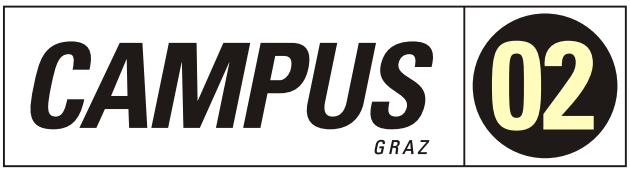
MASTERARBEIT

Technische Integration großer Sprachmodelle in Unternehmen

Entwicklung eines Frameworks zur technischen Vorbereitung von Unternehmen

auf die Nutzung von Sprachmodellen

ausgeführt am



Studiengang
Informationstechnologien und Wirtschaftsinformatik

Von: Christopher Haas
Pers. Kennz. 2010319003

SIGNATURINFORMATIONEN	**Signatory**	Christopher Franz Herbert Haas
	Date/Time-UTC	2025-03-16T20:09:27+0100
	Verification	Information about the verification of the electronic signature can be found at: https://www.signaturpruefung.gv.at
Note	This document is signed with a qualified electronic signature. According to Art. 25 para. 2 of the Regulation (EU) No 910/2014 of 23. July 2014 ("eIDAS-Regulation") it shall have the equivalent legal effect of a handwritten signature.	

Graz, am 16. März 2025

.....

Christopher Haas

Ehrenwörtliche Erklärung

Ich erkläre ehrenwörtlich, dass ich die vorliegende Arbeit selbständig und ohne fremde Hilfe verfasst, andere als die angegebenen Quellen nicht benutzt, die den Quellen wörtlich oder inhaltlich entnommenen Stellen als solche kenntlich gemacht, den Einsatz von generativen KI-Modellen (z.B. ChatGPT) kenntlich gemacht und mich sonst keiner unerlaubten Hilfsmittel bedient habe. Die Arbeit wurde bisher in gleicher oder ähnlicher Form keiner anderen Prüfungsbehörde vorgelegt und auch noch nicht veröffentlicht. Die vorliegende Fassung entspricht der eingereichten elektronischen Version.

SIGNATURINFORMATIONEN	**Signatory**	Christopher Franz Herbert Haas
	Date/Time-UTC	2025-03-16T20:10:37+0100
	Verification	Information about the verification of the electronic signature can be found at: https://www.signaturpruefung.gv.at
Note	This document is signed with a qualified electronic signature. According to Art. 25 para. 2 of the Regulation (EU) No 910/2014 of 23. July 2014 ("eIDAS-Regulation") it shall have the equivalent legal effect of a handwritten signature.	

Danksagung

An dieser Stelle möchte ich meinen aufrichtigen Dank aussprechen. Zunächst gilt meine Wertschätzung meiner Betreuerin Grobelscheg Lisa für ihre wertvolle fachliche Begleitung während des gesamten Entstehungsprozesses dieser Arbeit. Ihre konstruktiven Hinweise und ihre kontinuierliche Unterstützung waren für das Gelingen meines Vorhabens von unschätzbarem Wert. Besonders verbunden bin ich meinen Eltern, deren bedingungslose Zuneigung und fortwährende Ermutigung mich stets motiviert haben, nach Höchstleistungen zu streben. Ihr Rückhalt und ihre Wegweisung haben einen entscheidenden Beitrag zu meinem bisherigen Werdegang geleistet. Mit tiefer Zuneigung danke ich meiner Lebensgefährtin Irma Pezo, die mir in allen Phasen dieses akademischen Projekts mit Rat und Tat zur Seite stand. Ihre emotionale Unterstützung und ihr Verständnis für die damit verbundenen Herausforderungen haben mir die nötige Kraft gegeben, auch schwierige Phasen zu überwinden. Der erfolgreiche Abschluss dieser Arbeit wäre ohne das Mitwirken dieser bedeutenden Menschen in meinem Leben nicht möglich gewesen. Ihre Präsenz und Unterstützung erfüllen mich mit tiefer Dankbarkeit.

Christopher Haas

Graz, am 16. März 2025

Kurzfassung

Die zunehmende Verbreitung von Large Language Models (LLMs) in Unternehmenskontexten eröffnet neue Möglichkeiten zur Automatisierung von Wissensabruf, Inhaltserstellung und Entscheidungsunterstützung. Gleichzeitig stellt ihre Implementierung hohe technische Anforderungen an die bestehende IT-Infrastruktur und Datenarchitektur von Unternehmen. Diese Arbeit entwickelt ein strukturiertes Framework, das Unternehmen bei der technischen Vorbereitung auf die Integration von LLMs unterstützt. Im Fokus der Untersuchung stehen die zentralen technischen Anforderungen, die Unternehmen erfüllen müssen, um LLMs gezielt in spezifische Aufgaben innerhalb ihrer Geschäftsprozesse zu integrieren. Ein besonderer Schwerpunkt liegt auf der Analyse unternehmensinterner Datenformate und -strukturen sowie den notwendigen Vorverarbeitungsschritten zur Sicherstellung der Kompatibilität mit LLMs. Darüber hinaus werden Best Practices für die Archivierung und Bereitstellung von Daten identifiziert. Ergänzend erfolgt eine vergleichende Analyse von On-Premises- und Cloud-basierten Hosting-Lösungen sowie von selbst-gehosteten und extern-gehosteten LLMs, wobei Skalierbarkeit und Sicherheit als zentrale Bewertungskriterien herangezogen werden. Ein weiterer Fokus der Arbeit liegt auf Retrieval-Augmented Generation (RAG) als Methode zur Optimierung der LLM-Leistung durch die Nutzung unternehmensspezifischer Wissensbestände.

Die Arbeit folgt der Design Science Research Methodologie und kombiniert eine umfassende Literaturanalyse mit leitfadengestützten Experteninterviews, die primär zur Evaluation und Feinjustierung des entwickelten Frameworks dienen. Die Evaluation überprüft dessen Anwendbarkeit in verschiedenen Branchen und dessen Potenzial, IT-Abteilungen und Entscheidungsträger bei der strategischen Einführung von LLMs zu unterstützen.

Die Ergebnisse dieser Arbeit sollen praktische Empfehlungen für die technische Implementierung von LLMs in Unternehmen bieten. Sie berücksichtigen sowohl infrastrukturelle als auch datenspezifische Anforderungen und sollen eine Entscheidungsgrundlage für die Einführung und Nutzung von LLMs liefern.

Abstract

The increasing adoption of Large Language Models (LLMs) in corporate environments offers new opportunities for automating knowledge retrieval, content generation, and decision support. At the same time, their implementation imposes significant technical requirements on existing IT infrastructures and data architectures. This study develops a structured framework to assist companies in the technical preparation for LLM integration.

The research focuses on the key technical requirements that organizations must meet to effectively integrate LLMs into specific tasks within their business processes. A particular emphasis is placed on analyzing internal corporate data formats and structures, as well as the necessary preprocessing steps to ensure LLM compatibility. Additionally, best practices for data archiving and provisioning are identified. Furthermore, a comparative analysis of on-premises and cloud-based hosting solutions, as well as self-hosted and externally hosted LLMs, is conducted, with scalability and security serving as the primary evaluation criteria. Another focal point of the study is Retrieval-Augmented Generation (RAG) as a method to enhance LLM performance through the utilization of company-specific knowledge bases.

This study follows the Design Science Research methodology and combines an extensive literature review with semi-structured expert interviews, which primarily serve to evaluate and refine the developed framework. The evaluation assesses its applicability across different industries and its potential to support IT departments and decision-makers in the strategic implementation of LLMs.

The findings of this study aim to provide practical recommendations for the technical implementation of LLMs in businesses. They address both infrastructural and data-specific requirements and offer a well-founded decision-making basis for the introduction and utilization of LLMs.

Inhaltsverzeichnis

1. Einleitung **1**
1.1. Motivation . . . 1
1.2. Ziel und Forschungsfrage . . . 3
1.3. Aufbau der Arbeit . . . 4

2. Grundlagen **5**
2.1. Large Language Models . . . 5
2.1.1. Architektur und Funktionsweise . . . 6
2.1.2. Leistungsfähigkeit und Herausforderungen . . . 8
2.1.3. Integration in Unternehmen . . . 9
2.2. LLMs in Betrieben . . . 10
2.2.1. Chancen und Herausforderungen beim Einsatz von LLMs in Unternehmen . . . 10
2.3. Retrieval-Augmented Generation . . . 12
2.3.1. Prinzip von RAG: Kombination aus parametrischem und nicht-parametrischem Speicher . . . 12
2.3.2. Technische Architektur: Vektordatenbanken, Indexierung und Suchstrategien . . . 13
2.3.3. Vorteile und Herausforderungen von RAG . . . 13
2.3.4. Praxisbeispiel: RAG in Unternehmens-Wissensmanagementsystemen . . . 14

3. Methodik **15**
3.1. Design Science Research . . . 15
3.2. Literaturrecherche . . . 16
3.3. Inhaltsanalyse . . . 17

4. Technische Anforderungen **18**
4.1. Rechenleistung und Infrastruktur . . . 18
4.1.1. Speicherbedarf . . . 18
4.1.2. Netzwerk und Infrastruktur . . . 19
4.2. Datenformate und Archivierung . . . 19
4.3. Bereitstellung von LLMs: On-Premises vs. Cloud . . . 21
4.4. Vorbereitung von Daten für RAG . . . 22

5. LLM-Lösungen **24**
5.1. Überblick über LLM-Lösungen . . . 24
5.2. Vergleich Open-Source vs. kommerzielle Modelle . . . 27

6. Framework **30**
6.1. Dokumentation und Konzeption . . . 30
6.2. Darstellung des Frameworks . . . 32

7. Expert*inneninterviews **37**
7.1. Auswahl der Expert*innen und Begründung . . . 37
7.2. Interviewleitfaden und Durchführung . . . 39
7.3. Auswertung der Interviews . . . 43
7.3.1. Methodisches Vorgehen . . . 43
7.3.2. Interpretation der Kategorien . . . 45

8. Diskussion **52**
8.1. Validierung . . . 52
8.2. Herausforderungen bei der Implementierung . . . 53
8.3. Datenschutz- und Compliance-Herausforderungen . . . 55

9. Fazit **58**
9.1. Zusammenfassung der zentralen Ergebnisse . . . 58
9.2. Beantwortung der Forschungsfrage . . . 59
9.3. Kritische Reflexion und Limitationen . . . 60
9.4. Implikationen und Ausblick . . . 61

Abbildungsverzeichnis **63**

Tabellenverzeichnis **64**

Literatur **65**

A. Transkripte Expert*inneninterviews **74**

1. Einleitung

Unternehmen weltweit stehen derzeit an einem Wendepunkt, an dem die fortschreitende Digitalisierung und der Einsatz innovativer KI-Technologien, allen voran Large Language Models (LLMs), neue Maßstäbe für effiziente und intelligente Geschäftsprozesse setzen.

1.1. Motivation

Die rasante Entwicklung von Large Language Models (LLMs), großen Sprachmodellen, die darauf ausgelegt sind, menschliche Sprache zu verstehen und automatisch zu verarbeiten, hat die Art und Weise, wie Unternehmen mit Informationen umgehen, grundlegend verändert (Shanahan, 2023). Unternehmen setzen vermehrt auf Künstliche Intelligenz (KI), um Geschäftsprozesse zu optimieren, Wissensmanagement zu verbessern und Kundenanfragen effizienter zu bearbeiten. Insbesondere in Bereichen wie Personalwesen, Kundenservice und interner Kommunikation eröffnen LLMs dadurch neue Möglichkeiten zur Automatisierung und Unterstützung von Mitarbeitenden (Brynjolfsson et al., 2023; Budhwar et al., 2023; Cappelli et al., 2024).

Eine Studie von Gartner (2024) zeigt, dass generative KI (GenAI), eine Form der künstlichen Intelligenz, die in der Lage ist, neue Inhalte wie Texte, Bilder oder Musik zu erstellen, bereits die am häufigsten eingesetzte Form von KI in Unternehmen ist und noch vor regelbasierten Systemen sowie herkömmlichen maschinellen Lernverfahren zum Einsatz kommt. Ähnliche Ergebnisse liefert McKinsey (2024) aus dem Jahr 2024, demnach nutzen mittlerweile 65% der Unternehmen regelmäßig GenAI, fast doppelt so viele wie im Vorjahr.

Gleichzeitig zeigt die Untersuchung von Gartner (2024), dass viele Organisationen im Umgang mit KI-Projekten komplexen Herausforderungen gegenüberstehen. Eine Umfrage unter 632 Führungskräften, die stark in KI-Projekte eingebunden sind, identifiziert sechs Hauptprobleme, darunter die Abschätzung des wirtschaftlichen Nutzens, den Mangel an Fachkräften und ungenügende Datenqualität (siehe Abbildung 1.1).

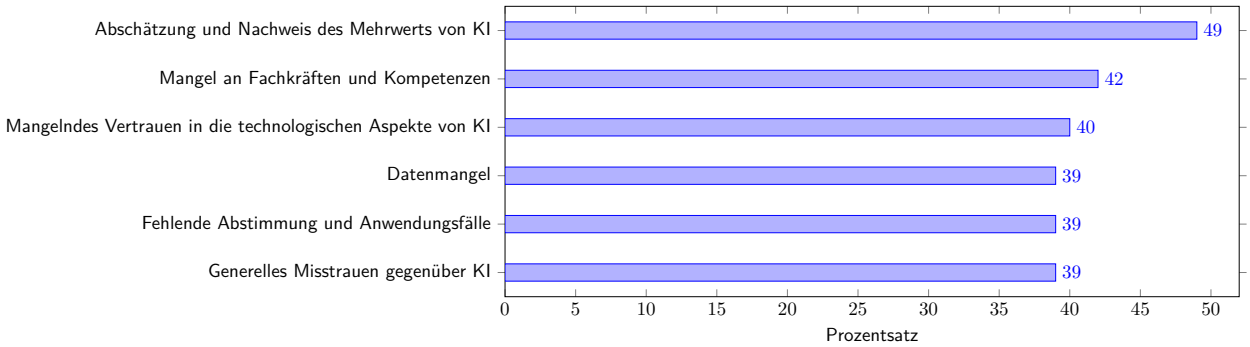


Abbildung 1.1.: Hauptbarrieren bei der Implementierung von KI-Techniken (in Anlehnung an Gartner (2024) n = 632, Führungskräfte, die stark in KI involviert sind; "unsicher" ausgeschlossen

Während die in Abbildung 1.1 gezeigten Werte bereits verdeutlichen, in welchen Bereichen Unternehmen bei der KI-Integration Unterstützung benötigen, werden die dahinterliegenden technischen und organisatorischen Aspekte in Unterabschnitt 2.1.2 ausführlicher diskutiert. Die Befragung von Gartner (2024) zeigt zudem, dass lediglich 48% der KI-Projekte überhaupt produktiv genutzt werden, ein weiteres Indiz dafür, wie hoch die praktischen Hürden bei der Umsetzung ausfallen.

Parallel dazu verdeutlicht eine Erhebung von Eurostat (2025), dass sich der Einsatz von KI-Technologien in europäischen Unternehmen zwar stetig ausweitet, aber regional sehr unterschiedlich fortschreitet. So nutzten im Jahr 2024 bereits 13,5% der Unternehmen in der EU mit mindestens zehn Mitarbeitenden KI-Technologien, ein Anstieg um 5,5 Prozentpunkte gegenüber dem Vorjahr. Besonders häufig kommen Text Mining (6,9%), Natural Language Generation (5,4%) und Speech Recognition (4,8%) zum Einsatz, die in direktem Zusammenhang mit LLMs stehen. Auf Länderebene variieren die Raten jedoch erheblich. Während Dänemark (27,6%), Schweden (25,1%), Belgien (24,7%) und Österreich (20,3%) eine hohe Adoptionsrate aufweisen, bleibt der KI-Einsatz in Rumänien (3,1%), Polen (5,9%), Bulgarien (6,5%) und Ungarn (7,4%) vergleichsweise gering (siehe Abbildung 1.2). Dies lässt vermuten, dass Unternehmen je nach Region vor unterschiedlichen infrastrukturellen und technologischen Herausforderungen stehen.

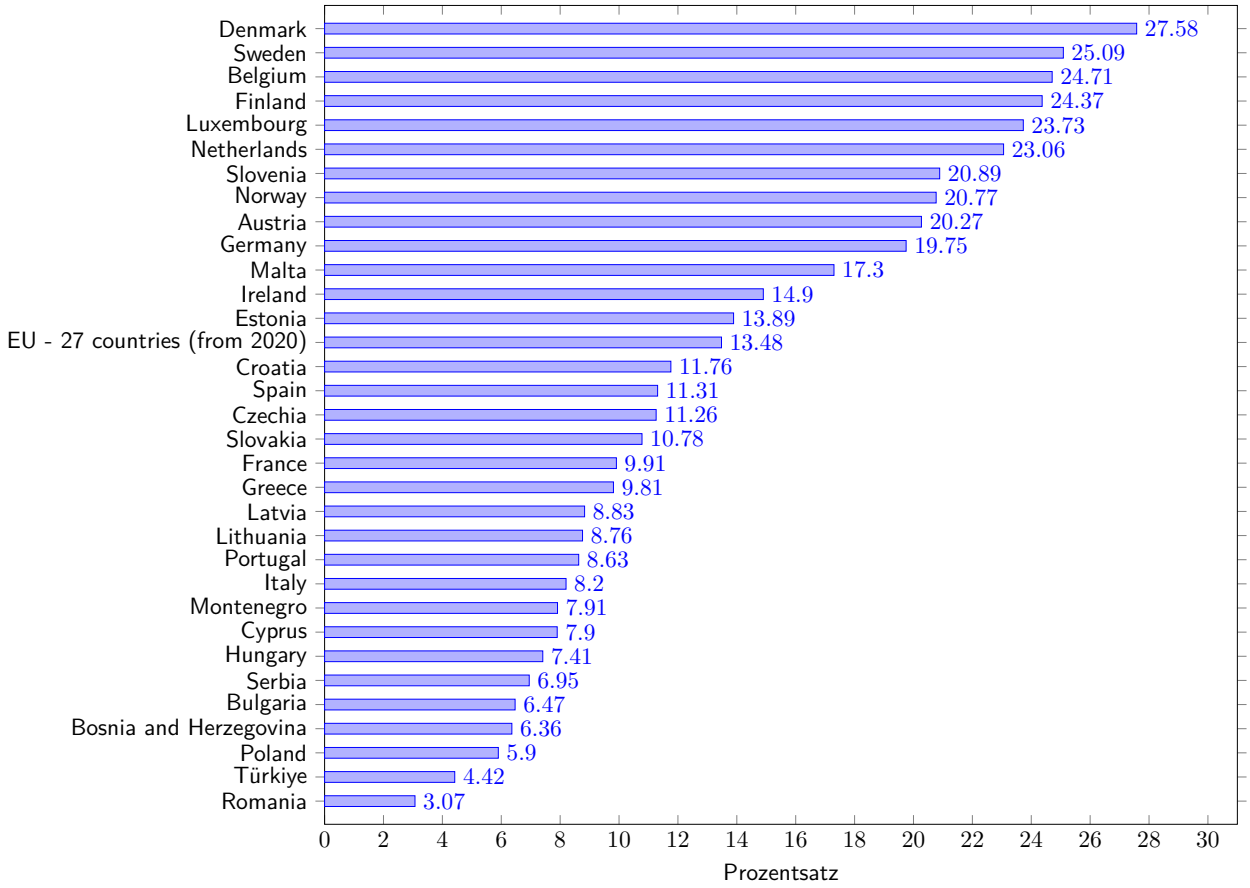


Abbildung 1.2.: Unternehmen, die KI-Technologien nutzen (vgl. Eurostat (2024))

Die geringe Erfolgsquote von KI-Projekten und die ungleiche Verbreitung von KI-Technologien innerhalb Europas unterstreichen die Notwendigkeit eines strukturierten Ansatzes, um LLMs nachhaltig in Unternehmensprozesse zu integrieren und bestehende Herausforderungen systematisch zu adressieren.

1.2. Ziel und Forschungsfrage

Die in Abschnitt 1.1 kurz angerissenen Herausforderungen verdeutlichen die Notwendigkeit einer systematischen Analyse der technischen Anforderungen, die für eine erfolgreiche Nutzung von LLMs in Unternehmensprozessen erfüllt sein müssen. Vor diesem Hintergrund ergibt sich die zentrale Forschungsfrage dieser Arbeit:

„Welche technischen Anforderungen müssen Unternehmen erfüllen, um LLMs erfolgreich in ihre Arbeitsprozesse zu integrieren?“

Die vorliegende Arbeit setzt genau an diesem Punkt an und entwickelt ein technisches Framework, das Unternehmen bei der Integration von LLMs unterstützt. Das Ziel besteht darin, die technischen Anforderungen zu identifizieren, die notwendig sind, um LLMs in bestehende Systeme zu integrieren und deren praktischen Nutzen für Unternehmen zu maximieren. Durch eine Analyse der technologischen Rahmenbedingungen soll ein strukturierter Leitfaden geschaffen werden, der Organisationen eine klare Orientierung bei der Implementierung von LLMs bietet.

Zur Beantwortung dieser Frage werden sowohl theoretische Grundlagen als auch praxisnahe Erkenntnisse aus Experteninterviews herangezogen. Die Untersuchung konzentriert sich auf die Analyse relevanter Datenformate, die Bewertung verschiedener Bereitstellungsmodelle (On-Premises vs. Cloud) sowie die Datenverarbeitung und den Vergleich bestehender LLM-Lösungen.

Zudem ist das Framework so angelegt, dass es sowohl für KMU als auch für Großkonzerne anwendbar ist. Während kleinere Betriebe meist mit begrenzten Ressourcen agieren, profitieren große Organisationen von umfangreichen IT-Kapazitäten und Budgets. Dieser Unterschied findet im Konzept Berücksichtigung, ohne Abstriche bei Skalierbarkeit oder Datensicherheit zu machen.

1.3. Aufbau der Arbeit

Die Struktur dieser Arbeit umfasst neun Kapitel, die in logischer Abfolge die wesentlichen Dimensionen der technischen Integration von LLMs in Unternehmenskontexten beleuchten.

In Kapitel 2 werden zunächst die theoretischen Grundlagen vermittelt. Es wird ein Überblick über Large Language Models, deren Einsatzmöglichkeiten in Unternehmen sowie über die Retrieval-Augmented Generation (RAG) als zentrale Methode gegeben.

Kapitel 3 beschreibt im Anschluss die methodische Vorgehensweise der Arbeit. Im Fokus steht der Design Science Research (DSR)-Ansatz, ergänzt durch eine Literaturrecherche und die qualitative Inhaltsanalyse der Experteninterviews.

Aufbauend darauf widmet sich Kapitel 4 den technischen Anforderungen, die für eine Integration von LLMs erfüllt sein müssen. Daraufhin vergleicht Kapitel 5 bestehende Open-Source- und kommerzielle LLM-Lösungen hinsichtlich ihrer Anwendbarkeit in Unternehmenskontexten. Basierend auf diesen Erkenntnissen wird in Kapitel 6 ein technisches Framework entwickelt, das Unternehmen bei der Implementierung von LLMs unterstützt.

Kapitel 7 stellt die durchgeführten Experteninterviews vor, die zur Evaluierung des entwickelten Frameworks herangezogen wurden. Daran anschließend werden in Kapitel 8 die Ergebnisse diskutiert, Herausforderungen analysiert und die praktische Umsetzbarkeit der vorgeschlagenen Lösungen bewertet. Abschließend fasst Kapitel 9 die zentralen Erkenntnisse zusammen, leitet praxisrelevante Implikationen ab und gibt einen Ausblick auf zukünftige Entwicklungen in diesem Bereich.

2. Grundlagen

Die in der Einleitung bereits erwähnte, zunehmende Verbreitung von LLMs in Unternehmen eröffnet neue Möglichkeiten für die Automatisierung und Optimierung von Geschäftsprozessen. Gleichzeitig erfordert die erfolgreiche Implementierung dieser Modelle ein tiefes Verständnis ihrer technischen Grundlagen und Herausforderungen.

Dieses Kapitel führt daher systematisch in die technologischen und betrieblichen Aspekte von LLMs ein. In Abschnitt 2.1 wird die Architektur und Funktionsweise moderner LLMs erläutert. Dabei stehen insbesondere die Transformer-Technologie, Skalierungsgesetze sowie die spezifischen Fähigkeiten der Modelle im Vordergrund.

Anschließend zeigt Abschnitt 2.2, wie LLMs in verschiedenen Unternehmensbereichen eingesetzt werden können. Dabei werden sowohl die Vorteile, etwa Prozessautomatisierung und Effizienzsteigerung, als auch die zentralen Herausforderungen behandelt, darunter Datenqualität, regulatorische Vorgaben und technische Integrationsanforderungen.

Abschließend stellt Abschnitt 2.3 mit Retrieval-Augmented Generation (RAG) einen vielversprechenden Ansatz vor, um Halluzinationen zu reduzieren und externe Wissensquellen gezielt einzubinden. Die in diesem Kapitel gewonnenen Erkenntnisse bilden das Fundament für ein umfassendes Framework zur Integration von LLM-Technologien in Unternehmensprozesse.

2.1. Large Language Models

LLMs sind eine spezialisierte Art von künstlicher Intelligenz (KI), die auf großen Mengen von Text trainiert wurde, um bestehende Inhalte zu verstehen und neue Inhalte zu generieren (Gartner, 2025b). Ihr Erfolg basiert maßgeblich auf der Transformer-Architektur, die 2017 von Vaswani et al. (2017) eingeführt wurde und in Unterabschnitt 2.1.1 näher erläutert wird (NVIDIA, n. d. b). Im Gegensatz zu früheren rekurrenten neuronalen Netzwerken (RNN) verzichten Transformer auf sequentielle Verarbeitung und nutzen stattdessen Self-Attention, um den Kontext aller Tokens in einer Sequenz parallel zu erfassen (Vaswani et al., 2023). Dies ermöglicht eine effizientere Parallelisierung bei Training und Anwendung, da das Modell Wortbeziehungen über beliebige Distanzen hinweg analysieren kann, ohne die Sequenz schrittweise verarbeiten zu müssen (IBM, 2025; NVIDIA, n. d. b).

Zusätzlich werden diese Modelle auf internetweiten Korpora mit rasant wachsenden Parametern trainiert (NVIDIA, n. d. b). Die Kombination aus fortschrittlicher Architektur und massiven Datenmengen ermöglicht es LLMs, menschenähnliche Texte zu verstehen und zu generieren, wie zum Beispiel überzeugende Antworten in Chatbots bis hin zu automatisch generiertem Programmcodes (NVIDIA, n. d. b). Aufgrund dieser herausragenden Eigenschaften haben LLMs in kurzer Zeit Spitzenleistungen in einer Vielzahl von Aufgaben der natürlichen Sprachverarbeitung (Natural Language Processing, NLP) erzielt und großes wirtschaftliches Interesse geweckt. Haar et al. (2024) beschreibt diese Aufgaben als zentrale Herausforderungen und Anwendungsfelder der NLP.

Um Aufgaben zu bewältigen, müssen LLMs eine Vielzahl spezifischer Fähigkeiten aufweisen. Eine Übersicht dieser Kompetenzen laut Minaee et al. (2024) ist in Abbildung Abbildung 2.1 dargestellt, während Kapitel 5 näher auf gängige Modelle mit diesen Eigenschaften eingeht. Minaee et al. (2024) unterteilen die Fähigkeiten in drei Grundbereiche. Zum Bereich Basic gehört beispielsweise die Eigenschaft "Coding", bei der LLMs von Entwickler*innen genutzt werden können, um automatisch Code-Snippets zur Verarbeitung von JSON-Daten zu generieren. Im Bereich Emerging findet sich die viel genutzte Eigenschaft "Reasoning", mit der LLMs das Projektmanagement unterstützen, indem sie Vor- und Nachteile verschiedener Handlungsoptionen logisch abwägen und Empfehlungen ableiten. Der letzte Bereich, Augmented, umfasst Eigenschaften wie Ïnteracting with users", bei der LLM-basierte Chatbots natürliche Dialoge führen, Kundenfragen in Echtzeit beantworten und bei Bedarf an menschliche Experten weiterleiten können.

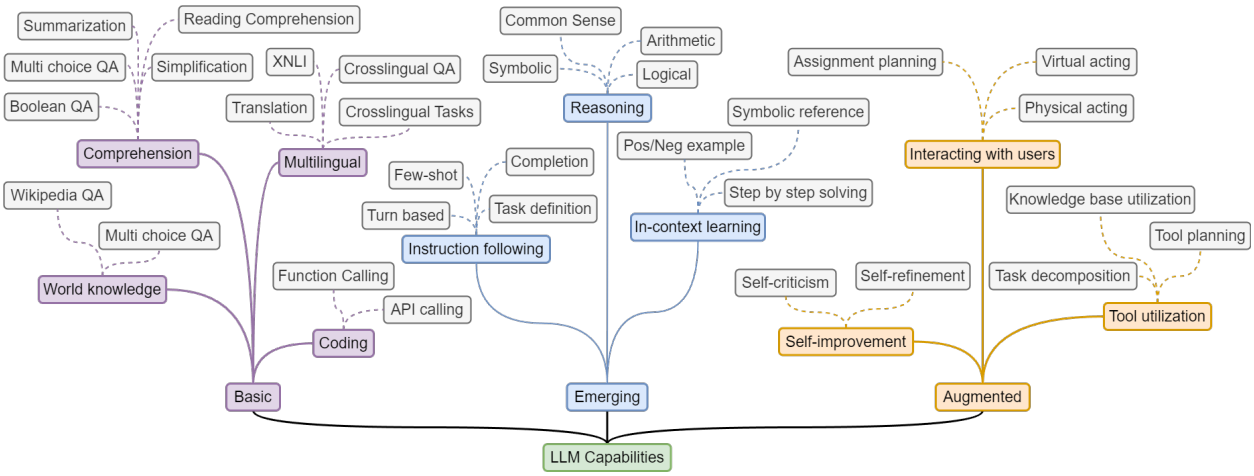


Abbildung 2.1.: LLM-Fähigkeiten (vgl. Minaee et al. (2024))

2.1.1. Architektur und Funktionsweise

Wie Eingangs bereits erwähnt, basieren die meisten modernen LLMs auf der Transformer-Architektur, die von Vaswani et al. (2023) entwickelt wurde. Abbildung 2.2 veranschaulicht schematisch den Aufbau dieses Modells mit einem Encoder-Decoder-Design, wie es ursprünglich für maschinelle Übersetzung entwickelt wurde (Vaswani et al., 2023). Der Encoder nimmt eine Eingabesequenz (z. B. einen Satz) entgegen und verarbeitet alle Tokens parallel durch mehrere hintereinandergeschaltete Schichten (Vaswani et al., 2023). Jede Encoderschicht besteht typischerweise aus zwei Hauptkomponenten, einer Multi-Head Self-Attention-Schicht, die für jedes Token berechnet, wie stark es auf andere Tokens der Sequenz achten sollte, und einem Feedforward Network, also einem vollvernetzten neuronalen Netz, das die durch Self-Attention angereicherten Token-Repräsentationen nicht-linear transformiert (NVIDIA, n. d. b). Zusätzlich werden pro Schicht Mechanismen wie Residual Connections und Normalisierung eingesetzt, um das Training tief gestapelter Schichten zu stabilisieren (NVIDIA, n. d. b). Der Encoder erzeugt so kontextuelle Vektorrepräsentationen der Eingabe (Vaswani et al., 2023).

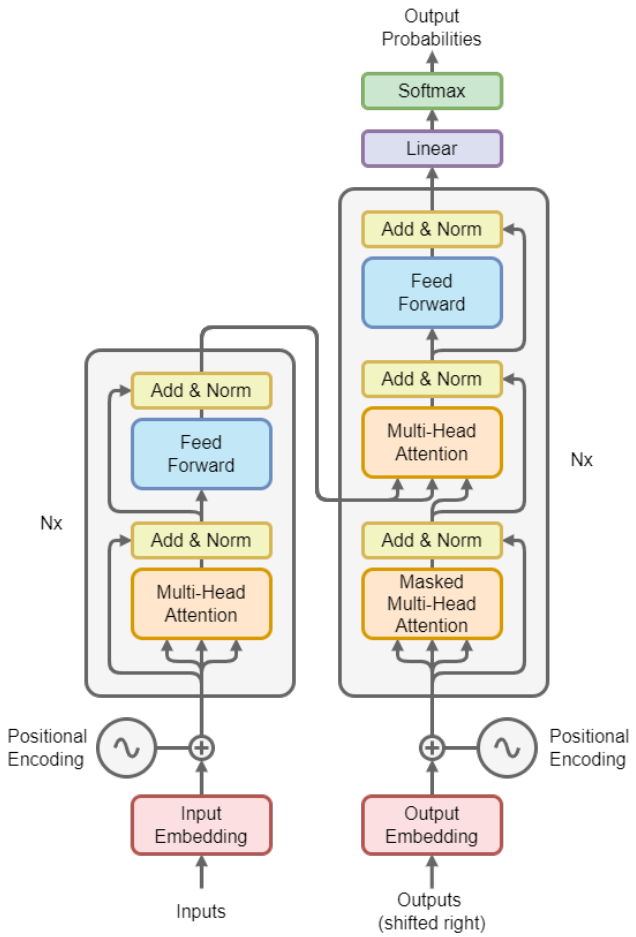


Abbildung 2.2.: Transformer Architektur (vgl. Vaswani et al. (2023))

Der Decoder nutzt diese Encoder-Ausgaben, um Schritt für Schritt eine Ausgabesequenz zu generieren (Vaswani et al., 2023). Ähnlich dem Encoder bestehen die Decoder-Schichten aus Self-Attention und Feedforward-Modulen (Vaswani et al., 2023). Zusätzlich kommt hier oft eine Encoder-Decoder Attention zum Einsatz, die es dem Decoder ermöglicht, auf die vom Encoder erzeugten Eingaberepräsentationen zuzugreifen (Vaswani et al., 2023). Während der Decoder jedes neue Wort generiert, bezieht er sich auf zuvor erzeugte Wörter, mittels masked Self-Attention und den encodierten Input, um konsistente und sinnvolle Fortsetzungen zu produzieren (Vaswani et al., 2023). Dieses autoregressive Vorgehen wird so lange fortgesetzt, bis eine vollständige Ausgabesequenz, etwa ein ganzer Satz, entstanden ist (Vaswani et al., 2023).

Eine Schlüsselrolle spielt dabei die Self-Attention, die jedem Wort der Eingabe einen gewichteten Einfluss relativ zu den anderen Wörtern zuweist (NVIDIA, n. d. b). Auf diese Weise lernt das Modell, welche Teile eines Satzes für die Vorhersage eines Wortes besonders relevant sind. Zum Beispiel kann in dem Satz "Die Bank hat gestern wegen Feiertag geschlossen" das Wort "Feiertag" die Bedeutung von "Bank" beeinflussen. Dank Multi-Head Attention geschieht dies in mehreren Projektionen gleichzeitig, sodass unterschiedliche Aspekte der Wortbeziehungen parallel erfasst werden (Vaswani et al., 2023). Eine weitere Innovation ist das Positional Encoding, das die Positionsinformationen der Tokens in der Sequenz einbettet, sodass die Wortreihenfolge auch ohne sequentielle Verarbeitung berücksichtigt wird (NVIDIA, n. d. b). Dadurch ermöglicht die Transformer-Architektur, komplexe Zusammenhänge in Texten über lange Distanzen hinweg zu modellieren (NVIDIA, n. d. b).

2.1.2. Leistungsfähigkeit und Herausforderungen

Wie in Abschnitt 1.1 bereits angedeutet, stehen Unternehmen bei der Nutzung von LLMs vor komplexen Anforderungen. Obwohl generative KI zunehmend Verbreitung findet, scheitern laut Gartner (2024) immer noch viele KI-Projekte an der praktischen Umsetzung.

LLMs haben in den vergangenen Jahren in Sprachverarbeitungsaufgaben beeindruckende Fortschritte erzielt. Ihre Qualität steigt mit zunehmender Modellgröße und Datenmenge, was empirisch in sogenannten Skalierungsgesetzen beschrieben wird (Kaplan et al., 2020). Erst Modelle mit Hunderten Milliarden Parametern (z. B. GPT-3 mit 175 Mrd.) erreichen eine fast schon menschliche Sprachqualität (NVIDIA, n. d. b). Daraus lässt sich schlussfolgern, dass diese Leistungszuwächse allerdings mit einem hohen technischen und organisatorischen Aufwand einhergehen.

Ein zentrales Thema ist hier die Skalierbarkeit von Systemen. Die Rechenressourcen, die für Training und Inferenz erforderlich sind, übersteigen häufig die Kapazitäten unternehmenseigener Hardware. Das Training von GPT-3 etwa umfasste geschätzt $3,14 \times 10^{23}$ Gleitkomma-Operationen (Casado et al., 2023). Für viele Firmen bedeutet dies hohe Investitionen oder die Auslagerung in Cloud-Infrastrukturen, was Datenschutz und Abhängigkeiten von Anbietern berührt (Developers, 2025). Zudem kann der Echtzeitbetrieb (z. B. für Chatbots) anspruchsvoll sein, da eine geringe Latenz meist Optimierungen oder Modellkompression erfordert.

Darüber hinaus hängt die Leistungsfähigkeit von LLMs stark von Datenqualität und -repräsentativität ab. Fehlen bestimmte Domänen oder aktuelle Ereignisse, generieren die Modelle veraltete oder unvollständige Informationen (Bisht, 2024). Geiger et al. (2021) bringen es in ihrer Arbeit treffend auf den Punkt, „Garbage in, garbage out“. Vorurteile im Trainingsmaterial spiegeln sich direkt in den Antworten wider. Unternehmen sollten daher ein gezieltes Fine-Tuning in Erwägung ziehen, was zusätzliche Aufbereitungsschritte und Fachpersonal erfordert.

Als kritische Aspekte erweisen sich ferner Bias und Halluzinationen. Da ein Großteil des Trainingsmaterials aus ungefilterten Internetquellen stammt, können soziale Vorurteile oder diskriminierende Inhalte entstehen (Gallegos et al., 2024). Hinzu kommt, dass LLMs faktenfreie, aber plausibel klingende Aussagen liefern können. Für den produktiven Einsatz bedeutet das zusätzliche Kontrollmechanismen, z. B. Filter- und Monitoring-Konzepte oder Retrieval-Augmented-Ansätze, um verlässliche Informationen aus Unternehmensdatenbanken einzubinden (Mialon et al., 2023). Letzteres wird näher im Abschnitt 2.3 besprochen.

Neben dem Modell selbst liegt eine Herausforderung in der Integration mit bestehenden IT-Systemen. LLMs müssen auf unternehmensinterne Daten (z. B. ERP-, CRM- oder DMS-Systeme) zugreifen können, um korrekte und kontextspezifische Ergebnisse zu liefern (Nahar et al., 2024). Dafür sind oft neue Schnittstellen und erweiterte Sicherheitskonzepte erforderlich, insbesondere wenn sensible Dokumente oder personenbezogene Daten ins Modelltraining einfließen (Kraus, 2024).

Nicht zuletzt erweist sich der Fachkräftemangel als Hemmschuh. Für LLM-Projekte werden sowohl KI-Expertinnen als auch IT-Architektinnen und Sicherheitsspezialistinnen benötigt, die eng mit den Fachabteilungen zusammenarbeiten (Ekuma, 2024). Ein durchdachtes Change Management ist ebenso unverzichtbar, da die Einführung von LLMs häufig Arbeitsabläufe verändert und einen Kulturwandel im Unternehmen erfordert. Hinzu kommt die Schwierigkeit, den geschäftlichen Nutzen zu belegen. Viele

LLM-Initiativen haben Pilot- oder Forschungscharakter, sodass konkrete Kennzahlen zum Return on Investment (ROI) erst in späteren Projektphasen verfügbar werden (Gartner, 2024).

2.1.3. Integration in Unternehmen

Die Integration von LLMs in Unternehmens-IT erfordert eine sorgfältige Planung der technischen Infrastruktur, der Datensicherheit sowie der operativen Prozesse. Viele führende Tech-Unternehmen haben LLM-Funktionen bereits in Produkte integriert, beispielsweise textgenerierende Assistenten in Office-Software (Nahar et al., 2024). Daraus lassen sich wichtige Anforderungen für Organisationen ableiten, die solche Modelle einsetzen möchten. Zunächst ist zu berücksichtigen, dass LLMs hohe Anforderungen an Rechenleistung und Systemarchitektur stellen. Unternehmen müssen im ersten Schritt entscheiden, ob sie ein Modell in einer eigenen Umgebung (On-Premises) oder über externe Dienste (Cloud) betreiben. Beide Ansätze bringen unterschiedliche Implikationen für Kosten, Kontrolle, Sicherheit und Flexibilität mit sich. Weitere Aspekte wie Netzwerk-Infrastruktur, Kapazitätsmanagement und Kühlung von GPU-Servern können hinzukommen, wenn das Unternehmen sich für einen lokalen Betrieb entscheidet (NVIDIA, n. d. b). Eine ausführliche Gegenüberstellung der Vor- und Nachteile von On-Premises- und Cloud-Lösungen erfolgt in Abschnitt 4.3.

Von ebenso großer Bedeutung ist der Umgang mit sensiblen Daten, da Unternehmen häufig vertrauliche Informationen wie Kundendaten oder interne Dokumente verarbeiten. Diese dürfen keinesfalls unkontrolliert in ein extern gehostetes Modell gelangen (Kraus, 2024). Kommen externe LLM-Services zum Einsatz, so muss vertraglich und technisch abgesichert werden, dass eingegebene Daten nicht missbräuchlich verwendet werden (Kraus, 2024).

Selbst bei On-Premises-Lösungen gilt es sicherzustellen, dass Logs oder gelernte Modellmuster keine vertraulichen Informationen preisgeben. Es ist leicht, dass Mitarbeiter versehentlich vertrauliche Unternehmensinformationen in LLM-Eingabeaufforderungen einfügen, was dazu führen kann, dass diese Informationen für den LLM-Anbieter zugänglich oder unbeabsichtigt der Öffentlichkeit zugänglich gemacht werden (Kraus, 2024). Datenschutz und Compliance sind daher integrale Bestandteile jeder LLM-Integrationsstrategie und werden bei der Wahl zwischen Cloud und On-Premises entsprechend berücksichtigt.

