deutschen Softwareunternehmen, das durch den Einsatz von Sprachmodellen seine Prozesse zur Bearbeitung von Kundenanfragen unterstützt hat. Diese praktischen Verbesserungen sollen illustrieren wie LLMs, nicht nur die Kommunikation erleichtern, sondern auch die Qualität des Kundenservices insgesamt steigern können.

### 3.1 Der ursprüngliche Prozess

Im ersten Schritt wollen wir den *ursprünglichen Prozess* vor Einführung von ChatGPT darstellen. Der ursprüngliche Serviceprozess beginnt damit, dass der Kunde oder die Kundin eine Kontaktanfrage an den Kundensupport formuliert, in der er sein oder sie ihr Problem beschreibt. Während der Kunde oder die Kundin seine oder ihre Anfrage formuliert, werden ihm oder ihr automatisch Überschriften potenziell passender FAQ-Artikel angezeigt. Um passende FAQ-Artikel zu finden, wird ein KI-gestütztes System eingesetzt, das die Ähnlichkeit zwischen der Kundenanfrage und den vorhandenen Artikeln misst. Entscheidet sich der Kunde oder die Kundin, die angebotenen Artikel zu ignorieren oder findet er oder sie in den Artikeln keine passende Lösung, so kann er oder sie die zuvor formulierte Anfrage abschicken. Dadurch wird ein Ticket eröffnet, das daraufhin von einer Mitarbeiterin bzw. einem Mitarbeiter des Kundensupports bearbeitet wird. Können die Supportmitarbeiterinnen und -mitarbeiter das Kundenproblem nicht ad hoc lösen, so besteht für sie die Möglichkeit, das Dokumentenarchiv des Softwareunternehmens selbstständig zu durchsuchen. Haben sie passende Dokumente gefunden, extrahieren sie die relevanten Informationen und formulieren eine Antwort. Ist das Kundenproblem mit der Antwort gelöst, so ist die Interaktion beendet. Andernfalls haben die Kunden

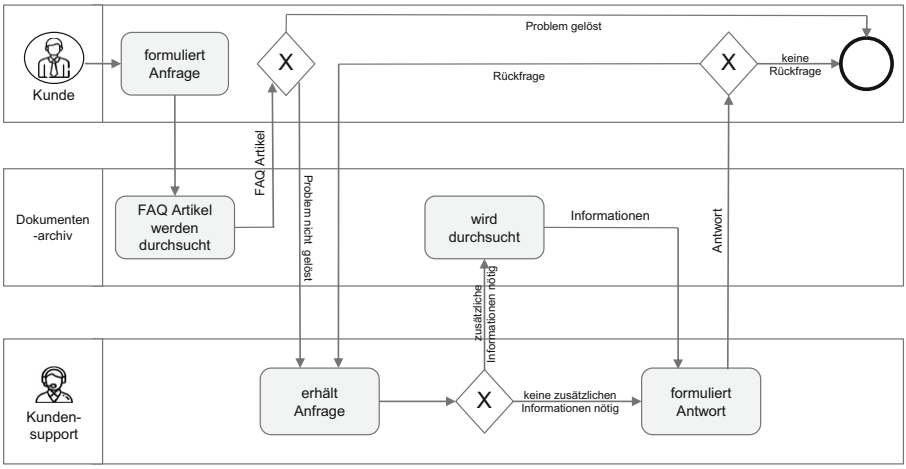


**Abb. 2** Schematische Darstellung des ursprünglichen Kundensupportprozesses vor Einführung von ChatGPT

oder Kundinnen die Möglichkeit, Rückfragen zu stellen. In diesem Fall muss der Kundensupport erneut reagieren und ggf. weitere Informationen aus dem Dokumentenarchiv abrufen. Dieser Prozess wiederholt sich, bis das Kundenproblem gelöst ist. Der beschriebene Ablauf vor Integration von ChatGPT ist in Abb. 2 schematisch dargestellt.