Die Einführung von LLMs in bestehende Prozesse ist außerdem aus Entwickler- und Betriebssicht anspruchsvoll. Zum einen verändern LLM-Komponenten den klassischen Software-Lifecycle, da probabilistische Modelle neue Fehlerbilder hervorbringen können (Chen et al., 2025). Beispielsweise kann es zu unvorhergesehenen falschen Antworten kommen, wodurch Entwickler und Qualitätssicherungsteams gefordert sind, sich mit Prompt Engineering auseinanderzusetzen und geeignete Methoden zu finden, um Modellantworten zu testen (Nahar et al., 2024). Darüber hinaus sind Monitoring und Qualitätssicherung komplexer, weil die Ausgaben von LLMs nicht einfach mit Unit-Tests validiert werden können (Ipek Ozkaya et al., 2023). Neue Metriken für Sprachqualität, Kosten oder Energieverbrauch rücken stärker in den Fokus. Gelingt die Integration, lässt sich das Potenzial von LLMs jedoch in Form höherer Produktivität oder neuer digitaler Dienstleistungen nutzen (NVIDIA, n. d. b).

2.2. LLMs in Betrieben

Die Integration von LLMs in Unternehmen bietet vielfältige Möglichkeiten zur Optimierung von Geschäftsprozessen. Kanabar et al. (2024) betonen, dass generative KI-Systeme, wie ChatGPT, Unternehmen unterstützen können, indem sie Arbeitsabläufe beschleunigen, Risiken bewerten und in Echtzeit zuverlässige Lösungen bereitstellen. Diese Systeme bieten über klassische regelbasierte Automatisierungen hinaus eine tiefere Kontextualisierung und ermöglichen eine natürlichere Interaktion mit Nutzer*innen (Kanabar et al., 2024). Die Einsatzmöglichkeiten erstrecken sich über zahlreiche Unternehmensbereiche.

Einer der bekanntesten Anwendungsbereiche ist das Personalwesen (HR). Hier ermöglichen LLMs die automatisierte Analyse von Bewerbungen, optimieren den Onboarding-Prozess neuer Mitarbeitender und unterstützen bei der Erstellung personalisierter Schulungsprogramme (Ekuma, 2024). Der Kundenservice profitiert von 24/7-Chatbots, die Kundenanfragen in Echtzeit bearbeiten und dadurch Wartezeiten reduzieren sowie die Kundenzufriedenheit verbessern (Abel Uzoka et al., 2024).

Ein weiteres wichtiges Anwendungsfeld ist das Wissensmanagement, in dem LLMs dazu beitragen, große Dokumentensammlungen zu strukturieren, relevante Informationen effizient zu extrahieren und die interne Suche innerhalb von Unternehmen zu optimieren (F. Jiang et al., 2024). Schließlich spielt auch der IT-Support eine zentrale Rolle. LLMs können bei der automatisierten Fehlerdiagnose unterstützen, Lösungen für häufig auftretende technische Probleme bereitstellen und Entwickler durch intelligente Code-Vervollständigungen unterstützen (Shareef, 2024).

2.2.1. Chancen und Herausforderungen beim Einsatz von LLMs in Unternehmen

Ein zentraler Vorteil von LLMs ist ihre Fähigkeit, natürliche Sprache zu verarbeiten und große Mengen an unstrukturierten Daten zu analysieren (NVIDIA, n. d. b). LLMs ermöglichen es, Marktdaten schnell zu analysieren, was zu beschleunigten Entscheidungsprozessen und einer höheren Produktivität führen kann. Sie können zudem die interne Kommunikation verbessern, indem sie präzisere Informationen bereitstellen, und Kundeninteraktionen optimieren (Kanabar et al., 2024; McKinsey, 2025). Somit liefern LLMs dem Management zuverlässige Entscheidungsgrundlagen, indem sie große Datenmengen verarbeiten (Xu et al., 2025).

Ein weiterer Vorteil von LLMs ermöglicht es, Arbeitsprozesse, insbesondere wiederkehrende Aufgaben, zu automatisieren, was nicht nur Zeit spart und die Effizienz in Unternehmensprozessen erhöht, sondern Unternehmen auch einen Wettbewerbsvorteil verschaffen kann (Kanabar et al., 2024; Safar, 2024). Laut Guan et al. (2023) können sich Mitarbeiter*innen durch die Automatisierung solcher Routineaufgaben stärker auf strategische Tätigkeiten konzentrieren, wodurch diese Prozesse effizienter abgewickelt werden und Unternehmen sich stärker auf strategische Entscheidungen und Wertschöpfung fokussieren können. Generative KI kann somit als zentraler Baustein für die Entwicklung moderner Unternehmenslösungen gesehen werden, da sie personalisierte Interaktionen ermöglicht und gleichzeitig organisatorische Ziele unterstützt (Kanabar et al., 2024).

Neben diesen Chancen bringt die Implementierung von LLMs in Unternehmen jedoch erhebliche Herausforderungen mit sich, die sowohl technische als auch organisatorische und regulatorische Aspekte betreffen. So stellen die Verfügbarkeit und Qualität der benötigten Daten eine zentrale Herausforderung

dar (Urlana et al., 2024; Zhao et al., 2024). LLMs benötigen große Mengen an strukturierten und unstrukturierten Daten, deren Konsistenz und Aktualität essenziell für präzise Ergebnisse sind (Liu et al., 2024). Wie Urlana et al. (2024) und Zhao et al. (2024) darlegen, können unvollständige oder verzerrte Datensätze die Leistung von LLMs erheblich beeinträchtigen. Zudem unterliegen viele branchenspezifische Daten regulatorischen Beschränkungen, was die Entwicklung spezialisierter Modelle erschwert. Eng damit verknüpft ist die Einhaltung von Datenschutz- und Compliance-Vorgaben, insbesondere in regulierten Branchen, in denen gesetzliche Anforderungen strikt zu erfüllen sind (Urlana et al., 2024). Zu diesen gehört unter anderem die Datenschutz-Grundverordnung (DSGVO) die den Schutz personenbezogener Daten und den freien Datenverkehr innerhalb der EU regelt (Europäische Union, 2016).

Ein weiterer wesentlicher Aspekt ist die technische Integration in bestehende IT-Systeme. Unternehmen verfügen häufig über heterogene Softwarelandschaften, die eine reibungslose Anbindung erfordern. Dabei müssen Schnittstellen (APIs) und Middleware so angepasst oder neu entwickelt werden, dass das LLM in bestehende Kernprozesse, etwa in ERP-, CRM- oder DMS-Systemen, eingebunden werden kann (Urlana et al., 2024; Zhao et al., 2024). In vielen Fällen empfiehlt sich ein integrierter Ansatz, der sowohl Development and Operations (DevOps) als auch Machine Learning Operations (MLOps) umfasst. Dieser Ansatz ermöglicht nicht nur eine effiziente Umsetzung von Updates und Wartungsarbeiten, unterstützt durch etablierte Monitoring- und Qualitätsprüfungen (beispielsweise mittels 4-Augen-Prinzip), sondern fördert auch die bereichsübergreifende Zusammenarbeit. Während DevOps darauf abzielt, die Kooperation zwischen Entwicklungs- und Betriebsteams zu stärken, sodass Software schneller und effizienter entwickelt und bereitgestellt wird, adaptiert MLOps diese Prinzipien für maschinelles Lernen, um die Prozesse der Modellentwicklung, des Deployments und der kontinuierlichen Überwachung von Machine Learning-Modellen zu optimieren (Gartner, 2025a; Tim Mucci & Cole Stryker, 2025).

Mit zunehmender Nutzung eines LLMs steigen außerdem die Anforderungen an Rechenkapazität und Budget. Ein plötzlich erhöhter Anfragestrom, wie er beispielsweise in Chatbot-Anwendungen auftreten kann, führt zu unerwartet hohen Cloud-Kosten oder einer Überlastung der On-Premises-GPU-Kapazitäten (Casado et al., 2023). Aus diesem Grund ist es unerlässlich, Last- und Kostenprognosen zu erstellen. Hierzu können auch Strategien wie Token-Limits, Caching, Modellkompression oder Batch-Verarbeitung beitragen, die in Kombination mit regelmäßiger Berichterstattung über KPI-Dashboards die Kostentransparenz und -kontrolle erhöhen. Schließlich stellt auch die langfristige Betriebskostenkontrolle eine Herausforderung dar. Das gewählte LLM muss sowohl in Bezug auf die technische Infrastruktur als auch in wirtschaftlicher Hinsicht den Budgetvorgaben entsprechen. Hierbei ist eine sorgfältige Prüfung aller Kostenfaktoren unabdingbar.

Organisatorische Herausforderungen ergeben sich darüber hinaus aus der veränderten Zusammenarbeit, die der Einsatz von LLMs mit sich bringt (Berretta et al., 2023). Mitarbeiter können aufgrund von Skepsis gegenüber automatischen Entscheidungshilfen oder Angst vor Arbeitsplatzverlust zurückhaltend agieren (Soulami et al., 2024). Um hier einer negativen Dynamik entgegenzuwirken, sind frühzeitige Change-Management-Maßnahmen essenziell (Soulami et al., 2024). Schulungen, Pilotphasen und klare Kommunikationsstrategien können dazu beitragen, den Mehrwert der KI-Anwendungen zu verdeutlichen (P. Jiang et al., 2024). Die Einrichtung von Change Agents oder KI-Botschaftern in den Fachbereichen kann zusätzlich die Akzeptanz fördern (Ekuma, 2024).

Nicht zuletzt muss auch die kontinuierliche Anpassung an sich ändernde regulatorische Rahmenbedingungen beachtet werden. Neben den Datenschutzanforderungen fordern neue Regelungen, wie beispielsweise der EU AI Act, eine flexible Governance-Struktur (European Parliament and Council of the European Union, 2024). Ein "AI Center of Excellence" oder ein vergleichbares Gremium kann dabei helfen, die Einhaltung von Compliance-Vorgaben sowie die Standardisierung von Freigabeprozessen und Monitoring-Verfahren institutionell zu verankern (Gallegos et al., 2024; Zimmergren, 2024).

Um einige der Herausforderungen, insbesondere die Problematik der Halluzinationen, zu adressieren, hat die Industrie vielversprechende Ansätze entwickelt, die im Kapitel 2.3 näher erläutert werden.

2.3. Retrieval-Augmented Generation

Wie von Lewis et al. (2021) dargelegt, speichern vortrainierte LLMs Wissen in ihren Parametern und erzielen beeindruckende Ergebnisse in NLP-Aufgaben. Allerdings weisen aktuelle Studien darauf hin, dass diese Modelle zu Halluzinationen neigen, indem sie nicht überprüfbare oder fehlerhafte Fakten generieren (Gao et al., 2024; Lewis et al., 2021; Merritt, 2025). Dies resultiert aus der Tatsache, dass ihr internes Wissen nicht aktiv aktualisiert werden kann.

Ein vielversprechender Ansatz zur Reduktion dieser Problematik ist die Retrieval-Augmented Generation (RAG). Diese Methode kombiniert ein LLM mit einer externen Wissensdatenbank, wodurch relevante Informationen während der Textgenerierung abgerufen werden können (Gao et al., 2024; Merritt, 2025). Dadurch wird es möglich, faktenbasierte und überprüfbare Antworten zu generieren, ohne ausschließlich auf den parametergestützten Speicher des Modells angewiesen zu sein (Lewis et al., 2021).

Während dieses Kapitels die theoretischen Grundlagen und architektonischen Konzepte von RAG erläutert werden, fokussiert sich Abschnitt 4.4 auf die praktische Vorbereitung und kontinuierliche Pflege der für RAG erforderlichen Datenbasis, um eine optimale Systemleistung sicherzustellen.

2.3.1. Prinzip von RAG: Kombination aus parametrischem und nicht-parametrischem Speicher

Das Eingangs von Lewis et al. (2021) erstellte Modell kombiniert einen parametrischen Speicher, also ein vortrainiertes LLM, das für die Textgenerierung optimiert wurde, mit einem nicht-parametrischen Speicher in Form einer externen Wissensdatenbank. Diese ist als Vektordatenbank organisiert und ruft relevante Dokumente auf Basis semantischer Ähnlichkeit ab (Lewis et al., 2021).

Diese hybride Architektur ermöglicht gezielte Wissensabfragen während der Generierung und trägt zur Erklärbarkeit und Transparenz bei, da die verwendeten Quellen explizit offengelegt werden können (Merritt, 2025). Zudem lassen sich Wissensaktualisierungen einfacher durchführen, da lediglich die externe Datenbank angepasst werden muss, ohne dass ein erneutes Training des LLM erforderlich ist (Lewis et al., 2021).

2.3.2. Technische Architektur: Vektordatenbanken, Indexierung und Suchstrategien

Das Herzstück eines RAG-Systems bildet die Vektordatenbank, wie in Abbildung 2.3 gezeigt, welche die effiziente Speicherung und Durchsuchung externer Wissensquellen ermöglicht (Merritt, 2025). Lewis et al. (2021) beschreiben hierzu das Konzept des Dense Passage Retrieval (DPR), das auf einem zweistufigen Verfahren basiert (Karpukhin, Oğuz et al., 2020). Zunächst werden externe Wissensquellen in kleinere Textabschnitte zerlegt und als Vektoren gespeichert. Anschließend erfolgt die semantische Suche mittels Maximum Inner Product Search (MIPS), wobei die relevantesten Dokumente zur weiteren Verarbeitung abgerufen werden (Lewis et al., 2021).

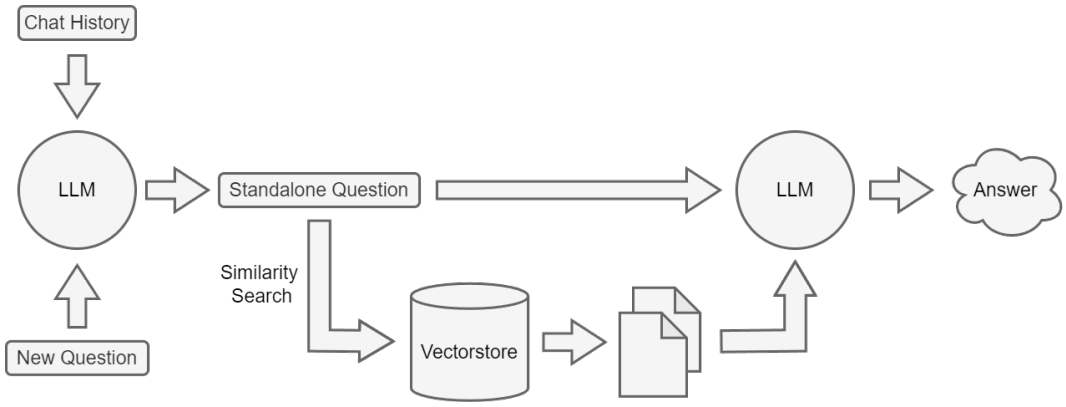


Abbildung 2.3.: RAG Architektur (vgl. Merritt (2025))

Innerhalb von RAG gibt es zwei unterschiedliche Strategien für die Dokumentenintegration (Gao et al., 2024). Während RAG-Sequence die gesamte Antwort auf einem einzigen abgerufenen Dokument basiert, erlaubt RAG-Token den Zugriff auf unterschiedliche Dokumente für jeden generierten Token, was flexiblere und präzisere Antworten ermöglicht (Lewis et al., 2021).

Zur technologischen Umsetzung kommen Vektordatenbanken wie FAISS oder Pinecone zum Einsatz, um eine schnelle und skalierbare Ähnlichkeitssuche zu ermöglichen (Merritt, 2025). Transformer-basierte Abfrage-Encoder wie BERT gewährleisten zudem eine effiziente semantische Repräsentation der Anfragen (Lewis et al., 2021).

2.3.3. Vorteile und Herausforderungen von RAG

RAG reduziert Halluzinationen, verbessert die Erklärbarkeit und ermöglicht eine einfache Wissensaktualisierung (Gao et al., 2024). Der Zugriff auf überprüfbare Quellen erhöht die Faktengenauigkeit, während Unternehmen ihre internen Wissensquellen besser nutzen können (Lewis et al., 2021).

Allerdings gibt es auch Herausforderungen bei der Implementierung. So kann man drauf schließen, dass der erhöhte Rechenaufwand für das Abrufen, Bewerten und Integrieren externer Dokumente die Antwortzeiten verlängern kann. Zudem ist die Qualität der generierten Antworten stark von der Struktur und Qualität der Vektordatenbank abhängig (Karpukhin, Oğuz et al., 2020). Fehlerhafte Indexierungen können zu irrelevanten oder fehlerhaften Ergebnissen führen. Darüber hinaus stellt die Skalierbarkeit eine Herausforderung dar, insbesondere wenn große Unternehmens-Wissensbestände effizient durchsucht werden müssen (Lewis et al., 2021).

2.3.4. Praxisbeispiel: RAG in Unternehmens-Wissensmanagementsystemen

Die Anwendung von RAG in Unternehmensumgebungen bietet erhebliche Potenziale, insbesondere im Bereich des Wissensmanagements (Gao et al., 2024). Unternehmen verwalten große Mengen an unstrukturierten Daten in internen Datenbanken, Dokumentenarchiven oder Support-Tickets. Ein praxisnahes Beispiel ist die automatisierte Beantwortung von Support-Anfragen:

Ein Mitarbeiter sucht Informationen zu einem firmeninternen Prozess. Die semantische Suche in der Unternehmensdatenbank wandelt die Anfrage in Vektorform um und vergleicht sie mit vorhandenen Dokumenten (Evidently AI, 2025). Anschließend werden die relevantesten Dokumente geladen, und das LLM kombiniert die abgerufenen Inhalte zu einer präzisen, faktenbasierten Antwort.

Hieraus ergeben sich zahlreiche Vorteile. Die Effizienz der internen Wissenssuche wird erheblich gesteigert, da Mitarbeiter schnell auf relevante Informationen zugreifen können (Evidently AI, 2025). Gleichzeitig können häufig wiederkehrende Anfragen automatisiert beantwortet werden, wodurch IT- und Support-Abteilungen entlastet werden. Die systematische Nutzung interner Wissensquellen führt zudem zu einer besseren internen Wissensvernetzung, was wiederum langfristig die Entscheidungsfindung unterstützt (Lewis et al., 2021).

3. Methodik

In diesem Kapitel wird näher auf die methodische Vorgehensweise der vorliegenden Arbeit eingegangen. Ziel ist es, die gewählte Methodik nachvollziehbar zu begründen und die einzelnen Schritte der Untersuchung systematisch darzustellen. Dabei wird insbesondere auf die Entwicklung eines wissenschaftlich begründeten und praxisorientierten Frameworks zur technischen Vorbereitung von Unternehmen auf LLMs eingegangen.

Im Rahmen dieser Arbeit wurde generative Künstliche Intelligenz als innovatives Instrument zur Ideengenerierung, Strukturierung und sprachlichen Optimierung der wissenschaftlichen Texte eingesetzt. Hierbei kamen die Versionen ChatGPT 4o, o1 und o3 zum Einsatz, um durch gezielte Abfragen die zu bearbeitende Forschungslücke präziser zu definieren. Während des gesamten Erstellungsprozesses unterstützte ChatGPT die Formulierung von Sätzen, Absätzen und inhaltlichen Aussagen. Es ist jedoch ausdrücklich hervorzuheben, dass der wesentliche inhaltliche Kern der Arbeit ausschließlich vom Autor stammt. Die KI diente dabei primär als Hilfsmittel, um die sprachliche Präzision und Konsistenz zu erhöhen. An ausgewählten Stellen wurden fremdsprachige Zitate mithilfe von ChatGPT ins Deutsche übersetzt und im Anschluss paraphrasiert, um eine einheitliche und verständliche Darstellung der zitierten Inhalte zu gewährleisten.

3.1. Design Science Research

Die Arbeit folgt dem Design Science Research (DSR) Ansatz, der sich besonders für die Entwicklung innovativer IT-Artefakte eignet und eine strukturierte Methodik bietet, um praxisnahe technische Lösungen zu entwerfen, zu evaluieren und iterativ zu verbessern. Dabei orientiert sich die Anwendung von DSR an einem sechsstufigen Prozess nach Peffers et al. (2007), wie in Abbildung 3.1 dargestellt.

Zunächst erfolgt die Problemidentifikation und Motivation (Kapitel 2 und Kapitel 4), bei der technische Herausforderungen für die LLM-Integration in Unternehmen analysiert und zentrale Anforderungen für Datenstrukturen und Skalierbarkeit identifiziert werden.

Darauf aufbauend wird in der Zieldefinition einer Lösung (Kapitel 6) ein Framework zur IT-Vorbereitung auf LLMs entwickelt, wobei bestehende Technologien und Standards berücksichtigt werden.

In der Phase Design & Entwicklung (Kapitel 6) erfolgt die Gestaltung und Entwicklung eines anwendungsorientierten Frameworks, welches auch für die technische Implementierung genutzt werden kann. Anschließend wird in der Demonstrationsphase das entwickelte Framework (Kapitel 6) auf praxisnahe Unternehmensszenarien angewendet und die Übertragbarkeit auf verschiedene IT-Infrastrukturen bewertet. Die Evaluationsphase (Kapitel 7) umfasst die Validierung durch Expert*inneninterviews sowie eine qualitative Inhaltsanalyse zur Bewertung der Praxistauglichkeit. Abschließend werden die Ergebnisse in der Phase der Kommunikation (Kapitel 8 und Kapitel 9) zusammengefasst, diskutiert und daraus Empfehlungen für Unternehmen abgeleitet, insbesondere hinsichtlich der Implikationen für IT-Abteilungen und Entscheidungsträger.

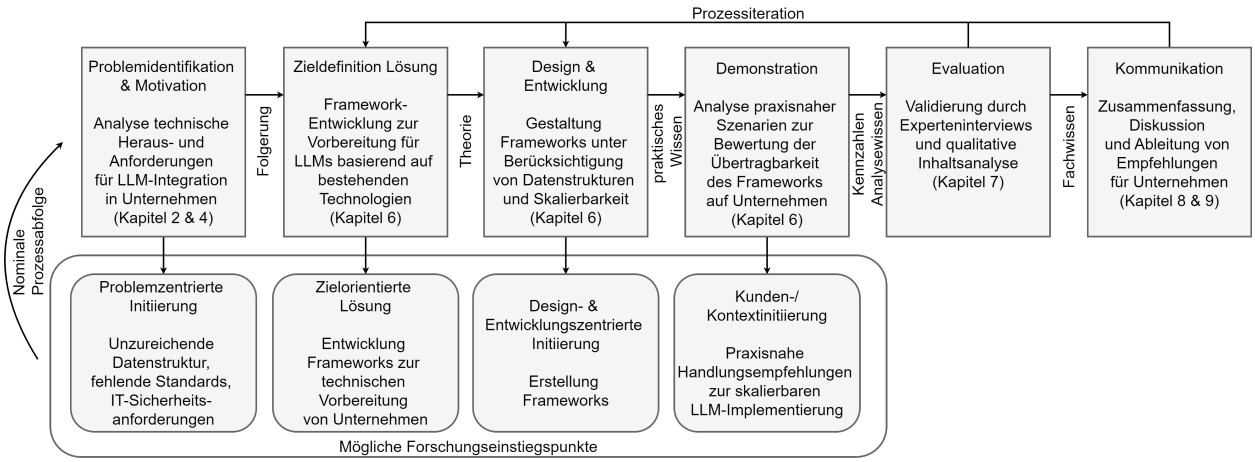


Abbildung 3.1.: DSRM Process (in Anlehnung an Peffers et al. (2007))

3.2. Literaturrecherche

Die vorliegende Arbeit stützt sich auf eine systematische Literaturrecherche, die sowohl der theoretischen Fundierung als auch der Ableitung spezifischer technischer Anforderungen dient. Ziel ist es, den aktuellen Stand der Forschung zur Integration von LLMs in Unternehmensumgebungen zu erfassen, bestehende Herausforderungen zu analysieren und Forschungslücken zu identifizieren. Insbesondere wird untersucht, welche infrastrukturellen und technischen Voraussetzungen Unternehmen erfüllen müssen, um LLMs effizient und sicher in bestehende IT-Landschaften zu integrieren.

Die Recherchemethodik folgt einem kombinierten Ansatz aus Backward und Forward Search. Während die Backward Search die Referenzen zentraler Publikationen analysiert, um die theoretischen Grundlagen systematisch zu erfassen, ermöglicht die Forward Search die Identifikation aktueller wissenschaftlicher Arbeiten, die auf etablierte Forschungsergebnisse referenzieren. Diese Vorgehensweise stellt sicher, dass sowohl grundlegende als auch gegenwärtige wissenschaftliche Erkenntnisse in die Untersuchung einfließen (Horn, 2024).

Die Auswahl relevanter Fachliteratur erfolgte anhand wissenschaftlicher Kriterien wie Peer-Review-Status, Zitationshäufigkeit und inhaltlicher Relevanz. Hierfür wurden etablierte wissenschaftliche Datenbanken wie SpringerLink, IEEE Xplore, arXiv und Google Scholar genutzt. Die inhaltliche Analyse fokussiert sich auf zentrale technische Aspekte der LLM-Integration. Dabei werden architektonische und infrastrukturelle Anforderungen untersucht, existierende Open-Source- und kommerzielle LLM-Lösungen verglichen und evaluiert sowie Fragen der Datenverarbeitung und -formate behandelt. Ergänzend wurden sicherheits- und datenschutzrelevante Aspekte sowie die strategischen Implikationen von On-Premises- und Cloud-Hosting-Lösungen analysiert.

Die Ergebnisse der Literaturrecherche bilden die Grundlage für die Ableitung der technischen Anforderungen, die in Kapitel 4 systematisch hergeleitet werden. Darüber hinaus liefert die Analyse bestehender wissenschaftlicher Arbeiten zentrale Impulse für die Konzeption des in Kapitel 6 vorgestellten Frameworks. Durch die Identifikation bestehender Forschungslücken konnte zudem ein gezielter methodischer Zugang für die empirische Untersuchung mittels leitfadengestützter Expert*inneninterviews (Kapitel 7) entwickelt werden.

3.3. Inhaltsanalyse

Die systematische Auswertung der Expert*inneninterviews erfolgt in Anlehnung an die qualitative Inhaltsanalyse nach Mayring und Fenzl (2019). Dabei werden zentrale Elemente dieser Methode verwendet, um eine strukturierte Analyse qualitativer Daten zu gewährleisten und tiefgehende Erkenntnisse über die technischen Anforderungen für die Integration von LLMs in Unternehmen zu gewinnen. Die methodische Umsetzung orientiert sich insbesondere an der Definition der Analyseeinheiten, der Festlegung eines Kategoriensystems sowie der Bestimmung der Richtung der Analyse. Eine vollständige Anwendung der Methode nach Mayring erfolgt jedoch nicht, sondern die Analyse wird flexibel an die spezifischen Anforderungen dieser Untersuchung angepasst.

Die Kategorisierung der Interviewdaten basiert auf einer Kombination induktiver und deduktiver Verfahren. Während induktive Kategorien aus dem erhobenen Datenmaterial heraus entwickelt werden, erfolgt die deduktive Kategorisierung auf Grundlage bestehender theoretischer Konzepte.

Um die Konsistenz und Nachvollziehbarkeit der Analyseprozesse sicherzustellen, wird ein detaillierter Kodierleitfaden erstellt, der die einheitliche Anwendung der Kategorien unterstützt (Mayring & Fenzl, 2019). Zusätzlich erfolgt eine kontinuierliche Reflexion der Kategorisierung während der Analyse, um sicherzustellen, dass die abgeleiteten Anforderungen die relevanten technischen Aspekte der LLM-Integration präzise widerspiegeln.

4. Technische Anforderungen

Die in den vorangegangenen Kapiteln beschriebenen Einsatzmöglichkeiten von LLMs veranschaulichen bereits, wie leistungsfähig diese Modelle im Unternehmenskontext sein können. Für einen reibungslosen und sicheren Betrieb sind jedoch umfangreiche technische Vorkehrungen nötig.

Abschnitt 4.1 beleuchtet die erforderlichen Rechenressourcen und Infrastrukturen, während Abschnitt 4.2 auf geeignete Datenformate und Strategien zur Archivierung großer Textmengen eingeht. Anschließend diskutiert Abschnitt 4.3 verschiedene Modelle zur Bereitstellung, von rein lokalen Lösungen bis hin zur Cloud. Abschließend widmet sich Abschnitt 4.4 den Aspekten bezüglich Vorbereitung von Daten damit diese in RAG verwendet werden kann.

4.1. Rechenleistung und Infrastruktur

Die Entwicklung und Nutzung großer Sprachmodelle erfordert erhebliche Rechenressourcen. Modernste LLMs werden auf spezialisierten Hardware trainiert. So berichteten Chowdhery et al. (2022) von Google, dass ihr PaLM-Modell mit 540 Milliarden Parametern auf 6144 TPU-v4-Chips parallel trainiert wurde. Ebenso nutzte Meta für OPT-175B (175 Mrd. Parameter) ein Cluster von 992 NVIDIA A100-GPUs (80 GB) im Verbund (Zhang et al., 2022a). Solche massive Parallelisierung verteilt das Training über tausende Recheneinheiten.

Metas LLaMA-Modell mit 65 Mrd. Parametern beispielsweise benötigt etwa 16 A100-GPUs für eine einzelne Anfrage, während das Training 2.000 GPUs verlangte (Yeluri, 2023).

Obwohl die meisten Industrieunternehmen vortrainierte LLMs und spezialisierte Dienstleister nutzen, um Kosten und Ressourcen zu sparen, verdeutlichen die in den folgenden Abschnitten dargestellten hohen Anforderungen an Rechenleistung, Speicher und Netzwerk die technischen Herausforderungen beim Einsatz von LLMs. Gleichzeitig zeigt die Unternehmenspraxis, dass vortrainierte Modelle, die durch gezieltes Fine-Tuning an spezifische Anwendungsfälle angepasst werden, eine kosteneffiziente Alternative zum eigenständigen Training darstellen, ein Ansatz, der den erheblichen Ressourcen- und Kostenaufwand deutlich reduziert (Maslej et al., 2023).

4.1.1. Speicherbedarf

Ein Sprachmodell mit 1 Billion Parametern benötigt in 16-Bit-Präzision während des Trainings rund 16 TiB GPU-Speicher, hauptsächlich bedingt durch die Modellgewichte, den Optimizer-Status und die Gradienten (Lockwood, 2025). Selbst kleinere Modelle benötigen bereits erhebliche Kapazitäten. Ein 7 Milliarden Parameter großes Modell beansprucht etwa 28 GB Speicher in FP32 (oder ca. 14 GB in FP16) nur für die Gewichte (Technologies, 2025).

Schätzungen zufolge verfügt GPT-4 über rund 1,8 Billionen Parameter, was etwa 7,2 TB Speicher erfordert, deutlich mehr als die 80 GB einer NVIDIA H100-GPU, und macht daher den Einsatz mehrerer GPUs unverzichtbar (NVIDIA, n. d. a; Walker, 2023).

Laskin (2023) beschreibt in ihrem Beitrag Techniken wie Model Parallelism und Sharding, die notwendig sind, um das Modell über den GPU-Speicher mehrerer Karten zu verteilen. Beim Model Parallelism wird das Modell in mehrere Segmente zerlegt, sodass jede GPU einen spezifischen Teil berechnet (Laskin, 2023). Dies ermöglicht es, größere Modelle zu trainieren, als es der Speicher einer einzelnen GPU zulassen würde (Laskin, 2023). Sharding hingegen verteilt den Speicherbedarf des Modells auf mehrere GPUs, sodass jede GPU nur einen Teil der Daten speichert und verarbeitet (Laskin, 2023).

4.1.2. Netzwerk und Infrastruktur

Da das Training großer LLMs auf verteilten Systemen erfolgt, sind Hochgeschwindigkeits-Netzwerke und eine optimierte Netzwerk-Topologie von zentraler Bedeutung (Russinovich, 2023). Insbesondere ermöglichen diese Technologien, die enormen Inter-GPU-Datenströme effizient zu bewältigen und Engpässe zu vermeiden, die zu längeren Trainingszeiten und reduzierter GPU-Auslastung führen könnten (Yeluri, 2023).

In diesem Zusammenhang spielen moderne Interconnect-Fabrics sowie verlustarme Netzwerke wie InfiniBand eine kritische Rolle in Bezug auf Performance und Kosten (Russinovich, 2023). Beispielsweise werden in aktuellen KI-Supercomputern GPUs innerhalb eines Servers über NVIDIA NVLink/NVSwitch mit Durchsatzraten von mehreren Terabyte pro Sekunde verbunden, während InfiniBand-Netze den Datenaustausch über Server-Grenzen hinweg ermöglichen (Brian et al., 2024; Russinovich, 2023).

Auch Google setzt bei seinen TPU-Pods auf maßgeschneiderte Netzwerk-Topologien, um tausende Chips effizient zu koppeln (Yeluri, 2023). Dank dieser schnellen Verbindungen können Gradienten- und Parameter-Updates nahezu in Echtzeit ausgetauscht werden, ohne dass die Skalierbarkeit des Trainings beeinträchtigt wird (Russinovich, 2023).

4.2. Datenformate und Archivierung

LLMs werden hauptsächlich mit großen unstrukturierten Textdaten aus verschiedenen Quellen wie Webseiten, Büchern, Code und Chats trainiert (Liu et al., 2024). Zunächst müssen solche Rohdaten in ein verarbeitbares Format überführt werden (Liu et al., 2024). In Unternehmen liegen Texte oft in diversen Formaten vor wie z. B. PDFs, HTML-Seiten, JSON/CSV-Dateien oder Office-Dokumente (Kruschwitz & Hull, 2017). Der reine Text wird extrahiert und nicht-textuelle Elemente wie HTML-Markup, CSS/JS-Scripts und Sonderzeichen werden entfernt oder normalisiert, bevor der Fließtext weiterverarbeitet wird (Simon Zamarin et al., 2024).

Der Text wird tokenisiert, also in diskrete Einheiten (Tokens) zerlegt, die das Modell benötigt. Moderne LLMs verwenden häufig Byte-Pair Encoding oder ähnliche Algorithmen; so nutzten die Entwickler von GPT-3 und OPT den GPT-2-BPE-Tokenizer für ihr gesamtes Korpus (Brown et al., 2020; Zhang et al., 2022b).

Im Falle von OPT-175B, ergab die Tokenisierung rund 180 Milliarden Tokens aus den gesammelten Textdaten (Zhang et al., 2022b).

Tokenisierte Datensätze sind deutlich kompakter als Rohtext und direkt vom Modell konsumierbar (Glenn K. Lockwood, 2025). Viele öffentliche LLM-Trainingsdaten werden in Form von JSON/JSONL-Dateien

mit Metadaten bereitgestellt, welche dann beim Laden ins Training in Tokens umgewandelt werden (Glenn K. Lockwood, 2025).

Für die effiziente Speicherung und Verarbeitung solcher Daten kommen optimierte Formate zum Einsatz. Während JSON-Lines, eine Textzeile pro Dokument und Beispiel, leicht lesbar und einfach zu handhaben ist, stoßen reine Textformate bei sehr großen Datenmengen an I/O-Grenzen (M. Ahmed, 2024).

Daher werden für Massendaten oft binäre Datenformate wie TensorFlows TFRecord verwendet, da sie sequentielle Binärblöcke enthalten und deutlich performanter beim Lesen/Schreiben großer Datenstreams sind (M. Ahmed, 2024).

Für kleinere bis mittlere Datensätze ist JSONL aufgrund seiner Einfachheit praktikabel, während bei riesigen Trainingscorpora (im Terabyte-Bereich) TFRecord oder ähnliche Binärformate klare Vorteile in Durchsatz und Speicherbedarf bieten (M. Ahmed, 2024). Letztere komprimieren die Daten und sind speziell auf das Streaming in Machine-Learning-Pipelines optimiert (M. Ahmed, 2024). In der Praxis werden Trainingsdaten häufig in Shards (Teildateien) organisiert, um paralleles Laden und Vorverarbeiten auf verteilten Rechenknoten zu ermöglichen (Wenzek et al., 2019).

Die Archivierung und langfristige Speicherung dieser enormen Datensätze stellt ebenfalls eine erhebliche Herausforderungen dar. Zum Beispiel umfasst das für LLaMA 2 verwendete Korpus etwa 2 Billionen Tokens, was in der Größenordnung von 8 TB aufbereiteten Textdaten liegt (Glenn K. Lockwood, 2025). OpenAIs GPT-3 verwendete 300 Mrd. Tokens (ca. 1,2 TB) und sogar öffentlich verfügbare Sammlungen wie The Pile (EleutherAI) enthalten 800 GB an Text (Glenn K. Lockwood, 2025). Entsprechend benötigen Unternehmen skalierbare Speicherlösungen (verteilte Dateisysteme, Object Stores oder Data Lakes), um diese Datenmengen vorzuhalten. Zur effizienten Indexierung und Wiederauffindbarkeit der Daten werden Metadaten und ggf. Datenbanken eingesetzt. Das Aufteilen großer Datenmengen in handhabbare Blöcke hat sich bewährt. So beschreiben Wenzek et al. (2019), dass 30 TB Webtext in 1.600 Shards à 5 GB mit jeweils 1,6 Mio. Dokumenten pro Shard zerlegt wurden (Simon Zamarin et al., 2024). Solche Aufteilungen ermöglichen parallele Verarbeitungsschritte wie Duplikaterkennung und Filtering auf vielen Knoten. Jeder Shard kann getrennt indiziert und mittels Hashing konsistent durchsucht werden (Simon Zamarin et al., 2024). Durch das Berechnen von Fingerprints, CCNet nutzt z. B. 64-bit SHA-1 Hashes pro Absatz, lassen sich Dubletten global identifizieren, indem gleiche Hashwerte innerhalb und über Shards hinweg gefunden und entfernt werden (Simon Zamarin et al., 2024). Über die Zeit ist es zudem wichtig, Versionen der Trainingsdaten zu archivieren (M. Ahmed, 2024). Eine sorgfältige Datenversionierung stellt sicher, dass bei erneuten Trainings oder Experimentiervarianten nachvollziehbar bleibt, welche Datenbasis verwendet wurde (M. Ahmed, 2024). Dies ist für Reproduzierbarkeit und Compliance (Nachweis der Datenherkunft) im Unternehmen essenziell.

4.3. Bereitstellung von LLMs: On-Premises vs. Cloud

Unternehmen stehen vor der Entscheidung, LLM-Infrastrukturen selbst zu betreiben (On-Premises) oder aus der Cloud als Service zu beziehen, oder Mischformen dazwischen zu nutzen. Beide Ansätze haben Vor- und Nachteile hinsichtlich Skalierbarkeit, Kosten und Regulierung.

On-Premises

On-Premises bietet maximale Kontrolle über Hardware, Daten und Sicherheit. Dies ist oft bevorzugt, wenn strikte Datenschutzanforderungen gelten oder sensible Daten nicht extern gelagert werden dürfen (Stichwort DSGVO). Bei lokaler Bereitstellung verbleiben alle Daten im eigenen Rechenzentrum, was volle Kontrolle über deren Verbleib und Schutz erlaubt (Dobosevych, Oles, 2025). Auch Sicherheitsmaßnahmen können vollständig selbst gestaltet werden (Dobosevych, Oles, 2025). Allerdings erfordert On-Premises einen hohen Initialaufwand an Hardwareanschaffung und Setup sowie kontinuierliche Wartung durch Experten (Convergence, 2024). Dieser Investitionsbedarf kann sich langfristig lohnen, da bei dauerhafter Nutzung die Gesamtkosten (TCO) oft niedriger ausfallen als bei kontinuierlichen Cloud-Gebühren (exxactcorp, n. d.).

Im On-Premises-Ansatz müssen Unternehmen zudem nicht notwendigerweise ein LLM von Grund auf selbst trainieren. Stattdessen können vortrainierte Modelle, mit bereits fertigen Parametern, auf den eigenen Servern betrieben werden (Dobur et al., 2024). Dadurch profitieren sie von den Vorteilen der lokalen Infrastruktur, wie vollständiger Datenkontrolle und individuellen Sicherheitsmaßnahmen, ohne den enormen Ressourcenaufwand für das Training großer Sprachmodelle stemmen zu müssen. Zudem ermöglicht das gezielte Fine-Tuning dieser vortrainierten Modelle eine Anpassung an unternehmensspezifische Anforderungen, wodurch der On-Premises-Ansatz eine praktikable Alternative für Industrieunternehmen darstellt (Dobur et al., 2024).

Cloud-basierte Lösungen

Cloud-LLM-Dienste sind sofort nutzbar und anfangs oft kostengünstiger (Pay-per-Use ohne hohe Kapitalkosten). Sie skalieren flexibel mit dem Bedarf, können jedoch bei intensivem Gebrauch auf Dauer teuer werden (Dobosevych, Oles, 2025). Cloud-Anbieter wie AWS, Google oder Azure stellen spezialisierte GPU/TPU-Ressourcen on-demand bereit, sodass auch kleine Teams große Modelle nutzen können. Dies geht allerdings mit einer Abhängigkeit vom Cloud-Anbieter einher, was sich auf Kosten und langfristige Flexibilität auswirken kann.

Hybrid-Modelle

Eine Mischform kombiniert Cloud- und In-House-Ressourcen. Unternehmen können sensible Daten und Kernanfragen lokal verarbeiten, für rechenintensive Aufgaben oder Spitzenlasten jedoch temporär auf die Cloud ausweichen (Convergence, 2024). Dies erhöht die Flexibilität und ermöglicht latenzkritische oder datenschutzrelevante Verarbeitung vor Ort, während weniger sensitive Teile in der Cloud skalieren.

Edge-Computing

Edge-Computing bezeichnet die Ausführung von Modellen direkt auf Endgeräten oder dezentralen Standorten (z. B. Industrie-PCs, Smartphones). Dies reduziert Latenzzeiten, ermöglicht Offline-Betrieb und schützt lokale Daten (Convergence, 2024). Allerdings sind Edge-Geräte in Rechenleistung und Speicher stark limitiert, weshalb dort nur kleinere, optimierte Modelle lauffähig sind.