Es zeigt sich, dass dieser Prozess häufig mit einem erheblichen Aufwand für die Supportmitarbeiterinnen und -mitarbeiter verbunden ist. Neben der Suche nach relevanten Dokumenten wird zusätzliche Zeit für das Extrahieren und Zusammenfassen von Informationen sowie für das Formulieren von Antworten aufgewendet. Darüber hinaus wird potenziell neues Wissen, das im Laufe der Konversation entstanden ist, nicht extrahiert, aufbereitet und gespeichert, wenn die Supportmitarbeiterinnen und -mitarbeiter dies nicht eigenmotiviert vornehmen.

### 3.2 Der neugestaltete Prozess

Der *transformierte Prozess* zielt darauf ab, Prozessverbesserungen durch die Integration von ChatGPT zu erreichen. Ebenso wie der ursprüngliche Prozess beginnt auch der transformierte Prozess mit der Formulierung einer Supportanfrage seitens der Kunden oder Kundinnen. Damit das LLM eine passende Antwort formulieren kann, müssen der Kunde oder die Kundin zunächst die notwendigen Informationen in Form von Dokumenten bereitgestellt werden. Dazu werden Vektordatenbanken verwendet (siehe Abschn. 2). Abb. 3 zeigt beispielhaft die drei notwendigen Schritte zur Integration zwischen Vektordatenbank und LLM:

1. Zunächst muss das Unternehmen eine Vektordatenbank erstellen, in der alle gewünschten Informationen (z. B. interne Dokumente, FAQ-Artikel, und die Support-Ticket-Historie) hochgeladen werden. Dabei wird für jedes Dokument durch

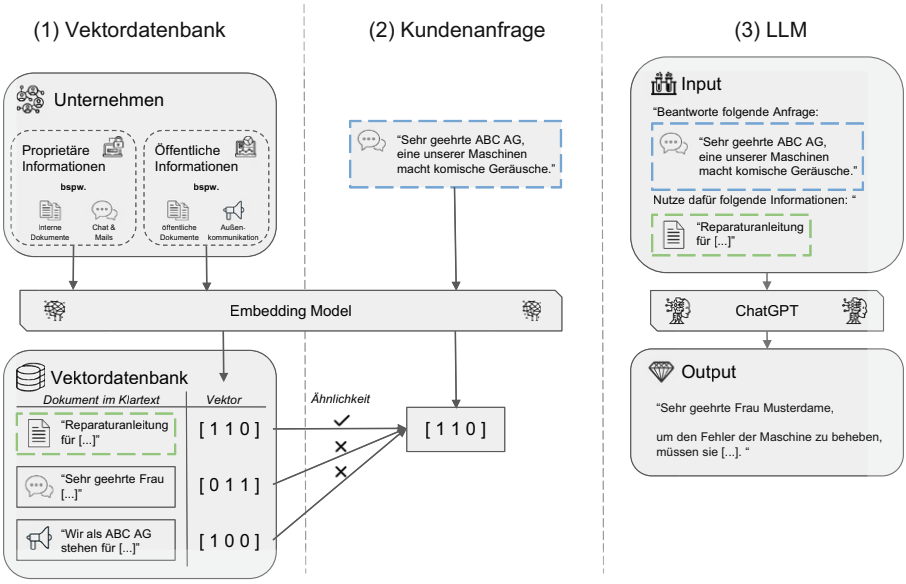


**Abb. 3** Zusammenspiel zwischen Vektordatenbank, Kundenanfrage und LLM

das Embedding Model ein repräsentativer Vektor berechnet und in der Vektordatenbank gespeichert.
2. Für die Kundenanfrage wird mit demselben Embedding Model ein repräsentativer Vektor berechnet. Daraufhin wird durch eine einfache mathematische Berechnung paarweise die Ähnlichkeit des Anfragevektors mit allen Dokumentvektoren aus der Vektordatenbank bestimmt.
3. Schließlich wird der Klartext der Kundenanfrage zusammen mit den Klartexten der ähnlichsten Dokumente in den Prompt des LLM aufgenommen, das auf Grundlage dieser Informationen eine Antwort formuliert.

Bevor die Antworten an die Kundinnen und Kunden verschickt werden, ist vorgesehen, dass der Kundensupport die von ChatGPT generierten Texte überprüft. Das bedeutet, dass die Mitarbeiterinnen und Mitarbeiter nicht ersetzt werden. Jedoch ändern sich ihre Aufgaben. Sie sollen zukünftig insbesondere die vom LLM vorformulierte Antwort prüfen und ggf. verbessern. Dabei besteht für die Mitarbeiterinnen und Mitarbeiter die Möglichkeit, mit dem LLM in einen Dialog zu treten, um den Textvorschlag iterativ zu verbessern und ggf. mit zusätzlichen Informationen anzureichern. Sobald die Supportmitarbeiterinnen und -mitarbeiter mit der Antwort zufrieden sind, leiten sie diese an den Kunden oder die Kundin weiter. Treten Rückfragen auf, so werden diese in einem weiteren Prozessdurchlauf mittels LLM-Unterstützung bearbeitet. Hat der Kunde oder die Kundin schließlich keine weiteren Fragen, so ist die Interaktion beendet. Im Gegensatz zum ursprünglichen Prozess stellt dies jedoch nicht das Ende dar, sondern löst einen weiteren Subprozess aus, da innerhalb dieser Interaktion neues Wissen entstanden sein kann. Dieses Wissen könnten die Supportmitarbeiterinnen oder -mitarbeiter beispielsweise im

Austausch mit Kollegen oder Kolleginnen oder anderen externen Informationsquellen erworben haben. Darüber hinaus könnte neues Wissen durch den Kunden oder die Kundin in die Interaktion eingebracht worden sein. In diesem abschließenden Subprozess werden die wesentlichen Informationen aus der Ticket-Historie durch das LLM extrahiert und aufbereitet. Bevor jedoch ein entsprechendes Dokument im Dokumentenarchiv abgelegt wird, bereitet der Supportmitarbeiter oder die Supportmitarbeiterin die Informationen auf, indem er beispielsweise deren Relevanz und Neuheit bewertet. Der beschriebene Prozess nach Integration von ChatGPT ist in Abb. 4 dargestellt.

### 3.3 Bewertung

Aus ökonomischer Perspektive zeigt sich, dass mit der vorgestellten Lösung für den Kundensupportprozess mehrere Vorteile erzielt werden können: So lassen sich potenziell Zeit und Kosten durch Nutzung der neuen Lösung einsparen. In Gesprächen mit dem Softwarehaus wurde insbesondere deutlich, dass sich das Management einen Wettbewerbsvorteil durch schnellere Antworten auf Kundenanfragen verspricht. Darüber hinaus können Unternehmen Kosten einsparen. Eine Beispielrechnung könnte so aussehen, dass pro Anfrage eine Minute eingespart werden könnte und die Kosten pro Minute 50 Cent betragen. Unter der Annahme, dass 10.000 Anfragen pro Tag in einer Serviceorganisation eingehen, würden die Kosteneinsparungen bei ca. 150.000 €/Monat liegen. Das Geschäftsmodell skaliert mit der Größe der Serviceorganisation. Darüber hinaus bietet die vorgestellte Lösung die Möglichkeit, bessere Antworten zu generieren und – wie oben dargestellt – die Wissensbasis zu verbessern.