Datenschutz und Compliance

Neben technischen und wirtschaftlichen Abwägungen spielen Datenschutz und Compliance eine entscheidende Rolle bei der Bereitstellungswahl. DSGVO fordert, dass personenbezogene Daten entsprechend geschützt werden und nur mit Rechtsgrundlage verarbeitet werden dürfen. Unternehmen mit sehr sensitiven Daten (z. B. im Gesundheitswesen oder Finanzsektor) tendieren deshalb zu On-Premises oder Private-Cloud-Lösungen, um die Datenhoheit zu behalten. Zwar sind große Cloud-Anbieter in der Regel nach ISO 27001 und ähnlichen Standards zertifiziert (Google Cloud, n. d.), dennoch müssen vertragliche Regelungen (z. B. Auftragsverarbeitung, Datenlokation, Drittstaatentransfer) genau geprüft werden (Dobosevych, Oles, 2025). So kann gewährleistet werden, dass eine DSGVO-konforme Datenverarbeitung stattfindet.

4.4. Vorbereitung von Daten für RAG

Die Vorbereitung von Daten für Retrieval-Augmented Generation (RAG) ist ein entscheidender Schritt, um sicherzustellen, dass Unternehmen das volle Potenzial dieser Technologie ausschöpfen können. Dazu gehören wesentliche Aspekte der Datenstruktur, des effizienten Retrievals, der Datenqualität sowie der Aktualisierung und Wartung der Wissensdatenbasis.

Datenstruktur

Damit RAG im Unternehmen effektiv arbeiten kann, müssen die Wissensbestände in strukturierter Form vorliegen. Große Dokumente werden hierzu in kleinere Textsegmente (z. B. ca. 100-Wort-Passagen) zerlegt, die als grundlegende Retrieval-Einheiten dienen (Karpukhin, Oguz et al., 2020). Jedes Segment wird anschließend mittels eines Embedding-Modells in einen hochdimensionalen Vektor überführt und in einer Vektordatenbank abgelegt (Lewis et al., 2021). Durch diese Darstellung können semantische Ähnlichkeiten zwischen Nutzeranfragen und Dokumentinhalten effizient berechnet werden. Wichtig ist dabei, dass jedem Vektor-Segment entsprechende Metadaten (wie Dokumentquelle oder Erstellungsdatum) zugeordnet werden, um die Nachvollziehbarkeit der Ergebnisse sicherzustellen (Hwang et al., 2025).

Effizientes Retrieval

Um in umfangreichen Knowledge Bases eine schnelle Suche zu ermöglichen, kommen spezielle Indexierungsverfahren zum Einsatz (Lewis et al., 2021). Ein optimierter FAISS-Index kann beispielsweise nahezu 1000 Anfragen pro Sekunde bearbeiten, dass entspricht um ein Vielfaches mehr als klassische keywordbasierte Suchverfahren (Merritt, 2025). Alle Dokumentvektoren werden dafür vorab offline berechnet

und indexiert, sodass Anfragen zur Laufzeit in Millisekunden die relevantesten Einträge liefern. Neben der Indexierung beeinflusst auch die Wahl des Embedding-Modells die Retrieval-Qualität: Ein auf die unternehmenseigene Domäne feinabgestimmtes Embedding (durch Finetuning) kann die semantische Treffergenauigkeit signifikant steigern und damit präzisere RAG-Antworten ermöglichen (Portes, Jacob et al., 2025).

Datenqualität

Die Qualität der Wissensdatenbasis ist ausschlaggebend für verlässliche generative Antworten. Die eingebundenen Informationen sollten autoritative und aktueller Herkunft sein, damit das LLM auf zuverlässiges und aktuelles Wissen zurückgreifen kann (AWS, n. d.). In der Praxis werden deshalb bevorzugt geprüfte interne Wissensquellen oder fachlich anerkannte Publikationen eingebunden (k2view, n. d.). Standard-RAG-Ansätze, die Dokumente allein nach inhaltlicher Relevanz abrufen, vernachlässigen mitunter Unterschiede in der Verlässlichkeit verschiedener Quellen, dies kann zur Übernahme von Fehlinformationen führen (Hwang et al., 2025). Daher gelten als Qualitätskriterien für RAG-Daten neben inhaltlicher Relevanz insbesondere die Vertrauenswürdigkeit und Aktualität der Quellen (AWS, n. d.; Hwang et al., 2025). Im Unternehmenskontext kommt hinzu, dass die Wissensbasis alle wichtigen Fachbegriffe und Dokumente der Domäne abdecken sollte, um eine hohe Antwortabdeckung zu gewährleisten (Hwang et al., 2025).

Aktualisierung und Wartung

RAG-Systeme erfordern kontinuierliche Pflege der Wissensdaten. Neue oder veränderte Informationen müssen zeitnah in die Datenbasis aufgenommen werden, damit das System stets mit dem neuesten Stand arbeitet (Merritt, 2025). Hierfür werden automatisierte Daten-Pipelines eingesetzt, die Änderungen in den Quellsystemen erkennen und die betroffenen Textsegmente neu extrahieren sowie inferieren (Embedding) (Merritt, 2025). Die Vektor-Datenbank wird im Hintergrund fortlaufend aktualisiert, indem neue Vektoren hinzugefügt und obsolete Einträge entfernt oder versioniert werden (Merritt, 2025). Durch diese dynamische Wartung bleibt die Retrieval-Komponente konsistent mit der realen Datenlage, ohne dass das Sprachmodell selbst neu trainiert werden muss (Merritt, 2025). Dies ist besonders im Unternehmensumfeld essenziell, da sich Vorschriften, Produkte oder Wissensinhalte häufig ändern und das System nur mit aktueller Wissensbasis vertrauenswürdige Auskünfte liefern kann (Merritt, 2025).

5. LLM-Lösungen

Der bereits mehrfach hervorgehobene Trend zur wachsenden Bedeutung von LLMs zeigt sich auch in der zunehmenden Anzahl verschiedener Modelle, wie Abbildung 5.1 verdeutlicht.

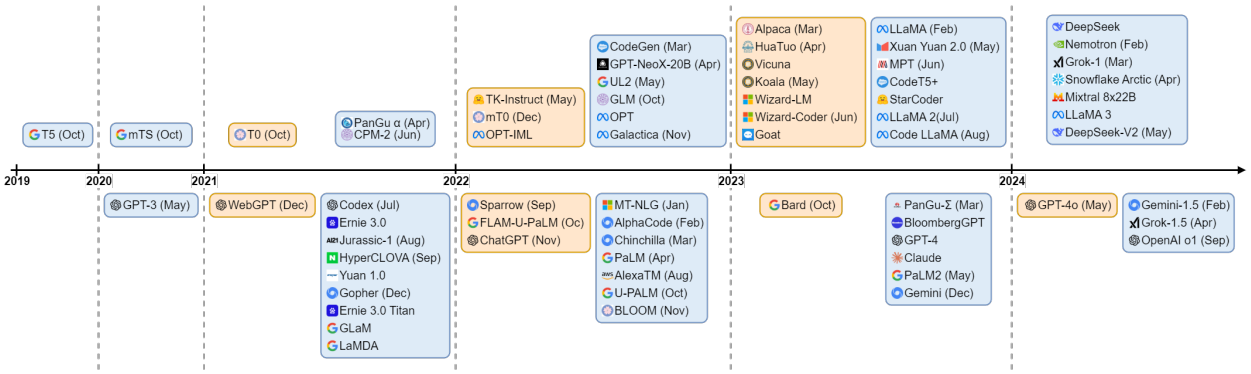


Abbildung 5.1.: Chronologische Darstellung der LLM-Veröffentlichungen (vgl. Naveed et al. (2024))

In dieser dargestellten Abbildung 5.1 teilen sich LLMs in zwei große Lager auf, die später in Abschnitt 5.2 gegenübergestellt werden. So beschreiben Naveed et al. (2024) in ihrer Arbeit, dass blaue Felder pre-trained Modelle repräsentieren, während orange Felder instruction-tuned Modelle darstellen. Modelle in der oberen Hälfte bedeuten Open-Source-Verfügbarkeit, während diejenigen in der unteren Hälfte Closed-Source sind (Naveed et al., 2024). Das Diagramm veranschaulicht den zunehmenden Trend zu instruction-tuned und Open-Source-Modellen und hebt die sich entwickelnde Landschaft und Trends in der Forschung zur Verarbeitung natürlicher Sprache hervor (Naveed et al., 2024).

Für Unternehmen stellt sich bei der Integration von LLMs in Geschäftsprozesse oft die zentrale Frage, ob ein Open-Source-Modell oder ein kommerzieller (proprietäre) Dienst genutzt werden soll. Eine einfache Antwort auf diese Frage ist schwer zu finden, da die in Kapitel 2 und Kapitel 4 bereits besprochenen Faktoren erheblichen Einfluss darauf haben.

Um die in Abbildung 5.1 dargestellte Vielfalt im Unternehmenskontext besser einordnen zu können, bietet Abschnitt 5.1 zunächst einen Überblick über die Merkmale verschiedener LLMs. Tabelle 5.1 fasst dabei fünf besonders häufig erwähnte Modelle zusammen. Abschließend stellt Abschnitt 5.2 die Unterschiede zwischen Open-Source- und kommerziellen Lösungen heraus, wobei unter anderem Datenschutz, Skalierbarkeit und Anbieterabhängigkeit beleuchtet werden.

5.1. Überblick über LLM-Lösungen

In ihrem Beitrag beschreibt Kelsie Anderson (2024), dass Open-Source-LLMs Modelle sind, deren Architektur öffentlich zugänglich ist. Dies ermöglicht es Dritten, diese Modelle selbst zu betreiben und anzupassen. Im Gegensatz dazu bieten KI-Anbieter kommerzielle LLMs als geschlossene Dienste an, wie Minaee et al. (2024) erläutert. Der Zugriff erfolgt hierbei typischerweise über Cloud-APIs, ohne dass ein direkter Einblick in das Modell gewährt wird. Tabelle 5.1 bietet eine umfassende Übersicht über die führenden LLMs zum aktuellen Stand der Forschung.

Eigenschaft	OpenAI – GPT-4	Anthropic – Claude	Google – Gemini	Meta – LLaMA	Mistral AI – Le Chat
Aktuelles Modell	GPT-4	Claude 3	Gemini 2.0	LLaMA 2	Le Chat (Beta)
Anbieter	OpenAI	Anthropic	Google DeepMind	Meta AI	Mistral AI (Frankreich)
Veröffentlichung	14.03.2023	14.03.2024	11.12.2024	Juli 2023	Feb 2024 (Beta-Launch)
Modellgröße (Parameter)	k.A.	k.A.	k.A.	7 Mrd., 13 Mrd., 70 Mrd.	7,3 Mrd.
Open Source?	Nein	Nein	Nein	Ja (Modelle & Gewichte frei verfügbar)	Ja (Apache 2.0-lizenziert)
Unterstützte Sprachen	Englisch (optimiert) & viele weitere	Englisch (Haupttraining); auch DE/FR	Über 40 Sprachen	Englisch, teils mehrsprachig	Mehrsprachig (EN, FR, DE, u.a.)
Multimodale Fähigkeiten	Text & Bild (GPT-4V seit 2023); Audio via separate Module	Text & Bild	Nativ multimodal (Text, Bild & Audio)	Nein (nur Text; separater Code-Modus)	Textbasiert; Bildverständnis (OCR) und -generierung integriert
Training Data Cutoff	Sept. 2021	Nov 2024	Aug 2024	2023	Okt 2023
API-Verfügbarkeit	Ja	Ja	Ja	Kein offizielles API (über Partner möglich)	Ja
Self-Hosting	Nein	Nein	Nein	Ja	Ja
Hardware-Anforderungen (Self-Hosting)	–	–	–	7B: 1 GPU (16 GB VRAM), 70B: mehrere High-End GPUs	7B läuft bereits auf Einzel-GPU, ggf. auch CPU
Lizenz / Kosten	Proprietär; API-Zugriff kostenpflichtig	Proprietär; API-Zugriff kostenpflichtig	Proprietär; API-Zugriff kostenpflichtig	Open Source (kostenfrei, keine Lizenzgebühr)	Open Source (kostenfrei); Basisfunktionen gratis; Pro kostenpflichtig
Datenschutz	API-Daten werden nicht zum Training genutzt (Speicherung max. 30 Tage)	API-Daten werden nicht zum Training genutzt (Löschung <30 Tage)	Bei bezahlter Nutzung kein Training auf Daten; unbezahlte Eingaben ggf. anonymisiert	Volle Datenhoheit bei Self-Hosting	Volle Datenkontrolle bei eigenem Hosting; Le Chat Enterprise ermöglicht On-Prem-Installation
Stärken	Vielschichtige Aufgaben, exzellentes Reasoning, hohe Kreativität, starke Coding-Fähigkeiten	Enormes Kontextfenster (100k Token), stark in Mathematik, Logik und Coding, freundlicher Dialogstil	State-of-the-Art in Benchmarks, native visuelle Fähigkeiten, exzellente Code-Generierung, agentisches Verhalten	Offen, anpassbar, solide Allround-NLP (mit Code Llama vergleichbar)	Schnelle Antworten, Integration aktueller Wissensquellen (Web, News), Vision+OCR, Code-Interpreter
RAG-Kompatibilität	Ja (bis 32k Token Kontext)	Ja (100k Token Kontext)	Ja (mit hohen Kontextlängen, $\sim 10^6$ Token angestrebt)	Eingeschränkt (nur 4k Token Kontext)	Ja (integrierte Websuche und Dokumenten-Upload)

Tabelle 5.1.: Übersicht über führende LLMs

erhoben aus: Anthropic (n. d. a, n. d. b, n. d. c, n. d. d, n. d. e), Bergmann, Dave (2023), DocsBot AI (2025), Dunenfeld, Emily (2024), Edwards (2023), Gillham, Jonathan (2024a, 2024b), Hassabis, Demis (2023), Hoblitzell, Andrew (2023), Joshi (2024), Meta (2024), Mistral AI Team (2023, 2024, 2025), OpenAI (2024, 2025), OpenAI et al. (2023), Portakal, Ertugrul (2024) und TypingMind (2025, n. d. a, n. d. b)

Die bisherigen Vergleiche in Tabelle 5.1 verdeutlichen die Unterschiede zwischen LLMs. Bereits frühzeitig wurden einige Modelle in der Forschung frei zugänglich gemacht. OpenAI veröffentlichte die ersten GPT-Versionen, etwa GPT-2, als Open-Source-Software, wohingegen neuere Generationen wie GPT-3 und GPT-4 initial ausschließlich über API-Zugänge verfügbar waren (Minaee et al., 2024). Ein wesentlicher

Meilenstein in der Entwicklung offener Modelle war die Veröffentlichung der LLaMA-Modellfamilie durch Meta AI im Februar 2023, bei der Gewichte mit Größen von 7 bis 65 Milliarden Parametern der Forschung zugänglich gemacht wurden (Minaee et al., 2024). Im Gegensatz zu GPT-4 ermöglichen LLaMA-Modelle, wenn auch unter bestimmten Lizenzauflagen, den eigenständigen Betrieb, was zahlreiche Innovationen anregte, da viele Teams LLaMA als Grundlage für offene Alternativen und spezialisierte Modelle nutzten (Minaee et al., 2024; Zhao et al., 2024).

Erste Vergleichstests zeigten sogar, dass das frei verfügbare LLaMA-13B-Modell (13 Milliarden Parameter) das proprietäre GPT-3-Modell (175 Milliarden Parameter) in vielen Benchmarks trotz seiner geringeren Größe übertreffen konnte (Minaee et al., 2024). Im Anschluss entstanden weitere offene LLMs, darunter BLOOM, Falcon, GPT-NeoX und StarCoder, die über Plattformen wie HuggingFace ohne Lizenzkosten verfügbar sind (Minaee et al., 2024; Zhao et al., 2024). Unternehmen haben hier die Möglichkeit, diese Modelle in eigener Umgebung zu hosten, was eine maximale Kontrolle über die Daten gewährleistet (Hugging Face, 2025a, 2025b). Zudem treibt eine aktive Community die kontinuierliche Weiterentwicklung offener Modelle voran (Hugging Face, 2025c). Als Randnotiz sei erwähnt, dass Modelle wie Stanford Alpaca oder Vicuna, die durch Feintuning von LLaMA-Modellen mit relativ geringem Aufwand erstellt wurden, in bestimmten Benchmarks bereits etwa 90% der Leistung von ChatGPT erreichen konnten (Minaee et al., 2024).

Zwischen diesen extrem leistungsfähigen Closed-Source-Modellen und den flexiblen, anpassbaren Open-Source-LLMs, kristallisieren sich unterschiedliche Lösungsansätze heraus.

Angesichts der enormen Ressourcen, die für die Entwicklung großer Modelle benötigt werden, darunter Datensätze, Rechenleistung und Expertise, erweist sich für die Mehrheit der Unternehmen der Rückgriff auf vorhandene Open-Source-Modelle oder den Einsatz kommerzieller LLM-Services als praktikablere Option, um von der LLM-Technologie zu profitieren.

5.2. Vergleich Open-Source vs. kommerzielle Modelle

Die Wahl zwischen einem offenen und einem kommerziellen LLM wirkt sich auf verschiedene Dimensionen aus. Tabelle 5.2 bietet einen strukturierten Vergleich der wichtigsten Kriterien, insbesondere in Bezug auf Datenschutz, Kosten, Skalierbarkeit, Anpassbarkeit, Support, Transparenz und Anbieterabhängigkeit. Anschließend werden diese Punkte näher erläutert.

Kriterium	Open-Source LLMs	Kommerzielle LLMs
Datenschutz & Kontrolle	Volle Datenhoheit: Modellbetrieb in eigener Infrastruktur, dadurch verlassen sensible Daten nicht das Unternehmen. Hohe Kontrolle über Zugriffe und Verwendung der Daten.	Datenweitergabe: Verarbeitung extern beim Anbieter (Cloud). Mögliche Datenschutzrisiken, da vertrauliche Informationen das Unternehmen verlassen und man dem Anbieter vertrauen muss. Kein direkter Einblick, wie Daten intern genutzt werden.
Kosten	Keine Lizenzkosten: Modelle frei verfügbar (Open Source). Allerdings Hardware-Investitionen nötig (GPUs/Server) plus laufende Betriebs- und Wartungskosten. Längerfristig bei hoher Nutzung oft kosteneffizienter.	Nutzungsentgelte: Bezahlung nach Verbrauch (z. B. API-Kosten pro Anfrage/Token). Keine eigenen Hardwarekosten, aber laufende Gebühren können bei hohem Volumen beträchtlich sein. Anfangshürde gering, Kosten skalieren linear mit Nutzung.
Skalierbarkeit & Leistung	Eigenverantwortliche Skalierung: Unternehmen muss ausreichende Rechenressourcen bereitstellen. Volle Kontrolle über Performance-Tuning (z. B. Modellkompression, spezialisierte Hardware). Allerdings limitiert durch eigene Infrastruktur; ganz große Modelle (>100B) u.U. kaum on-prem einsetzbar.	Anbietergetragene Skalierung: Provider stellt dynamisch Rechenleistung bereit, um Lastspitzen abzufangen. Zugriff auf state-of-the-art Modelle (höchste Leistung wie GPT-4) sofort möglich. Dafür Abhängigkeit von externen Limits (Rate Limits, Verfügbarkeit) und Netzwerklatenz.
Anpassbarkeit	Maximale Flexibilität: Quellcode und Modellgewichte verfügbar – erlaubt Feintuning auf eigene Daten, Änderungen an der Architektur und Einbau firmenspezifischer Funktionen. Dadurch genaue Anpassung an Geschäftsbedürfnisse möglich.	Begrenzt anpassbar: Modell ist vorgegeben; Anpassung meist nur indirekt über Prompt-Engineering oder vom Anbieter angebotene Einstellungen/Feintuning-Optionen. Kein Zugriff auf innere Modellstruktur, daher eingeschränkte Individualisierung.
Support & Wartung	Community-Support: Unterstützung durch eine offene Entwicklergemeinschaft, Foren, etc. Qualität und Verlässlichkeit des Supports variieren. Interne Experten nötig für Betrieb, Troubleshooting und regelmäßige Updates/Patches des Modells.	Professioneller Support: Anbieter bietet technische Unterstützung, Wartung und Service Level Agreements (SLAs). Kontinuierliche Modellverbesserungen und Sicherheitsupdates erfolgen durch den Anbieter. Geringerer interner Aufwand, dafür Abhängigkeit von dessen Support-Qualität.
Transparenz	Hohe Nachvollziehbarkeit: Offenlegung von Code und Trainingsdaten (teilweise) ermöglicht es, das Modellverhalten zu auditieren. Etwaige Verzerrungen oder Fehler können zumindest erkannt und ggf. durch Retraining behoben werden.	Black-Box-Modell: Interne Funktionsweise und Trainingsdaten sind intransparent. Erklärbarkeit der Ergebnisse reduziert, was z. B. im regulierten Umfeld (Compliance-Prüfung) problematisch sein kann. Vertrauen in den Anbieter ersetzt eigene Kontrolle.
Anbieterabhängigkeit	Keine Lock-in-Effekte: Open-Source-Modelle können frei gewechselt werden; es bestehen keine Vertragsbindungen. Volle Unabhängigkeit von Dritten – langfristig strategischer Vorteil.	Vendor-Lock-in-Risiko: Starke Abhängigkeit vom Anbieter (technologisch und vertraglich). Wechsel des Dienstes oft mit Migrationsaufwand verbunden. Änderungen der Preismodell- oder Nutzungsbedingungen durch den Anbieter können Einfluss auf das eigene Geschäft haben.

Tabelle 5.2.: Vergleich von Open-Source und kommerziellen LLMs
erhoben aus: T. Ahmed et al. (2024), Jones (2024), Malec, Melissa (2025), Network (2024), Uspenskyi, Serhii (2024) und Yu et al. (2023)

Die Wahl zwischen Open-Source-LLMs und kommerziellen LLMs wirkt sich auf verschiedene Dimensionen aus, wobei Datenschutz, Kosten, Skalierbarkeit, Anpassbarkeit, Support, Transparenz und Anbieterabhängigkeit eine zentrale Rolle spielen. Beim Datenschutz ermöglicht der Betrieb von LLMs in unternehmenseigener Infrastruktur eine eigenverantwortliche Umsetzung strenger Richtlinien, da sensible Daten im Unternehmen verbleiben, technische Privacy-Mechanismen können zusätzlich sicherstellen, dass keine Daten unkontrolliert das Unternehmen verlassen (Kelbert et al., 2024; Microsoft, 2024). Im Gegensatz dazu erfolgt bei kommerziellen LLMs die Verarbeitung in der Cloud, was potenzielle Risiken birgt, da vertrauliche Informationen an externe Anbieter übermittelt werden und weniger direkte Kontrolle möglich ist (Microsoft, 2024).

Kostenmäßig bieten Open-Source-Modelle den Vorteil, dass sie lizenzkostenfrei verfügbar sind und so den Einstieg in Experimente und Pilotprojekte erleichtern (Fiza Fatima, 2024). Allerdings müssen Unternehmen hier in eigene Infrastruktur investieren, etwa in leistungsfähige GPUs, Server und den dazugehörigen Betrieb, was insbesondere bei intensiver Nutzung langfristig kosteneffizient sein kann. Im Gegensatz dazu werden kommerzielle LLM-Services meist nach einem Pay-per-Use-Modell abgerechnet, wobei API-Gebühren anfallen; diese Kosten können bei hohem Anfragevolumen beträchtlich werden, auch wenn eigene Hardwareinvestitionen entfallen (Fiza Fatima, 2024).

Hinsichtlich der Skalierbarkeit bieten kommerzielle Anbieter dynamische Rechenressourcen, die Lastspitzen abfangen und eine sofortige Hochleistung gewährleisten, während der Betrieb von Open-Source-LLMs eine eigenverantwortliche Bereitstellung und Optimierung der Infrastruktur erfordert. Moderne Software-Optimierungen wie verteiltes Serving, Quantisierung oder GPU-Batching können die Performance erheblich verbessern, wobei der damit verbundene Engineering-Aufwand jedoch höher ist (Fiza Fatima, 2024; Marshall, 2024).

Die Anpassbarkeit stellt einen weiteren entscheidenden Unterschied dar. Da bei Open-Source-LLMs Quellcode und Modellgewichte offen zugänglich sind, ermöglichen sie ein detailliertes Feintuning auf unternehmensspezifische Daten und Anforderungen, wobei die Anpassungen über standardisierte Parameter hinausgehen (Fiza Fatima, 2024). Im Gegensatz dazu sind kommerzielle Modelle weitgehend vorgegeben, sodass Anpassungen meist nur indirekt über Prompt-Engineering oder vom Anbieter bereitgestellte Optionen erfolgen können (Fiza Fatima, 2024).

Bei Support und Wartung profitieren Unternehmen von kommerziellen LLM-Diensten, da diese professionellen Support, regelmäßige Sicherheitsupdates und Wartung im Rahmen von Service Level Agreements bieten, was den internen Aufwand erheblich reduziert (Fiza Fatima, 2024). Hingegen erfordert der Betrieb von Open-Source-LLMs den Aufbau interner Kompetenzen in Bereichen wie MLOps und DevOps, um den reibungslosen Betrieb, Troubleshooting und regelmäßige Updates sicherzustellen (Fiza Fatima, 2024).

Transparenz ist ein weiterer wesentlicher Aspekt, Open-Source-LLMs zeichnen sich durch eine hohe Nachvollziehbarkeit aus, da der offene Zugang zu Quellcode und teilweise auch Trainingsdaten unabhängige Audits und eine detaillierte Analyse des Modellverhaltens ermöglicht (ITMAGINATION, 2024). Bei Closed-Source-Modellen bleibt die interne Funktionsweise hingegen weitgehend intransparent, was in regulierten Umfeldern zu Compliance-Herausforderungen führen kann (ITMAGINATION, 2024).

Schließlich spielt die Anbieterabhängigkeit eine wichtige Rolle. Open-Source-LLMs bieten den Vorteil, dass sie ohne vertragliche Bindungen genutzt werden können, was langfristig zu strategischer Unabhän-

gigkeit führt (Fiza Fatima, 2024). Im Gegensatz dazu besteht bei kommerziellen LLMs ein erhebliches Risiko des Vendor-Lock-in, da Änderungen in den Preismodellen oder Nutzungsbedingungen direkten Einfluss auf das Unternehmen haben können (Marshall, 2024).

Unternehmen müssen die verschiedenen Dimensionen wie Datenschutz, Kosten, Skalierbarkeit, Anpassbarkeit, Support, Transparenz und Anbieterabhängigkeit sorgfältig abwägen, um die für ihre spezifischen Geschäftsprozesse am besten geeignete LLM-Lösung zu identifizieren (Fiza Fatima, 2024; ITMAGINATION, 2024; Marshall, 2024).

6. Framework

Die Integration von LLMs in Unternehmensprozesse erfordert technisches Verständnis und einen strukturierten Ansatz zur systematischen Bewertung und Umsetzung der Anforderungen. Dieses Kapitel präsentiert als Artefakt der DSR ein umfassendes Framework, das als praxisorientierte Entscheidungsgrundlage für IT-Abteilungen und Führungskräfte dient. Es vereint theoretische Erkenntnisse und praktische Erfahrungen und deckt alle relevanten Bereiche ab, von technischer Infrastruktur über Datenbereitstellung und Deployment-Modelle bis hin zu Sicherheit, Betrieb und unternehmensspezifischen Anpassungen.

Im ersten Abschnitt 6.1 werden die konzeptionellen Grundlagen und methodischen Überlegungen erläutert, die der Entwicklung des Frameworks zugrunde liegen. Hierbei wird dargestellt, wie die unterschiedlichen Komponenten miteinander verknüpft sind, um eine ganzheitliche Sicht auf die Herausforderungen und Chancen bei der Integration von LLMs zu ermöglichen. Der zweite Abschnitt 6.2 visualisiert das entwickelte Modell anhand detaillierter Checklisten, die den aktuellen Status der IT-Infrastruktur und -Prozesse erfassbar machen. So erhalten Unternehmen ein praktisches Instrument zur Bewertung ihrer Voraussetzungen und zur Ableitung gezielter Maßnahmen.

6.1. Dokumentation und Konzeption

Das Framework aus Abschnitt 6.2 gliedert sich in mehrere Bereiche, die wesentlichen technischen und organisatorischen Anforderungen abdecken. Im ersten Bereich, der technischen Infrastruktur, werden die notwendigen Voraussetzungen wie ausreichende Rechenleistung (z. B. GPU/TPU), eine flexible Skalierbarkeit der Infrastruktur, spezialisierte Hardware, schnelle Netzwerkverbindungen sowie skalierbare Speichersysteme (z. B. Data Lakes) betrachtet. Diese Aspekte bilden die Basis, da ohne leistungsfähige Hardware weder ein effektives Training noch eine reibungslose Inferenz möglich ist.

Im zweiten Bereich liegt der Fokus auf der Datenbereitstellung und -qualität. Hier wird die Standardisierung der Datenformate (beispielsweise JSONL oder TFRecord) durch den Einsatz automatisierter ETL-Pipelines (wie Apache Airflow) thematisiert. Darüber hinaus werden Maßnahmen zur Qualitätssicherung, wie etablierte Prozesse zur Duplikat-Erkennung und -Bereinigung, sowie Strategien zur Archivierung und Versionierung von Daten beschrieben. Eine regelmäßige und automatisierte Überprüfung der Datenaktualität ist hierbei essenziell.

Der dritte Bereich behandelt das Deployment-Modell. Unternehmen müssen entscheiden, ob sie LLM-Lösungen On-Premises, in der Cloud oder in hybriden Modellen implementieren. Dabei sind Datenschutzrichtlinien, die Evaluierung der Anbieterabhängigkeit und eine langfristige Kosten-Nutzen-Bewertung zentrale Faktoren, die individuell bewertet werden müssen.

Im vierten Bereich erfolgt die Modellwahl, bei der ein Vergleich zwischen Open-Source- und kommerziellen LLMs vorgenommen wird. Wichtige Kriterien sind dabei die Datensicherheitsanforderungen, die Gesamtkosten (TCO), die Skalierbarkeit und Performance sowie die Möglichkeit zur individuellen Anpassung (Feintuning). Auch der Support und die Wartungsfähigkeit der Modelle spielen eine entscheidende Rolle.

Der fünfte Bereich widmet sich der Integration in bestehende IT-Systeme. Hierbei werden relevante Schnittstellen und APIs, die Anbindung an ERP-, CRM- oder DMS-Systeme sowie die Kompatibilität mit vorhandenen Sicherheitslösungen und etablierten Monitoring- und Logging-Strategien in den Blick genommen.

Der sechste Bereich umfasst den Betrieb und die Wartung der Modelle. Neben dem Ressourcenmanagement und der Skalierung werden Strategien zur Modellversionierung und automatisierte Prozesse zur Modellvalidierung als wichtige Elemente hervorgehoben. Zudem ist die regelmäßige Überprüfung der Betriebskosten, etwa durch den Einsatz von Monitoring-Dashboards und KPIs, von Bedeutung.

Im siebten Bereich wird die Sicherheit und Compliance, insbesondere im Hinblick auf IT-Sicherheit und Datenschutz, behandelt. Hierzu zählen Schutzmechanismen für sensible Daten, die Umsetzung einer effektiven Zugriffskontrolle, die gesetzeskonforme Datenverarbeitung sowie Maßnahmen zur Verhinderung von Datenlecks, wie Datenmaskierung und Verschlüsselung.

Der achte Bereich konzentriert sich auf die Evaluierung und kontinuierliche Verbesserung der LLM-Implementierung. Durch regelmäßige Evaluierungen mittels Monitoring-Dashboards, A/B-Tests und Benchmarking werden Zielwerte definiert und anhand von KPIs (wie Antwortzeit, Genauigkeit und Betriebskosten) die Effektivität des Systems gemessen. Ein iterativer Verbesserungsprozess, der auf regelmäßigem Feedback basiert, ermöglicht es, das Framework stetig an neue Anforderungen und technologische Entwicklungen anzupassen.

Abschließend werden im neunten Bereich unternehmensspezifische Anpassungen vorgenommen. Dabei fließen bestehende IT-Landschaften (z. B. ERP, CRM, DMS), individuelle Datenschutz- und Compliance-Richtlinien sowie die Evaluierung der internen Ressourcen und MLOps-/DevOps-Kapazitäten ein, um sicherzustellen, dass das Framework optimal in die vorhandene Systemumgebung integriert wird.

Die Implementierung des Frameworks erfolgt in mehreren Phasen. In der Analysephase wird zunächst die IST-Situation der IT-Infrastruktur und der Datenbereitstellung umfassend untersucht. Anschließend folgt die Entwicklungsphase, in der auf Basis der gewonnenen Erkenntnisse ein Kriterienkatalog erstellt wird, der Zielwerte und Gewichtungen für die einzelnen Prüfkriterien definiert. Die daraus resultierenden Checklisten dienen als praktische Instrumente zur Bewertung der aktuellen Situation und zur kontinuierlichen Überwachung der Fortschritte. In der Implementierungsphase wird das Framework zunächst in ausgewählten Pilotprojekten erprobt, IT-Abteilungen und Entscheidungsträger werden geschult und notwendige Systemanpassungen vorgenommen. Schließlich wird in der Evaluationsphase die Systemleistung kontinuierlich überwacht und das Framework mithilfe von Feedback und aktuellen Entwicklungen iterativ verbessert.

6.2. Darstellung des Frameworks

1. Technische Anforderungen an die Infrastruktur

Prüfkriterien/Anforderungen	**erfüllt**	**teilweise**	**nicht erfüllt**	**Kommentar**
Rechenleistungsbedarf ermittelt und Investitionsbereitschaft bewertet				Grundvoraussetzung für LLM-Integration – ohne leistungsfähige Hardware ist weder Training noch Inferenz möglich.
Skalierbarkeit der Infrastruktur				Infrastruktur muss flexibel skalierbar sein, um Lastspitzen abzudecken.
Verfügbarkeit spezialisierter Hardware (z. B. GPUs)				Essenziell für effiziente Inferenz; spezifische Priorisierungsmechanismen werden nicht im Detail behandelt.
Interne Netzwerkgeschwindigkeit und Latenzoptimierung				Eine hohe Netzwerkgeschwindigkeit ist entscheidend für schnelle Datenübertragungen – eine Analyse der bestehenden Infrastruktur ist empfehlenswert.
Verfügbarkeit von Hochgeschwindigkeits-Interconnects (InfiniBand, NVLink)				Für verteiltes Training und Inferenz wichtig – Unternehmen sollten prüfen, welche Interconnect-Technologien bereits vorliegen und ob ein Upgrade sinnvoll ist.
Speicherung großer Modell- und Datenmengen (mehrere TB)				Nutzung skalierbarer Speichersysteme, wie Data Lakes, wird empfohlen.

2. Anforderungen an Datenbereitstellung und -qualität

Prüfkriterien/Anforderungen	erfüllt	teilweise	nicht erfüllt	Kommentar
Einheitliche Datenformate (z. B. JSONL, TFRecord)				Daten sollten über automatisierte ETL-Pipelines (z. B. Apache Airflow) standardisiert werden; gängiger Einsatz von Data-Quality-Tools (z. B. Great Expectations).
Maßnahmen zur Datenvorverarbeitung und Qualitätssicherung				Automatisierte ETL-Prozesse sichern Datenvollständigkeit, Konsistenz und Aktualität.
Etablierte Prozesse zur Duplikat-Erkennung und -Bereinigung				Standardisierte Verfahren (z. B. Hashing-basierte Filterung) werden eingesetzt.
Sicherstellung der Aktualität und Vollständigkeit der Daten				Regelmäßige, automatisierte Überprüfungen und Updates sind notwendig.
Archivierungs- und Versionierungsstrategien etabliert				Nutzung versionierter Speichersysteme wird empfohlen; unternehmensspezifische Richtlinien sollten definiert werden.

3. Bereitstellung (Deployment-Modell)

Prüfkriterien/Anforderungen	erfüllt	teilweise	nicht erfüllt	Kommentar
Festlegung auf On-Premises, Cloud oder Hybridmodell				Hybridmodelle werden häufig gewählt, wenn lokale Verarbeitung aus Datenschutzgründen erforderlich ist und Cloud-Ressourcen für Lastspitzen genutzt werden.
Datenschutzrichtlinien klar definiert und erfüllt				Praxisübliche Maßnahmen: vertragliche Klauseln, regelmäßige Backups, Exportmechanismen als Exit-Strategie.
Evaluierung der Anbieterabhängigkeit				Wird als dynamischer, unternehmensspezifischer Prozess betrachtet – eine aktuelle Erhebung ist notwendig.
Kosten-Nutzen-Bewertung langfristig durchgeführt				Kostenvarianten müssen aktuell erhoben werden; diese Bewertung erfolgt individuell.

4. Modellwahl (Open Source vs. kommerziell)

Prüfkriterien/Anforderungen	erfüllt	teilweise	nicht erfüllt	Kommentar
Entscheidung auf Basis der Datensicherheitsanforderungen				Bewertung erfolgt dynamisch – spezifische Datenschutz- und Compliance-Richtlinien müssen geprüft werden.
Vergleich der Gesamtkosten (TCO)				Da sich Kosten kontinuierlich ändern, sollte eine aktuelle Erhebung erfolgen.
Erfüllung der Anforderungen an Skalierbarkeit und Performance				Soll anhand aktueller Benchmarks und unternehmensspezifischer Anforderungen geprüft werden.
Möglichkeit zur Anpassung/Feintuning				Open-Source-Modelle bieten hier höhere Flexibilität, während kommerzielle Modelle oft standardisierte Optionen bieten.
Support und Wartungsfähigkeit				Wird individuell bewertet – professioneller Support versus interner Expertise.

5. Integration in bestehende IT-Systeme

Prüfkriterien/Anforderungen	erfüllt	teilweise	nicht erfüllt	Kommentar
Vorhandensein relevanter Schnittstellen/APIs				Umsetzung erfolgt unternehmensspezifisch; IT-Landschaft (ERP, CRM, DMS) sowie vorhandene API-Gateways sollten berücksichtigt werden.
Anbindung an bestehende ERP-, CRM- oder DMS-Systeme				Integration wird individuell erarbeitet.
Kompatibilität mit bestehenden Sicherheitslösungen				Anpassungen erfolgen je nach vorhandener IT-Sicherheitsarchitektur.
Monitoring- und Logging-Strategien etabliert				Erforderlich, wird in Abstimmung mit IT und Security umgesetzt.

6. Betrieb und Wartung der Modelle

Prüfkriterien/Anforderungen	erfüllt	teilweise	nicht erfüllt	Kommentar
Ressourcenmanagement und Skalierungsstrategien vorhanden				Prozesse sind stark unternehmensspezifisch – universelle Vorgaben können hier nicht definiert werden.
Strategien zur Modellversionierung etabliert				Wird individuell erarbeitet, da Modelle sich schnell ändern.
Automatisierte Prozesse zur Modellvalidierung				Umsetzung erfolgt je nach IT-Umgebung und internen Anforderungen.
Betriebskosten regelmäßig überprüft und optimiert				Gängige Methode: Einsatz von Monitoring-Dashboards und KPIs (z. B. Antwortzeit, Genauigkeit, Kosten pro Anfrage).

7. Sicherheit und Compliance (IT-Sicherheit und Datenschutz)

Prüfkriterien/Anforderungen	erfüllt	teilweise	nicht erfüllt	Kommentar
Schutzmechanismen für sensible Daten implementiert				Maßnahmen hängen von internen Policies ab – eigene Datenschutzvorgaben (z. B. DSGVO) sind zu berücksichtigen.
Zugriffskontrolle und Rechteverwaltung etabliert				Umsetzung erfolgt über unternehmensspezifische Identity-Management-Systeme.
Datenverarbeitung gesetzeskonform				Muss an geltende gesetzliche Vorgaben (z. B. DSGVO) angepasst werden.
Maßnahmen zur Verhinderung von Datenlecks umgesetzt				Praxisübliche Maßnahmen: Datenmaskierung, Verschlüsselung, regelmäßige Sicherheitsüberprüfungen.

8. Evaluierung und kontinuierliche Verbesserung

Prüfkriterien/Anforderungen	erfüllt	teilweise	nicht erfüllt	Kommentar
Regelmäßige Evaluierung der Modellergebnisse				Gängige Methode: Einsatz von Monitoring-Dashboards, A/B-Tests, Benchmarking (z. B. Antwortzeit, Genauigkeit, Kosten pro Anfrage).
Iterativer Verbesserungsprozess auf Basis von Feedback etabliert				Regelmäßige Reviews und Anpassungen werden empfohlen.
Messung der Effektivität und Effizienz der LLM-Implementierung				Verwendung von KPIs (Antwortzeit, Genauigkeit, Betriebskosten) als Bewertungsmaßstab.
Umsetzung von Verbesserungsmaßnahmen dokumentiert				Interne Prozesse zur kontinuierlichen Optimierung sollten vorhanden sein.

9. Unternehmensspezifische Anpassungen

Prüfkriterien/Anforderungen	erfüllt	teilweise	nicht erfüllt	Kommentar
Bestehende IT-Landschaft (ERP, CRM, DMS) analysiert und dokumentiert				Unternehmen sollten ihre vorhandenen Schnittstellen und Systeme prüfen und diese in das Framework einbinden.
Eigene Datenschutz- und Compliance-Richtlinien definiert und implementiert				Unternehmensspezifische Vorgaben (z. B. DSGVO) müssen geprüft und integriert werden.
Vorhandene Sicherheitslösungen (Zugriffsmanagement, Audit-Trails) überprüft				Sicherstellen, dass bestehende Systeme mit dem LLM-Deployment kompatibel sind.
Interne Ressourcen und MLOps-/DevOps-Kapazitäten evaluiert				Bewertung, ob eigene Ressourcen für Support und Wartung vorhanden sind oder extern eingebunden werden müssen.

7. Expert*inneninterviews

Im Rahmen des DSR bildet die Evaluationsphase einen wesentlichen Baustein der vorliegenden Arbeit. Dieses Kapitel widmet sich der empirischen Validierung des in Kapitel Kapitel 6 entwickelten Frameworks. Durch leitfadengestützte Expert*inneninterviews werden nicht nur die in der Theorie postulierten Anforderungen an die technische Integration von LLMs in Unternehmensprozesse überprüft, sondern auch neue Perspektiven und praxisrelevante Erkenntnisse gewonnen.