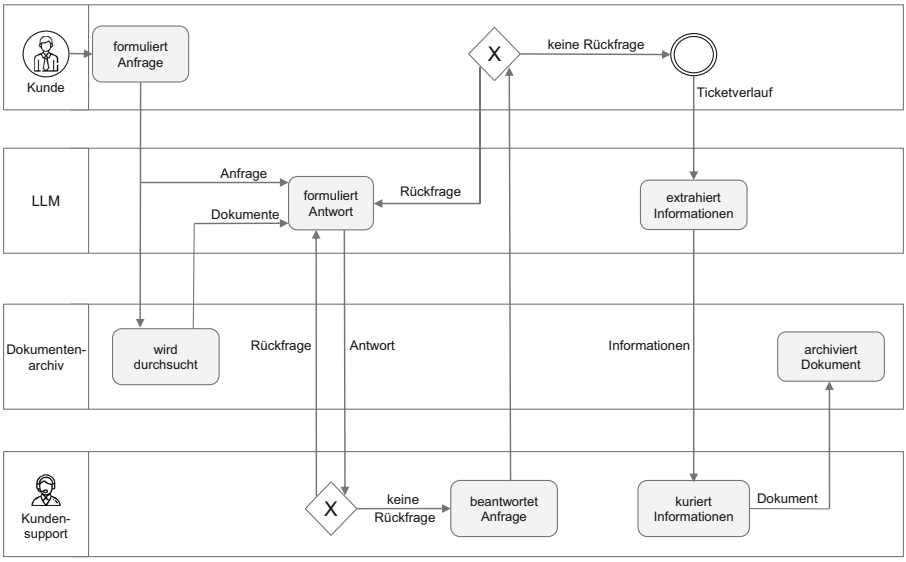


**Abb. 4** Schematische Darstellung des transformierten Kundensupportprozesses nach Einführung von ChatGPT

Neben den klar ersichtlichen wirtschaftlichen Vorteilen ergibt sich für die Supportmitarbeiterinnen und -mitarbeiter eine signifikante Transformation ihrer Arbeitsweise. Die Integration von ChatGPT befähigt sie, ihre Aufgaben mit gesteigerter Effizienz und Effektivität zu bewältigen. Im Einklang mit bestehender Literatur ist das Ziel dieses Systems nicht, den Menschen zu ersetzen, sondern vielmehr seine Arbeit zu ergänzen (Brynjolfsson 2023; Bankins et al. 2023). Daher ist keine Reduktion von Arbeitsstellen zu erwarten, sondern vielmehr eine Umorientierung der Tätigkeiten: Weg von zeitaufwendigen Routinearbeiten, wie der Recherche und Zusammenführung von Informationen sowie der Formulierung von Antworten, hin zu einer verstärkten Fokussierung auf die Gewährleistung qualitativ hochwertiger Antworten. Parallel dazu profitieren die Kunden und Kundinnen von reduzierten Antwortzeiten und einer erhöhten Qualität sowie Konsistenz der Serviceleistungen. Aufgrund der Integration in bestehende Systeme und Prozesse, dem Fokus auf dem Menschen als finalen Entscheidungsträger und der hohen Transparenz, geschaffen durch die Dialogfähigkeit und das Hinterlegen von Quellen aus dem Dokumentenarchiv, kann eine hohe Akzeptanz und Zufriedenheit seitens der Supportmitarbeiterinnen und -mitarbeiter vermutet werden (Chatterjee et al. 2021; Bankins et al. 2023; Bedué und Fritzsche 2022).

Dennoch ist es entscheidend, die bereits erörterten Risiken wie Halluzinationen und Bias in die Bewertung mit einzubeziehen. Wirtschaftliche Vorteile stehen potenziellen Fehlinformationen gegenüber, die das Vertrauen der Kunden und Kundinnen beeinträchtigen könnten. Eine fundierte Risikoanalyse vor der Implementierung und etablierte Korrekturmechanismen sind notwendig, um eine dauerhafte und verlässliche Qualitätssicherung im Kundensupport zu gewährleisten.

## 4 Fazit und Ausblick

Die Fallstudie zu Serviceprozessen zeigt, zum einen, wie es grundsätzlich möglich ist, ChatGPT in Unternehmensprozesse zu integrieren. Zum anderen wird deutlich, welche ökonomischen Effekte für Unternehmen erzielt werden können. Außerdem wurde anhand bestehender Literatur diskutiert welche Einflüsse sich auf die Akzeptanz und Zufriedenheit der Nutzerinnen und Nutzer ergeben könnten.