Die Interviews dienen primär dazu, die Praxistauglichkeit des Frameworks zu belegen und dessen kontinuierliche Weiterentwicklung voranzutreiben. Sie ermöglichen es, Einblicke in den tatsächlichen Umgang mit LLM-Anwendungen in unterschiedlichen Unternehmenskontexten zu erhalten und etwaige Diskrepanzen zwischen theoretischen Annahmen und praktischen Herausforderungen zu identifizieren. Darüber hinaus fließen die gewonnenen Ergebnisse in einen iterativen Verbesserungsprozess ein, der dazu beiträgt, das Framework noch besser an die realen Anforderungen und Rahmenbedingungen in der Wirtschaft anzupassen.

Im Folgenden werden zunächst in Abschnitt 7.1, die Auswahl der Expert*innen sowie die Begründung dieser Entscheidung dargestellt. Anschließend folgt in Abschnitt 7.2 eine detaillierte Beschreibung des Interviewleitfadens und der durchgeführten Interviews. Abschließend wird in Abschnitt 7.3 die Auswertung der Interviews mittels qualitativer Inhaltsanalyse präsentiert, die als Grundlage für die abschließende Validierung und Optimierung des Frameworks dient.

7.1. Auswahl der Expert*innen und Begründung

Im Rahmen der Validierung des entwickelten Frameworks zur Integration von LLMs in Unternehmensprozesse wurden Expert*inneninterviews durchgeführt. Dabei kamen Fach- und Führungskräfte aus unterschiedlichen Bereichen, wie IT-Infrastruktur, Softwareentwicklung, Cloud- und Automationslösungen sowie Projektmanagement im KI-Umfeld, zu Wort. Die Auswahl der Interviewpartner*innen spiegelt die Diversität der Unternehmenslandschaft wider, wobei die befragten Organisationen von kleinen Unternehmen mit rund 15 Mitarbeiter*innen bis hin zu Großkonzernen mit über 7500 Beschäftigten reichen. Diese Bandbreite unterstreicht, dass die Relevanz von LLMs sowohl für kleine als auch für große Organisationen gegeben ist und diverse Anwendungsszenarien sowie individuelle Herausforderungen bedient.

Um den vielfältigen Perspektiven gerecht zu werden, wurden die Expert*innen anonymisiert, etwa durch die Änderung von Namen und die Verwendung des Personalstands anstelle der Firmennamen, da einige Interviewpartnerinnen den Wunsch äußerten, ihre Identität und die Zugehörigkeit zu bestimmten Unternehmen nicht preiszugeben. Die praktischen Erfahrungen reichen dabei von der experimentellen Nutzung von ChatGPT und spezifischen LLM-Lösungen über den Einsatz von KI-gestützten Tools zur Unterstützung der Softwareentwicklung bis hin zur Implementierung von LLMs in komplexe IT-Infrastrukturen und deren Integration in bestehende Geschäftsprozesse.

Eine tabellarische Übersicht der Expert*innen, ihrer Tätigkeitsbereiche, bisherigen Erfahrungen mit LLMs und weiterer relevanter Informationen ist in Tabelle 7.1 dargestellt.

Tabelle 7.1.: Übersicht der Expert*innenprofile

Name	Unternehmen	Bereich	Erfahrung mit LLM	Weitere Relevante Infos
Anton (anonymisiert)	Personalstand ca. 3000	IT-Infrastruktur (Administration, Containerisierung, Softwareentwicklungsprozess, Ressourcen- und Data Center Hardwareplanung)	Selbstnutzer von ChatGPT; experimenteller Einsatz des chinesischen LLM „DeepSeek“ (lokale Ressourcen- und Performancetests)	Praxisnahe Erfahrung in der Planung und Verwaltung komplexer IT-Infrastrukturen; Verständnis für Herausforderungen bei der Integration moderner KI-Technologien in unternehmenskritische Umgebungen
Eric (anonymisiert)	Personalstand ca. 7500	IT-Teamleiter für Cloud & Automation sowie Solution Architekt (Fokus auf Requirements Engineering)	Erweiterte Erfahrung in der Nutzung und Implementierung von LLMs sowie Begleitung der Einführung weiterer KI-Lösungen	Setzt sich kontinuierlich mit neuen Technologien auseinander; beteiligt an einem KI-Gremium im Konzern zur strategischen Integration von KI-Technologien
Mia (anonymisiert)	Personalstand ca. 3200	IT-Management, insbesondere als Projektleitung im KI-Zirkel	Einführung von BingChat for Enterprise/CoPilot in Edge; Erfahrungen bisher durchwachsen (unzureichende Schulungen)	Starke Fokussierung auf Datenschutz (Präferenz für On-Premises-Lösungen bei LLMs); Workshops und Use Case-Erhebungen zur Identifikation von KI-Potenzialen in den Fachbereichen
Marcus (anonymisiert)	Personalstand ca. 15	Senior Software Architect – verantwortlich für Software-Implementierung sowie Auswahl und Beschaffung der IT-Infrastruktur (Netzwerk, Server, NAS, Datensicherung, etc.)	Nutzung kommerzieller LLMs (OpenAI) zur Erstellung von Code-Snippets, Textumformulierung und Textbausteinen; keine In-House-Lösung aufgrund fehlender dedizierter Hardware/Software-Infrastruktur	Stets auf der Suche nach neuen Technologien zur Optimierung der Firmenprozesse; legt großen Wert auf Datensicherheit und die Vermeidung von Vendor Lock-In; der Einsatz von LLMs erfolgt aktuell primär im Bereich Textunterstützung
Ralf (anonymisiert)	Personalstand ca. 450	Software Development & Operations – IT-Infrastruktur und Integration von KI-Technologien in Unternehmensprozesse	Umfassende praktische Erfahrungen mit LLMs und KI-gestützten Tools: Einsatz von ChatGPT für kreative Prozesse, Konzeptentwicklung und Planung; Nutzung von Coding-Assistenten (z. B. JetBrains AI Assistant, GPT-4o-mini via OpenAI-API); Aktiv in Evaluierungsprogrammen (z. B. Vergleich von Google Gemini und ChatGPT)	Arbeitet in einem Umfeld mit hybriden Cloud-Lösungen (Private Cloud bei Hetzner, Public Cloud-Ressourcen von Google Cloud); Integration von LLM-Anwendungen in operative Systeme via RESTful-APIs und OpenAPI; Thematisiert Shadow-IT und betont den Mehrwert zentraler KI-Initiativen sowie die Einhaltung von Datenschutz (DSGVO, vertragliche Regelungen)
Theo (anonymisiert)	Personalstand ca. 4500	Software Engineering / Software Design – 11 Jahre Branchenerfahrung	Privat und in Forschung: Entwicklung von Wrappern für Language Modelle; Im Unternehmen: Einsatz von Microsoft Copilot (vor allem im Coding); Beteiligung als Testnutzer und in Feedbackgruppen zur Evaluierung von KI-Lösungen	Arbeitet im Automotive-Sektor mit Fokus auf Softwarelösungen; Nutzt Azure-Cloud-Ressourcen sowie On-Premise Hardware; Erfahrung mit klassischen Datenformaten (PDF, Excel, Word, JSON, CSV) und DMS (MS Sharepoint); Organisatorische Hürden (z. B. langsame Ressourcenfreigaben) als kritischer Faktor; Interesse an zukünftigen Entwicklungen wie RAG-Systemen

7.2. Interviewleitfaden und Durchführung

Die Expert*inneninterviews wurden in unterschiedlichen Formaten durchgeführt, um den variierenden zeitlichen und organisatorischen Anforderungen der Interviewpartner*innen gerecht zu werden. Die meisten Gespräche fanden online statt und dauerten etwa 35 bis 50 Minuten. Alle Interviews wurden aufgezeichnet und systematisch ausgewertet.

Bereits in einem Vorgespräch wurden die Expert*innen in das Forschungsprojekt eingeführt. Dabei wurden die Rahmenbedingungen, etwa Ablauf, Dauer und der vertrauliche Umgang mit den erhobenen Daten, erläutert. Anschließend kam der semi-strukturierte Interviewleitfaden zum Einsatz, der sich an den zentralen Komponenten des bestehenden Frameworks aus Abschnitt 6.2 orientierte. Der Leitfaden, wie in Tabelle 7.2 dargestellt, ermöglichte es, den beruflichen Hintergrund, die jeweilige Rolle sowie die bisherigen Erfahrungen der Teilnehmerinnen mit LLMs und anderen KI-Anwendungen umfassend zu erfassen.

Im ersten inhaltlichen Block wurden Fragen zu grundlegenden technischen Infrastrukturen gestellt, so etwa zu vorhandener Hardware, Netzwerkleistung und Speichersystemen. Darüber hinaus lag ein Schwerpunkt auf der Datenqualität, der Standardisierung, Vorverarbeitung und Archivierung von Daten sowie auf den bevorzugten Deployment-Modellen (On-Premises, Cloud, Hybrid). Ein besonderes Augenmerk galt den Kriterien, die die Modellwahl beeinflussen, und der Integration von LLM-Lösungen in bestehende Systeme wie ERP, CRM oder DMS, einschließlich der relevanten Schnittstellen und Sicherheitskonzepte.

Im zweiten inhaltlichen Block stand der Betrieb und die Wartung der LLM-Anwendungen im Mittelpunkt. Hierzu wurden Fragen zum Ressourcenmanagement, zu Strategien der Modellversionierung und des Rollbacks sowie zu Maßnahmen zur kontinuierlichen Optimierung, beispielsweise durch Monitoring, A/B-Tests und Benchmarking, gestellt. Ergänzend wurden die Einhaltung von Sicherheits- und Compliance-Anforderungen (wie DSGVO, Datenmaskierung und Zugriffskontrolle) sowie unternehmensspezifische Anpassungen, etwa die Analyse der bestehenden IT-Landschaft und die Integration interner Sicherheitslösungen, thematisiert.

Aufgrund von Zeitbeschränkungen wurden einige Interviews auch schriftlich per E-Mail durchgeführt. Dank des flexiblen, semi-strukturierten Leitfadens konnten in diesen Fällen Rückfragen problemlos per E-Mail geklärt werden, sodass auch dieses Format alle relevanten Informationen vollständig erfasste.

Den Abschluss bildeten offene Fragen, die nun den zusätzlichen Bereich „Offene Fragen und Ausblick“ umfassen. In diesem Abschnitt hatten die Interviewpartner*innen die Möglichkeit, weitere kritische Aspekte, Empfehlungen sowie zukünftige Entwicklungen im Bereich der LLM-Integration zu äußern.

Tabelle 7.2.: Darstellung des Interviewleitfadens

Bereich	Fragen
1. Allgemeine Fragen	
Hintergrund und Rolle	▪ „Kannst du kurz deinen beruflichen Hintergrund und deine Rolle im Unternehmen beschreiben?“ ▪ „Wie bist du in den Entscheidungsprozess bzw. die Implementierung von KI-Technologien eingebunden?“
Erfahrungen mit LLMs und KI	▪ „Welche Erfahrungen hast du persönlich mit der Integration oder Nutzung von LLMs bzw. anderen KI-Lösungen gemacht?“ ▪ „Gibt es bei euch im Unternehmen bereits LLM-Lösungen, die ihr verwendet?“ ▪ „Welche Veränderungen oder Trends im Bereich KI beobachtest du in deinem Unternehmen?“
2. Technische Anforderungen und Infrastruktur	
Technische Infrastruktur	▪ „Welche Hardware-Ressourcen (z. B. GPUs, TPUs) stehen in deinem Unternehmen für KI-Anwendungen zur Verfügung?“ ▪ „Wie schätzt du die Skalierbarkeit eurer aktuellen Infrastruktur ein, gerade im Hinblick auf mögliche Lastspitzen?“ ▪ „Wie beurteilst du die Netzwerkgeschwindigkeit und Latenz in eurem Unternehmen? Gibt es schon Optimierungen, z. B. durch InfiniBand oder NVLink?“ ▪ „Welche Speichersysteme (z. B. Data Lakes) nutzt ihr für große Datenmengen und wie skaliert ihr diese?“
Datenbereitstellung und -qualität	▪ „Welche Datenformate (z. B. JSONL, TFRecord) und Datenstrukturen verwendet ihr aktuell?“ ▪ „Welche automatisierten Prozesse (z. B. ETL-Pipelines mit Apache Airflow) kommen zum Einsatz, um die Datenqualität sicherzustellen?“ ▪ „Wie geht ihr mit der Duplikaterkennung und -bereinigung in großen Datenbeständen um?“ ▪ „Welche Prozesse sorgen dafür, dass die Daten immer aktuell und vollständig sind?“ ▪ „Wie organisiert ihr die Archivierung und Versionierung der Daten?“

Bereich	Fragen
Deployment-Modell	▪ „Welche Deployment-Modelle (On-Premises, Cloud, Hybrid) nutzt ihr aktuell oder plant ihr für den Einsatz von LLMs, und warum?“ ▪ „Wie definiert und implementiert ihr eure Datenschutzrichtlinien im Zusammenhang mit der LLM-Bereitstellung?“ ▪ „Welche Kriterien fließen in eure Evaluierung der Anbieterabhängigkeit ein und wie überprüft ihr diese regelmäßig?“ ▪ „Wie erfolgt bei euch die langfristige Kosten-Nutzen-Bewertung des gewählten Deployment-Modells?“
3. Modellwahl und Integration	
Modellwahl (Open Source vs. kommerziell)	▪ „Welche Kriterien (z. B. Datensicherheit, Anpassungsfähigkeit, Support) sind für dich ausschlaggebend, wenn es um die Wahl eines LLMs geht?“ ▪ „Welche Erfahrungen hast du mit Open-Source-Modellen im Vergleich zu kommerziellen LLMs gemacht?“ ▪ „Wie bewertest du die Flexibilität und den Support der von euch genutzten Modelle?“
Integration in bestehende IT-Systeme	▪ „Wie bindet ihr LLM-Anwendungen in eure bestehenden Systeme (ERP, CRM, DMS) ein?“ ▪ „Welche Schnittstellen und API-Gateways nutzt ihr, um die Kommunikation zwischen den Systemen sicherzustellen?“ ▪ „Wie stellt ihr sicher, dass die LLM-Integration mit euren bestehenden Sicherheitslösungen kompatibel ist?“ ▪ „Welche Monitoring- und Logging-Strategien habt ihr etabliert, um den Betrieb zu überwachen?“
4. Betrieb, Wartung, Sicherheit und Evaluierung	
Betrieb und Wartung der Modelle	▪ „Wie organisiert ihr das Ressourcenmanagement und die Skalierung eurer LLM-Anwendungen im laufenden Betrieb?“ ▪ „Welche Strategien zur Modellversionierung und zum Rollback habt ihr in eurem Unternehmen implementiert?“ ▪ „Welche automatisierten Prozesse zur Modellvalidierung und Qualitätskontrolle sind bei euch im Einsatz?“ ▪ „Wie überprüft und optimiert ihr regelmäßig die Betriebskosten, z. B. mithilfe von Monitoring-Dashboards und KPIs?“

Bereich	Fragen
Sicherheit und Compliance	▪ „Welche Maßnahmen setzt ihr ein, um die Datensicherheit und den Datenschutz bei der Integration von LLMs zu gewährleisten?“ ▪ „Wie regelt ihr den Zugriff auf sensible Daten, und welche Identity-Management-Systeme nutzt ihr dabei?“ ▪ „Wie stellt ihr sicher, dass eure Datenverarbeitung den gesetzlichen Vorgaben (z. B. DSGVO) entspricht?“ ▪ „Welche Maßnahmen (z. B. Datenmaskierung, Verschlüsselung, regelmäßige Sicherheitsüberprüfungen) habt ihr implementiert, um Datenlecks zu vermeiden?“
Evaluierung und kontinuierliche Verbesserung	▪ „Welche KPIs und Metriken (z. B. Antwortzeit, Genauigkeit, Kosten pro Anfrage) nutzt ihr, um die Effektivität eurer LLM-Implementierung zu messen?“ ▪ „Wie gestaltet ihr den iterativen Verbesserungsprozess, etwa durch regelmäßige Reviews, A/B-Tests oder Benchmarking?“ ▪ „Wie dokumentiert und kommuniziert ihr die Umsetzung von Verbesserungsmaßnahmen innerhalb eures Unternehmens?“
5. Unternehmensspezifische Anpassungen	
	▪ „Wie habt ihr eure bestehende IT-Landschaft (ERP, CRM, DMS) analysiert und dokumentiert, um die Integration von LLMs zu erleichtern?“ ▪ „Welche unternehmensspezifischen Datenschutz- und Compliance-Richtlinien wurden definiert und in den Integrationsprozess einbezogen?“ ▪ „Wie wurden eure bestehenden Sicherheitslösungen (z. B. Zugriffsmanagement, Audit-Trails) überprüft und an die Anforderungen der LLM-Integration angepasst?“ ▪ „Wie schätzt du eure internen MLOps-/DevOps-Kapazitäten im Hinblick auf Support und Wartung von LLM-Anwendungen ein?“
6. Offene Fragen und Ausblick	
Zusätzliche Aspekte	▪ „Gibt es weitere technische oder organisatorische Aspekte, die du als kritisch für die erfolgreiche Integration von LLMs siehst?“ ▪ „Welche zukünftigen Entwicklungen oder Trends im Bereich LLMs findest du besonders relevant – sowohl technologisch als auch in Bezug auf die Datenbereitstellung?“
Empfehlungen und Best Practices	▪ „Welche Empfehlungen würdest du anderen Unternehmen geben, die den Einsatz von LLMs planen?“ ▪ „Gibt es Best Practices aus deinem Unternehmen, die du als besonders effektiv empfindest?“

7.3. Auswertung der Interviews

Die Auswertung der geführten Expert*inneninterviews erfolgte mittels qualitativer Inhaltsanalyse angelehnt an Mayring und Fenzl (2019), wie bereits in Abschnitt 3.3 beschrieben. Auch die im Interviewleitfaden Abschnitt 7.2 definierten Themenblöcke wurden auf diese Weise analysiert. Aus den groben Themenbereichen und ihren Fragen wurden feinere Kategorien abgeleitet, welche durch präzise Definitionen und repräsentative Aussagen aus den Interviews, sogenannte Ankerbeispiele, ergänzt wurden (Mayring & Fenzl, 2019).

7.3.1. Methodisches Vorgehen

Zunächst wurden auf Basis des in Abschnitt 6.2 entwickelten Frameworks sowie der Literatur deduktive Kategorien vorab definiert und in Tabelle 7.3 dargestellt, welche die erwarteten Themenbereiche (z. B. Infrastruktur, Datenqualität, Deployment) repräsentieren sollten. Gleichzeitig blieb Raum für induktive Kategorien, um neu auftretende Aspekte aus den Interviews heraus zu erfassen, die im Vorfeld nicht betrachtet wurden und in Tabelle 7.5 dargestellt.

Im Kodierprozess wurden die Transkripte der Interviews mehrfach durchgearbeitet. Relevante Textstellen wurden zunächst offenen Codes zugeordnet und anschließend den Oberkategorien aus dem Kodierleitfaden zugewiesen (Mayring & Fenzl, 2019). Zur Sicherstellung der Reliabilität der Auswertung wurden die kodierten Textstellen in mehreren Durchläufen geprüft und gegebenenfalls die Kategoriendefinitionen geschärft.

Schließlich resultierte der Kodierprozess in einem Kategoriensystem, das alle bedeutsamen Themen aus den Interviews abbildet. Tabelle 7.3 bietet einen Überblick über dieses Kategoriensystem, gegliedert in Hauptkategorien und feinere Unterkategorien, jeweils orientiert an den im Interviewleitfaden behandelten Themenbereichen.

Tabelle 7.3.: Darstellung deduktive Kategorien

1. Allgemeine Informationen und Rolle
1.1. Hintergrund und berufliche Rolle
1.2. Einbindung in Entscheidungsprozesse und KI-Implementierung

2. Erfahrungen mit LLMs und KI
2.1. Persönliche Erfahrungen und Nutzung von LLMs/KI-Lösungen
2.2. Aktueller Einsatz und Beobachtungen zu Trends in Unternehmen

3. Technische Infrastruktur
3.1. Hardware und Skalierbarkeit
  3.1.1. Verfügbare Ressourcen (z. B. GPUs, TPUs)
  3.1.2. Skalierbarkeit der Infrastruktur und Lastspitzenmanagement
  3.1.3. Netzwerkleistung, Latenz und Interconnect-Technologien (InfiniBand, NVLink)
3.2. Speichersysteme und Datenmanagement
  3.2.1. Genutzte Speichersysteme (z. B. Data Lakes)
  3.2.2. Archivierung, Versionierung und Skalierung großer Datenmengen

4. Datenbereitstellung und -qualität
4.1. Verwendete Datenformate und -strukturen (z. B. JSONL, TFRecord)
4.2. Automatisierte Prozesse zur Datenqualitätssicherung (z. B. ETL-Pipelines)
4.3. Maßnahmen zur Duplikat-Erkennung und Aktualitätskontrolle

5. Deployment und Datenschutz
5.1. Deployment-Modelle
  5.1.1. Entscheidungskriterien: On-Premises, Cloud oder Hybrid
  5.1.2. Evaluierung der Anbieterabhängigkeit und Kosten-Nutzen-Bewertung
5.2. Datenschutz und Compliance
  5.2.1. Definition und Implementierung von Datenschutzrichtlinien
  5.2.2. Maßnahmen zur Sicherstellung gesetzlicher Vorgaben (z. B. DSGVO)

6. Modellwahl und Integration
6.1. Auswahlkriterien für LLMs
  6.1.1. Kriterien wie Datensicherheit, Anpassungsfähigkeit und Support
  6.1.2. Vergleich: Open-Source vs. kommerzielle Modelle
6.2. Integration in bestehende Systeme
  6.2.1. Anbindung an ERP-, CRM- oder DMS-Systeme
  6.2.2. Nutzung von Schnittstellen und API-Gateways
  6.2.3. Kompatibilität mit bestehenden Sicherheitslösungen und Monitoring

7. Betrieb, Wartung und Evaluierung
7.1. Betrieb und Ressourcenmanagement
  7.1.1. Skalierung, Modellversionierung und Rollback-Strategien
  7.1.2. Überwachung der Betriebskosten (z. B. durch KPIs und Dashboards)
7.2. Evaluierung und kontinuierliche Verbesserung
  7.2.1. Nutzung von KPIs (Antwortzeit, Genauigkeit, Kosten)
  7.2.2. Iterativer Verbesserungsprozess (Reviews, A/B-Tests, Benchmarking)

8. Sicherheit und Compliance
8.1. Maßnahmen zur Datensicherheit (z. B. Datenmaskierung, Verschlüsselung)
8.2. Zugriffskontrolle und Identity Management
8.3. Sicherstellung der Einhaltung gesetzlicher Vorgaben

9. Unternehmensspezifische Anpassungen
9.1. Analyse und Dokumentation der bestehenden IT-Landschaft (ERP, CRM, DMS)
9.2. Integration unternehmensspezifischer Datenschutz- und Compliance-Richtlinien
9.3. Bewertung der internen MLOps-/DevOps-Kapazitäten für Support und Wartung

10. Offene Fragen und Zukunftsperspektiven
10.1. Identifikation weiterer kritischer technischer oder organisatorischer Aspekte
10.2. Empfehlungen und Best Practices für den LLM-Einsatz
10.3. Einschätzung zukünftiger Entwicklungen und Trends in Bezug auf LLMs

7.3.2. Interpretation der Kategorien

Im Rahmen der qualitativen Inhaltsanalyse wurden alle Kategorien präzise definiert und jeweils durch exemplarische Zitate, sogenannten Ankerbeispiele verknüpft, um die Zuordnung der Aussagen nachvollziehbar zu gestalten(Mayring & Fenzl, 2019). Hierzu wurden die Kernaussagen der befragten Expert*innen für jede Kategorie zusammengefasst und den Forschungsfragen bzw. den Zielsetzungen dieser Arbeit zugeordnet. Das abschließend entwickelte Kategoriensystem, das im methodischen Vorgehen Unterabschnitt 7.3.1 zur Auswertung der Interviews herangezogen wurde, wird im Folgenden präsentiert. Zur Veranschaulichung werden für die jeweiligen Kategorien konkrete Aussagen aus den Interviewprotokollen als Musterbeispiele ausgewählt, die ein tieferes Verständnis ermöglichen. Gleichzeitig wurden spezifische Kodierregeln definiert, um etwaigen Problemen bei der Abgrenzung zwischen den Kategorien entgegenzuwirken und eine eindeutige Zuordnung sicherzustellen (Mayring & Fenzl, 2019).

In Tabelle 7.4 sind die identifizierten Kategorien mit ihren Definitionen sowie passenden Ankerzitaten der Interviewpartner*innen dargestellt. Anschließend zeigt Tabelle 7.5 die zusätzlich induktiv ermittelten Kategorien, die in der Theorievorbetrachtung nicht explizit erwartet wurden.

Tabelle 7.4.: Übersicht der Ankerzitate

Kategorienummer	Definition der Kategorie	Ankerbeispiel (Zitat)	Name
1.1	Hintergrund und berufliche Rolle	"Als ehemaliger Lehrling und in weiterer Folge fertig studierter Dipl.-Ing. in Wirtschaftsinformatik habe ich mein Berufsleben immer mit neuen Technologien beschäftigt. Als IT-Teamleiter für Cloud und Automation liegen nicht nur die Cloud Services in meinem Team, sondern auch die Implementierung von Automatisierungen, wo in diesem Fall die KI eine große Rolle spielt."	Eric
1.2	Einbindung in Entscheidungsprozesse und KI-Implementierung	"Ich bin unter anderem die Projektleitung des KI-Zirkels, um Use-Cases und Potenziale für den Einsatz von KI festzustellen und für die Analyse mit den Fachbereichen zuständig und begleite diese."	Mia
2.1	Persönliche Erfahrungen und Nutzung von LLMs/KI-Lösungen	"Ich habe sehr gute Erfahrungen mit LLMs und KI-gestützten Tools gemacht. Besonders für kreative Prozesse, Konzeptentwicklung und Planung finde ich ChatGPT sowie vergleichbare Chatbots hilfreich. Auch KI-gestützte Coding-Assistenten wie GitHub Copilot halte ich für äußerst produktivitätssteigernd bei Coding-Aufgaben."	Ralf
2.2	Aktueller Einsatz und Beobachtungen zu Trends in Unternehmen	"Es ist ein sehr gegenwärtiges Thema, da die Nachfrage der Fachbereiche in Bezug auf KI steigt. Besonders CoPilot für M365 ist ein sehr aufkommendes Thema, sowie Predictive Maintenance."	Mia

Kategorienummer	Definition der Kategorie	Ankerbeispiel (Zitat)	Name
3.1.1	Verfügbare Ressourcen (z. B. GPUs, TPUs)	"Wir haben dedizierte Grafikkartencluster mit Nvidia T4, basierend auf Kubernetes, wo containerbasierte Services (wie Ollama-Instanzen) betrieben werden können. Beim aktuellen On-Premise-Setting verwenden wir Nvidia T4 GPUs."	Anton
3.1.2	Skalierbarkeit der Infrastruktur und Lastspitzenmanagement	"Aufgrund der Auslagerung der Berechnungen in die Microsoft Azure Cloud schätze ich die Skalierbarkeit als sehr gut ein."	Eric
3.1.3	Netzwerkleistung, Latenz und Interconnect-Technologien (InfiniBand, NVLink)	"Es gibt keine Probleme mit der Netzwerkgeschwindigkeit oder Latenzen. Auch die Kommunikation mit Cloud-Services funktioniert zuverlässig und performant. Optimierungen sind unsererseits nicht erforderlich."	Ralf
3.2.1	Genutzte Speichersysteme (z. B. Data Lakes)	"Geschäftsdaten aus dem operativen Betrieb werden in SQL-Datenbanken gespeichert. Für die Analyse großer Datenmengen nutzen wir Google BigQuery in Kombination mit Google Cloud Storage, wodurch eine hohe Skalierbarkeit sichergestellt ist."	Ralf
3.2.2	Archivierung, Versionierung und Skalierung großer Datenmengen	"Regelmäßige automatische Backups in der Cloud sind vorhanden. Eine spezielle Archivierung oder Versionierung von Daten über die Backups hinaus erfolgt nicht."	Ralf
4.1	Verwendete Datenformate und -strukturen (z. B. JSONL, TFRecord)	"Hauptsächlich: PDF, Excel, Word, was die meisten Leute verwenden. Im Hintergrund würde ich JSON und CSV sagen."	Theo
4.2	Automatisierte Prozesse zur Datenqualitätssicherung (z. B. ETL-Pipelines)	"Es existieren keine expliziten Prozesse zur Kontrolle der Datenqualität."	Ralf
4.3	Maßnahmen zur Duplikat-Erkennung und Aktualitätskontrolle	"Die Daten, die in unserem DMS vorhanden sind, haben natürlich eine Duplikaterkennung und werden über das Dokumentenmanagementsystem gelöst."	Theo

Kategorienummer	Definition der Kategorie	Ankerbeispiel (Zitat)	Name
5.1.1	Entscheidungskriterien: On-Premises, Cloud oder Hybrid	"Wir nutzen definitiv alles – On-Premise, Cloud und Hybridmodelle. On-Premise wird genutzt, wenn große Datenmengen im eigenen Data Center liegen und der Aufwand, mehrere Terabyte in die Cloud zu schieben, zu hoch wäre. Cloud nutzen wir für kleine Datenmengen, sofern die Datenschutzgrundverordnung dies zulässt. Strictly Confidential-Daten sollten das eigene Data Center nicht verlassen. Hybridmodelle kommen zum Einsatz, wenn beispielsweise die Daten On-Premise liegen, aber die GPUs in der Cloud vorhanden sind."	Anton
5.1.2	Evaluierung der Anbieterabhängigkeit und Kosten-Nutzen-Bewertung	"Bei neuen Lösungen werden alle möglichen Optionen geprüft. Bestehende Lösungen werden regelmäßig auf Basis von Kosten und Qualität reevaluiert."	Ralf
5.2	Datenschutz und Compliance	"Die Datensicherheit hat in unserem Arbeitsumfeld höchste Priorität; somit ist eine Cloud-Lösung für uns aus rechtlichen Gründen derzeit nicht möglich."	Marcus
5.2.1	Datenschutzrichtlinien: Definition und Implementierung	"Die Datenschutzrichtlinien für LLMs werden analog zu anderen Drittanbieter-Softwarelösungen gehandhabt, die Unternehmensdaten speichern oder verarbeiten (z. B. Google Workspace und Google Cloud). Vertragsregelungen stellen sicher, dass Daten weder an Dritte weitergegeben noch für Trainingszwecke genutzt werden dürfen."	Ralf
5.2.2	Maßnahmen zur Sicherstellung gesetzlicher Vorgaben (z. B. DSGVO)	"Es werden regelmäßig Audits durchgeführt. Wir haben auch jährlich verpflichtende Schulungen."	Theo
6.1.1	Auswahlkriterien für LLMs (Datensicherheit, Anpassungsfähigkeit, Support)	"Zum einen kommt es darauf an, welche Daten ich mit dem Modell verarbeiten will – wenn es um personenbezogene Daten geht, ist Datensicherheit relevanter. Im Engineering-Bereich würde ich Modelle bevorzugen, die anpassbar sind, also adaptiver auf meinen Use Case zugeschnitten."	Theo
6.1.2	Vergleich: Open-Source vs. kommerzielle Modelle	"Die Qualität und die Performance des LLMs ist meist nicht so gut wie die kommerziellen LLMs. Weiters ist die Implementierung von kommerziellen LLMs leichter, da sie fertige APIs und SDKs zur Verfügung stellen."	Eric

Kategorienummer	Definition der Kategorie	Ankerbeispiel (Zitat)	Name
6.2.1	Integration in bestehende Systeme (ERP, CRM, DMS)	"Unsere LLM-Anwendungen sind derzeit nur teilweise in bestehende Unternehmenssysteme integriert. Der für Kunden verfügbare KI-Assistent, der auf OpenAI GPT-4o-mini basiert, kann beispielsweise Daten zu Bestellungen, Anlieferungen und Produkten abrufen und verarbeiten. Dadurch wird der Support für Fulfillment-Kunden effizienter gestaltet."	Ralf
6.2.2	Nutzung von Schnittstellen und API-Gateways	"Die Kommunikation zwischen den Systemen erfolgt über standardisierte RESTful-HTTP-APIs. Wir nutzen OpenAPI-Spezifikationen, um diese einfach in OpenAI zu integrieren."	Ralf
6.2.3	Kompatibilität mit bestehenden Sicherheitslösungen und Monitoring	"Das erfolgt in enger Zusammenarbeit mit der IT-Security und den Datenschutzverantwortlichen. Externe Penetrationstests werden beauftragt, um das System sowohl als White-Box als auch als Black-Box zu prüfen."	Anton
7.1.1	Skalierung, Modellversionierung und Rollback-Strategien	"Die Containerwelt und Kubernetes ermöglichen feste Versionen von Container-Images, die mittels Semantic Versioning aktualisiert oder zurückgesetzt werden können."	Anton
7.1.2	Überwachung der Betriebskosten (z. B. KPIs, Dashboards)	"Das erfolgt im Rahmen von FinOps. Die Ressourcenauslastung und Cloud-Kosten werden kontinuierlich überwacht und den Entwicklungsteams berichtet, sodass mögliche Kosten-Spikes frühzeitig erkannt werden."	Anton
7.2.1	Nutzung von KPIs (Antwortzeit, Genauigkeit, Kosten)	"Wir nutzen Kundenfeedback und G-Eval."	Ralf
7.2.2	Iterativer Verbesserungsprozess (Reviews, A/B-Tests, Benchmarking)	"Die kontinuierliche Verbesserung erfolgt durch Logging und die Evaluierung anonymisierter Prompts mit Modellantworten, wodurch die Instruktionen regelmäßig optimiert werden."	Ralf
8.1	Maßnahmen zur Datensicherheit (z. B. Datenmaskierung, Verschlüsselung)	"Kundendaten werden gemäß den DSGVO-Vorgaben anonymisiert, um das Risiko von Datenlecks zu minimieren."	Ralf
8.2	Zugriffskontrolle und Identity Management	"Primär: Azure AD für die Authentifizierung und Autorisierung, OneIdentity + SAP SuccessFactor im Hintergrund für die Benutzerverwaltung."	Eric
8.3	Sicherstellung der Einhaltung gesetzlicher Vorgaben	"Interner Meldungsprozess an den Datenschutzbeauftragten + Freigabe vom Security-Team."	Eric

Kategorienummer	Definition der Kategorie	Ankerbeispiel (Zitat)	Name
9.1	Analyse und Dokumentation der bestehenden IT-Landschaft (ERP, CRM, DMS)	"Datenstrukturen wurden analysiert und ggf. vereinheitlicht, um die Vektoren-Datenbank für das RAG-Modell vorzubereiten."	Eric
9.2	Integration unternehmensspezifischer Datenschutz- und Compliance-Richtlinien	"Es gibt eine klare Klassifizierung der Daten – ob öffentliche, interne oder vertrauliche Daten – und diese sind mit dem bestehenden Rollen- und Rechtemanagement versehen."	Anton
9.3	Bewertung der internen MLOps-/DevOps-Kapazitäten (Support/Wartung)	"Intern verfügen wir bereits über das notwendige Know-how durch unsere Data Analysts, die entsprechende Prozesse begleiten können. Auch DevOps-Ressourcen sind ausreichend vorhanden."	Ralf
10.1	Weitere kritische technische oder organisatorische Aspekte	"Bias und Ethik, Angriffsszenarien, Modellanpassung. Einführung von KI bedarf Changemanagement, Verantwortung, Schulung der Mitarbeiter, fehlende Data Governance und Datenmanagement."	Mia
10.2	Empfehlungen und Best Practices für den LLM-Einsatz	"Man sollte die Mitarbeiter frühzeitig abholen, klare Ziele definieren und Sicherheits- sowie Datenschutzbedenken nicht außer Acht lassen. Zudem muss man sich darüber im Klaren sein, dass diese Technologien hohe Investitionen in Infrastruktur erfordern – und abwägen, ob der Return on Investment den Aufwand rechtfertigt."	Anton
10.3	Zukunftsperspektiven: Entwicklungen und Trends in Bezug auf LLMs	"Zukünftig sehe ich insbesondere die Veränderung bestehender Arbeitsweisen als relevanten Trend, da LLMs zunehmend Prozesse automatisieren und neue Formen der Zusammenarbeit ermöglichen. Die Sammlung großer spezialisierter Datenmengen für das Training und Fine-Tuning von LLMs wird auch wichtig sein, um qualitativ zufriedenstellende und für den Einsatzbereich optimierte Ergebnisse zu erhalten."	Ralf

Tabelle 7.5.: Darstellung induktive Kategorien

Kategorie	Definition	Ankerbeispiel (Zitat)	Name
Management-unterstützung und KI-Strategie	Die Notwendigkeit einer klaren KI-Strategie und der Unterstützung durch das Top-Management, um LLM-Projekte erfolgreich umzusetzen. Ohne strategischen Rückhalt und Priorisierung durch die Führungsebene geraten Implementierungen ins Stocken.	"Ja, es ist essenziell, dass das Top-Level-Management die Aktivitäten unterstützt..."	Anton
Mangel an KI-Fachkräften und Schulungs-bedarf	Unerwartetes Defizit an internem Know-how und spezialisierten Fachkräften für LLM-Implementierungen. Unternehmen sehen sich gezwungen, Mitarbeiter weiterzubilden und externe KI-Expertise aufzubauen, um LLM-Projekte realisieren und betreiben zu können.	"... weiters müssen Personen richtig geschult werden sowie eigene KI-Spezialisten eingestellt werden die sich mit dem Thema gut auskennen."	Eric
Fehlende Daten-Governance und Datenqualitäts-prozesse	Das Fehlen etablierter Prozesse und Richtlinien für Datenmanagement und -qualität. Es zeigen sich Lücken bei der Sicherstellung konsistenter, aktueller und bereinigter Daten sowie bei der übergreifenden Daten-Governance, was die Vorbereitung von Trainingsdaten für LLMs erschwert.	"... fehlende Data Governance und Datenmanagement.."	Mia
Bedeutung offener Datenformate für Kompatibilität	Die Erkenntnis, dass die Verwendung offener, standardisierter Datenformate langfristig die Integration von LLMs erleichtert. Durch frühzeitige Festlegung auf zukunftssichere Formate können aufwändige und fehleranfällige Datenkonvertierungen in späteren Projektphasen vermieden werden.	"Eine grundsätzliche Firmen-Entscheidung für offene und zukunftssichere Datenformate ist in allen Fällen sinnvoll, da offene Datenformate eine Integration in jedweden Software-Systemen vereinfachen... Eine nachträgliche Daten Konvertierung ist nicht nur zeit- und kostenaufwändig, sondern auch fehleranfällig."	Marcus
Bedeutung offener Datenformate für Kompatibilität	Die Verwendung offener, standardisierter Datenformate erleichtert langfristig die Integration von LLMs und vermeidet durch frühzeitige Festlegung auf zukunftssichere Formate aufwändige, fehleranfällige Datenkonvertierungen in späteren Projektphasen.	"Eine grundsätzliche Firmen-Entscheidung für offene und zukunftssichere Datenformate ist in allen Fällen sinnvoll, da offene Datenformate eine Integration in jedweden Software-Systemen vereinfachen... Eine nachträgliche Daten Konvertierung ist nicht nur zeit- und kostenaufwändig, sondern auch fehleranfällig."	Marcus
Eingeschränkte Integration und Pilotcharakter	Viele LLM-Anwendungen befinden sich noch in der Pilotphase und sind nicht voll in die bestehende IT-Landschaft eingebunden. Statt einer tiefen Integration in ERP-, CRM- oder DMS-Systeme laufen erste LLM-Lösungen oft isoliert oder als Prototypen, was auf einen unerwartet hohen Implementierungsaufwand und Zurückhaltung hinweist.	"Unsere LLM-Anwendungen sind derzeit nur teilweise in bestehende Unternehmenssysteme integriert... Interne Tools sind hingegen bislang nicht direkt an unsere Systeme angebunden und werden derzeit ausschließlich als Standalone-Lösungen genutzt."	Ralf

Kategorienummer	Definition der Kategorie	Ankerzitat	Name
Neue Sicherheits- und Compliance-Herausforderungen	Durch den Einsatz von LLM-Diensten entstehen neuartige Sicherheitsanforderungen. Unternehmen müssen z. B. vertraglich sicherstellen, dass vertrauliche Daten nicht vom Anbieter zu Trainingszwecken genutzt werden, und bestehende Sicherheitskonzepte auf potenzielle Angriffsflächen im LLM-Kontext ausweiten. Diese zusätzlichen Maßnahmen waren zuvor nicht offensichtlich.	"Alle Tools unterliegen NDAs und Opt-Out-Klauseln, um eine Nutzung unserer Daten zu Trainingszwecken auszuschließen. Die Verträge werden von unserer Legal Abteilung geprüft."	Ralf
Changemanagement und Nutzerakzeptanz	Die Einführung von LLM-Technologie erfordert begleitendes Change Management und Maßnahmen zur Nutzerakzeptanz. Unvorhergesehen zeigt sich, dass ohne passende Schulungen, Einbindung der Mitarbeiter und schrittweises Heranführen an KI-Tools entweder die Nutzung ausbleibt oder Mitarbeitende auf inoffizielle Tools ausweichen (Schatten-IT).	"Einführung von KI bedarf Changemanagement, Verantwortung, Schulung der Mitarbeiter, ..."	Mia

8. Diskussion

In diesem Kapitel werden die Ergebnisse der Expert*inneninterviews den theoretischen Erkenntnissen gegenübergestellt und das entwickelte Framework kritisch reflektiert. Dabei zeigt sich, inwiefern die in der Theorie identifizierten Anforderungen in der Praxis Bestand haben und welche zusätzlichen Aspekte bei der Implementierung von LLM-Lösungen beachtet werden müssen. Die Diskussion gliedert sich in drei Teile. Zunächst wird in Abschnitt 8.1 die Validierung des Frameworks durch die Expert*innen beleuchtet. Anschließend werden Herausforderungen bei der Implementierung von LLM-Lösungen in der Unternehmenspraxis Abschnitt 8.2 diskutiert. Abschließend widmet sich Abschnitt 8.3 den Datenschutz- und Compliance-Herausforderungen, die im Zusammenhang mit der Einführung von LLMs auftreten.