Der Artikel konzentriert sich primär auf die Vorstellung eines praxisnahen Schemas zur potenziellen Transformation wissensbasierter Unternehmensprozesse durch Sprachmodelle. Es ist jedoch zu beachten, dass zum jetzigen Zeitpunkt keine Daten zur Interaktionsqualität, Akzeptanz und Nutzerzufriedenheit vorliegen. Diese Aspekte sind wesentlich für die Beurteilung der Praktikabilität und Wirksamkeit des vorgestellten Frameworks. Außerdem wurde in diesem Artikel die Qualität der LLM-basierten Prozessunterstützung nicht evaluiert. Während eine Expertenbewertung der Antwortqualität zweifellos von Interesse wäre, liegt eine solche Bewertung außerhalb des aktuellen Untersuchungsbereichs. Zukünftige Forschung, angereichert mit empirischen Daten, könnten diese Lücke schließen und einen Mehrwert für die Anwendung von LLMs in Unternehmensprozessen bieten.

In unserem Anwendungsbeispiel konnten wir zeigen, dass die Integration von LLMs mit Unternehmensdaten mit Hilfe von Vektordatenbanken erfolgen kann.

Diese Anbindung verstehen wir als das *erste Level* der LLM-Integration in Unternehmen, deren Ziel es ist, Wissen in Echtzeit aufbereitet zur Verfügung zu stellen und gleichzeitig die Dialogfähigkeiten von LLMs nutzen zu können.

Das *zweite Level* bezieht sich auf die Anbindung verschiedener Werkzeuge an LLMs: D. h. LLMs erhalten die Möglichkeit, verschiedene Anwendungen über APIs selbständig zu nutzen. Wie Bubeck et al. (2023) zeigen sind aktuelle LLMs – wie bspw. GPT-4 – bereits in der Lage, mit minimalen Anweisungen Werkzeuge zu nutzen. Eine Beispielanwendung für dieses Level sind die von OpenAI vorgestellten ChatGPT Plug-ins (OpenAI 2023), die es ChatGPT ermöglichen eigenständig im Internet zu suchen, Code auszuführen oder auf Drittanbieterdienste zuzugreifen. Mit dieser zusätzlichen Handlungsfähigkeit können LLMs selbstständig auch komplexere Routineaufgaben übernehmen und damit den Menschen weiter entlasten.

Das *dritte Level* bezieht sich auf Anwendungen wie AutoGPT (GitHub 2023a), MetaGPT (GitHub 2023b) und HuggingGPT (Shen et al. 2023). Auf Basis vorgegebener Ziele können diese LLM-Agenten eigenständig Aufgaben festlegen, priorisieren und abarbeiten. Dabei werden die autonom definierten Aufgaben in einem rekursiven Prozess in durchführbare Pläne heruntergebrochen, erstellte Pläne reflektiert, geeignete Werkzeuge identifiziert und ausgeführt. Als Beispiel dient MetaGPT: Ein Tool, das aus einer Anforderung in einer Zeile umfassende Softwareentwicklungsoutputs wie Benutzergeschichten und API-Entwürfe erzeugt. Es simuliert ein Softwareunternehmen mit Large Language Models in Rollen wie Produktmanagern und -managerinnen oder Ingenieuren und Ingenieurinnen, die standardisierte Prozesse anwenden. Durch diese Fähigkeit könnte die vollständige Integration solcher LLM-Agenten in Unternehmensprozesse transformative Auswirkungen haben, deren Charakter – utopisch oder dystopisch – noch zu definieren ist. Zukünftige Untersuchungen sollten daher die Auswirkungen der LLM-Integration auf die Veränderung der Arbeitstätigkeiten und die Entwicklung neuer Qualifikationsprofile und Fähigkeiten der Arbeitnehmenden in den Fokus stellen, um eine gezielte Anpassung an die sich wandelnde Arbeitswelt zu ermöglichen.