8.1. Validierung

Die Praxistauglichkeit des entwickelten Frameworks wurde systematisch durch Experteninterviews validiert. Mehrere Fachexpert*innen aus unterschiedlichen Branchen und Unternehmensgrößen, von kleinen Betrieben (15 Mitarbeitende) bis hin zu Großkonzernen (>7500 Mitarbeitende) wurden befragt. Diese breite Auswahl gewährleistet, dass das Framework in diversen Kontexten anwendbar ist. Die leitfadengestützten Interviews prüften, ob die im Framework identifizierten technischen Anforderungen und Schritte mit den Realitäten in Unternehmen übereinstimmen. Dabei zeigte sich eine hohe Übereinstimmung zwischen Theorie und Praxis: „Die Kernkomponenten des Frameworks, von Infrastruktur über Daten und Deployment bis hin zu Betrieb, wurden durch die Expertenmeinungen größtenteils bestätigt."

Insbesondere betonten alle Befragten die im Framework hervorgehobenen technischen Grundvoraussetzungen. So wurde eine robuste IT-Infrastruktur als essenzielle Basis für den erfolgreichen LLM-Einsatz bestätigt. Ein Interviewpartner erklärte: „Die Containerwelt und Kubernetes sind bei uns die Basis, um LLM-Services bedarfsgerecht hoch- und runterzufahren“ (Anton). Diese Aussage deckt sich mit dem Framework-Baustein zur Skalierbarkeit der Infrastruktur, der Containerisierung und Orchestrierung als Schlüsseltechnologien empfiehlt. Ebenso fanden die Aspekte Datenbereitstellung und -qualität starke Resonanz. Mehrere Experten hoben hervor, dass gründliche Datenvorbereitung und Qualitätssicherung unverzichtbar sind, um “Garbage in, Garbage out”-Effekte zu vermeiden (Geiger et al., 2021). Damit bestätigen sie die im Framework geforderte Standardisierung von Datenformaten (z.B. JSONL, TFRecord) und den Einsatz automatisierter ETL-Pipelines zur Datenbereinigung.

Neben technischen Anforderungen wurden auch die im Framework vorgesehenen strategischen Entscheidungsfelder validiert. So war in den Interviews der Umgang mit verschiedenen Bereitstellungsmodellen ein zentrales Thema, genau wie im Framework, das zwischen On-Premises-, Cloud- und Hybrid-Ansätzen differenziert. Mehrere Expert*innen aus Großunternehmen berichteten, derzeit Cloud-LLM-Dienste zu pilotieren, da diese einen schnellen Einstieg ermöglichen (Julie). Gleichzeitig betonten sie aber, dass bei sensiblen Daten On-Premises- oder private Cloud-Lösungen bevorzugt werden (Marcus). Eine der befragten Expert*innen merkte an, dass in ihrem Unternehmen aus Datenschutzgründen „LLM-Funktionen vorerst nur in einer privaten Cloud-Umgebung getestet“ (Julie). Dieses Spannungsfeld zwischen Cloud-

Komfort und Datenschutz spiegelt sich im Framework wider, das entsprechende Entscheidungsfaktoren (Datenschutzrichtlinien, Anbieterabhängigkeit, Kosten-Nutzen) als Prüfpunkte aufführt.

Auch die Auswahl des Modelltyps wurde durch die Interviews bestätigt. Einige Interviewpartner bevorzugen Open-Source-LLMs für mehr Kontrolle und potenzielle Kosteneinsparungen, während andere den Support und die Stabilität kommerzieller Anbieter schätzen (Marcus). Diese Perspektiven entsprechen dem Framework-Bereich, der die Modellwahl (Open-Source vs. kommerziell) anhand von Kriterien wie Anpassbarkeit, Performance und Support evaluiert.

Darüber hinaus lieferten die Interviews wertvolles Feedback zur Verbesserung des Frameworks. Wo die Theorie gewisse Aspekte nur am Rande behandelte, betonten Praktiker*innen zusätzliche Erfolgsfaktoren. Beispielsweise ergab sich, dass Change Management und Mitarbeiter*innenschulungen für die Einführung von LLM-Lösungen entscheidend sind, was auch durch Julie bekräftigt wird: „Einführung von KI bedarf Changemanagement“. Diese Aspekte stehen zwar im Framework weniger im Vordergrund, sind jedoch für den Erfolg der Implementierung unverzichtbar. Solche Rückmeldungen sind hilfreich bei einer Weiterentwicklung des Frameworks. Konkret wurde empfohlen, frühzeitig alle Stakeholder, inklusive Datenschutz- und Sicherheitsabteilungen, in LLM-Projekte einzubinden, um Hürden proaktiv abzubauen (Anton). Ein weiterer Punkt, welcher in eine mögliche Optimierung des Frameworks einfließen könnte, indem Maßnahmen des Stakeholder- und Change-Managements verankert werden.

8.2. Herausforderungen bei der Implementierung

Trotz eines klaren Frameworks offenbaren sich in der Praxis vielfältige Herausforderungen bei der Implementierung von LLM-Lösungen. Die Interviews und Literatur zeigen übereinstimmend, dass Unternehmen sowohl technische Hürden als auch organisatorische Barrieren überwinden müssen, um LLM-Initiativen erfolgreich zu realisieren. Im technischen Bereich beginnen die Herausforderungen bei der Infrastruktur. Ohne ausreichende Rechenleistung (GPUs/TPUs) und eine skalierbare Architektur stoßen selbst mittelgroße Sprachmodelle schnell an Leistungsgrenzen. Viele Bestands-IT-Umgebungen sind nicht von vornherein auf die hohen Anforderungen kontinuierlicher LLM-Inferenz ausgelegt. Unternehmen müssen daher in Hardware-Upgrades und Cloud-Ressourcen investieren, was speziell für kleinere Betriebe eine Hürde darstellen kann. So berichtete ein Experte, dass sein Unternehmen für große Datenmengen bewusst auf eigene Rechenzentren setzt, da „der Aufwand, mehrere Terabyte in die Cloud zu schieben, zu hoch wäre“, während für geringere Datenvolumina auch Cloud-Dienste genutzt werden (Anton) Dieses Beispiel illustriert, dass Datenmenge und Infrastrukturkapazität die Deployment-Entscheidung beeinflussen. Eine Herausforderung, vor der besonders datenintensive Projekte stehen.

Ein weiteres zentrales technisches Hindernis ist die Datenvorbereitung. LLMs erfordern strukturierte, qualitativ hochwertige Daten, doch viele Unternehmen kämpfen mit fragmentierten oder unbereinigten Datenbeständen. Die Experten betonten unisono, dass erhebliche Aufwände in ETL-Pipelines, Datenbereinigung und -Standardisierung fließen müssen, bevor ein LLM produktiv nutzbaren Output liefern kann (Ralf).

Werden inkonsistente oder unvollständige Daten verwendet, drohen „Garbage in, Garbage out“-Probleme, wie Geiger et al. (2021) warnen. In der Praxis müssen daher Prozesse zur Duplikaterkennung, Datenanreicherung und kontinuierlichen Aktualisierung etabliert werden, was sowohl personelle Ressourcen

als auch technische Lösungen (z.B. automatisierte Data Pipelines) erfordert. Besonders anspruchsvoll ist die Vorbereitung von Wissensdatenbanken für RAG. Hier müssen Dokumente vektorisiert, Indexe aufgebaut und Suchalgorithmen integriert werden, um unternehmensspezifisches Wissen für das LLM nutzbar zu machen. Die Implementierung solcher RAG-Pipelines erfordert spezialisiertes Know-how in NLP und Datenbanken. Eine Hürde, die nicht alle Unternehmen intern überwinden können.

Neben den rein technischen Aufgaben treten organisatorische Barrieren deutlich zutage. Fast alle Expert*innen wiesen darauf hin, dass menschliche Faktoren und Prozesse den Erfolg von LLM-Projekten maßgeblich beeinflussen. Häufig fehlt es an internem Know-how im Umgang mit großen KI-Modellen. Qualifizierte KI-Entwicklerinnen, Data Engineers oder MLOps-Spezialist*innen sind rar und heiß umkämpft. Dieser Fachkräftemangel kann Projekte ausbremsen oder dazu führen, dass externe Dienstleister einbezogen werden müssen, was wiederum Abstimmungsaufwand bedeutet. Zudem können Widerstände in der Belegschaft auftreten, wenn neue, KI-basierte Arbeitsweisen eingeführt werden. Änderungen von etablierten Prozessen und die Angst vor Arbeitsplatzveränderungen oder -verlusten lösen mitunter Skepsis aus. Ohne begleitendes Change Management besteht das Risiko, dass selbst gut konzipierte LLM-Lösungen nicht angenommen werden. Ein Praxisbeispiel lieferte hier eine Interviewpartnerin. In ihrem Unternehmen blieb die Nutzung eines eingeführten KI-Assistenzsystems deutlich hinter den Erwartungen zurück, „weil es an begleitenden Schulungen mangelte und die Belegschaft das Potenzial nicht einschätzen konnte“ (Julie). Dieses Zitat verdeutlicht, dass fehlende Weiterbildung und interne Kommunikation eine erhebliche Implementierungsbarriere darstellen. Entsprechend empfehlen sowohl Experten als auch die Literatur, frühzeitig umfassende Schulungsprogramme, Pilotphasen und transparente Kommunikationsstrategien einzusetzen, um Akzeptanz für LLM-Initiativen zu schaffen. Ohne das Mitnehmen der Mitarbeiterinnen und eine klare Zuteilung von Verantwortlichkeiten (Wer betreut das LLM? Wer überwacht die Outputs?) drohen Verzögerungen oder Fehlentwicklungen im Projekt.

Auch die IT-Sicherheit und bestehende Unternehmensstrukturen beeinflussen die Umsetzung. LLM-Systeme müssen nahtlos in eine oft vielschichtige IT-Landschaft integriert werden, welche aus Legacy-Systemen, Datenbanken und Anwendungen besteht. Die Integration in bestehende Systeme erfordert die Entwicklung von Schnittstellen oder die Nutzung von API-Gateways, damit LLM-Funktionen beispielsweise in Intranet-Plattformen oder ERP-Systeme eingebettet werden können.

Eine Herausforderung liegt darin, diese Integrationen stabil und sicher zu gestalten. Bestehende Sicherheitskonzepte (Firewalls, Identity Management, Monitoring) müssen auf LLM-Dienste ausgedehnt werden, um unbefugten Zugriff und Datenabfluss zu verhindern. Die Interviews machten deutlich, dass strikte Zugriffskontrollen, Authentifizierungsmechanismen und Verschlüsselung bei LLM-APIs unerlässlich sind. In der Umsetzung stoßen Unternehmen hier jedoch auf komplexe Fragen. Wie lässt sich ein externer Cloud-Dienst ins interne Identity Management einbinden? Wie verhindert man, dass sensible Daten überhaupt die Umgebung des LLM verlassen? Solche Sicherheitsaspekte erfordern enge Abstimmung zwischen Entwicklungs-, IT-Sicherheits- und Compliance-Teams. Insbesondere in stark regulierten Branchen (Finanzwesen, Gesundheitssektor) entstehen organisatorische Silos und Freigabeprozesse, die die Einführung neuer Technologien verlangsamen können. Eine Expertin berichtete etwa, dass ihr Unternehmen aus Sorge vor Datenlecks LLM-Funktionen zunächst nur in abgeschotteten Umgebungen testet. Die Unternehmensgröße und -struktur spielt hierbei eine Rolle. Größere Konzerne verfügen zwar über dedizierte Sicherheitsabteilungen und klare Prozesse, diese können aber auch zu mehr Bürokratie

führen. Kleine Unternehmen sind agiler, haben jedoch oft weniger ausgearbeitete Sicherheitsrichtlinien. Sie müssen teils erst geeignete Policies entwickeln, was Zeit kostet. In beiden Fällen gilt es, Technik und Organisation in Einklang zu bringen, damit Sicherheitsauflagen die technische Integration nicht lähmen.

Schließlich sind wirtschaftliche Faktoren nicht zu vernachlässigen. Die Kosten für die Entwicklung, den Betrieb und die Skalierung von LLM-Anwendungen können beträchtlich sein. Hardwarebeschaffung, Cloud-Nutzung, Lizenzgebühren für Modelle oder APIs sowie laufende Wartung summieren sich zu einem Investment, das gut begründet werden will. Mehrere Interviewpartner warnten, dass ohne fortlaufendes Kosten-Monitoring die betrieblichen Aufwände aus dem Ruder laufen können. Ein Experte erläuterte dazu, dass in seinem Unternehmen die Cloud-Kosten für LLM-Services im Rahmen von FinOps genau beobachtet und optimiert werden, um wirtschaftlich tragfähig zu bleiben. Hier zeigt sich eine Herausforderung. Der Nutzen (ROI) von LLM-Projekten ist häufig erst mittel- bis langfristig quantifizierbar, da viele Initiativen experimentellen Charakter haben. Diese Verzögerung bei der Erfolgsmessung erschwert intern die Rechtfertigung weiterer Investitionen. Management und Fachbereiche müssen von den Mehrwerten der KI-Lösungen überzeugt werden, bevor diese vollumfänglich nachweisbar sind, eine klassische Hürde in Innovationsprojekten (Gartner, 2024).

8.3. Datenschutz- und Compliance-Herausforderungen

Bei allen technischen Möglichkeiten von LLMs darf die Bedeutung von Datenschutz und Compliance nicht in den Hintergrund treten. Die Gewährleistung regulatorischer Vorgaben, allen voran der Datenschutz-Grundverordnung (DSGVO), wurde von allen Expert*innen als zentrales Erfolgskriterium genannt. In den Interviews zeigte sich ein klares Bild. Unternehmen stehen vor dem Spagat, einerseits innovative KI-Lösungen nutzen zu wollen, andererseits aber strikte Datenschutzrichtlinien einhalten zu müssen. „Die Datensicherheit hat in unserem Arbeitsumfeld höchste Priorität; somit ist eine Cloud-Lösung für uns aus rechtlichen Gründen derzeit nicht möglich“ (Marcus), erklärte ein Experte aus einem hochregulierten Umfeld. Dieses Zitat bringt die Kernproblematik auf den Punkt, insbesondere bei öffentlichen Cloud-Angeboten herrscht oft Unsicherheit, ob und wie personenbezogene oder vertrauliche Daten dort geschützt sind. Entsprechend tendieren Organisationen mit sehr sensiblen Daten (z.B. im Gesundheitswesen oder Finanzsektor) dazu, On-Premises- oder Private-Cloud-Lösungen zu favorisieren, um die Datenhoheit zu bewahren. Hochsicherheitsbereiche werden intern gehalten, während Cloud-Dienste, wenn überhaupt, nur für unkritische Anwendungsfälle oder anonymisierte Daten genutzt werden.

Um datenschutzkonform zu arbeiten, sollten Unternehmen eine Reihe von Maßnahmen ergreifen, die bereits bei der technischen Konzeption ansetzen. Ein zentrales Prinzip ist Privacy by Design. Schon beim Aufbau von LLM-Integrationen müssen Mechanismen implementiert werden, die den Schutz personenbezogener Daten sicherstellen. Dazu zählt beispielsweise, dass sensible Daten vor der Verarbeitung durch ein LLM maskiert oder pseudonymisiert werden und dass nur notwendige Informationen an das Modell gelangen. Technisch können hier Datenmaskierung, Anonymisierung und Ende-zu-Ende-Verschlüsselung zum Einsatz kommen. Mehrere Experten betonten außerdem die Wichtigkeit von Zugriffskontrollen. Nicht jeder Mitarbeitende sollte ungefiltert LLM-Abfragen mit Unternehmensdaten durchführen dürfen. Durch ein gestuftes Berechtigungskonzept und Logging aller LLM-Anfragen lässt sich nachvollziehen, welche Daten verwendet wurden und ob ggf. gegen Richtlinien verstoßen wurde. Solche technischen

Schutzvorkehrungen, im Framework als Teil des Sicherheits- und Compliance-Bereichs verankert, schaffen die Grundlage dafür, dass LLM-Systeme sicher in bestehende Datenschutzkonzepte eingebettet werden können.

Neben technischen Maßnahmen sind organisatorische und vertragliche Vorkehrungen unverzichtbar, um Compliance-Hürden zu bewältigen, ohne die technische Integration zu behindern. Die Interviews ergaben, dass viele Unternehmen LLM-Dienste analog zu anderen Drittanbieter-Services behandeln. Datenschutzrichtlinien für LLMs orientieren sich an bestehenden Vorgaben für SaaS-Lösungen (z.B. Office-Cloud-Dienste). Es werden detaillierte Vertragsregelungen (AVV/Datenschutzverträge) mit Anbietern geschlossen, um sicherzustellen, dass übermittelte Daten weder weitergegeben noch zu eigenen Zwecken (wie dem Training des Anbietermodells) genutzt werden dürfen. Ein Experte erklärte, man stelle vertraglich klar, dass Daten beim Cloud-LLM-Anbieter „weder an Dritte weitergegeben noch für Trainingszwecke genutzt werden dürfen“ (Ralf). Solche Vereinbarungen zur Auftragsverarbeitung und zur Datenlokation (Speicherung nur in bestimmten Rechtsräumen) sind essenziell, wenn externe Services eingebunden werden. Unternehmen sollten diese Verträge gründlich prüfen und regelmäßig aktualisieren, insbesondere im Hinblick auf neue Gesetzgebungen (etwa zukünftige KI-Regulierungen) und internationale Datentransfers. Nur so kann gewährleistet werden, dass auch bei Nutzung von Cloud-LLMs eine DSGVO-konforme Datenverarbeitung stattfindet.

Die interne Compliance-Perspektive ist ebenso wichtig. Ein funktionierendes LLM-Integrationsprojekt erfordert die Einbindung der Compliance- und Rechtsabteilung von Beginn an. Die Experten empfahlen, Datenschutzbeauftragte und IT-Sicherheitsverantwortliche frühzeitig ins Boot zu holen, um gemeinsam passende Lösungen zu erarbeiten (Anton). So lassen sich mögliche rechtliche Bedenken oder Prüfprozesse bereits während der Entwicklung berücksichtigen, anstatt in späten Projektphasen böse Überraschungen zu erleben. Diese proaktive Einbindung kann beispielsweise dazu führen, dass bestimmte Daten gar nicht erst in das LLM eingespeist werden oder dass ein Kontrollmechanismus eingebaut wird, der jede Antwort des LLM auf vertrauliche Inhalte scannt. Durch solch vorausschauende Planung bleibt die technische Integration im Fluss, da Compliance-Restriktionen nicht als Blockade, sondern als gestaltbarer Parameter verstanden werden. Mehrere Unternehmen führen zudem regelmäßige Audits und Schulungen durch, um die Einhaltung aller Vorgaben sicherzustellen. Ein Experte berichtete von jährlich verpflichtenden Datenschutz-Schulungen für Mitarbeitende im Umgang mit sensiblen Daten (Theo). Diese Sensibilisierung der Belegschaft sorgt dafür, dass auf operativer Ebene keine unbeabsichtigten Datenschutzverstöße passieren. Ein wichtiger Aspekt, da menschliches Fehlverhalten oft das schwächste Glied in der Sicherheitskette ist.

Insgesamt wurde deutlich, dass Datenschutz und technische Innovation Hand in Hand gehen müssen. Anstatt LLM-Projekte aus Angst vor Compliance-Problemen zu bremsen, sollten Unternehmen einen integrierten Ansatz verfolgen. Die LLM-Integration ist so zu gestalten, dass Security & Compliance by Design gewährleistet sind, und parallel müssen Organisationsstrukturen geschaffen werden, die diese neuen Technologien verantwortungsvoll nutzen. Die identifizierten Datenschutz-Herausforderungen erklären mit, warum viele KI-Projekte nur zögerlich vorankommen. Werden sie jedoch durch geeignete Maßnahmen adressiert, steht einer erfolgreichen und rechtskonformen Nutzung von LLMs im Unternehmen nichts im Wege. Das vorgestellte Framework trägt dem Rechnung, indem es Datenschutz als Querschnittsthema begreift und an mehreren Stellen, von der Architekturplanung bis zum Betrieb, konkrete

Prüfpunkte für Compliance integriert. Unternehmen, die diese Empfehlungen aufgreifen, können LLM-Technologien einsetzen, ohne in Konflikt mit geltenden Datenschutzgesetzen zu geraten, und damit die Vorteile der KI nutzen, während sie zugleich das Vertrauen von Kunden, Partnern und Aufsichtsbehörden bewahren.

9. Fazit

Ausgangspunkt dieser Arbeit war die Beobachtung, dass LLMs enorme Potenziale für die Automatisierung von Wissensarbeit bieten, gleichzeitig aber hohe Anforderungen an Daten und IT-Infrastruktur stellen. Vor diesem Hintergrund wurde als zentrale Forschungsfrage formuliert: „Welche technischen Anforderungen müssen Unternehmen erfüllen, um LLMs erfolgreich in ihre Arbeitsprozesse zu integrieren?“ Die Motivation lag darin, Unternehmen einen Leitfaden an die Hand zu geben, der sie bei der Vorbereitung auf den Einsatz von LLM-Technologien unterstützt. Entsprechend zielte die Arbeit darauf ab, ein Framework zu entwickeln, das diese Frage beantwortet und praxisnahe Handlungsempfehlungen bietet.

9.1. Zusammenfassung der zentralen Ergebnisse

Im Verlauf der Arbeit wurden grundlegende Eigenschaften und Voraussetzungen von LLMs sowie konkrete Unternehmensanforderungen systematisch untersucht. Zunächst wurden die Grundlagen zu LLMs erarbeitet. Alle modernen LLMs basieren auf Transformer-Netzwerken und werden mit sehr großen Textmengen trainiert. Diese Trainingsdaten umfassen typischerweise Milliarden von Wörtern aus Internetquellen, Büchern, Code und weiteren Domänen (Brown et al., 2020). So verwendete OpenAIs GPT-3 etwa 300 Milliarden Tokens (1,2 TB Textdaten), und öffentliche Datensammlungen wie "The Pile" umfassen rund 800 GB an Text (Glenn K. Lockwood, 2025). Dadurch wurde deutlich, welche Art von Daten LLMs benötigen, nämlich gewaltige, vielfältige Korpora an überwiegend von Menschen verfassten Inhalten. In der Arbeit wurde auch aufgezeigt, dass alle LLM-Modelle trotz unterschiedlicher Architekturen und Größen in ihrer Leistungsfähigkeit stark von der Qualität und Diversität dieser Trainingsdaten abhängen. Eine allgemeine Aussage, die für alle LLMs gilt, lautet daher „Ein LLM ist immer nur so gut wie seine Trainingsdaten“. Damit einher geht die Feststellung, dass aktuelle Spitzenmodelle vor allem durch enormen Rechenaufwand und Datenmengen herausragen (Walker, 2023).

Aufbauend auf diesen Grundlagen wurden die technischen Anforderungen für die Unternehmensintegration analysiert. Hier kristallisierten sich drei zentrale Bereiche heraus: Dateninfrastruktur, IT-Infrastruktur und Integrationsarchitektur. Erstens muss eine geeignete Dateninfrastruktur vorhanden sein, was bedeutet, dass Unternehmensdaten in angemessenen Formaten und hoher Qualität vorliegen und zugänglich sind. Dazu zählen strukturierte Datenspeicher, effiziente Datenpipelines sowie Mechanismen zur Gewährleistung von Datenschutz und Compliance.

Zweitens benötigt es eine skalierbare IT-Infrastruktur. LLMs erfordern erhebliche Rechenressourcen (GPU-/TPU-Hardware oder Cloud-Services) und Speicherkapazitäten, sodass Unternehmen entsprechende Hardware bereitstellen oder Cloud-Strategien nutzen müssen.

Drittens ist eine durchdachte Integrationsarchitektur nötig. Etwa über APIs oder Containerisierung, um LLM-Dienste nahtlos in bestehende Softwarelandschaften einzubetten. Diese Ergebnisse wurden im entwickelten Framework gebündelt, das als Kriterienkatalog alle identifizierten technischen Voraussetzungen übersichtlich darstellt. Das Framework wurde universell konzipiert, sodass es unabhängig von Branche

oder LLM-Plattform angewendet werden kann, und dient Entscheidungsträgern als Orientierung bei der Implementierung.

Des Weiteren widmete sich die Arbeit einem Vergleich verschiedener LLM-Lösungen. Untersucht wurden sowohl führende proprietäre Modelle als auch Open-Source-Alternativen. Die Analyse zeigte, dass insbesondere GPT-4 (OpenAI) und Googles Gemini zu den aktuell leistungsstärksten kommerziellen LLMs zählen, während auf Open-Source-Seite Meta's LLaMA-2 sowie neuere Modelle wie Mistral Le Chat hervorstechen.

Diese Modelle repräsentieren den State of the Art und erreichten auf diversen Benchmarks herausragende Ergebnisse (Mistral AI Team, 2023; OpenAI, 2024). Gleichzeitig konnte anhand der Literatur gezeigt werden, dass Open-Source-Modelle mit deutlich weniger Parametern auf bestimmten Aufgaben bereits mit größeren proprietären Modellen konkurrieren können. So übertrifft etwa ein LLaMA-Modell (13B Parameter) das ältere GPT-3 (175B) auf vielen Benchmarks (Minaee et al., 2024).

Dennoch bleiben Top-Modelle wie GPT-4 in ihrer allgemeinen Leistungsfähigkeit vorerst unerreicht. Insgesamt unterstrich der Modellvergleich: Alle LLMs weisen gemeinsame Grundprinzipien auf, differenzieren sich aber in Aspekten wie Modellgröße, Trainingsdaten, Kosten und Zugänglichkeit. Für Unternehmensentscheider ist diese Gegenüberstellung wertvoll, um abzuwägen, ob etwa ein flexibles Open-Source-Modell genügen kann oder ob die Mehrleistung eines proprietären Dienstes den höheren Integrationsaufwand rechtfertigt.

Zur Evaluierung des Frameworks und der herausgearbeiteten Anforderungen wurden Expert*inneninterviews durchgeführt. Die befragten Expert*innen bestätigten weitgehend die in der Literatur identifizierten technischen Erfolgsfaktoren. Insbesondere hoben sie die Bedeutung von Datenqualität und -sicherheit hervor sowie die Notwendigkeit, interne Wissensbestände für LLMs nutzbar zu machen (etwa durch Ansätze wie Retrieval-Augmented Generation). Ihre Praxisperspektive lieferte zusätzliche Einsichten, zum Beispiel dass neben technischen auch organisatorische Hürden (Schulungen, Change Management) relevant sind. Die Diskussion der Ergebnisse in Kapitel 7 zeigte eine hohe Übereinstimmung zwischen Theorie und Praxis. Die im Rahmen der Literaturreview abgeleiteten Anforderungen decken sich größtenteils mit den Aussagen der Expert*innen. Dieses Ergebnis stärkt die Aussagekraft des entwickelten Frameworks deutlich.

9.2. Beantwortung der Forschungsfrage

Durch die gewonnenen Erkenntnisse lässt sich die Forschungsfrage nun konkret beantworten. „Welche technischen Anforderungen müssen Unternehmen erfüllen, um LLMs erfolgreich in ihre Arbeitsprozesse zu integrieren?“ Die vorliegende Arbeit hat gezeigt, dass Unternehmen im Wesentlichen drei Kernanforderungen erfüllen müssen: Verfügbarkeit hochwertiger LLM-kompatibler Daten, Bereitstellung geeigneter Rechen- und Systemressourcen und eine integrative Systemarchitektur zur Einbindung der Modelle. Konkret bedeutet dies erstens, dass Unternehmensdaten in ausreichender Menge und Qualität vorliegen und aufbereitet sein müssen (z.B. Bereinigung, einheitliche Formate, Metadaten). Zweitens müssen entsprechende IT-Ressourcen vorhanden sein, um vom skalierbaren Storage für große Modelle und Datensätze bis hin zu leistungsfähigen GPUs bzw. Cloud-Diensten, die das Training oder zumindest das

Inference (Nutzung) der Modelle bewältigen können. Drittens bedarf es einer technischen Integrationsstrategie, etwa über Schnittstellen (APIs) oder Container, um die LLM-Funktionalitäten in bestehende Anwendungen (wie Wissensmanagementsysteme, Chatbots oder Prozesse) einzubetten.

Mit diesen Ergebnissen wurde die Zielsetzung der Arbeit erreicht. Die Forschungsfrage ist beantwortet und das Ziel, ein universell einsetzbares Framework als Entscheidungsgrundlage bereitzustellen, wurde erfolgreich umgesetzt. Das Framework bietet einen klaren Mehrwert, da es Unternehmen erlaubt, systematisch Lücken in ihrer technischen Vorbereitung zu identifizieren. Beispielsweise kann ein Unternehmen damit feststellen, ob seine Datenlandschaft bereits "LLM-ready" ist oder ob Investitionen in bestimmte Technologien (etwa ein Data Lake oder ein Vector-Datenbank-System für RAG) nötig sind. Indem die Arbeit die technischen Anforderungen transparent gemacht hat, trägt sie dazu bei, die Schwelle für den Einsatz von LLMs in der Unternehmenspraxis zu senken. Insgesamt leistet die Beantwortung der Forschungsfrage sowohl wissenschaftlich einen Beitrag, indem sie die Schnittstelle von LLM-Technologie und Unternehmens-IT beleuchtet, als auch praktisch, indem sie handlungsorientierte Empfehlungen formuliert.

9.3. Kritische Reflexion und Limitationen

Trotz des Erfolgs bei der Beantwortung der Forschungsfrage ist eine kritische Reflexion der Vorgehensweise und Ergebnisse geboten. Methodisch setzte die Arbeit auf einen (DSR)-Ansatz, kombiniert mit Literaturrecherche und qualitativen Experteninterviews. Ein Vorteil dieses Ansatzes war die iterative Entwicklung des Frameworks. So erwiesen sich die leitfadengestützten Interviews als wertvoller Evaluierungsschritt im DSR-Zyklus, um das konzipierte Kriteriengerüst an realen Bedürfnissen zu spiegeln. Tatsächlich zeigte die hohe Übereinstimmung zwischen den Interviewaussagen und den theoretisch abgeleiteten Anforderungen, dass die gewählte Methode geeignet war, um valide Ergebnisse zu erzielen. Die Expert*innenvalidierung untermauerte die Relevanz der identifizierten Kriterien und half, das Framework praxisnah zu schärfen.

Dennoch sind auch Schwächen und Limitationen der Methoden zu berücksichtigen. Die Literaturrecherche war bewusst auf neueste Quellen (ab 2023) fokussiert, um den rasanten Fortschritten im LLM-Bereich Rechnung zu tragen. Dadurch könnten allerdings ältere, möglicherweise ebenfalls relevante Grundlagenarbeiten unterrepräsentiert sein. Zudem stellt die enorme Dynamik des Themas eine Herausforderung dar. Zwischen Recherche und Fertigstellung der Arbeit erschienen bereits neue Modelle und Technologien, was das Risiko birgt, dass einige Angaben schnell überholt sein könnten. Dieses Aktualitätsproblem ist ein generelles Charakteristikum von Forschungsarbeiten in schnelllebigen Technologiefeldern und wurde durch eine fortlaufende Aktualisierung der Quellen bestmöglich adressiert.

Ein weiterer limitierender Faktor liegt in der Stichprobe der Expert*inneninterviews. Die Anzahl der Interviews war aus Zeit- und Ressourcengründen begrenzt, und die Auswahl der Expert*innen (alle aus dem IT- bzw. KI-nahen Umfeld) könnte zu einem gewissen Bias geführt haben, etwa indem bestimmte Branchenperspektiven weniger berücksichtigt wurden. Qualitative Interviews bieten tiefe Einblicke, sind aber nicht statistisch generalisierbar. Die Ergebnisse spiegeln also primär die Erfahrungen der befragten Fachleute wider. Möglicherweise hätten zusätzliche Interviews mit diverseren Rollen (z.B. Datenschutzbeauftragte, Endanwender) weitere Aspekte zutage gefördert.

Auch inhaltlich gibt es Grenzen. Das entwickelte Framework konzentriert sich auf technische Aspekte und klammert z.B. rechtliche Analysen (Datenschutzgesetzgebung) oder detaillierte betriebswirtschaftliche Implikationen aus. Diese Fokussetzung war notwendig, lässt aber Raum für weiterführende Betrachtungen außerhalb des engen technischen Spektrums. Darüber hinaus wurde RAG als Schwerpunkt beleuchtet, was zwar ein aktuelles und relevantes Thema ist, jedoch auch nicht die einzige Methode darstellt, LLMs mit externem Wissen zu koppeln. Andere Verfahren (wie z.B. spezialisierte Fine-Tunings oder Prompt-Engineering-Strategien) konnten nur am Rande diskutiert werden.

Abschließend muss betont werden, dass die schnelle Weiterentwicklung im LLM-Sektor eine inhärente Limitation darstellt. Jedes heutige Fazit ist eine Momentaufnahme. Unternehmen, die die Ergebnisse dieser Arbeit nutzen, sollten die empfohlenen Lösungen und Technologien stets vor dem Hintergrund der neuesten Entwicklungen neu bewerten.

9.4. Implikationen und Ausblick

Trotz der genannten Limitationen liefern die Ergebnisse der Arbeit wertvolle Implikationen für die Praxis. Für Unternehmen bedeutet dies konkret, dass sie frühzeitig in den Auf- und Ausbau ihrer technischen Grundlagen investieren sollten, um für den Einsatz von LLMs gewappnet zu sein. Insbesondere die Datenqualität und -bereitstellung erweisen sich als Schlüsselfaktor. Unternehmen sollten ihre unternehmensinternen Datenbestände prüfen und aufbereiten (Stichwort Datenhaltung in einem geeigneten Format, Bereinigung von Dubletten, Anreicherung mit Metadaten). Ebenso wichtig ist es, entsprechende Infrastrukturen bereitzustellen, sei es durch Anschaffung moderner Hardware oder durch das Nutzen von Cloud-Angeboten, um die Rechenintensität von LLM-Anwendungen bedienen zu können. Dabei dürfen IT-Sicherheit und Datenschutz nicht vernachlässigt werden. In vielen Fällen empfiehlt es sich, Datenschutzrichtlinien früh in LLM-Projekte zu integrieren oder auf Lösungen wie Private-Cloud- oder On-Premises-Modelle zu setzen, wenn sensible Daten involviert sind.

Ein zentrales praktisches Anwendungsfeld ergibt sich im Wissensmanagement von Unternehmen. Hier können LLMs in Kombination mit bestehenden Wissensdatenbanken erhebliche Effizienzgewinne ermöglichen. Speziell der RAG-Ansatz bietet sich an, um unternehmenseigenes Wissen gezielt nutzbar zu machen. Durch die Verknüpfung eines LLM mit einer internen Dokumenten- oder Datenbank kann das Modell auf aktuelle, unternehmensspezifische Informationen zurückgreifen, anstatt ausschließlich auf seinem statischen Trainingswissen zu basieren (Lewis et al., 2021). Diese Integration verringert das Risiko von Halluzinationen und erhöht die Relevanz der vom LLM generierten Antworten (Mialon et al., 2023). In der Praxis bedeutet das z. B., dass ein Support-Chatbot mit RAG-Unterstützung konkrete Antworten aus den Handbüchern und Richtlinien des Unternehmens liefern kann, anstatt generische Antworten zu geben. Für Unternehmen ist dies ein vielversprechender Weg, LLMs in bestehende Wissensmanagementsysteme einzubetten und somit die Stärken beider Welten zu kombinieren, die Sprachkompetenz der KI und die Verlässlichkeit der eigenen Datenbasis.

Weiterführende Forschung kann an mehrere Punkte anknüpfen. Zum einen wäre eine Validierung des Frameworks in der Breite wünschenswert. Künftige Studien könnten das vorgestellte Kriterien-Framework in unterschiedlichen Unternehmenskontexten erproben, um dessen Allgemeingültigkeit zu testen und ggf. branchenspezifische Anpassungen vorzunehmen. Auch eine quantitative Bewertung der Umsetzung

der Kriterien (etwa durch Metriken oder Reifegradmodelle) könnte anschließen, um Unternehmen noch konkretere Fahrpläne an die Hand zu geben. Zum anderen bieten die rasanten technischen Fortschritte selbst stetig neue Forschungsfragen. Beispielsweise könnte untersucht werden, wie sich neue Modellgenerationen wie GPT-5 oder multimodale LLMs in das bestehende Framework einfügen lassen und ob zusätzliche Anforderungen (z.B. bezüglich Bild-/Videodaten) entstehen. Auch der Vergleich von Integrationsstrategien, etwa Fine-Tuning eines Modells vs. Nutzung von RAG vs. Prompt Engineering, ist ein spannendes Feld, um herauszufinden, welche Methode unter welchen Bedingungen die besten Ergebnisse liefert.

Ein besonders aktueller Aspekt für die Zukunft ist die Frage der Trainingsdatenqualität im Zeitalter wachsender KI-generierter Inhalte. KI lernt zunehmend von sich selbst, d.h. zukünftige Modelle werden möglicherweise auf Datensätzen trainiert, die bereits KI-generierte Texte enthalten. Dies birgt das Risiko eines Model Collapse, bei dem die Modellleistung degeneriert, wenn vorwiegend von KI erzeugte (und potenziell fehlerbehaftete oder einseitige) Daten als Trainingsgrundlage dienen (Seddik et al., 2024). Für die Weiterentwicklung von LLMs bedeutet das, dass Datenkurationsstrategien immer wichtiger werden. Künftige Forschung sollte untersuchen, wie sich ein hoher Anteil synthetischer Daten auf LLMs auswirkt und wie man dem entgegenwirken kann (etwa durch Filtermethoden oder den kontinuierlichen Einbezug menschlicher Originaldaten). Dieses Problem der Datenverarmung könnte langfristig alle LLMs betreffen und ist daher sowohl wissenschaftlich als auch praktisch von Bedeutung. Unternehmen und Entwickler sind gefordert, hier proaktiv gegenzusteuern, um die Leistungsfähigkeit der Modelle zu erhalten.

Abbildungsverzeichnis

1.1. Hauptbarrieren bei der Implementierung von KI-Techniken (in Anlehnung an Gartner (2024) 1
1.2. Unternehmen, die KI-Technologien nutzen (vgl. Eurostat (2024)) 2

2.1. LLM-Fähigkeiten (vgl. Minaee et al. (2024)) 6
2.2. Transformer Architektur (vgl. Vaswani et al. (2023)) 7
2.3. RAG Architektur (vgl. Merritt (2025)) 13

3.1. DSRM Process (in Anlehnung an Peffers et al. (2007)) 16

5.1. Chronologische Darstellung der LLM-Veröffentlichungen (vgl. Naveed et al. (2024)) . . . 24

Tabellenverzeichnis

5.1. Übersicht über führende LLMs erhoben aus: Anthropic (n. d. a, n. d. b, n. d. c, n. d. d, n. d. e), Bergmann, Dave (2023), DocsBot AI (2025), Dunenfeld, Emily (2024), Edwards (2023), Gillham, Jonathan (2024a, 2024b), Hassabis, Demis (2023), Hoblitzell, Andrew (2023), Joshi (2024), Meta (2024), Mistral AI Team (2023, 2024, 2025), OpenAI (2024, 2025), OpenAI et al. (2023), Portakal, Ertugrul (2024) und TypingMind (2025, n. d. a, n. d. b) 25
5.2. Vergleich von Open-Source und kommerziellen LLMs erhoben aus: T. Ahmed et al. (2024), Jones (2024), Malec, Melissa (2025), Network (2024), Uspenskyi, Serhii (2024) und Yu et al. (2023) . . 27

7.1. Übersicht der Expert*innenprofile 38
7.2. Darstellung des Interviewleitfadens 40
7.3. Darstellung deduktive Kategorien 44
7.4. Übersicht der Ankerzitate 45
7.5. Darstellung induktive Kategorien 50

Literatur

Abel Uzoka, Emmanuel Cadet & Pascal Ugochukwu Ojukwu. (2024). Leveraging AI-Powered chatbots to enhance customer service efficiency and future opportunities in automated support. *Computer Science & IT Research Journal*, *5*(10), 2485–2510. https://doi.org/10.51594/csitrj.v5i10.1676

Ahmed, M. (2024, 27. März). *LLMOps Course: Data Preparation for AI Developers*. Verfügbar 28. Februar 2025 unter https://aifordevelopers.io/llmops-part-2-data-preparation/

Ahmed, T., Bird, C., Devanbu, P., & Chakraborty, S. (2024, 23. Februar). *Studying LLM Performance on Closed- and Open-source Data*. arXiv: 2402.15100 [cs]. https://doi.org/10.48550/arXiv.2402.15100

Anthropic. (n. d. a). *Can I use Claude in different languages? | Anthropic Help Center*. Verfügbar 6. März 2025 unter https://support.anthropic.com/en/articles/7996851-can-i-use-claude-in-different-languages

Anthropic. (n. d. b). *Claude 2*. Verfügbar 6. März 2025 unter https://www.anthropic.com/news/claude-2

Anthropic. (n. d. c). *Claude 3.7 Sonnet and Claude Code*. Verfügbar 6. März 2025 unter https://www.anthropic.com/news/claude-3-7-sonnet

Anthropic. (n. d. d). *How up-to-date is Claude's training data? | Anthropic Help Center*. Verfügbar 6. März 2025 unter https://support.anthropic.com/en/articles/8114494-how-up-to-date-is-claude-s-training-data

Anthropic. (n. d. e). *Introducing the next generation of Claude*. Verfügbar 6. März 2025 unter https://www.anthropic.com/news/claude-3-family

AWS. (n. d.). *What is RAG? - Retrieval-Augmented Generation AI Explained - AWS*. Amazon Web Services, Inc. Verfügbar 4. März 2025 unter https://aws.amazon.com/what-is/retrieval-augmented-generation/

Bergmann, Dave. (2023, 18. Dezember). *What Is Llama 2? | IBM*. Verfügbar 6. März 2025 unter https://www.ibm.com/think/topics/llama-2

Berretta, S., Tausch, A., Ontrup, G., Gilles, B., Peifer, C., & Kluge, A. (2023). Defining Human-AI Teaming the Human-Centered Way: A Scoping Review and Network Analysis. *Frontiers in Artificial Intelligence*, *6*, 1250725. https://doi.org/10.3389/frai.2023.1250725

Bisht, P. (2024). *Large language models for enterprises: Key challenges and advantages*. Verfügbar 26. Februar 2025 unter https://www.kellton.com/kellton-tech-blog/large-language-models-challenges-benefits

Brian, S., Nick, C., Ashraf, E., Joe, D., & Shivam, R. (2024, August). *NVIDIA NVLink and NVIDIA NVSwitch Supercharge Large Language Model Inference* [Accessed: 2025-02-28]. https://developer.nvidia.com/blog/nvidia-nvlink-and-nvidia-nvswitch-supercharge-large-language-model-inference/

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020, 22. Juli). *Language*

Models are Few-Shot Learners. arXiv: 2005.14165 [cs]. https://doi.org/10.48550/arXiv.2005.14165

Brynjolfsson, E., Li, D., & Raymond, L. R. (2023, April). *Generative AI at Work*. 31161. https://doi.org/10.3386/w31161

Budhwar, P., Chowdhury, S., Wood, G., Aguinis, H., Bamber, G. J., Beltran, J. R., Boselie, P., Lee Cooke, F., Decker, S., DeNisi, A., Dey, P. K., Guest, D., Knoblich, A. J., Malik, A., Paauwe, J., Papagiannidis, S., Patel, C., Pereira, V., Ren, S., ... Varma, A. (2023). Human Resource Management in the Age of Generative Artificial Intelligence: Perspectives and Research Directions on ChatGPT. *Human Resource Management Journal*, *33*(3), 606–659. https://doi.org/10.1111/1748-8583.12524

Cappelli, P., Tambe, P. (, & Yakubovich, V. (2024, 4. März). *Will Large Language Models Really Change How Work Is Done?* MIT Sloan Management Review. Verfügbar 2. März 2025 unter https://sloanreview.mit.edu/article/will-large-language-models-really-change-how-work-is-done/

Casado, M., Bornstein, M., & Appenzeller, G. (2023, 27. April). *Navigating the High Cost of AI Compute*. Andreessen Horowitz. Verfügbar 26. Februar 2025 unter https://a16z.com/navigating-the-high-cost-of-ai-compute/

Chen, X., Gao, C., Chen, C., Zhang, G., & Liu, Y. (2025, 18. Februar). *An Empirical Study on Challenges for LLM Application Developers*. arXiv: 2408.05002 [cs]. https://doi.org/10.48550/arXiv.2408.05002

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., Schuh, P., Shi, K., Tsvyashchenko, S., Maynez, J., Rao, A., Barnes, P., Tay, Y., Shazeer, N., Prabhakaran, V., ... Fiedel, N. (2022, 5. Oktober). *PaLM: Scaling Language Modeling with Pathways*. arXiv: 2204.02311 [cs]. https://doi.org/10.48550/arXiv.2204.02311

Convergence, I. T. (2024, 1. Juli). *Generative AI LLMs Deployment Options*. IT Convergence. Verfügbar 28. Februar 2025 unter https://www.itconvergence.com/blog/what-are-the-different-deployment-options-for-generative-ai-llms/

Developers, G. (2025). *Machine Learning Glossary | Google for Developers*. Verfügbar 28. Februar 2025 unter https://developers.google.com/machine-learning/glossary

Dobosevych, Oles. (2025, 3. Februar). *Why local LLMs are the future of enterprise AI*. Geniusee. Verfügbar 28. Februar 2025 unter https://geniusee.com/single-blog/local-llm-models

Dobur, B., Bıçakcı, E., Terim, A., & Arık, C. (2024). Building an On-Premises Knowledge Repository with Large Language Models for Instant Information Access. *Orclever Proceedings of Research and Development*, *5*(1), 261–273. https://doi.org/10.56038/oprd.v5i1.545

DocsBot AI. (2025). *GPT-4o vs Gemini 2.0 Flash - Detailed Performance & Feature Comparison*. DocsBot AI. Verfügbar 2. März 2025 unter https://docsbot.ai/models/compare/gpt-4o/gemini-2-0-flash

Dunenfeld, Emily. (2024). *Claude vs OpenAI: Pricing Considerations*. Verfügbar 6. März 2025 unter https://www.vantage.sh/blog/aws-bedrock-claude-vs-azure-openai-gpt-ai-cost

Edwards, B. (2023, 18. Juli). *Meta launches Llama 2, a source-available AI model that allows commercial applications [Updated]*. Ars Technica. Verfügbar 6. März 2025 unter https://arstechnica.com/

information-technology/2023/07/meta-launches-llama-2-an-open-source-ai-model-that-allows-commercial-applications/

Ekuma, K. (2024). Artificial Intelligence and Automation in Human Resource Development: A Systematic Review. *Human Resource Development Review*, *23*(2), 199–229. https://doi.org/10.1177/15344843231224009

Europäische Union. (2016). *VERORDNUNG (EU) 2016/ 679 DES EUROPÄISCHEN PARLAMENTS UND DES RATES - vom 27. April 2016 - zum Schutz natürlicher Personen bei der Verarbeitung personenbezogener Daten, zum freien Datenverkehr und zur Aufhebung der Richtlinie 95/ 46/ EG (Datenschutz-Grundverordnung)*. https://eur-lex.europa.eu/legal-content/DE/TXT/PDF/?uri=CELEX:32016R0679

European Parliament and Council of the European Union. (2024). *Regulation - EU - 2024/1689 - EN - EUR-Lex*. Verfügbar 8. März 2025 unter https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng
Doc ID: 32024R1689
Doc Sector: 3
Doc Title: Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (Text with EEA relevance)
Doc Type: R
Usr_lan: en.

Eurostat. (2024, 12. Dezember). *Statistics | Eurostat*. Verfügbar 1. März 2025 unter https://ec.europa.eu/eurostat/databrowser/view/isoc_eb_ai__custom_15030348/bookmark/table?lang=en&bookmarkId=13d72414-8c41-43d8-bcad-fdb4b3f12c31

Eurostat. (2025, 23. Januar). *Usage of AI technologies increasing in EU enterprises*. Verfügbar 18. Februar 2025 unter https://ec.europa.eu/eurostat/web/products-eurostat-news/w/ddn-20250123-3

Evidently AI. (2025). *10 RAG examples and use cases from real companies*. Verfügbar 8. März 2025 unter https://www.evidentlyai.com/blog/rag-examples

exxactcorp. (n. d.). *On-premise vs Cloud vs Hybrid Storage | Enterprise Data Storage | Exxact Blog*. Verfügbar 28. Februar 2025 unter https://www.exxactcorp.com/blog/storage/on-premises-vs-cloud-vs-hybrid-storage

Fiza Fatima. (2024). *Open-Source LLM vs Closed-Source LLM: Best for Enterprises?* Verfügbar 1. März 2025 unter https://datasciencedojo.com/blog/open-source-llm/

Gallegos, I. O., Rossi, R. A., Barrow, J., Tanjim, M. M., Kim, S., Dernoncourt, F., Yu, T., Zhang, R., & Ahmed, N. K. (2024, 12. Juli). *Bias and Fairness in Large Language Models: A Survey*. arXiv: 2309.00770 [cs]. https://doi.org/10.48550/arXiv.2309.00770

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, M., & Wang, H. (2024, 27. März). *Retrieval-Augmented Generation for Large Language Models: A Survey*. arXiv: 2312.10997 [cs]. https://doi.org/10.48550/arXiv.2312.10997

Gartner. (2024). *Gartner survey finds generative ai is now the most frequently deployed ai solution in organizations*. Gartner. Verfügbar 18. Februar 2025 unter https://www.gartner.com/en/

newsroom/press-releases/2024-05-07-gartner-survey-finds-generative-ai-is-now-the-most-frequently-deployed-ai-solution-in-organizations

Gartner. (2025a). *Definition of DevOps - Gartner Information Technology Glossary*. Gartner. Verfügbar 8. März 2025 unter https://www.gartner.com/en/information-technology/glossary/devops

Gartner. (2025b). *Definition of Large Language Models (LLMs) - Gartner Information Technology Glossary*. Gartner. Verfügbar 26. Januar 2025 unter https://www.gartner.com/en/information-technology/glossary/large-language-models-llm

Geiger, R. S., Cope, D., Ip, J., Lotosh, M., Shah, A., Weng, J., & Tang, R. (2021). "Garbage in, Garbage out" Revisited: What Do Machine Learning Application Papers Report about Human-Labeled Training Data? *Quantitative Science Studies*, *2*(3), 795–827. https://doi.org/10.1162/qss_a_00144

Gillham, Jonathan. (2024a). *Meta Llama 2: Statistics on Meta AI and Microsoft's Open Source LLM – Originality.AI*. Verfügbar 6. März 2025 unter https://originality.ai/blog/meta-llama-2-statistics

Gillham, Jonathan. (2024b). *Meta Llama 2: Statistics on Meta AI and Microsoft's Open Source LLM – Originality.AI*. Verfügbar 6. März 2025 unter https://originality.ai/blog/meta-llama-2-statistics

Glenn K. Lockwood. (2025, 26. Januar). *LLM training datasets*. Glenn's Digital Garden. Verfügbar 28. Februar 2025 unter https://glennklockwood.com/garden/LLM-training-datasets

Google Cloud. (n. d.). *ISO/IEC 27001 - Compliance*. Google Cloud. Verfügbar 28. Februar 2025 unter https://cloud.google.com/security/compliance/iso-27001

Guan, Y., Wang, D., Chu, Z., Wang, S., Ni, F., Song, R., Li, L., Gu, J., & Zhuang, C. (2023, 4. Dezember). *Intelligent Virtual Assistants with LLM-based Process Automation*. arXiv: 2312.06677 [cs]. https://doi.org/10.48550/arXiv.2312.06677

Haar, M., Sonntagbauer, M., & Kluge, S. (2024). Stellenwert von Natural Language Processing und chatbasierten Generative Language Models. *Medizinische Klinik - Intensivmedizin und Notfallmedizin*, *119*(3), 181–188. https://doi.org/10.1007/s00063-023-01098-5

Hassabis, Demis. (2023, 6. Dezember). *Introducing Gemini: Our largest and most capable AI model*. Google. Verfügbar 6. März 2025 unter https://blog.google/technology/ai/google-gemini-ai/

Hoblitzell, Andrew. (2023). *Google Launches New Multi-Modal Gemini AI Model*. InfoQ. Verfügbar 6. März 2025 unter https://www.infoq.com/news/2023/12/google-launches-gemini/

Horn, J. (2024, 12. September). *LibGuides: Guide to Science Information Resources: Backward and Forward Reference Searching*. Verfügbar 9. Februar 2025 unter https://libguides.fau.edu/science-resources/reference-searching

Hugging Face. (2025a, 26. Februar). *Enterprise Hub - Hugging Face*. Verfügbar 2. März 2025 unter https://huggingface.co/enterprise

Hugging Face. (2025b, 26. Februar). *Hugging Face – The AI Community Building the Future.* Verfügbar 2. März 2025 unter https://huggingface.co/

Hugging Face. (2025c, 26. Februar). *Hugging Face Forums - Hugging Face Community Discussion*. Hugging Face Forums. Verfügbar 2. März 2025 unter https://discuss.huggingface.co/

Hwang, J., Park, J., Park, H., Park, S., & Ok, J. (2025). Retrieval-Augmented Generation with Estimation of Source Reliability. *International Conference on Learning Representations (ICLR)*. https://openreview.net/forum?id=J3xRByRqOz

IBM. (2025, 18. Februar). *What is self-attention? | IBM*. Verfügbar 24. Februar 2025 unter https://www.ibm.com/think/topics/what-is-self-attention

Ipek Ozkaya, Anita Carleton, John E. Robert & Douglas Schmidt. (2023, 2. Oktober). *Application of Large Language Models (LLMs) in Software Engineering: Overblown Hype or Disruptive Change?* Verfügbar 8. März 2025 unter https://insights.sei.cmu.edu/blog/application-of-large-language-models-llms-in-software-engineering-overblown-hype-or-disruptive-change/

ITMAGINATION. (2024). *Closed-source vs Open-source Large Language Models (LLMs) in Enterprise - Which Should You Choose?* Verfügbar 1. März 2025 unter https://www.itmagination.com/blog/closed-source-vs-open-source-large-language-models-llms-in-enterprise-which-should-you-choose

Jiang, F., Qin, C., Yao, K., Fang, C., Zhuang, F., Zhu, H., & Xiong, H. (2024). Enhancing Question Answering for Enterprise Knowledge Bases using Large Language Models. In M. Onizuka, J.-G. Lee, Y. Tong, C. Xiao, Y. Ishikawa, S. Amer-Yahia, H. V. Jagadish & K. Lu (Hrsg.), *Database Systems for Advanced Applications* (S. 273–290). Springer Nature. https://doi.org/10.1007/978-981-97-5562-2_18

Jiang, P., Niu, W., Wang, Q., Yuan, R., & Chen, K. (2024). Understanding Users' Acceptance of Artificial Intelligence Applications: A Literature Review. *Behavioral Sciences*, *14*(8), 671. https://doi.org/10.3390/bs14080671

Jones, J. (2024, 31. Dezember). *Should I Choose an Open Source or Closed Source Large Language Model (LLM)? Advantages of Each*. JLytics. Verfügbar 8. März 2025 unter https://jlytics.com/2024/12/should-i-choose-an-open-source-or-closed-source-large-language-model-llm-advantages-of-each/

Joshi, P. (2024, 11. Juli). *Gamechanger: ChatGPT update introduces browsing tool with full access to internet. No more knowledge cutoff limit to database.* Advanced Ads. Verfügbar 2. März 2025 unter https://wpadvancedads.com/chatgpt-provides-current-data/

k2view. (n. d.). *What is Retrieval-Augmented Generation (RAG)? A Practical Guide*. Verfügbar 4. März 2025 unter https://www.k2view.com/what-is-retrieval-augmented-generation

Kanabar, V., Wong, J., & Vargas, R. V. (2024). *The AI revolution in project management: Elevating productivity with generative AI*. Pearson.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., & Amodei, D. „Scaling Laws for Neural Language Models". https://arxiv.org/abs/2001.08361

Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., & Yih, W.-t. (2020). Dense Passage Retrieval for Open-Domain Question Answering. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 6769–6781. https://doi.org/10.18653/v1/2020.emnlp-main.550

Karpukhin, V., Oğuz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., & Yih, W.-t. (2020, 30. September). *Dense Passage Retrieval for Open-Domain Question Answering*. arXiv: 2004.04906 [cs]. https://doi.org/10.48550/arXiv.2004.04906

Kelbert, J., Siebert, P., & Thorsten, H. (2024, 19. Januar). *Open Source Large Language Models selbst betreiben - Blog des Fraunhofer IESE*. Fraunhofer IESE. Verfügbar 28. Februar 2025 unter https://www.iese.fraunhofer.de/blog/open-source-large-language-models-selbst-betreiben/

Kelsie Anderson. (2024, 8. April). *What is an open-source LLM? Definition and applications*. Telnyx. Verfügbar 2. März 2025 unter https://telnyx.com/resources/what-is-open-source-llm

Kraus, C. (2024, 15. Mai). *How to Protect Your Company Data When Using LLMs*. Krista AI. Verfügbar 26. Februar 2025 unter https://krista.ai/how-to-protect-your-company-data-when-using-llms/

Kruschwitz, U., & Hull, C. (2017). Searching the Enterprise. *Foundations and Trends® in Information Retrieval*, *11*, 1–142. https://doi.org/10.1561/1500000053

Laskin, M. (2023, 4. März). *Misha Laskin website*. Misha Laskin website. Verfügbar 28. Februar 2025 unter https://www.mishalaskin.com/posts/tensor_parallel

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., & Kiela, D. (2021, 12. April). *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks*. arXiv: 2005.11401 [cs]. https://doi.org/10.48550/arXiv.2005.11401

Liu, Y., Cao, J., Liu, C., Ding, K., & Jin, L. (2024, 28. Februar). *Datasets for Large Language Models: A Comprehensive Survey*. arXiv: 2402.18041 [cs]. https://doi.org/10.48550/arXiv.2402.18041

Lockwood, G. K. (2025, 12. Februar). *LLM training*. Glenn's Digital Garden. Verfügbar 28. Februar 2025 unter https://glennklockwood.com/garden/LLM-training

Malec, Melissa. (2025). *Open-Source LLMs vs Closed: Unbiased Guide for Innovative Companies [2025]*. Verfügbar 4. März 2025 unter https://hatchworks.com/blog/gen-ai/open-source-vs-closed-llms-guide/

Marshall, M. (2024, 24. Oktober). *The enterprise verdict on AI models: Why open source will win*. VentureBeat. Verfügbar 1. März 2025 unter https://venturebeat.com/ai/the-enterprise-verdict-on-ai-models-why-open-source-will-win/

Maslej, N., Fattorini, L., Brynjolfsson, E., Etchemendy, J., Ligett, K., Lyons, T., Manyika, J., Ngo, H., Niebles, J. C., Parli, V., Shoham, Y., Wald, R., Clark, J., & Perrault, R. (2023, 5. Oktober). *Artificial Intelligence Index Report 2023*. arXiv: 2310.03715 [cs]. https://doi.org/10.48550/arXiv.2310.03715

Mayring, P., & Fenzl, T. (2019). Qualitative Inhaltsanalyse. In N. Baur & J. Blasius (Hrsg.), *Handbuch Methoden der empirischen Sozialforschung* (S. 633–648). Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-21308-4_42

McKinsey. (2024, 30. Mai). *The State of AI in Early 2024 | McKinsey*. Verfügbar 26. Januar 2025 unter https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai

McKinsey. (2025, 28. Januar). *AI in the Workplace: A Report for 2025 | McKinsey*. Verfügbar 18. Februar 2025 unter https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/superagency-in-the-workplace-empowering-people-to-unlock-ais-full-potential-at-work#/

Merritt, R. (2025, 31. Januar). *What Is Retrieval-Augmented Generation aka RAG?* NVIDIA Blog. Verfügbar 9. Februar 2025 unter https://blogs.nvidia.com/blog/what-is-retrieval-augmented-generation/

Meta. (2024, 6. Dezember). *Meta-Llama/Llama-2-7b · Hugging Face*. Verfügbar 6. März 2025 unter https://huggingface.co/meta-llama/Llama-2-7b

Mialon, G., Dessì, R., Lomeli, M., Nalmpantis, C., Pasunuru, R., Raileanu, R., Rozière, B., Schick, T., Dwivedi-Yu, J., Celikyilmaz, A., Grave, E., LeCun, Y., & Scialom, T. (2023, 15. Februar).

Augmented Language Models: A Survey. arXiv: 2302.07842 [cs]. https://doi.org/10.48550/arXiv.2302.07842

Microsoft. (2024, 18. Dezember). *Data, privacy, and security for Azure OpenAI Service - Azure AI services*. Verfügbar 1. März 2025 unter https://learn.microsoft.com/en-us/legal/cognitive-services/openai/data-privacy

Minaee, S., Mikolov, T., Nikzad, N., Chenaghlu, M., Socher, R., Amatriain, X., & Gao, J. (2024, 20. Februar). *Large Language Models: A Survey*. arXiv: 2402.06196 [cs]. https://doi.org/10.48550/arXiv.2402.06196

Mistral AI Team. (2023). *Mistral 7B | Mistral AI*. Verfügbar 6. März 2025 unter https://mistral.ai/news/announcing-mistral-7b

Mistral AI Team. (2024). *Le Chat | Mistral AI*. Verfügbar 6. März 2025 unter https://mistral.ai/news/le-chat-mistral

Mistral AI Team. (2025, 6. Februar). *The all new le Chat: Your AI assistant for life and work | Mistral AI*. Verfügbar 6. März 2025 unter https://mistral.ai/news/all-new-le-chat

Nahar, N., Kästner, C., Butler, J., Parnin, C., Zimmermann, T., & Bird, C. (2024, 4. Dezember). *Beyond the Comfort Zone: Emerging Solutions to Overcome Challenges in Integrating LLMs into Software Products*. arXiv: 2410.12071 [cs]. https://doi.org/10.48550/arXiv.2410.12071

Naveed, H., Khan, A. U., Qiu, S., Saqib, M., Anwar, S., Usman, M., Akhtar, N., Barnes, N., & Mian, A. (2024, 17. Oktober). *A Comprehensive Overview of Large Language Models*. arXiv: 2307.06435 [cs]. https://doi.org/10.48550/arXiv.2307.06435

Network, S. (2024, 31. Juli). *2024 Comparison of Open-Source Vs Closed-Source LLMs*. Spheron's Blog. Verfügbar 4. März 2025 unter https://blog.spheron.network/choosing-the-right-llm-2024-comparison-of-open-source-vs-closed-source-llms

NVIDIA. (n. d. a). *NVIDIA H100 Tensor Core GPU Datasheet*. NVIDIA. Verfügbar 3. März 2025 unter https://resources.nvidia.com/en-us-tensor-core/nvidia-tensor-core-gpu-datasheet

NVIDIA. (n. d. b). *What are Large Language Models? | NVIDIA Glossary*. NVIDIA. Verfügbar 24. Februar 2025 unter https://www.nvidia.com/en-us/glossary/large-language-models/

OpenAI. (2024, 12. Januar). *GPT-4*. OpenAI. Verfügbar 2. März 2025 unter https://openai.com/index/gpt-4-research/

OpenAI. (2025). *ChatGPT Pricing*. Verfügbar 2. März 2025 unter https://openai.com/chatgpt/pricing/

OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., ... many others. (2023). GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774*. https://cdn.openai.com/papers/gpt-4.pdf

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A Design Science Research Methodology for Information Systems Research. *Journal of Management Information Systems*, *24*(3), 45–77. https://doi.org/10.2753/MIS0742-1222240302

Portakal, Ertugrul. (2024). *How to Access Claude 2? [2024 Guide]*. Verfügbar 6. März 2025 unter https://textcortex.com/post/how-to-access-claude-2

Portes, Jacob, Drozdov, Andrew, Yuen, Erica Ji, Chen, Vincent, Kulinski, Sean, Cress, Milo, Peltier, Colton, Havens, Sam & Carbin, Michael. (2025). *Improving Retrieval and RAG with Embedding*

Model Finetuning. Databricks. Verfügbar 4. März 2025 unter https://www.databricks.com/blog/improving-retrieval-and-rag-embedding-model-finetuning

Russinovich, M. (2023, Mai). *What runs ChatGPT? Inside Microsoft's AI supercomputer* [Accessed: 2025-02-28]. https://techcommunity.microsoft.com/blog/microsoftmechanicsblog/what-runs-chatgpt-inside-microsofts-ai-supercomputer--featuring-mark-russinovich/3830281

Safar, M. (2024, 8. August). *Komplexe Prozesse mit RPA und LLMS effizient automatisieren*. Weissenberg. Verfügbar 3. März 2025 unter https://weissenberg-group.de/komplexe-prozesse-mit-rpa-und-llms-effizient-automatisieren/

Seddik, M. E. A., Chen, S.-W., Hayou, S., Youssef, P., & Debbah, M. (2024, 7. April). *How Bad is Training on Synthetic Data? A Statistical Analysis of Language Model Collapse*. arXiv: 2404.05090 [cs]. https://doi.org/10.48550/arXiv.2404.05090

Shanahan, M. (2023, 16. Februar). *Talking About Large Language Models*. arXiv: 2212.03551 [cs]. https://doi.org/10.48550/arXiv.2212.03551

Shareef, F. (2024). Enhancing Conversational AI with LLMs for Customer Support Automation. *2024 2nd International Conference on Self Sustainable Artificial Intelligence Systems (ICSSAS)*, 239–244. https://doi.org/10.1109/ICSSAS64001.2024.10760403

Simon Zamarin, David Ping, Vikram Elango, Graham Horwood, Vikesh Pandey, Qingwei Li & Vinayak Arannil. (2024, 19. Dezember). *An introduction to preparing your own dataset for LLM training | AWS Machine Learning Blog*. Verfügbar 28. Februar 2025 unter https://aws.amazon.com/blogs/machine-learning/an-introduction-to-preparing-your-own-dataset-for-llm-training/

Soulami, M., Benchekroun, S., & Galiulina, A. (2024). Exploring How AI Adoption in the Workplace Affects Employees: A Bibliometric and Systematic Review. *Frontiers in Artificial Intelligence*, *7*, 1473872. https://doi.org/10.3389/frai.2024.1473872

Technologies, D. (2025). *Infrastructure considerations | Technical White Paper–Generative AI in the Enterprise – Model Training | Dell Technologies Info Hub*. Verfügbar 28. Februar 2025 unter https://infohub.delltechnologies.com/nl-nl/l/technical-white-paper-generative-ai-in-the-enterprise-model-training/infrastructure-considerations-13/

Tim Mucci & Cole Stryker. (2025, 5. April). *Was ist MLOps? | IBM*. Verfügbar 8. März 2025 unter https://www.ibm.com/de-de/topics/mlops

TypingMind. (2025). *Data Usage and Model Training Policy*. Verfügbar 2. März 2025 unter https://docs.typingmind.com/security-and-compliance/data-usage-and-model-training-policy

TypingMind. (n. d. a). *Data Usage and Model Training Policy*. Verfügbar 6. März 2025 unter https://docs.typingmind.com/security-and-compliance/data-usage-and-model-training-policy

TypingMind. (n. d. b). *Data Usage and Model Training Policy*. Verfügbar 6. März 2025 unter https://docs.typingmind.com/security-and-compliance/data-usage-and-model-training-policy

Urlana, A., Kumar, C. V., Singh, A. K., Garlapati, B. M., Chalamala, S. R., & Mishra, R. (2024, 22. Februar). *LLMs with Industrial Lens: Deciphering the Challenges and Prospects – A Survey*. arXiv: 2402.14558 [cs]. https://doi.org/10.48550/arXiv.2402.14558

Uspenskyi, Serhii. (2024). *Open Source vs Closed Source LLMs. Pros and Cons - Springs*. Verfügbar 4. März 2025 unter https://springsapps.com/knowledge/open-source-vs-closed-source-llms-pros-and-cons,%20https://springsapps.ai/blog/open-source-vs-closed-source-llms-pros-and-cons

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention Is All You Need. *CoRR*, *abs/1706.03762*. http://arxiv.org/abs/1706.03762

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2023, 2. August). *Attention Is All You Need*. arXiv: 1706.03762 [cs]. https://doi.org/10.48550/arXiv.1706.03762

Walker, S. M. (2023, 1. September). *Everything We Know About GPT-4 — Klu*. Verfügbar 28. Februar 2025 unter https://klu.ai/blog/gpt-4-llm

Wenzek, G., Lachaux, M.-A., Conneau, A., Chaudhary, V., Guzmán, F., Joulin, A., & Grave, E. (2019, 15. November). *CCNet: Extracting High Quality Monolingual Datasets from Web Crawl Data*. arXiv: 1911.00359 [cs]. https://doi.org/10.48550/arXiv.1911.00359

Xu, F., Hao, Q., Zong, Z., Wang, J., Zhang, Y., Wang, J., Lan, X., Gong, J., Ouyang, T., Meng, F., Shao, C., Yan, Y., Yang, Q., Song, Y., Ren, S., Hu, X., Li, Y., Feng, J., Gao, C., & Li, Y. (2025, 23. Januar). *Towards Large Reasoning Models: A Survey of Reinforced Reasoning with Large Language Models*. arXiv: 2501.09686 [cs]. https://doi.org/10.48550/arXiv.2501.09686

Yeluri, S. (2023, Oktober). *Large language models: The hardware connection* [Accessed: 2025-02-28]. https://community.juniper.net/blogs/sharada-yeluri/2023/10/03/large-language-models-the-hardware-connection

Yu, H., Yang, Z., Pelrine, K., Godbout, J. F., & Rabbany, R. (2023, 19. August). *Open, Closed, or Small Language Models for Text Classification?* arXiv: 2308.10092 [cs]. https://doi.org/10.48550/arXiv.2308.10092

Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X. V., Mihaylov, T., Ott, M., Shleifer, S., Shuster, K., Simig, D., Koura, P. S., Sridhar, A., Wang, T., & Zettlemoyer, L. (2022a, 21. Juni). *OPT: Open Pre-trained Transformer Language Models*. arXiv: 2205.01068 [cs]. https://doi.org/10.48550/arXiv.2205.01068

Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X. V., Mihaylov, T., Ott, M., Shleifer, S., Shuster, K., Simig, D., Koura, P. S., Sridhar, A., Wang, T., & Zettlemoyer, L. (2022b, 21. Juni). *OPT: Open Pre-trained Transformer Language Models*. arXiv: 2205.01068 [cs]. https://doi.org/10.48550/arXiv.2205.01068

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., … Wen, J.-R. (2024, 13. Oktober). *A Survey of Large Language Models*. arXiv: 2303.18223 [cs]. https://doi.org/10.48550/arXiv.2303.18223

Zimmergren. (2024, 1. November). *Establish an AI Center of Excellence - Cloud Adoption Framework*. Verfügbar 8. März 2025 unter https://learn.microsoft.com/en-us/azure/cloud-adoption-framework/scenarios/ai/center-of-excellence

A. Transkripte Expert*inneninterviews

Transkript Anton

Datum: 07.03.2025 **Name:** Anton **Unternehmen:** ca. 3000 Mitarbeiter*innen

Christopher Haas: OK, bist du einverstanden, damit das Gespräch aufgezeichnet wird?
Anton: Sicher.
Christopher Haas: Gibt es Bereiche aus dem Interview, die du gerne anonymisiert haben möchtest?
Anton: Ich möchte mein Unternehmen nicht nennen.
Christopher Haas: Perfekt, gut, dann beginnen wir jetzt offiziell mit dem Experteninterview. Anton, bitte stell dich kurz vor und erkläre, wer du bist.
Anton: Ja, mein Name ist Andreas. Ich arbeite in einem größeren Betrieb in der IT-Infrastruktur als Administrator und bin dort für die IT-Infrastruktur-Teile für Containerisierung und den Softwareentwicklungsprozess zuständig.
Christopher Haas: OK, wie bist du in den Entscheidungsprozess beziehungsweise die Implementierung von KI-Technologien eingebunden?
Anton: Ich bin dahingehend eingebunden, dass ich Ressourcenplanung, Kapazitätsplanung und die Data Center Hardwareplanung durchführe.
Christopher Haas: OK, interessant. Welche Erfahrungen hast du persönlich mit der Integration oder Nutzung von Large Language Models beziehungsweise anderen KI-Lösungen gemacht?
Anton: Ich nutze Large Language Models vor allem als Benutzer, also klassisch ChatGPT. Ich habe aber auch bereits ein chinesisches Modell verwendet. **Christopher Haas:** Sprichst du von DeepSeek?
Anton: Genau, DeepSeek habe ich lokal bei mir ausprobiert und quasi Ressourcen-Performancetests durchgeführt – auf eigener Hardware.
Christopher Haas: Und wie sind die ausgefallen?
Anton: Das ist mit CPU Only unverwendbar, auch wenn die CPUs dementsprechend leistungsstark sind. Sobald man eine oder mehrere GPUs zur Verfügung hat, ist es wesentlich performanter und erreicht an die Geschwindigkeiten, die man in der Cloud in Form von Tokens per Second herankommt.
Christopher Haas: OK, interessant. Und im Unternehmenskontext, welche Erfahrungen hast du da sammeln können?
Anton: Noch keine.
Christopher Haas: Noch keine? Gibt es bei euch im Unternehmen bereits LLM-Lösungen, die ihr verwendet?
Anton: Ja, es gibt welche, aber alles befindet sich noch im Entwicklungsstadium. Wenn wir jetzt darüber reden, dass Services für den Endkunden zur Verfügung gestellt werden sollen – also Service-Chatbots, Dokumentationen und ähnliches – wird noch entwickelt.
Christopher Haas: Mhm, welche Veränderungen oder Trends im Bereich KI beobachtest du in deinem Unternehmen?
Anton: Viele Leute wollen es verwenden. Das Interesse ist sehr hoch, vor allem im Servicebereich im Customer-Segment. Es ergänzt sehr gut, um häufig gestellte Fragen oder Rückfragen auf Informationen möglichst performant und automatisierbar bereitzustellen.
Christopher Haas: Und welche Hardware-Ressourcen (z. B. GPUs, TPUs) stehen in deinem Unternehmen für KI-Anwendungen zur Verfügung?
Anton: Wir haben dedizierte Grafikkartencluster mit Nvidia T4, basierend auf Kubernetes, wo containerbasierte Services (wie Ollama-Instanzen) betrieben werden können. Beim aktuellen On-Premise-Setting verwenden wir Nvidia T4 GPUs.
Christopher Haas: Wie schätzt du die Skalierbarkeit eurer aktuellen Infrastruktur ein, gerade im Hinblick auf mögliche Lastspitzen?
Anton: On-Premise ist sehr statisch. Skalierungsmöglichkeiten gibt es, aber sie erfolgen nicht schnell – zusätzliche GPU-Ressourcen können relativ einfach hinzugefügt werden, müssen aber bestellt und eingebaut werden, was mit einer Durchlaufzeit von Monaten verbunden ist.
Christopher Haas: OK, das ist dann schon beträchtlich.
Anton: Ja, aber in der Cloud würde man wesentlich schneller skalieren können – auch wenn man sich wegen der Kosten noch zögert.
Christopher Haas: Apropos schneller: Wie beurteilst du die Netzwerkgeschwindigkeit und Latenz in eurem Unternehmen? Gibt es schon Optimierungen?
Anton: Für die aktuellen Use Cases ausreichend! Allerdings gilt das nicht, sobald man große Modelle erzeugen möchte oder wenn Inter-GPU- bzw. Interhost-Kommunikation erforderlich ist und das Netzwerk als Bottleneck wirkt.
Christopher Haas: Ausreichend, aber ausbaufähig?
Anton: Ausreichend, aber wenn weiter skaliert wird, muss man frühzeitig über entsprechende Optimierungen nachdenken.
Christopher Haas: Welches Speichersystem nutzt ihr für große Datenmengen und wie skaliert ihr diese?
Anton: Der Speicher wird zentral über eine NetApp Appliance geregelt, die semiautomatisch skaliert. Bei Erreichen bestimmter Schwellenwerte werden weitere Storage-Instanzen hinzugefügt – die Anbindung erfolgt über mehrere 100 Gigabit.
Christopher Haas: Na OK, danke für die Antwort. Welche Datenformate (z. B. JSON) und Datenstrukturen verwendet ihr aktuell?

Anton: Dazu möchte ich mich nicht äußern.
Christopher Haas: Welche automatisierten Prozesse kommen zum Einsatz, um die Datenqualität sicherzustellen?
Anton: Dazu möchte ich mich nicht äußern.
Christopher Haas: Wie geht ihr mit Duplikaterkennung und Bereinigung in großen Datenbeständen um?
Anton: Dazu möchte ich mich nicht äußern.
Christopher Haas: Welche Prozesse sorgen dafür, dass die Daten immer aktuell und vollständig sind?
Anton: Dazu möchte ich mich nicht äußern.
Christopher Haas: Wie organisiert ihr die Archivierung und Versionierung der Daten?
Anton: Dazu möchte ich mich nicht äußern.
Christopher Haas: Kann man abschließend sagen, dass du hinsichtlich Datenbereitstellung und Qualität in deinem Bereich keine ausreichenden Aussagen treffen darfst?
Anton: Genau. Ich kann nur sagen, dass wir die Plattform zur Verfügung stellen und alles Weitere dann die Entwicklungsteams übernehmen.
Christopher Haas: Dann kommen wir jetzt zum Deployment-Modell. Welche Deploymentmodelle (On-Premise, Cloud, Hybrid) nutzt ihr aktuell oder plant ihr, und warum?
Anton: Wir nutzen definitiv alles – On-Premise, Cloud und Hybridmodelle. On-Premise wird genutzt, wenn große Datenmengen im eigenen Data Center liegen und der Aufwand, mehrere Terabyte in die Cloud zu schieben, zu hoch wäre. Cloud nutzen wir für kleine Datenmengen, sofern die Datenschutzgrundverordnung dies zulässt. Strictly Confidential-Daten sollten das eigene Data Center nicht verlassen. Hybridmodelle kommen zum Einsatz, wenn beispielsweise die Daten On-Premise liegen, aber die GPUs in der Cloud oder in einem Azure Blob Store vorhanden sind.
Christopher Haas: Und wie definiert und implementiert ihr eure Datenschutzrichtlinien im Zusammenhang mit der LLM-Bereitstellung?
Anton: Es gibt Datenschutzbeauftragte und Infrastruktur-Teams, die klären, wie und wo die Daten liegen und welche Wege sie nehmen. Allerdings weiß ich nicht, was mit den Informationen letztlich gemacht wird.
Christopher Haas: Welche Kriterien fließen in eure Evaluierung der Anbieterabhängigkeit ein, und wie überprüft ihr diese regelmäßig?
Anton: Dazu möchte ich mich nicht äußern.
Christopher Haas: Wie erfolgt bei euch die langfristige Kosten-Nutzen-Bewertung des gewählten Deployment-Modells?
Anton: Das erfolgt über eine klassische Kostenrechnung. Wir wissen, was unsere On-Premise-Infrastruktur kostet. Zwar ist On-Premise aktuell in der Experimentierphase oft günstiger als Cloud-Umgebungen, da in der Cloud die Kosten pro Stunde oder Minute abgerechnet werden.
Christopher Haas: Kommen wir jetzt zum nächsten Bereich: Modellwahl und Integration. Welche Kriterien, etwa Datensicherheit, Anpassungsfähigkeit und Support, sind für dich ausschlaggebend, wenn es um die Wahl eines Large Language Models geht?
Anton: Support ist extrem wichtig. Wenn bei Problemstellungen nicht geholfen werden kann oder es träge ist, ist das ein großes No-Go. Datenschutz ist besonders wichtig, wenn es um Forschungs- und unternehmenskritische Daten geht.
Christopher Haas: Gut, welche Erfahrungen hast du mit Open-Source-Modellen im Vergleich zu kommerziellen Lösungen gemacht?
Anton: Keine.
Christopher Haas: Wie bewertest du die Flexibilität und den Support der von euch genutzten Modelle?
Anton: Darüber möchte ich keine Aussage treffen.
Christopher Haas: Wie bindet ihr Large Language Model-Anwendungen in eure bestehenden Systeme (z. B. ERP, CRM, DMS) ein?
Anton: Darüber darf ich keine Aussage treffen.
Christopher Haas: OK, dann auch keine Aussage zu den Schnittstellen und API-Gateways, die ihr nutzt, um die Kommunikation zwischen den Systemen sicherzustellen?
Anton: Genau, darüber möchte ich mich nicht äußern.
Christopher Haas: Wie stellt ihr sicher, dass die LLM-Integration mit euren bestehenden Sicherheitslösungen kompatibel ist?
Anton: Das erfolgt in enger Zusammenarbeit mit der IT-Security und den Datenschutzverantwortlichen. Externe Penetrationstests werden beauftragt, um das System sowohl als White-Box als auch als Black-Box zu prüfen.
Christopher Haas: Welche Monitoring- und Logging-Strategien habt ihr etabliert, um den Betrieb zu überwachen?
Anton: Wir nutzen unser Standard-Monitoring und Logging, das den Entwicklungsteams zur freien Nutzung zur Verfügung steht. Es handelt sich um klassisches Applikationsmonitoring, ohne LLM-spezifische Metriken – die Umsetzung liegt letztlich in der Verantwortung der Entwicklungsteams.
Christopher Haas: OK.
Anton: Wir stellen die Möglichkeiten zur Verfügung, aber die Nutzung obliegt den Entwicklungsteams.
Christopher Haas: Kommen wir nun zum Betrieb, zur Wartung, Sicherheit und Evaluierung. Wie organisiert ihr das Ressourcenmanagement und die Skalierung eurer LLM-Anwendungen im laufenden Betrieb?
Anton: Wir setzen auf Kubernetes als Orchestrierungsplattform, wodurch in der Cloud dynamisch skaliert werden kann – abhängig von der Ressourcenauslastung können zusätzliche Computeworker und GPUs hinzugefügt werden. On-Premise ist das schwieriger, daher obliegt es dem Service Manager, Engpässe frühzeitig zu erkennen und entsprechende Maßnahmen zu ergreifen.
Christopher Haas: Welche Strategien zur Modellversionierung und zum Rollback habt ihr implementiert?
Anton: Die Containerwelt und Kubernetes ermöglichen feste Versionen von Container-Images, die mittels Semantic Versioning aktualisiert oder zurückgesetzt werden können.