Neben den aufgezeigten Potenzialen ist es wichtig, sich auch der Risiken der Anwendung von LLMs bewusst zu sein. Dazu zählen insbesondere die bereits angesprochenen Probleme, wie Halluzinationen und Bias. Verschärft wird diese Problematik dadurch, dass LLMs und damit auch deren Antworten intransparent sind. So ist z. B. unklar, aus welchen Gründen LLMs manchmal falsche und seltsam anmutende Texte erstellen. Dies gilt nicht nur für die Nutzerinnen und Nutzer. Selbst die Entwicklerinnen und Entwickler von GPT-4 können die Ergebnisse zum Teil nicht erklären, wie folgende Aussage von OpenAI verdeutlicht:

> „Language models have become more capable and more widely deployed, but we do not understand how they work“ (Bills et al. 2023, S. 1, Satz 1).

**Funding** Open Access funding enabled and organized by Projekt DEAL.

## Literatur

Bankins S, Ocampo AC, Marrone M, Restubog SLD, Woo SE (2023) A multilevel review of artificial intelligence in organizations: Implications for organizational behavior research and practice. J Organ Behav. https://doi.org/10.1002/job.2735

Bass D (2023) Microsoft Invests $ 10 Billion in ChatGPT Maker OpenAI. Bloomberg. https://www.bloomberg.com/news/articles/2023-01-23/microsoft-makes-multibillion-dollar-investment-in-openai#xj4y7vzkg. Zugegriffen: 30. Juni 2023

Bedué P, Fritzsche A (2022) Can we trust AI? An empirical investigation of trust requirements and guide to successful AI adoption. J Enterp Inf Manag 35(2):530–549. https://doi.org/10.1108/JEIM-06-2020-0233

Bender E, Gebru T, Major AM, Shmitchell S (2021) On the dangers of stochastic parrots: Can language models be too big? FAccT 21:610–623. https://doi.org/10.1145/3442188.3445922

Bengio Y, Ducharme R, Vincent P, Janvin C (2003) A neural probabilistic language model. J Machin Learn 2003:1137–1155

Bills S, Cammarata N, Mossing D et al (2023) Language models can explain neurons in language models. https://openaipublic.blob.core.windows.net/neuron-explainer/paper/index.html

Bommasani R, Hudson D, Adeli E et al (2021) On the opportunities and risks of foundation models. arXiv:2108.07258 (https://arxiv.org/abs/2108.07258)

Bouschery S, Blazevic V, Piller F (2023) Augmenting human innovation teams with artificial intelligence: Exploring transformer-based language models. J Prod Innov Manag 40(2):139–153. https://doi.org/10.1111/jpim.12656

Bryan K (2023) A User's Guide to GPT and LLMs for Economic Research. markus academy. https://bcf.princeton.edu/wp-content/uploads/2023/05/A_User_s_Guide_to_GPT_and_LLMs_for_Economic_Research.pdf. Zugegriffen: 30. Juni 2023

Brynjolfsson E (2023) A call to augment—not automate—workers. In Generative AI: Perspectives from Stanford HAI. Stanford University Human-Centered Artificial Intelligence. https://hai.stanford.edu/sites/default/files/2023-03/Generative_AI_HAI_Perspectives.pdf. Zugegriffen: 5. Nov. 2023

Bubeck S, Chandrasekaran V, Eldan R et al (2023) Sparks of Artificial General Intelligence: Early experiments with GPT-4 (https://arxiv.org/abs/2303.12712)

Chatterjee S, Rana NP, Dwivedi YK, Baabdullah AM (2021) Understanding AI adoption in manufacturing and production firms using an integrated TAM-TOE model. Technol Forecast Soc Change 170:120880. https://doi.org/10.1016/j.techfore.2021.120880

Choudrie J, Patil S, Kotecha K et al (2021) Applying and understanding an advanced, novel deep learning approach: a Covid 19, text based, emotions analysis study. Inf Syst Front 23:1431–1465. https://doi.org/10.1007/s10796-021-10152-6

Fridman L, Altman S (2023) Sam Altman: openAI CEO on GPT-4, chatGPT, and the future of AI. Youtube Podcast. https://www.youtube.com/watch?v = L_Guz73e6fw. Zugegriffen: 30. Juni 2023

Garske V (2023) AI / ML / LLM / Transformer Models Timeline and List. https://ai.v-gar.de/ml/transformer/timeline/. Zugegriffen: 30. Juni 2023

GitHub (2023a) AutoGPT: An Autonomous GPT-4 Experiment. https://github.com/Significant-Gravitas/Auto-GPT. Zugegriffen: 30. Juni 2023

GitHub (2023b) MetaGPT. https://github.com/geekan/MetaGPT. Zugegriffen: 30. Juni 2023

Graves A (2014) Generating Sequences with Recurrent Neural Networks. arXiv1308:0850v5 (https://arxiv.org/abs/1308.0850v5)

Hu K (2023) ChatGPT sets record for fastest-growing user base—analyst note. Reuters. https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/. Zugegriffen: 30. Juni 2023

Kaplan J, McCandlish S, Henighan T et al (2020) Scaling laws for neural language models. arXiv:2001.08361 (https://arxiv.org/abs/2001.08361)
Kumar A, Telang R (2012) Does the web reduce customer service cost? Empirical evidence from a Call center. Inform Syst Res 23(3):721–737. https://doi.org/10.1287/isre.1110.0390
Li H (2022) Language models: past, present, and future. Commun ACM 65(7):56–63. https://doi.org/10.1145/3490443
Lu Q, Zhu L, Xu X et al (2023) A framework for designing foundation model based systems. arXiv:2305.05352v4 (https://arxiv.org/pdf/2305.05352.pdf)
Luo B, Lau RYK, Li C, Si YW (2021) A critical review of state-of-the-art chatbot designs and Applications. WIREs DMKD. https://doi.org/10.1002/widm.1434
Nestor GV, Gonzalez-Abascal J (1997) Intelligent word-prediction to enhance text input rate. IUI. https://doi.org/10.1145/238218.238333
Open AI (2023) ChatGPT plugins. https://openai.com/blog/chatgpt-plugins. Zugegriffen: 30. Juni 2023
Ray G, Muhanna WA, Barney JB (2005) Information technology and the performance of the customer service process: a resource-based analysis. MIS Q 29(4):625–652
Shen Y, Song K, Tan X, Li D, Lu W, Zhuang Y (2023) HuggingGPT: solving AI tasks with chatGPT and its friends in huggingface. arXiv:2303.17580v3 (https://arxiv.org/pdf/2303.17580.pdf)
Subramani M, Wagle M, Ray G et al (2021) Capability development through just-in-time access to knowledge in document repositories: A longitudinal examination of technical problem solving. MIS Q 45(3):1287–1308. https://doi.org/10.25300/MISQ/2021/15635
Sun C, Qiu X, Xu Y, Huang X (2019) How to Fine-Tune BERT for Text Classification? Chinese Computational Linguistics. CCL 2019. Lecture Notes in Computer Science, Bd. 11856. Springer, Cham https://doi.org/10.1007/978-3-030-32381-3_16
Vaswani A, Shazeer N, Pramar N et al (2017) Attention is all you need. arXiv:1706.03762 (https://arxiv.org/abs/1706.03762)
Wei J, Tay Y, Bommasani R et al (2022) Emergent abilities of large language models. arXiv:2206.07682 (https://arxiv.org/abs/2206.07682)
Xu S, Barbosa SE, Hong D (2020) BERT feature based model for predicting the helpfulness scores of online customers reviews. Adv Inf Commun 1130:270–281. https://doi.org/10.1007/978-3-030-39442-4_21
Zhao W, Zhou K, Li J et al (2023) A survey of large language models. arXiv:2303.18223 (https://arxiv.org/abs/2303.18223)