Christopher Haas: Welche automatisierten Prozesse zur Modellvalidierung und Qualitätskontrolle sind bei euch im Einsatz?
Anton: Dazu möchte ich mich nicht äußern.
Christopher Haas: Wie überprüft und optimiert ihr regelmäßig die Betriebskosten, z. B. mithilfe von Monitoring, Dashboards und KPIs?
Anton: Das erfolgt im Rahmen von FinOps. Die Ressourcenauslastung und Cloud-Kosten werden kontinuierlich überwacht und den Entwicklungsteams berichtet, sodass mögliche Kosten-Spikes frühzeitig erkannt werden.
Christopher Haas: Danke für deine Antwort. Kommen wir jetzt zum Bereich Sicherheit und Compliance: Welche Maßnahmen setzt ihr ein, um die Datensicherheit und den Datenschutz bei der Integration von LLMs zu gewährleisten?
Anton: Klassisch – es gibt keine speziellen LLM-Sicherheitsmechanismen. Es gelten die üblichen Unternehmensrichtlinien, die auch für normale Mitarbeiter Anwendung finden.
Christopher Haas: Wie regelt ihr den Zugriff auf sensible Daten und welche Identity-Management-Systeme nutzt ihr?
Anton: Der Zugriff erfolgt klassisch über rollenbasierte Systeme. Die Daten liegen in den jeweiligen Storage-Systemen, die ihr eigenes Identity Management besitzen, und die Rechtevergabe wird zentral geregelt.
Christopher Haas: Und wie stellt ihr sicher, dass eure Datenverarbeitung den gesetzlichen Vorgaben (z. B. DSGVO) entspricht?
Anton: Das ist primär die Aufgabe des Datenschutzverantwortlichen.
Christopher Haas: Welche Maßnahmen (z. B. Datenmaskierung, Verschlüsselung, regelmäßige Sicherheitsüberprüfungen) setzt ihr ein, um Datenlecks zu vermeiden?
Anton: Zum einen gibt es Firewalls, die ungewöhnliche Datenflüsse melden, und zum anderen verschiedene IT-Security-Tools (wie Sentinel), die das System überwachen. Diese Aufgaben werden vom IT-Security-Team übernommen.
Christopher Haas: Kommen wir nun zur Evaluierung und kontinuierlichen Verbesserung. Welche KPIs und Metriken (z. B. Antwortzeit, Genauigkeit, Kosten pro Anfrage) nutzt ihr, um die Effektivität eurer LLM-Implementierung zu messen?
Anton: Darüber habe ich keine Informationen.
Christopher Haas: Wie gestaltet ihr den iterativen Verbesserungsprozess, etwa durch regelmäßige Reviews, A/B-Tests oder Benchmarking?
Anton: Auch darüber habe ich keine Informationen.
Christopher Haas: Wie dokumentiert und kommuniziert ihr die Umsetzung von Verbesserungsmaßnahmen innerhalb eures Unternehmens?
Anton: Über einen allgemeinen Verbesserungsprozess, der existiert, aber keine spezifischen LLM-Prozesse umfasst.
Christopher Haas: Wie habt ihr eure bestehende IT-Landschaft (ERP, CRM, DMS) analysiert und dokumentiert, um die Integration von LLMs zu erleichtern?
Anton: Darüber darf ich keine Aussage tätigen.
Christopher Haas: Welche unternehmensspezifischen Datenschutz- und Compliance-Richtlinien wurden definiert und in den Integrationsprozess einbezogen?
Anton: Es gibt eine klare Klassifizierung der Daten – ob öffentliche, interne oder vertrauliche Daten – und diese sind mit dem bestehenden Rollen- und Rechtemanagement versehen.
Christopher Haas: Und wie wurden eure bestehenden Sicherheitslösungen (z. B. Zugriffsmanagement, Audit Trails) überprüft und an die Anforderungen der LLM-Integration angepasst?
Anton: Das wurde noch nicht spezifisch an LLM-Anforderungen angepasst; es gibt die Sicherheitslösungen, aber bisher wurde nicht daran gedacht.
Christopher Haas: Wie schätzt du eure internen DevOps-Kapazitäten im Hinblick auf Support und Wartung von LLM-Anwendungen ein?
Anton: Es gibt DevOps-Teams, aber diese sind oft schlecht organisiert und arbeiten ohne übergeordnete Struktur. Die Ressourcen sind zwar vorhanden, aber ineffizient eingesetzt.
Christopher Haas: Gibt es weitere technische oder organisatorische Aspekte, die du als kritisch für die erfolgreiche Integration von LLMs siehst?
Anton: Ja, es ist essenziell, dass das Top-Level-Management die Aktivitäten unterstützt und genügend qualifiziertes Personal vorhanden ist. Zudem sollten Security- und Datenschutz-Teams frühzeitig und lösungsorientiert eingebunden werden.
Christopher Haas: Welche zukünftigen Entwicklungen oder Trends im Bereich LLM findest du besonders relevant – sowohl technologisch als auch in Bezug auf die Datenbereitstellung?
Anton: MLOps ist ein spannendes Thema, vor allem, wie man die Prozesse automatisiert, Modelle entwickelt und sie effizient zum Kunden bringt. Meiner Meinung nach steckt das noch in den Kinderschuhen und wird eine zukunftsverändernde Technologie sein.
Christopher Haas: Welche Empfehlungen würdest du anderen Unternehmen geben, die den Einsatz von Large Language Models planen?
Anton: Man sollte die Mitarbeiter frühzeitig abholen, klare Ziele definieren und Sicherheits- sowie Datenschutzbedenken nicht außer Acht lassen. Zudem muss man sich darüber im Klaren sein, dass diese Technologien hohe Investitionen in Infrastruktur erfordern – und abwägen, ob der Return on Investment den Aufwand rechtfertigt.
Christopher Haas: Gibt es Best Practices aus deinem Unternehmen, die du als besonders effektiv empfindest?
Anton: Noch keine Best Practices.

Transkript Eric

Datum: 07.03.2025 **Name:** Eric **Unternehmen:** ca. 7000 Mitarbeiter*innen

Christopher Haas: „Kannst du kurz deinen beruflichen Hintergrund und deine Rolle im Unternehmen beschreiben?“
Eric: Als ehemaliger Lehrling und in weiterer Folge fertig studierter Dipl.-Ing. in Wirtschaftsinformatik habe ich mich in meinem Berufsleben immer mit neuen Technologien beschäftigt. Als IT-Teamleiter für Cloud und Automation liegen nicht nur die Cloud Services in meinem Team, sondern auch die Implementierung von Automatisierungen, wo in diesem Fall die KI eine große Rolle spielt.
Christopher Haas: „Wie bist du in den Entscheidungsprozess bzw. die Implementierung von KI-Technologien eingebunden?“
Eric: Ein KI-Gremium innerhalb des Konzerns wird derzeit evaluiert und die IT darauf ausgerichtet. Als Solution Architekt bin ich hier eingesetzt, wobei mein primäres Aufgabengebiet im Requirements Engineering sowohl im organisatorischen als auch im technischen Sinne liegt
Christopher Haas: „Welche Erfahrungen hast du persönlich mit der Integration oder Nutzung von LLMs bzw. anderen KI-Lösungen gemacht?“
Eric: Aufgrund meiner kürzlich geschriebenen Masterarbeit zu diesem Thema habe ich erweiterte Erfahrung in der Nutzung sowie in der Implementierung von LLMs. Andere KI-Lösungen habe ich innerhalb der Firma bei der Einführung begleitet.
Christopher Haas: „Gibt es bei euch im Unternehmen bereits LLM- Lösungen, die ihr verwendet?“
Eric: Microsoft Copilot befindet sich in der Pilotphase, und Azure AI Foundry wird genutzt, um LLMs in der Cloud anzumieten. Dies dient sowohl der Erweiterung von Produkten als auch ersten Tests im IT-Servicemanagement.
Christopher Haas: „Welche Veränderungen oder Trends im Bereich KI beobachtest du in deinem Unternehmen?“
Eric: Diverse Entwicklungen von Machine-Learning-Algorithmen sowie die Nutzung von LLMs, um Programme mit menschenähnlichem Verhalten auszustatten.
Christopher Haas: „Welche Hardware-Ressourcen (z. B. GPUs, TPUs) stehen in deinem Unternehmen für KI-Anwendungen zur Verfügung?“
Eric: Derzeit wurde keine spezielle Hardware für KI-Anwendungen gekauft, da die aktuellen Server mit GPUs ausreichen. Die primären Berechnungen der Prototypen finden derzeit in der Cloud statt
Christopher Haas: „Wie schätzt du die Skalierbarkeit eurer aktuellen Infrastruktur ein – gerade im Hinblick auf mögliche Lastspitzen?“
Eric: Aufgrund der Auslagerung der Berechnungen in die Microsoft Azure Cloud schätze ich die Skalierbarkeit als sehr gut ein.
Christopher Haas: „Wie beurteilst du die Netzwerkgeschwindigkeit und Latenz in eurem Unternehmen? Gibt es schon Optimierungen, z. B. durch InfiniBand oder NVLink?“
Eric: Eine direkte Expressroute-Anbindung zu Azure ist vorhanden. Intern wurden bisher keine Netzwerkadaptierungen vorgenommen
Christopher Haas: „Welche Speichersysteme (z. B. Data Lakes) nutzt ihr für große Datenmengen und wie skaliert ihr diese?“
Eric: Azure DataLakes, Azure Vector Store, Azure Blob Container, OnPremise: Netapp S3
Christopher Haas: „Welche Datenformate (z. B. JSONL, TFRecord) und Datenstrukturen verwendet ihr aktuell?“
Eric: Vectore Store (Json based)
Christopher Haas: „Welche automatisierten Prozesse (z. B. ETL-Pipelines mit Apache Airflow) kommen zum Einsatz, um die Datenqualität sicherzustellen?“
Eric: Aktuell noch in evaluierung
Christopher Haas: „Wie geht ihr mit der Duplikaterkennung und -bereinigung in großen Datenbeständen um?“
Eric: Aktuell noch in evaluierung
Christopher Haas: „Welche Prozesse sorgen dafür, dass die Daten immer aktuell und vollständig sind?“
Eric: Aktuell noch in evaluierung
Christopher Haas: „Wie organisiert ihr die Archivierung und Versionierung der Daten?“
Eric: Aktuell noch in evaluierung
Christopher Haas: „Welche Deployment-Modelle (On-Premises, Cloud, Hybrid) nutzt ihr aktuell oder plant ihr für den Einsatz von LLMs, und warum?“
Eric: Hybrid
Christopher Haas: „Wie definiert und implementiert ihr eure Datenschutzrichtlinien im Zusammenhang mit der LLM-Bereitstellung?“
Eric: ISO27001 + DSGVO und Hosting in Europa.
Christopher Haas: „Welche Kriterien fließen in eure Evaluierung der Anbieterabhängigkeit ein und wie überprüft ihr diese regelmäßig?“
Eric: Zertifizierung des Anbieters (ISO27001), Kein AWS wenn möglich, Diverse weitere Kriterien Wie Datensicherheit, Backups, Log Management Wenn gewünscht kann ich hier die Headlines nachliefern unseren Kriterienkatalogs für generelle Freigaben einer APP.
Christopher Haas: „Wie erfolgt bei euch die langfristige Kosten-Nutzen-Bewertung des gewählten Deployment-Modells?“
Eric: Hierzu wird intern keine Bewertung durchgeführt.
Christopher Haas: „Welche Kriterien (z. B. Datensicherheit, Anpassungsfähigkeit, Support) sind für dich ausschlaggebend, wenn es um die Wahl eines LLMs geht?“
Eric: Datensicherheit: Wichtig das die Daten nicht für Trainingszwecke verwendet werden. Anpassungfähigkeit: Meistens verwenden wir RAG Modelle, sprich wir nutzen fertige Modelle und adaptieren diese nicht. Support: Aktuelle keine Kriterien.
Christopher Haas: „Welche Erfahrungen hast du mit Open-Source-Modellen im Vergleich zu kommerziellen LLMs gemacht?“
Eric: Die Qualität und die Performance des LLMs ist meist nicht so gut wie die kommerziellen LLMs. Weiters ist die implementierung

von kommerziellen LLMs leichter da sie fertige APIs und SDKs zur Verfügung stellen
Christopher Haas: „Wie bewertest du die Flexibilität und den Support der von euch genutzten Modelle?“
Eric: Flexibillität ist durch Microsoft Azure gegeben da wir relativ schnell das LLM Model wechseln können.
Christopher Haas: „Wie bindet ihr LLM-Anwendungen in eure bestehenden Systeme (ERP, CRM, DMS) ein?“
Eric: Aktuell nur als Prototyp.
Christopher Haas: „Welche Schnittstellen und API-Gateways nutzt ihr, um die Kommunikation zwischen den Systemen sicherzustellen?“
Eric: Rest API Calls
Christopher Haas: „Wie stellt ihr sicher, dass die LLM-Integration mit euren bestehenden Sicherheitslösungen kompatibel ist?“
Eric: Gutes Architekturschaubild + durchsetzung des Least Privilege Prinzip
Christopher Haas: „Welche Monitoring- und Logging-Strategien habt ihr etabliert, um den Betrieb zu überwachen?“
Eric: Application Insights von Microsoft Azure angebunden in unser zentralles Log Management
Christopher Haas: „Wie organisiert ihr das Ressourcenmanagement und die Skalierung eurer LLM-Anwendungen im laufenden Betrieb?“
Eric: Automatische Skalierung der Azure Platform
Christopher Haas: „Welche Strategien zur Modellversionierung und zum Rollback habt ihr in eurem Unternehmen implementiert?“
Eric: Aktuell noch keine
Christopher Haas: „Welche automatisierten Prozesse zur Modellvalidierung und Qualitätskontrolle sind bei euch im Einsatz?“
Eric: Aktuell noch keine
Christopher Haas: „Wie überprüft und optimiert ihr regelmäßig die Betriebskosten – z. B. mithilfe von Monitoring-Dashboards und KPIs?“
Eric: Azure Cost Management.
Christopher Haas: „Welche Maßnahmen setzt ihr ein, um die Datensicherheit und den Datenschutz bei der Integration von LLMs zu gewährleisten?“
Eric: Grundlegende Maßnahmen die in die ISO27001 fallen sowie die DSGVO
Christopher Haas: „Wie regelt ihr den Zugriff auf sensible Daten, und welche Identity-Management-Systeme nutzt ihr dabei?“
Eric: Primär: Azure AD für die Authentifizierung und Autorisierung, OneIdentity + SAP SuccessFactor im Hintergrund für die Benutzerverwaltung
Christopher Haas: „Wie stellt ihr sicher, dass eure Datenverarbeitung den gesetzlichen Vorgaben (z. B. DSGVO) entspricht?“
Eric: Interner Meldungsprozess an den Datenschutzbeauftragten + Freigabe von Security Team
Christopher Haas: „Welche Maßnahmen (z. B. Datenmaskierung, Verschlüsselung, regelmäßige Sicherheitsüberprüfungen) habt ihr implementiert, um Datenlecks zu vermeiden?“
Eric: Aktuell wird das noch evaluiert.
Christopher Haas: „Welche KPIs und Metriken (z. B. Antwortzeit, Genauigkeit, Kosten pro Anfrage) nutzt ihr, um die Effektivität eurer LLM-Implementierung zu messen?“
Eric: Aktuell noch keine
Christopher Haas: „Wie gestaltet ihr den iterativen Verbesserungsprozess – etwa durch regelmäßige Reviews, A/B-Tests oder Benchmarking?“
Eric: Wird noch evaluiert
Christopher Haas: „Wie dokumentiert und kommuniziert ihr die Umsetzung von Verbesserungsmaßnahmen innerhalb eures Unternehmens?“
Eric: Aktuell noch nicht da es noch im Labor betrieben ist.
Christopher Haas: „Wie habt ihr eure bestehende IT-Landschaft (ERP, CRM, DMS) analysiert und dokumentiert, um die Integration von LLMs zu erleichtern?“
Eric: Datenstrukturen wurden analysiert und ggf. vereinheitlicht um die Vectoren Datenbank für das RAG Modell vorzubereiten.
Christopher Haas: „Welche unternehmensspezifischen Datenschutz- und Compliance-Richtlinien wurden definiert und in den Integrationsprozess einbezogen?“
Eric: Sozusagen das "Grounding"des Microsoft Copilots muss auch bei der Verwendung der KI möglich sein. Sprich der User darf nur mit seinen Zugriffsberechtigungen auf Daten Zugriff haben auf die er aktuell auch zugriff hat. Durch die KI darf kein Datenleck entstehen.
Christopher Haas: „Wie wurden eure bestehenden Sicherheitslösungen (z. B. Zugriffsmanagement, Audit-Trails) überprüft und an die Anforderungen der LLM-Integration angepasst?“
Eric: Aufgrund der Verwendung von Azure AD und einer Single Identity innerhalb des Unternehmens kann das Zugriffsmanagement auch der KI weiter vererbt werden. Logs werden für Admin Tätigkeiten im zentrallen Log Mangement erfasst, werden aufgrund der Prototy Phase noch nicht ausgewertet.
Christopher Haas: „Wie schätzt du eure internen MLOps-/DevOps-Kapazitäten im Hinblick auf Support und Wartung von LLM-Anwendungen ein?“
Eric: Sehr gering. Innerhalb des Unternehmens gibt es nur 4-5 Leute die sich mit diesen Theman beschäftigen. Hier wird gerade aktiv an der Rückendeckung des Managements gearbeitet.

Christopher Haas: „Gibt es weitere technische oder organisatorische Aspekte, die du als kritisch für die erfolgreiche Integration von LLMs siehst?“
Eric: Eine KI Strategie muss seitens der IT sowieso des Unternehmens erstellt werden. Weiters müssen Personen richtig geschult werden sowie eigene KI Spezialisten eingestellt werden die sich mit den Thema gut auskennen.
Christopher Haas: „Welche zukünftigen Entwicklungen oder Trends im Bereich LLMs findest du besonders relevant – sowohl technologisch als auch in Bezug auf die Datenbereitstellung?“
Eric: Das RAG Model war ein Meilenstein das Unternehmen auch ohne speziellen Training, Firmen Daten verwenden können.
Weitere Entwicklungen sehe ich in Agent Systemen die eine Kombination aus RAG und Schnittstellen darstellen wird um Tasks direkt von der KI erledigen zu lassen.
Christopher Haas: „Welche Empfehlungen würdest du anderen Unternehmen geben, die den Einsatz von LLMs planen?“
Eric: KI Personen einstellen, KI Stratgie in der IT sowie im Unternehmen verankern. Security Teams drauf zu schulen und auch in weiterer Folge die KI Nutzung nicht verhindern sondern mit Schulungen und gezielten Einführen an die Personen im Unternehmen bringen.
Christopher Haas: „Gibt es Best Practices aus deinem Unternehmen, die du als besonders effektiv empfindest?“
Eric: Leider gibt es aufgrund des aktuellen Stands innerhalb der Firma noch keine Best Practices.

Transkript Marcus

Datum: 04.03.2025 **Name: Marcus:** **Unternehmen:** ca. 15 Mitarbeiter*innen

Christopher Haas: „Kannst du kurz deinen beruflichen Hintergrund und deine Rolle im Unternehmen beschreiben?“
Marcus: Senior Software Architect, zuständing für die Software-Implementierung sowie Auswahl des Basis Systems (OS, Libraries, GUI Frameworks, . . .), Design, Implementierung und beschaffung der IT Infrastruktur (Netzwerk, Internet, Firewall, Server-Landschaft, NAS, Datenspeicherung, Datensicherung)
Christopher Haas: „Wie bist du in den Entscheidungsprozess bzw. die Implementierung von KI-Technologien eingebunden?“
Marcus: Als Betreiber der IT Infrastruktur und Technologie-Enthusiast sind neue Technologien, welche einen Erleichterung oder Hilfestellung der Arbeit in der Firma bringen können, bin ich immer auf der Suche nach neuen Möglichkeiten, diese Technologien auch auszuprobieren bzw. in den Arbeitsablauf der Firmen-Prozesse zu integrieren.
Christopher Haas: „Welche Erfahrungen hast du persönlich mit der Integration oder Nutzung von LLMs bzw. anderen KI-Lösungen gemacht?“
Marcus: Im Bereich der Software-Entwicklung zur schnellen Erstellung von Code-Snippets, zur genaueren Erklärung von Sachverhalten, sowie zur Umformulierung von Texten.
Christopher Haas: „Gibt es bei euch im Unternehmen bereits LLM- Lösungen, die ihr verwendet?“
Marcus: Im Moment werden Hauptsächlich kommerzielle Anbieter (OpenAI) zur Unterstützung von Text-Formulierungen bzw. Erstellung von Textbausteinen verwendet. Es gibt jedoch im Moment keine In-House LLM Lösung, da im Moment keine benötigte Hardware / Software Infrastruktur vorhanden ist.
Christopher Haas: „Welche Veränderungen oder Trends im Bereich KI beobachtest du in deinem Unternehmen?“ **Marcus:** Der Einsatz von KI zur Unterstützung für das Verfassen von Textbausteinen sowie die erstellung von Dokument-Grundgerüsten ist normal geworden und wird eigentlich nicht mehr als etwas besonderes gesehen.
Der Einsatz von KI in anderen Gebieten (Lokale Prozesse) bedarf jedoch einer viel genaueren Vorbereitung, Festlegung des Einsatz-Gebietes, Use-Case Erstellung, Daten Aufbereitung zu Maschinellen Verarbeitung der Daten. Weiters ist in userem Arbeitsumfeld die Datensicherheit höchste priorität, somit ist eine Cloud-Lösung für uns rein Rechtlich im Moment noch nicht möglich.
Christopher Haas: „Welche Hardware-Ressourcen (z. B. GPUs, TPUs) stehen in deinem Unternehmen für KI-Anwendungen zur Verfügung?“
Marcus: Im Moment werden eingentlich nur Cloud-Lösungen verwendet, dabei aber auch nur hauptsächlich LLM zur Unterstützung der Text-Verfassung, Formulierungen, Dokument-Erstellungen, sowie Code-Snippet Erstellungen. Eine lokale KI Infrastruktur ist im Moment noch nicht vorhanden.
Christopher Haas: „Wie schätzt du die Skalierbarkeit eurer aktuellen Infrastruktur ein – gerade im Hinblick auf mögliche Lastspitzen?“
Marcus: Keine Infrastruktur vorhanden
Christopher Haas: „Wie beurteilst du die Netzwerkgeschwindigkeit und Latenz in eurem Unternehmen? Gibt es schon Optimierungen, z. B. durch InfiniBand oder NVLink?“
Marcus: Im Moment gibt es im Unternehmen eine klassiche Office-IT Infrastruktur mit Gigabit-Ethernet sowie Wifi-7 für mobile Geräte. Es ist kein Performance-Network für Server-Infrastruktur vorhanden.
Christopher Haas: „Welche Speichersysteme (z. B. Data Lakes) nutzt ihr für große Datenmengen und wie skaliert ihr diese?“
Marcus: Im Moment gibt es In-House eine NAS – System Lösung sowie ein automatisches Backup aller Office-Daten bzw. anfallender Firmen-Daten. Weiters werden die Daten verschlüsselt in eine Off-Site Daten-Cloud gespeichert, welche ein österreichischer Cloud-Anbieter zu Verfügung stellt.
Christopher Haas: „Welche Datenformate (z. B. JSONL, TFRecord) und Datenstrukturen verwendet ihr aktuell?“
Marcus: Im Moment wird noch keine automatsiche Datenverarbeitung durchgeführt, es wird jedoch schon an eine maschinelle Datenverarbeitung gedacht, deswegen werden Daten schon in einem lesbaren Format gespeichert (CSV).
Christopher Haas: „Welche automatisierten Prozesse (z. B. ETL-Pipelines mit Apache Airflow) kommen zum Einsatz, um die Datenqualität sicherzustellen?“
Marcus: Keine Automatische Verarbeitung, deswegen auch keine Pipelines
Christopher Haas: „Wie geht ihr mit der Duplikaterkennung und -bereinigung in großen Datenbeständen um?“
Marcus: Durch Automatische Abläufe (Programmierte Tasks) in denen kein manuelles Datenhandling notwendig bzw. auch nicht mehr möglich ist, werden und können Daten nicht mehr dupliziert werden. Die Daten liegen in einem verschlüsseltem Backup, welches nicht mehr geändert werden kann. Auf dieses Backup kann jederzeit jedoch nur lesend zugegriffen werden.
Christopher Haas: „Welche Prozesse sorgen dafür, dass die Daten immer aktuell und vollständig sind?“
Marcus: Wie oben schon beschrieben, die Daten werden automatisch gesichert und gegen Veränderungen geschützt. Diese Funktion wird durch das verwendete Software-System sicher gestellt.
Christopher Haas: „Wie organisiert ihr die Archivierung und Versionierung der Daten?“
Marcus: Das verwendete Sofware-System unterstützt ein Rolling-Backup, d.h. es gibt tägliche Backups bzw. auf Wunsch auch Backups in kürzeren Zeit-Spannen.
Christopher Haas: „Welche Deployment-Modelle (On-Premises, Cloud, Hybrid) nutzt ihr aktuell oder plant ihr für den Einsatz von LLMs, und warum?“
Marcus: Keine KI Infrastuktur

Christopher Haas: „Wie definiert und implementiert ihr eure Datenschutzrichtlinien im Zusammenhang mit der LLM-Bereitstellung?“
Marcus: Keine Verwendung von Firmen Daten in der Cloud, kein Datenabfluss in die Cloud. Volle Zugang bzw. Rechte Management in der Firma, nur Berechtigte können die Daten lesen.
Christopher Haas: „Welche Kriterien fließen in eure Evaluierung der Anbieterabhängigkeit ein und wie überprüft ihr diese regelmäßig?“
Marcus: Verwendung von offenen Systemen, kein Vendor-LockIn, Kontrolle und Risko-Bewertung bei ausfall des Anbieters, sowie Möglichkeiten des Daten-Exports aus den verwendeten Systemen.
Christopher Haas: „Wie erfolgt bei euch die langfristige Kosten-Nutzen-Bewertung des gewählten Deployment-Modells?“
Marcus: Im Moment werden noch keine finanziellen Mittel in eine KI Infrastrutur bzw. KI Lösung investiert, deswegen eauch keine Kosten / Nutzen Rechnung aufgestellt.
Christopher Haas: „Welche Kriterien (z. B. Datensicherheit, Anpassungsfähigkeit, Support) sind für dich ausschlaggebend, wenn es um die Wahl eines LLMs geht?“
Marcus: Natürlich wird auch bei der Auswahl von LLMs eine Risiko-Analyse durchgeführt, um die Abhängigkeit an einen Anbieter möglichst gering zu halten. Offene Systeme bzw. freie Systeme werden hier bevorzugt, wenn es möglich ist. Sollte dies nicht der Fall sein, werden genaue Risko-Analysen durchgeführt.
Christopher Haas: „Welche Erfahrungen hast du mit Open-Source-Modellen im Vergleich zu kommerziellen LLMs gemacht?“
Marcus: Zurzeit wird hauptsächlich die Lösung von OpenAI verwedet, für den Betrieb von anderen Modellen bzw. LLMs steht keine IT-Infrastruktur zu Verfügung.
Christopher Haas: „Wie bewertest du die Flexibilität und den Support der von euch genutzten Modelle?“
Marcus: Noch keine Erfahrungen von Support gemacht.
Christopher Haas: „Wie bindet ihr LLM-Anwendungen in eure bestehenden Systeme (ERP, CRM, DMS) ein?“
Marcus: Kein KI System im Moment in der Firmen-Infrastruktur
Christopher Haas: „Welche Schnittstellen und API-Gateways nutzt ihr, um die Kommunikation zwischen den Systemen sicherzustellen?“
Marcus: Kein KI System im Moment in der Firmen-Infrastruktur, im Moment nur Interaktion mit Systemen über Web-Interface
Christopher Haas: „Wie stellt ihr sicher, dass die LLM-Integration mit euren bestehenden Sicherheitslösungen kompatibel ist?“
Marcus: Sollte es eine KI-Lösung in unserer Firma geben, so wird diese nur Lokal betrieben werden, um den möglichen Daten-Abfluss aus dem Unternehmen so gering wie möglich zu halten. Aufgrund der sensiblen Daten ist eine Cloud-KI im Moment rechtlich nicht umsetzbar.
Christopher Haas: „Welche Monitoring- und Logging-Strategien habt ihr etabliert, um den Betrieb zu überwachen?“
Marcus: Kein KI-System installiert, deswegen auch kein Monitoring.
Christopher Haas: „Wie organisiert ihr das Ressourcenmanagement und die Skalierung eurer LLM-Anwendungen im laufenden Betrieb?“
Marcus: Kein KI-System installiert
Christopher Haas: „Welche Strategien zur Modellversionierung und zum Rollback habt ihr in eurem Unternehmen implementiert?“
Marcus: Kein KI-System installiert
Christopher Haas: „Welche automatisierten Prozesse zur Modellvalidierung und Qualitätskontrolle sind bei euch im Einsatz?“
Marcus: Kein KI-System installiert
Christopher Haas: „Wie überprüft und optimiert ihr regelmäßig die Betriebskosten – z. B. mithilfe von Monitoring-Dashboards und KPIs?“
Marcus: Kein KI-System installiert
Christopher Haas: „Welche Maßnahmen setzt ihr ein, um die Datensicherheit und den Datenschutz bei der Integration von LLMs zu gewährleisten?“
Marcus: Kein KI-System installiert
Christopher Haas: „Wie regelt ihr den Zugriff auf sensible Daten, und welche Identity-Management-Systeme nutzt ihr dabei?“
Marcus: Im Moment eine zentralle Benutzerverwaltung (LDAP), an welche die lokalten Dienste angeschlossen sind.
Christopher Haas: „Wie stellt ihr sicher, dass eure Datenverarbeitung den gesetzlichen Vorgaben (z. B. DSGVO) entspricht?“
Marcus: Keine automatische Datenverabeitung, Kein KI-System installiert
Christopher Haas: „Welche Maßnahmen (z. B. Datenmaskierung, Verschlüsselung, regelmäßige Sicherheitsüberprüfungen) habt ihr implementiert, um Datenlecks zu vermeiden?“
Marcus: Alle Daten sind nur über ein Rechte-Management System einsehbar und nur jene Personen , welche den Zugang zu den Daten benötigen, erhalten den Zugang.
Christopher Haas: „Welche KPIs und Metriken (z. B. Antwortzeit, Genauigkeit, Kosten pro Anfrage) nutzt ihr, um die Effektivität eurer LLM-Implementierung zu messen?“
Marcus: Kein KI-System installiert
Christopher Haas: „Wie gestaltet ihr den iterativen Verbesserungsprozess – etwa durch regelmäßige Reviews, A/B-Tests oder Benchmarking?“
Marcus: Kein KI-System installiert
Christopher Haas: „Wie dokumentiert und kommuniziert ihr die Umsetzung von Verbesserungsmaßnahmen innerhalb eures Unter-

nehmens?“
Marcus: Kein KI-System installiert, Erfassung von Unternehmens-Kennzahlten und Re-Evaluierung der Werte in regelmäsigen Abständen. (ISO9001 Zertifizierung)
Christopher Haas: „Wie habt ihr eure bestehende IT-Landschaft (ERP, CRM, DMS) analysiert und dokumentiert, um die Integration von LLMs zu erleichtern?“
Marcus: Keine Anpassungen im Moment für die Installation bzw. Einführung von KI-Systemen in das Unternehmen geplant.
Christopher Haas: „Welche unternehmensspezifischen Datenschutz- und Compliance-Richtlinien wurden definiert und in den Integrationsprozess einbezogen?“
Marcus: Keine Anpassungen im Moment für die Installation bzw. Einführung von KI-Systemen in das Unternehmen geplant.
Christopher Haas: „Wie wurden eure bestehenden Sicherheitslösungen (z. B. Zugriffsmanagement, Audit-Trails) überprüft und an die Anforderungen der LLM-Integration angepasst?“
Marcus: Keine Anpassungen im Moment für die Installation bzw. Einführung von KI-Systemen in das Unternehmen geplant.
Christopher Haas: „Wie schätzt du eure internen MLOps-/DevOps-Kapazitäten im Hinblick auf Support und Wartung von LLM-Anwendungen ein?“
Marcus: Keine Anpassungen im Moment für die Installation bzw. Einführung von KI-Systemen in das Unternehmen geplant. Für die Integration von KI-Lösungen wird sicher ein Aufbau von Kern-Kompetenz in diesem Bereich von Nöten sein. Im Moment sind nur sehr Basic-Kompetenzen vorhanden.
Christopher Haas: „Gibt es weitere technische oder organisatorische Aspekte, die du als kritisch für die erfolgreiche Integration von LLMs siehst?“
Marcus: Für die Integration von LLMs in die Firmen-Prozesse bzw. in die Firmen-Infrastruktur ist eine genaue Anforderungsanalyse Notwendig. Was soll das System erreichen, welche Anforderungen an die zu bereitstellenden Daten sind von Nöten, was soll das System an Output liefern, kann das System diese Outputs überhaupt liefern, stimmt die Kosten-Nutzen Rechnung. Im Moment ist der Anwendungsfall bzw. diese Analyse in unserem Unternehmen noch nicht vorhanden.
Christopher Haas: „Welche zukünftigen Entwicklungen oder Trends im Bereich LLMs findest du besonders relevant – sowohl technologisch als auch in Bezug auf die Datenbereitstellung?“
Marcus: Im Hinblick auf KI-Systeme bzw. LLMs und den Einsatz solcher Systeme ist schon bei der Erstellung von Daten auf eine möglichst gute maschinelle Datenverarbeitung zu Achten. Weiters ist auch der Punkt Datensicherheit sowie Datenschutz ein Thema, welches in diese Thematik fällt.
Christopher Haas: „Welche Empfehlungen würdest du anderen Unternehmen geben, die den Einsatz von LLMs planen?“
Marcus: Erstellung einer möglichst detailierten Beschreibung, was wird von solch einem System erwartet, können solche Systeme diese Anforderungen überhaupt erfüllen ? Wie schaut es im Bezug auf Daten, Daten-Bereitstellung sowie Datenschutz aus ? Wie ist die Kosten-Nutzen Rechung, Lohnt sich der finanzielle Mehraufwand ?
Christopher Haas: „Gibt es Best Practices aus deinem Unternehmen, die du als besonders effektiv empfindest?“
Marcus: Eine grundsätzliche Firmen-Entscheidung für offene und zukunftssichere Datenformate ist in allen Fällen sinnvoll, da offene Datenformate eine Integration in jedweigen Software-Systemen vereinfachen und somit Kosten-Effizient sind. Eine nachträgliche Daten Konvertierung ist nicht nur Zeit und Kostenaufwending, sondern auch Fehleranfällig.

Transkript Mia

Datum: 10.03.2025 **Name:** Mia **Unternehmen:** ca. 3200 Mitarbeiter*innen

Christopher Haas: „Kannst du kurz deinen beruflichen Hintergrund und deine Rolle im Unternehmen beschreiben?“
Mia: Mitarbeiterin im IT-Management
Christopher Haas: „Wie bist du in den Entscheidungsprozess bzw. die Implementierung von KI-Technologien eingebunden?“
Mia: Ich bin unteranderem die Projektleitung des KI-Zirkels, um Use-Cases und Potenziale für den Einsatz von KI festzustellen und für die Analyse mit den Fachbereichen zuständig und begleite diese.
Christopher Haas: „Welche Erfahrungen hast du persönlich mit der Integration oder Nutzung von LLMs bzw. anderen KI-Lösungen gemacht?“
Mia: Bisher haben wir erst ein Tool eingeführt, welches zwar kommuniziert wurde,jedoch nicht überall bekannt war. Insgesamt sind es durchwachsene Erfahrungen, da man das Tool zwar eingeführt hat, aber die Schulungen nicht für den Arbeitsalltag geeignet sind. Deswegen wurde das Tool nicht umfänglich genutzt.
Christopher Haas: „Gibt es bei euch im Unternehmen bereits LLM- Lösungen, die ihr verwendet?“
Mia: Bisher haben wir nur BingChat for Enterprise/CoPilot in Edge.
Christopher Haas: „Welche Veränderungen oder Trends im Bereich KI beobachtest du in deinem Unternehmen?“
Mia: Es ist ein sehr gegenwärtiges Thema, da die Nachfrage der Fachbereiche in Bezug auf KI steigt. Meistens hören sie von etwas oder nutzen im Privatbereich schon einzelne Tools, welche sie im Arbeitsalltag nicht missen möchten. Besonders CoPilot für M365 ist ein sehr aufkommendes Thema, sowie Predictive Maintenance
Christopher Haas: „Welche Hardware-Ressourcen (z. B. GPUs, TPUs) stehen in deinem Unternehmen für KI-Anwendungen zur Verfügung?“
Mia: keine Aussage kann getroffen werden
Christopher Haas: „Wie schätzt du die Skalierbarkeit eurer aktuellen Infrastruktur ein – gerade im Hinblick auf mögliche Lastspitzen?“
Mia: keine Aussage kann getroffen werden
Christopher Haas: „Wie beurteilst du die Netzwerkgeschwindigkeit und Latenz in eurem Unternehmen? Gibt es schon Optimierungen, z. B. durch InfiniBand oder NVLink?“
Mia: keine Aussage kann getroffen werden
Christopher Haas: „Welche Speichersysteme (z. B. Data Lakes) nutzt ihr für große Datenmengen und wie skaliert ihr diese?“
Mia: keine Aussage kann getroffen werden
Christopher Haas: „Welche Datenformate (z. B. JSONL, TFRecord) und Datenstrukturen verwendet ihr aktuell?“
Mia: keine Aussage kann getroffen werden
Christopher Haas: „Welche automatisierten Prozesse (z. B. ETL-Pipelines mit Apache Airflow) kommen zum Einsatz, um die Datenqualität sicherzustellen?“
Mia: keine Aussage kann getroffen werden
Christopher Haas: „Wie geht ihr mit der Duplikaterkennung und -bereinigung in großen Datenbeständen um?“
Mia: Derzeit wird noch an diesem Punkt gearbeitet bzw. wird in neuen Systemen darauf vermehrt geachtet.
Christopher Haas: „Welche Prozesse sorgen dafür, dass die Daten immer aktuell und vollständig sind?“
Mia: Derzeit ist kein Prozess hierfür implementiert, jedoch wird daran gerade gearbeitet.
Christopher Haas: „Wie organisiert ihr die Archivierung und Versionierung der Daten?“
Mia: keine Aussage kann getroffen werden
Christopher Haas: „Welche Deployment-Modelle (On-Premises, Cloud, Hybrid) nutzt ihr aktuell oder plant ihr für den Einsatz von LLMs, und warum?“
Mia: Der Datenschutz hat bei uns besondere Bedeutung. Wir beziehen bzw. versuchen bei der Auswahl von LLM-Tools Varianten zu wählen, die on premises funktionieren
Christopher Haas: „Wie definiert und implementiert ihr eure Datenschutzrichtlinien im Zusammenhang mit der LLM-Bereitstellung?“
Mia: Hier sind wir noch in der Definition.
Christopher Haas: „Welche Kriterien fließen in eure Evaluierung der Anbieterabhängigkeit ein und wie überprüft ihr diese regelmäßig?“
Mia: Hier sind wir noch in der Definition
Christopher Haas: „Wie erfolgt bei euch die langfristige Kosten-Nutzen-Bewertung des gewählten Deployment-Modells?“
Mia: keine Aussage kann getroffen werden
Christopher Haas: „Welche Kriterien (z. B. Datensicherheit, Anpassungsfähigkeit, Support) sind für dich ausschlaggebend, wenn es um die Wahl eines LLMs geht?“
Mia: Datensicherheit, Support, Leistungsfähigkeit, Datenbasis, Integration, Dokumentation
Christopher Haas: „Welche Erfahrungen hast du mit Open-Source-Modellen im Vergleich zu kommerziellen LLMs gemacht?“
Mia: keine Aussage kann getroffen werden
Christopher Haas: „Wie bewertest du die Flexibilität und den Support der von euch genutzten Modelle?“
Mia: Es passt für die ersten Schritte in Richtung KI.
Christopher Haas: „Wie bindet ihr LLM-Anwendungen in eure bestehenden Systeme (ERP, CRM, DMS) ein?“
Mia: Derzeit noch gar nicht

Christopher Haas: „Welche Schnittstellen und API-Gateways nutzt ihr, um die Kommunikation zwischen den Systemen sicherzustellen?“
Mia: derzeit noch keine
Christopher Haas: „Wie stellt ihr sicher, dass die LLM-Integration mit euren bestehenden Sicherheitslösungen kompatibel ist?“
Mia: keine Aussage kann getroffen werden
Christopher Haas: „Welche Monitoring- und Logging-Strategien habt ihr etabliert, um den Betrieb zu überwachen?“
Mia: Keine Aussage kann getroffen werden
Christopher Haas: „Wie organisiert ihr das Ressourcenmanagement und die Skalierung eurer LLM-Anwendungen im laufenden Betrieb?“
Mia: keine Aussage kann getroffen werden
Christopher Haas: „Welche Strategien zur Modellversionierung und zum Rollback habt ihr in eurem Unternehmen implementiert?“
Mia: keine Aussage kann getroffen werden
Christopher Haas: „Welche automatisierten Prozesse zur Modellvalidierung und Qualitätskontrolle sind bei euch im Einsatz?“
Mia: keine Aussage kann getroffen werden
Christopher Haas: „Wie überprüft und optimiert ihr regelmäßig die Betriebskosten – z. B. mithilfe von Monitoring-Dashboards und KPIs?“
Mia: keine Aussage kann getroffen werden
Christopher Haas: „Welche Maßnahmen setzt ihr ein, um die Datensicherheit und den Datenschutz bei der Integration von LLMs zu gewährleisten?“
Mia: Sind derzeit im Aufbau der Maßnahmen. Ein großes Thema ist ein technisches und organisatorisches Berechtigungskonzept sowie Datenklassifizierung
Christopher Haas: „Wie regelt ihr den Zugriff auf sensible Daten, und welche Identity-Management-Systeme nutzt ihr dabei?“
Mia: keine Aussage kann getroffen werden
Christopher Haas: „Wie stellt ihr sicher, dass eure Datenverarbeitung den gesetzlichen Vorgaben (z. B. DSGVO) entspricht?“
Mia: regelmäßige Überprüfungen
Christopher Haas: „Welche Maßnahmen (z. B. Datenmaskierung, Verschlüsselung, regelmäßige Sicherheitsüberprüfungen) habt ihr implementiert, um Datenlecks zu vermeiden?“
Mia: keine Aussage kann getroffen werden
Christopher Haas: „Welche KPIs und Metriken (z. B. Antwortzeit, Genauigkeit, Kosten pro Anfrage) nutzt ihr, um die Effektivität eurer LLM-Implementierung zu messen?“
Mia: derzeit keine
Christopher Haas: „Wie gestaltet ihr den iterativen Verbesserungsprozess – etwa durch regelmäßige Reviews, A/B-Tests oder Benchmarking?“
Mia: Noch nicht vorhanden
Christopher Haas: „Wie dokumentiert und kommuniziert ihr die Umsetzung von Verbesserungsmaßnahmen innerhalb eures Unternehmens?“
Mia: Innerhalb des Projektteams werden Verbesserungsmaßnahmen in einer Datei dokumentiert und regelmäßig kommuniziert.
Christopher Haas: „Wie habt ihr eure bestehende IT-Landschaft (ERP, CRM, DMS) analysiert und dokumentiert, um die Integration von LLMs zu erleichtern?“
Mia: Dokumentation durch einen Servicekatalog sowie einer visuellen Übersicht der IT-Landschaft, jedoch gab es noch keine korrekte Analyse zur Integration von LLMs
Christopher Haas: „Welche unternehmensspezifischen Datenschutz- und Compliance-Richtlinien wurden definiert und in den Integrationsprozess einbezogen?“
Mia: keine Aussage kann getroffen werden
Christopher Haas: „Wie wurden eure bestehenden Sicherheitslösungen (z. B. Zugriffsmanagement, Audit-Trails) überprüft und an die Anforderungen der LLM-Integration angepasst?“
Mia: keine Aussage kann getroffen werden
Christopher Haas: „Wie schätzt du eure internen MLOps-/DevOps-Kapazitäten im Hinblick auf Support und Wartung von LLM-Anwendungen ein?“
Mia: keine Aussage kann getroffen werden
Christopher Haas: „Gibt es weitere technische oder organisatorische Aspekte, die du als kritisch für die erfolgreiche Integration von LLMs siehst?“
Mia: Bias und Ethik, Angriffsszenarien, Modellanpassung
Einführung von KI bedarf Changemanagement, Verantwortung, Schulung der Mitarbeiter, fehlende Data Governance und Datenmanagement
Christopher Haas: „Welche zukünftigen Entwicklungen oder Trends im Bereich LLMs findest du besonders relevant – sowohl technologisch als auch in Bezug auf die Datenbereitstellung?“
Mia: Agentenbasierte LLM
Christopher Haas: „Welche Empfehlungen würdest du anderen Unternehmen geben, die den Einsatz von LLMs planen?“

Mia: Erhebung des Bedarfes der Mitarbeiter sowie schrittweise Heranführung an das Thema
Christopher Haas: „Gibt es Best Practices aus deinem Unternehmen, die du als besonders effektiv empfindest?“
Mia: Workshop mit den Fachbereichen, um Use-Cases zu erheben

Transkript Ralf

Datum: 07.03.2025 **Name:** Ralf **Unternehmen:** ca. 450 Mitarbeiter*innen

Christopher Haas: „Kannst du kurz deinen beruflichen Hintergrund und deine Rolle im Unternehmen beschreiben?“ **Ralf:** Ich arbeite als Software Development & Operations Engineer in einem österreichischen E-Commerce-Unternehmen mit etwa 300 Mitarbeitern. Dabei entwickle ich Softwarelösungen für Online-Shops und Fulfillment-Prozesse als Teil eines agilen Softwareentwicklungsteams (7 Mitarbeiter) innerhalb der Unternehmens-IT (40 Mitarbeiter).
Christopher Haas: „Wie bist du in den Entscheidungsprozess bzw. die Implementierung von KI-Technologien eingebunden?“
Ralf: Ich bin an der Entwicklung übergeordneter strategischer Ziele für die Teams innerhalb der IT beteiligt. Zusätzlich treffe ich gemeinsam mit anderen Teammitgliedern unmittelbare Technologie- und Implementierungsentscheidungen – auch in Bezug auf KI-Technologien – innerhalb meines Teams.
Christopher Haas: „Welche Erfahrungen hast du persönlich mit der Integration oder Nutzung von LLMs bzw. anderen KI-Lösungen gemacht?“
Ralf: Ich habe sehr gute Erfahrungen mit LLMs und KI-gestützten Tools gemacht. Besonders für kreative Prozesse, Konzeptentwicklung und Planung finde ich ChatGPT sowie vergleichbare Chatbots hilfreich. Auch KI-gestützte Coding-Assistenten wie GitHub Copilot halte ich für äußerst produktivitätssteigernd bei Coding-Aufgaben.
Christopher Haas: „Gibt es bei euch im Unternehmen bereits LLM- Lösungen, die ihr verwendet?“
Ralf: Ja, wir setzen sowohl intern als auch extern verschiedene LLM-Lösungen ein. Google Gemini (ehemals Bard) ist als Teil von Google Workspace für alle Mitarbeiter verfügbar. Ein ChatGPT-Team-Plan steht ausgewählten Nutzern aus verschiedenen Unternehmensbereichen zur Verfügung, um das Tool im Vergleich zu Gemini zu evaluieren (ich bin ebenfalls Teil dieses Programms). Für Entwickler ist der JetBrains AI Assistant als Coding-Assistent in der PhpStorm-Lizenz enthalten und kann genutzt werden. Ein spezialisierter KI-Assistent dient als First-Level-Support für spezifische Anfragen von Fulfillment-Kunden. Dabei wurde GPT-4o-mini über die OpenAI-API in unser UI eingebunden und kann über Funktionsaufrufe mit OpenAPI-Spezifikation mit unserem System kommunizieren, um relevante Daten abzurufen.
Christopher Haas: „Welche Veränderungen oder Trends im Bereich KI beobachtest du in deinem Unternehmen?“
Ralf: Ich nehme ein zunehmendes Interesse an der Nutzung von KI-Tools wahr. Gleichzeitig gibt es jedoch auch Kritik daran, dass das Unternehmen (noch) nicht alle gewünschten KI-Tools zentral zur Verfügung stellt. Vereinzelt existieren Shadow-IT-Lösungen, bei denen Mitarbeiter eigenständig KI-Tools wie ChatGPT oder GitHub Copilot nutzen, ohne dass diese offiziell vom Unternehmen bereitgestellt werden. Das Unternehmen versucht aktiv, diesem Trend entgegenzuwirken – beispielsweise durch die Einführung einer KI-CoP (Community of Practice) sowie durch Evaluierungsphasen für neue KI-Tools, bevor sie unternehmensweit ausgerollt werden (wie aktuell bei ChatGPT).
Christopher Haas: „Welche Hardware-Ressourcen (z. B. GPUs, TPUs) stehen in deinem Unternehmen für KI-Anwendungen zur Verfügung?“
Ralf: In unserer Private Cloud bei Hetzner sind keine speziell für KI-Anwendungen angemieteten Ressourcen vorhanden. Allerdings stehen bei Bedarf flexible Ressourcen über die Public Cloud bei Google zur Verfügung – einschließlich GPUs und TPUs für rechenintensive KI-Prozesse.
Christopher Haas: „Wie schätzt du die Skalierbarkeit eurer aktuellen Infrastruktur ein – gerade im Hinblick auf mögliche Lastspitzen?“
Ralf: Für Online-Shops gibt es in der Private Cloud keine dynamische Skalierung, sie ist jedoch auf Zuruf möglich. Für Fulfillment-Lösungen in der Public Cloud setzen wir Kubernetes mit Horizontal Pod Autoscaling (HPA) ein. Das System skaliert bei höherer Last automatisch anhand vordefinierter Metriken je nach Service – beispielsweise basierend auf der Anzahl der Hintergrundjobs, der CPU-Auslastung oder dem RAM-Verbrauch.
Christopher Haas: „Wie beurteilst du die Netzwerkgeschwindigkeit und Latenz in eurem Unternehmen? Gibt es schon Optimierungen, z. B. durch InfiniBand oder NVLink?“
Ralf: Es gibt keine Probleme mit der Netzwerkgeschwindigkeit oder Latenzen. Auch die Kommunikation mit Cloud-Services funktioniert zuverlässig und performant. Optimierungen sind unsererseits nicht erforderlich.
Christopher Haas: „Welche Speichersysteme (z. B. Data Lakes) nutzt ihr für große Datenmengen und wie skaliert ihr diese?“
Ralf: Geschäftsdaten aus dem operativen Betrieb werden in SQL-Datenbanken gespeichert. Für die Analyse großer Datenmengen nutzen wir Google BigQuery in Kombination mit Google Cloud Storage, wodurch eine hohe Skalierbarkeit sichergestellt ist.
Christopher Haas: „Welche Datenformate (z. B. JSONL, TFRecord) und Datenstrukturen verwendet ihr aktuell?“
Ralf: Wir verwenden verschiedene Datenformate. Sowohl strukturierte Formate wie XML und JSON, als auch unstrukturierte Daten.
Christopher Haas: „Welche automatisierten Prozesse (z. B. ETL-Pipelines mit Apache Airflow) kommen zum Einsatz, um die Datenqualität sicherzustellen?“
Ralf: Es existieren keine expliziten Prozesse zur Kontrolle der Datenqualität.
Christopher Haas: „Wie geht ihr mit der Duplikaterkennung und -bereinigung in großen Datenbeständen um?“
Ralf: Daten, die aus SQL-Datenbanken synchronisiert werden, sind durch referenzielle Integrität abgesichert. Für andere Datenbestände existiert keine spezielle Lösung zur Duplikaterkennung oder -bereinigung.
Christopher Haas: „Welche Prozesse sorgen dafür, dass die Daten immer aktuell und vollständig sind?“
Ralf: Die Synchronisierung erfolgt je nach Art der Daten in regelmäßigen Intervallen (stündlich, täglich, wöchentlich oder monatlich). Wenn unmittelbare Aktualisierungen erforderlich sind, setzen wir Messaging-Systeme wie Google Pub/Sub ein.

Christopher Haas: „Wie organisiert ihr die Archivierung und Versionierung der Daten?“
Ralf: Regelmäßige automatische Backups in der Cloud sind vorhanden. Eine spezielle Archivierung oder Versionierung von Daten über die Backups hinaus erfolgt nicht.
Christopher Haas: „Welche Deployment-Modelle (On-Premises, Cloud, Hybrid) nutzt ihr aktuell oder plant ihr für den Einsatz von LLMs, und warum?“
Ralf: Alles läuft in der Cloud, da das für kleinere Unternehmen mit geringerem Bedarf kosteneffizienter ist. LLMs für Kunden sind derzeit über die OpenAI-API implementiert. Intern genutzte LLM-Lösungen stammen ausschließlich von Drittanbietern. Alle Tools unterliegen NDAs und Opt-Out-Klauseln, um eine Nutzung unserer Daten zu Trainingszwecken auszuschließen. Die Verträge werden von unserer Legal Abteilung geprüft.
Christopher Haas: „Wie definiert und implementiert ihr eure Datenschutzrichtlinien im Zusammenhang mit der LLM-Bereitstellung?“
Ralf: Die Datenschutzrichtlinien für LLMs werden analog zu anderen Drittanbieter-Softwarelösungen gehandhabt, die Unternehmensdaten speichern oder verarbeiten (z. B. Google Workspace und Google Cloud). Vertragsregelungen stellen sicher, dass Daten weder an Dritte weitergegeben noch für Trainingszwecke genutzt werden dürfen.
Christopher Haas: „Welche Kriterien fließen in eure Evaluierung der Anbieterabhängigkeit ein und wie überprüft ihr diese regelmäßig?“
Ralf: Bei neuen Lösungen werden alle möglichen Optionen geprüft. Bestehende Lösungen werden regelmäßig auf Basis von Kosten und Qualität reevaluiert.
Christopher Haas: „Wie erfolgt bei euch die langfristige Kosten-Nutzen-Bewertung des gewählten Deployment-Modells?“
Ralf: Einfache Make-or-Buy-Analyse mit Berechnung der laufenden und langfristigen Kosten, um den Betrieb von KI-Modellen in verschiedenen Szenarien zu vergleichen.
Christopher Haas: „Welche Kriterien (z. B. Datensicherheit, Anpassungsfähigkeit, Support) sind für dich ausschlaggebend, wenn es um die Wahl eines LLMs geht?“
Ralf: Der Anbieter muss vertrauenswürdig sein und eine hohe Qualität der generierten Inhalte gewährleisten. Auch die Kosteneffizienz spielt eine entscheidende Rolle. Idealerweise erfolgt die Abrechnung nach einem Pay-per-Use-Modell, um eine flexible Skalierung zu ermöglichen. Unternehmensdaten dürfen nicht für Trainingszwecke genutzt werden, weshalb wir bei der Auswahl auf entsprechende vertragliche Regelungen achten, die durch unsere Legal Abteilung geprüft werden.
Christopher Haas: „Welche Erfahrungen hast du mit Open-Source-Modellen im Vergleich zu kommerziellen LLMs gemacht?“
Ralf: Meine Erfahrungen mit Open-Source-Modellen wie LLaMA und DeepSeek im Vergleich zu kommerziellen Lösungen waren bisher eher durchwachsen. Open-Source-Modelle sind stark von ihrer Umgebung und den verfügbaren Ressourcen abhängig. LLMs skalieren sehr gut vertikal – eine höhere Rechenleistung führt in der Regel zu besseren Ergebnissen. Ohne die richtige Infrastruktur und gezieltes Fine-Tuning ergibt die Anwendung von LLMs meiner Meinung nach wenig Sinn.
Christopher Haas: „Wie bewertest du die Flexibilität und den Support der von euch genutzten Modelle?“
Ralf: Einen Support für die von uns genutzten Modelle haben wir bisher nicht in Anspruch nehmen müssen. OpenAI z.B. bietet eine umfangreiche und gut strukturierte Dokumentation, die es uns als Entwicklern erleichtert, die Modelle via API zu integrieren.
Christopher Haas: „Wie bindet ihr LLM-Anwendungen in eure bestehenden Systeme (ERP, CRM, DMS) ein?“
Ralf: Unsere LLM-Anwendungen sind derzeit nur teilweise in bestehende Unternehmenssysteme integriert. Der für Kunden verfügbare KI-Assistent, der auf OpenAI GPT-4o-mini basiert, kann beispielsweise Daten zu Bestellungen, Anlieferungen und Produkten abrufen und verarbeiten. Dadurch wird der Support für Fulfillment-Kunden effizienter gestaltet. Interne Tools sind hingegen bislang nicht direkt an unsere Systeme angebunden und werden derzeit ausschließlich als Standalone-Lösungen genutzt.
Christopher Haas: „Welche Schnittstellen und API-Gateways nutzt ihr, um die Kommunikation zwischen den Systemen sicherzustellen?“
Ralf: Die Kommunikation zwischen den Systemen erfolgt über standardisierte RESTful-HTTP-APIs. Wir nutzen OpenAPI-Spezifikationen, um diese einfach in OpenAI zu integrieren.
Christopher Haas: „Wie stellt ihr sicher, dass die LLM-Integration mit euren bestehenden Sicherheitslösungen kompatibel ist?“
Ralf: Wir setzen auf moderne Sicherheitsstandards und halten unsere Systeme regelmäßig auf dem neuesten Stand.
Christopher Haas: „Welche Monitoring- und Logging-Strategien habt ihr etabliert, um den Betrieb zu überwachen?“
Ralf: Das Monitoring und Logging aller unserer Anwendungen erfolgt über Graylog und Google Cloud Logs.
Christopher Haas: „Wie organisiert ihr das Ressourcenmanagement und die Skalierung eurer LLM-Anwendungen im laufenden Betrieb?“
Ralf: Der Betrieb unserer LLM-Anwendungen erfolgt vollständig über Drittanbieter, weshalb derzeit keine eigene Skalierungslösung implementiert ist.
Christopher Haas: „Welche Strategien zur Modellversionierung und zum Rollback habt ihr in eurem Unternehmen implementiert?“
Ralf: Eine systematische Modellversionierung oder Rollback-Strategie ist nicht vorhanden.
Christopher Haas: „Welche automatisierten Prozesse zur Modellvalidierung und Qualitätskontrolle sind bei euch im Einsatz?“
Ralf: Zur Qualitätskontrolle validieren wir regelmäßig anonymisierte Prompts und Modellantworten auf der OpenAI-Plattform und passen die Instruktionen entsprechend an.
Christopher Haas: „Wie überprüft und optimiert ihr regelmäßig die Betriebskosten – z. B. mithilfe von Monitoring-Dashboards und KPIs?“
Ralf: Da unsere LLM-Lösungen entweder über eine Flat-Subscription oder ein Pay-per-Use-Modell mit geringer Nutzung abgerechnet werden, entstehen derzeit kaum relevante Betriebskosten.

Christopher Haas: „Welche Maßnahmen setzt ihr ein, um die Datensicherheit und den Datenschutz bei der Integration von LLMs zu gewährleisten?“
Ralf: Datensicherheit und Datenschutz werden über vertragliche Regelungen mit den Drittanbietern sichergestellt, die von unserer Legal Abteilung geprüft werden.
Christopher Haas: „Wie regelt ihr den Zugriff auf sensible Daten, und welche Identity-Management-Systeme nutzt ihr dabei?“
Ralf: Der Zugriff auf sensible Daten wird durch ein unternehmensintern entwickeltes Authentifizierungs- und Autorisierungssystem geregelt, das moderne Sicherheitsstandards einhält (z.B. OAuth2). Zugriffe werden individuell pro Nutzer eingeschränkt.
Christopher Haas: „Wie stellt ihr sicher, dass eure Datenverarbeitung den gesetzlichen Vorgaben (z. B. DSGVO) entspricht?“
Ralf: Die Einhaltung gesetzlicher Vorgaben, wie der DSGVO, wird regelmäßig von unserer Legal Abteilung geprüft.
Christopher Haas: „Welche Maßnahmen (z. B. Datenmaskierung, Verschlüsselung, regelmäßige Sicherheitsüberprüfungen) habt ihr implementiert, um Datenlecks zu vermeiden?“
Ralf: Kundendaten werden gemäß den DSGVO-Vorgaben anonymisiert, um das Risiko von Datenlecks zu minimieren.
Christopher Haas: „Welche KPIs und Metriken (z. B. Antwortzeit, Genauigkeit, Kosten pro Anfrage) nutzt ihr, um die Effektivität eurer LLM-Implementierung zu messen?“
Ralf: Wir nutzen Kundenfeedback und G-Eval.
Christopher Haas: „Wie gestaltet ihr den iterativen Verbesserungsprozess – etwa durch regelmäßige Reviews, A/B-Tests oder Benchmarking?“
Ralf: Die kontinuierliche Verbesserung erfolgt durch Logging und die Evaluierung anonymisierter Prompts mit Modellantworten, wodurch die Instruktionen regelmäßig optimiert werden.
Christopher Haas: „Wie dokumentiert und kommuniziert ihr die Umsetzung von Verbesserungsmaßnahmen innerhalb eures Unternehmens?“
Ralf: Änderungen und Verbesserungsmaßnahmen werden dokumentiert und in einem Changelog erfasst.
Christopher Haas: „Wie habt ihr eure bestehende IT-Landschaft (ERP, CRM, DMS) analysiert und dokumentiert, um die Integration von LLMs zu erleichtern?“
Ralf: Unsere bestehende IT-Landschaft ist durch Architekturdiagramme und Domain-Driven Design (DDD) dokumentiert.
Christopher Haas: „Welche unternehmensspezifischen Datenschutz- und Compliance-Richtlinien wurden definiert und in den Integrationsprozess einbezogen?“
Ralf: Unternehmensspezifische Datenschutz- und Compliance-Richtlinien gehen nicht über die gesetzlichen Vorgaben hinaus – die Einhaltung externer Regularien wie der DSGVO ist bereits aufwendig genug.
Christopher Haas: „Wie wurden eure bestehenden Sicherheitslösungen (z. B. Zugriffsmanagement, Audit-Trails) überprüft und an die Anforderungen der LLM-Integration angepasst?“
Ralf: Die Überprüfung und Anpassung bestehender Sicherheitslösungen erfolgt im Rahmen externer Audits.
Christopher Haas: „Wie schätzt du eure internen MLOps-/DevOps-Kapazitäten im Hinblick auf Support und Wartung von LLM-Anwendungen ein?“
Ralf: Intern verfügen wir bereits über das notwendige Know-how durch unsere Data Analysts, die entsprechende Prozesse begleiten können. Auch DevOps Ressourcen sind ausreichend vorhanden.
Christopher Haas: „Gibt es weitere technische oder organisatorische Aspekte, die du als kritisch für die erfolgreiche Integration von LLMs siehst?“
Ralf: Meiner Meinung nach sind weitere entscheidende Faktoren für die erfolgreiche Integration von LLMs externe Erwartungen und Anforderungen, der tatsächliche Bedarf im Unternehmen sowie der wirtschaftliche Wandel, der die Prioritäten beeinflussen kann.
Christopher Haas: „Welche zukünftigen Entwicklungen oder Trends im Bereich LLMs findest du besonders relevant – sowohl technologisch als auch in Bezug auf die Datenbereitstellung?“
Ralf: Zukünftig sehe ich insbesondere die Veränderung bestehender Arbeitsweisen als relevanten Trend, da LLMs zunehmend Prozesse automatisieren und neue Formen der Zusammenarbeit ermöglichen. Die Sammlung großer spezialisierter Datenmengen für das Training und Fine-Tuning von LLMs wird auch wichtig sein, um qualitativ zufriedenstellende und für den Einsatzbereich optimierte Ergebnisse zu erhalten.
Christopher Haas: „Welche Empfehlungen würdest du anderen Unternehmen geben, die den Einsatz von LLMs planen?“
Ralf: Unternehmen, die den Einsatz von LLMs planen, empfehle ich, klein zu starten, mit Prototyping zu arbeiten und verschiedene Lösungen schrittweise zu evaluieren.
Christopher Haas: „Gibt es Best Practices aus deinem Unternehmen, die du als besonders effektiv empfindest?“
Ralf: Eine der wichtigsten Best Practices in unserem Unternehmen ist es, den Mitarbeitern zuzuhören und sie aktiv in den Prozess einzubeziehen, anstatt neue Lösungen einfach vorzugeben. Initiativen sollten nicht unterdrückt, sondern wahrgenommen und strukturiert bewertet werden.

Transkript Theo

Datum: 05.03.2025 **Name:** Theo **Unternehmen:** ca. 4500 Mitarbeiter*innen

Christopher Haas: Dann fangen wir an mit dem allgemeinen Teil. Kannst du kurz deinen beruflichen Hintergrund und deine Rolle im Unternehmen beschreiben?
Theo: Ja, also ich bin in einem Unternehmen tätig und dort im Software Engineering, Software Design und das seit 11 Jahren mittlerweile.
Christopher Haas: Hast du? Beziehungsweise, wie bist du in Entscheidungsprozessen, bei der Implementierung von KI-Technologien in deinem Bereich, involviert?
Theo: Entscheidungsprozesse bin ich involviert, im Sinne, dass ich hin und wieder als Testnutzer und in einer Feedbackgruppe dabei bin und Empfehlungen abgebe, ob Lösungen Sinn machen oder nicht, für unsere Abteilung oder Segment.
Christopher Haas: OK. Und welche Erfahrungen hast du persönlich mit der Integration oder Nutzung von LLMs bzw. anderen KI-Lösungen gemacht?
Theo: Persönlich hab ich einige Projekte privat beziehungsweise in Forschungseinrichtungen durchgeführt und "Wrapper"für Language-Modelle gebaut. Um da drüber ein höheres Abstraktionslevel einzubinden, um ein Language Model für einen spezifischen Use Case zu entwickeln oder damit es für einen spezifischen Use Case verwendet werden kann.
Christopher Haas: OK, im beruflichen Kontext gibt es bei euch im Unternehmen bereits Lösungen von Large Language Models, die verwendet werden.
Theo: Ja, wir haben in der Firma Microsoft Copilot im Einsatz. Und das ist in verschiedenen Bereichen bei uns integriert, sei es beim Coding (das ist mein Hauptanwendungsgebiet). Es gibt auch weitere lokale Instanzen, für die ich, glaube, deshalb wenig Erfahrung oder wenig zu tun gehabt habe in der Firma. Mit Document-Retrieval-Systemen gibt es auch mittlerweile im Unternehmen; bin ich mir aber nicht ganz sicher.
Christopher Haas: Okay, habt ihr auch diese klassische Microsoft M365-Integration von Copilot in sämtlichen Microsoft-Produkten?
Theo: Ah, was ich gesehen habe, ja; ich versuche mich aber aktuell von Microsoft fernzuhalten.
Christopher Haas: OK. Und warum ist das so?
Theo: Mehr in der Hinsicht, dass ich für meine Themen eher Atlassian verwende, zum Beispiel Confluence und Jira. Fernzuhalten im Sinne davon, dass die Tools von Microsoft etwas überladen sind und mich von der eigentlichen Arbeit abhalten. Aber in gewissen Bereichen ist die M365-Integration vorhanden, soweit ich weiß.
Christopher Haas: OK, welche Veränderungen oder Trends im Bereich KI beobachtest du in deinem Unternehmen?
Theo: Es kommt aktuell immer mehr darauf an, dass es in diese Richtung geht oder es Umfragen gibt, ok? Welche Daten haben wir? Welche Daten könnten wir verwenden, um KI-Modelle damit zu füttern und Insights für zukünftige Projekte oder in diesem Bereich zu gewinnen? Aktuell liegt ein starker Fokus darauf, welche Daten wir haben und wie wir diese in Zukunft mit KI-Tools verwenden können.
Christopher Haas: Also, bist du der Meinung, dass die Firma in dem Bereich schon versucht, Fahrtwind aufzunehmen?
Theo: Ah, versucht ja, meiner persönlichen Meinung nach, etwas zu spät und zu langsam, mit zu wenig Ressourcen. Aktuell ist es bei uns so, dass studentische Mitarbeiterinnen und Mitarbeiter Umfragen durchführen, mit Leuten reden, die sich mit der Thematik auskennen, und dann untersuchen, ok, welche Daten wir haben und wie wir diese verwenden können. Man könnte aber durchaus mehr in diesen Bereich untersuchen.
Christopher Haas: OK. Danke für deine Antworten. Dann würde ich den allgemeinen Teil jetzt beenden und in den technischen Bereich, vom Framework her, gehen. Welche Hardware-Ressourcen stehen in deinem Unternehmen für KI-Anwendungen zur Verfügung?
Theo: Zum Unternehmen: Zum einen Azure-Ressourcen, also was jetzt dann nicht on-premise wäre. Da haben wir, was auch immer Azure zu bieten hat, zur Verfügung, und auf der anderen Seite haben wir einiges an on-premise Ressourcen, und in Bezug auf KI gibt es eine relativ große Anzahl an GPUs und auch CPUs, die für KI-Anwendungen nutzbar sind. Oder sind?
Christopher Haas: OK, und wie schätzt du die Skalierbarkeit eurer aktuellen Infrastruktur ein, gerade im Hinblick auf mögliche Lastspitzen?
Theo: Skalierbarkeit? Glaub ich ziemlich gut, da hätten wir eigentlich ja ein Hybrid-System. Genau im Bezug auf KI kann ich das jetzt leider nicht beantworten, weil ich die genauen Ressourcen nicht kenne. Wenn aber die KI auf Azure laufen sollte, sollten eigentlich die Lastspitzen, die für unsere Firma relevant sind, abgefangen werden.
Christopher Haas: OK, ßum Abfangenïst ein gutes Stichwort. Wie beurteilst du die Netzwerkgeschwindigkeit und die Latenz in eurem Unternehmen?
Theo: Sehr gut.
Christopher Haas: Sehr gut, OK. Dann kommen wir zur nächsten Frage: Welche Speichersysteme, zum Beispiel Data Lakes, nutzt ihr für große Datenmengen und wie skaliert ihr diese?
Theo: Bei dieser Frage tue ich mir persönlich schwer, sie zu beantworten. Ich glaube, die Infrastruktur dafür ist vorhanden. Ich persönlich habe aber leider keine Verwendung dafür, beziehungsweise ich bin nicht involviert.
Christopher Haas: OK, also kann man für diesen Punkt keine Aussage treffen.
Theo: Leider nicht.
Christopher Haas: OK, ja, da kommen wir zum Bereich Datenbereitstellung und -qualität. Welche Datenformate und Datenstrukturen verwendet ihr aktuell?

Theo: Hauptsächlich: PDF, Excel, Word, was die meisten Leute verwenden. Im Hintergrund würde ich JSON und CSV sagen.
Christopher Haas: Okay.
Theo: Das ist wiederum aber themenspezifisch zu betrachten bei uns.
Christopher Haas: OK, ja. Welche automatisierten Prozesse, zum Beispiel ETL-Pipelines mit Apache Airflow, kommen zum Einsatz, um die Datenqualität sicherzustellen?
Theo: Automatisierte Prozesse? Kann ich leider auch nicht beantworten, wie weit diese vorhanden sind.
Christopher Haas: OK.
Theo: Manuelle Prüfungen sind natürlich vorhanden.
Christopher Haas: Wie geht ihr dann mit der Duplikaterkennung und -bereinigung von großen Datenbeständen um?
Theo: Die Daten, die in unserem DMS vorhanden sind, also im Dokumentenmanagementsystem, haben natürlich eine Duplikaterkennung und werden über das Dokumentenmanagementsystem gelöst.
Christopher Haas: Welches DMS habt ihr da im Einsatz?
Theo: MS SharePoint.
Christopher Haas: SharePoint? OK. Welche Prozesse sorgen dafür, dass die Daten immer aktuell und vollständig sind?
Theo: Prozesse – [Unklar: Äls Datenaktuell und vollständig sind"].
Christopher Haas: Oder liegt es dann bei jedem Mitarbeiter selbst einfach?
Theo: Ich glaube, aktuell liegt es bei jedem Mitarbeiter selbst. Ja, wirklich, dass es offizielle Prozesse dazu gibt, ist mir aktuell nicht bekannt.
Christopher Haas: OK, und hinsichtlich des aktuell Haltens der Daten – können wir dann auch schon zur nächsten und letzten Frage für diesen Bereich kommen: Wie organisiert ihr die Archivierung und die Versionierung der Daten?
Theo: Das wäre wieder über das Dokumentenmanagementsystem dann erledigt.
Christopher Haas: OK. Dann kommen wir zur nächsten Frage: Welche Deployment-Modelle, On-Premises, Cloud, Hybrid, nutzt ihr aktuell oder plant ihr für den Einsatz von LLMs und warum?
Theo: Genau, also wir nutzen aktuell unseren Haupt-Cloud-Anbieter Azure. Und je nach Anforderungen wäre natürlich auch lokales Hosting möglich.
Christopher Haas: Und wie definiert und implementiert ihr eure Datenschutzrichtlinien im Zusammenhang mit der Bereitstellung?
Theo: Das kann ich jetzt leider nicht mit Garantie sagen, wie es genau gehandhabt wird. Ich kann nur vermuten, dass kundenspezifische und personenbezogene Daten nicht über die Cloud verarbeitet bzw. anonymisiert werden und falls relevant auf einem lokal gehosteten System verarbeitet werden. Ist aber nur eine Vermutung.
Christopher Haas: OK, danke für deine Antwort. Welche Kriterien fließen in eure Evaluierung der Anbieterabhängigkeit ein, und wie überprüft ihr diese regelmäßig?
Theo: Das kann ich leider jetzt nicht beantworten.
Christopher Haas: Gut, dann kommen wir zur nächsten Frage: Wie erfolgt bei euch die langfristige Kosten-Nutzen-Bewertung des gewählten Deployment-Modells?
Theo: Das kann ich auch nicht beantworten.
Christopher Haas: War das nicht auch bei dem Umfang dabei, was du vorher gesagt hast, mit den studentischen Mitarbeiterinnen und Mitarbeitern, die da Umfragen machen?
Theo: Die Umfragen sind aktuell aber nur in Richtung, was möglich ist oder was wäre möglich. Aber du hast natürlich recht, es wäre eine Möglichkeit, dass man eine Art Snapshot-Messung macht, vor Implementierung versus nach der Implementierung, um dann eine Kosten-Nutzen-Bewertung durchzuführen.
Christopher Haas: Ah, okay? Welche Kriterien, zum Beispiel in Sachen Datensicherheit, Anpassungsfähigkeit und Support, sind für dich ausschlaggebend, wenn es um die Wahl von Large Language Models geht?
Theo: Kommt darauf an, für was ich es verwenden will. Zum einen kommt es darauf an, welche Daten ich mit dem Modell verarbeiten will – wenn es um personenbezogene Daten geht, ist Datensicherheit relevanter. Im Engineering-Bereich würde ich Modelle bevorzugen, die anpassbar sind, also adaptiver auf meinen Use Case zugeschnitten.
Christopher Haas: OK, und welche Erfahrungen hast du mit Open-Source-Modellen im Vergleich zu kommerziellen Large Language Models gemacht?
Theo: Mit Open-Source-Modellen ist es mehr Aufwand bei der Implementierung und beim Nutzen. Bei kommerziellen Tools hat man mehr oder weniger sehr wenig Aufwand, aber man weiß auch nicht, was mit den Daten passiert. Wenn ich Open Source selbst implementiere und selbst hoste, habe ich komplette Kontrolle, soweit das möglich ist, darüber, was das Modell tut und was mit den Daten passiert. Bei den kommerziellen Produkten, die ich aktuell kenne, lässt sich nicht immer sagen, was passiert und warum etwas passiert.
Christopher Haas: OK, und wie bewertest du die Flexibilität und den Support der von euch genutzten Modelle?
Theo: Also, ich habe aktuell noch keinen Support gebraucht, da es eigentlich immer funktioniert hat.
Christopher Haas: Kennst du jemanden im näheren Arbeitsumfeld, bei dem es relevant war?
Theo: Nein, das fällt mir im Moment nicht ein.
Christopher Haas: Dann kommen wir zu Punkt [4.2.1]: Integration in bestehende IT-Systeme. Wie bindet ihr Large Language Model-Anwendungen in eure bestehenden Systeme, also ERP, CRM, DMS, ein, sofern du dazu eine Aussage treffen kannst?
Theo: Zu ERP und CRM kann ich keine Aussage treffen. Zu DMS eigentlich auch nicht, leider.

Christopher Haas: OK. Welche Schnittstellen und API-Gateways nutzt ihr, um die Kommunikation zwischen den Systemen sicherzustellen?
Theo: Moment, leider habe ich dazu auch keine Ahnung.
Christopher Haas: OK. Wie stellt ihr sicher, dass die Integration von Large Language Models mit euren bestehenden Sicherheitslösungen kompatibel ist?
Theo: Dafür habe ich zu wenig Einblick in die IT-Seite.
Christopher Haas: Ja, dann noch die letzte Frage in dem Bereich: Welche Monitoring- und Logging-Strategien habt ihr etabliert, um den Betrieb zu überwachen?
Theo: Mir ist nur bewusst, dass wir DevOps einsetzen und über das Logging und Monitoring verfügen.
Christopher Haas: OK. Gut, danke. Dann kommen wir zur nächsten Frage: Wie organisiert ihr das Ressourcenmanagement und die Skalierung eurer LLM-Anwendungen im laufenden Betrieb?
Theo: Der Copilot skaliert automatisch über die Azure-Cloud, meines Wissens.
Christopher Haas: Mhm.**Theo:** Die lokale Lösung läuft, glaube ich, über Docker und Kubernetes, sodass sie automatisch dahingehend skaliert.
Christopher Haas: OK. Welche Strategien zur Modellversionierung und zum Rollback habt ihr in eurem Unternehmen implementiert?
Theo: Ich glaube, ich kann leider nicht mit Sicherheit sagen, ob und inwieweit es bei uns eine Modellversionierung gibt.
Christopher Haas: Mhm, aber die besteht vermutlich vom GitHub Copilot?
Theo: Genau, und ich gehe davon aus, dass diese entweder automatisch upgedated werden beziehungsweise immer der Letztstand aktiv ist, wiederum nur eine Vermutung.
Christopher Haas: Ja, in dem Fall würde ich dann die nächsten beiden Fragen überspringen, wenn du bezüglich der Prozesse, der Modellvalidierung und auch der Kosten vermutlich keine Aussage treffen kannst.
Theo: Nein, kann ich nicht.
Christopher Haas: Ja, dann gehen wir zum Punkt 5.2 Sicherheit und Compliance. Welche Maßnahmen setzt ihr ein, um die Datensicherheit und den Datenschutz bei der Integration von LLMs zu gewährleisten?
Theo: Ich glaube, es geht darum, dass personenbezogene Daten nicht in die Modelle einfließen dürfen, aber wie der Datenschutz direkt sichergestellt wird, kann ich intern nicht beantworten.
Christopher Haas: OK, wie regelt ihr den Zugriff auf sensible Daten, und welche Identity-Management-Systeme nutzt ihr dabei?
Theo: Die Zugriffskontrolle auf unsere Daten ist prinzipiell über User-Gruppen und abgesicherte Bereiche definiert. Damit ist in unseren ERP- und CRM-Systemen geregelt, dass man auf keine Kundendaten zugreifen kann, die nicht zu den eigenen Projekten gehören. Und wie heißt das in unserer Firma? Mir ist der Name leider entfallen.
Christopher Haas: OK, das heißt, ihr habt aber ein gängiges System in der Firma etabliert.
Theo: Ja, ein Standardsystem, wie die meisten Firmen haben, tut mir leid, mir ist der Name davon entfallen.
Christopher Haas: OK. Wie stellt ihr sicher, dass eure Datenverarbeitung den gesetzlichen Vorgaben, zum Beispiel der DSGVO, entspricht?
Theo: Es werden regelmäßig Audits durchgeführt.
Christopher Haas: Mhm, habt ihr da auch Schulungen?
Theo: Wir haben auch jährlich verpflichtende Schulungen.
Christopher Haas: Ah, OK. Welche Maßnahmen, zum Beispiel Datenmaskierung, Verschlüsselung, regelmäßige Sicherheitsüberprüfungen, habt ihr implementiert, um Datenlecks zu vermeiden?
Theo: Ich bin mir leider nicht hundertprozentig sicher, ob das exakt auf die Frage zutrifft. Es werden teilweise Tests durchgeführt, um zu testen, ob Datenlecks provoziert werden können. Genauso, ich weiß, dass auch Praxistests von der IT kontrolliert durchgeführt werden, um zu testen, ob Mitarbeiter Daten freigeben.
Christopher Haas: Ja, das wären so quasi Penetrationstests, die überprüfen...
Theo: Genau, ja.
Christopher Haas: Ob es da eben zu Datenlecks kommt?
Christopher Haas: Ja, super, danke. Dann kommen wir zur nächsten Frage: Welche KPIs und Metriken, zum Beispiel Antwortzeit, Genauigkeit, Kosten pro Anfrage, nutzt ihr, um die Effektivität eurer LLM-Implementierung zu messen?
Theo: Kann ich nicht beantworten.
Christopher Haas: Da geht es um Antwortzeiten, Genauigkeit – das sind Punkte, die dich im Speziellen jetzt nicht betreffen oder interessieren, nehme ich an.
Theo: Genau, also welche offiziellen Tests von unserer kontrollierten IT durchgeführt werden, kann ich nicht beantworten; es fließt nur persönliche Erfahrung ein. Es geht darum, dass die Antwortzeit und die Genauigkeit der Antworten persönlich angeschaut beziehungsweise evaluiert und dann Feedback gegeben wird. An offiziellen Tests bin ich nicht beteiligt.
Christopher Haas: Ja.
Christopher Haas: OK, gut, somit wird es dann, glaube ich, für die nächsten beiden Fragen der Fall sein, dass du da keine Aussage treffen kannst.
Theo: Nein, kann ich nicht.
Christopher Haas: Ja. Kommen wir dann zu den unternehmensspezifischen Anpassungen. Da geht es um die erste Frage: Wie habt ihr eure bestehende IT-Landschaft analysiert und dokumentiert, um die Integration von LLMs zu erleichtern?

Theo: Unsere eigene IT, also die komplette IT-Landschaft und Prozesslandschaft, ist dahingehend einfach dokumentiert und digital verfügbar. Ah, und mit allen Schnittstellen des Prozesses, definiert mit den Inputs und Outputs.
Christopher Haas: OK, und welche unternehmensspezifischen Datenschutz- und Compliance-Richtlinien wurden definiert und in den Integrationsprozess einbezogen?
Theo: Das sind die DSGVO und die ISO 27001.
Christopher Haas: Habt ihr auch ein firmeninternes Dokument hinsichtlich dieser Richtlinien, das euch vorschreibt, zum Beispiel welche Daten für Large Language Models verwendet werden dürfen?
Theo: Genau, wir haben intern noch die Compliance-Richtlinien und auch Schulungen für diese, die wir durchführen müssen beziehungsweise die verpflichtend sind.
Christopher Haas: OK. Wie wurden eure bestehenden Sicherheitslösungen, zum Beispiel Zugriffsmanagement und Audit-Trails, überprüft und an die Anforderungen der LLM-Integration angepasst?
Theo: Zugriffsmanagement? Wieder über die User-Gruppen und Test-User beziehungsweise wer dann für die Modelle freigeschaltet ist. Weiteres kann ich leider nicht beantworten.
Christopher Haas: OK, dann die abschließende Frage für den Bereich: Wie schätzt du eure internen MLOps-/DevOps-Kapazitäten im Hinblick auf Support und Wartung von LLM-Anwendungen ein?
Theo: Ressourcen in Bezug auf Personen, oder?
Christopher Haas: In Bezug auf Personen, ja.
Theo: Sind schlechter geworden, da aktuell die Auftragslage unserer Firma schlechter ist und es zu keinem Personenaufbau kommt. Wie viele Personen wir genau für den DevOps-Bereich haben, kann ich leider auch nicht als fixe Zahl beantworten.
Christopher Haas: OK. Gut, dann kommen wir zur nächsten Frage: Gibt es weitere technische oder organisatorische Aspekte, die du als kritisch für die erfolgreiche Integration von LLMs siehst?
Theo: Der kritischste Punkt ist, glaube ich, in Bezug auf den Datenschutz, also was will ich jetzt wirklich mit den Daten machen? Zum einen, mit personenbezogenen Daten, was bei unserer Firma eher zweitrangig ist. Kannst du kurz nochmal die Frage wiederholen? Tschuldige?
Christopher Haas: Ja, gerne: Gibt es weitere technische oder organisatorische Aspekte, die du als kritisch für die erfolgreiche Integration von LLMs siehst?
Theo: Organisatorisch, in dem Engineering-Bereich, in dem ich tätig bin: Wir haben eine relativ hohe Hierarchie, bis dahin Entscheidungen getroffen, Daten freigegeben beziehungsweise auch Ressourcen freigegeben werden, damit Daten aufbereitet werden können. Das ist bei uns die größere beziehungsweise zeitaufwendigste Hürde.
Christopher Haas: OK, also ein Zeitfaktor ist für dich ein kritischer Aspekt?
Theo: Definitiv, in dem Bereich, in dem ich tätig bin, hat die Thematik aktuell keine wirkliche Priorität.
Christopher Haas: OK.
Theo: Themenspezifische LLM-Applikationen zu entwickeln oder zu implementieren.
Christopher Haas: Und welche zukünftigen Entwicklungen oder Trends im Bereich von LLMs findest du besonders relevant, sowohl technologisch als auch in Bezug auf die Datenbereitstellung?
Theo: Aktuell sind RAG-Systeme definitiv am spannendsten und auch für Firmen vermutlich am relevantesten, da man bestehende Language-Modelle verwenden kann und dann auf existierende Daten zugreifen kann, ohne das Modell komplett neu zu trainieren.
Christopher Haas: Empfehlungen und Best Practices: Welche Empfehlungen würdest du anderen Unternehmen geben, die den Einsatz von LLMs planen?
Theo: Frühzeitig über die Daten nachdenken, wie Daten aufbereitet werden können; welche Daten und natürlich welche Probleme mit dem Modell gelöst werden sollen. Und zum anderen, welche Lösungen verwendet werden, ob es jetzt eine on-premise Lösung ist oder eine existierende Lösung in der Cloud, und wie weit das Modell auf einen spezifischen Use Case trainiert werden muss oder sollte.
Christopher Haas: Gibt es Best Practices aus deinem Unternehmen, die du als besonders effektiv empfindest?
Theo: Fehlt mir im Moment nicht, nein.