HAW HAMBURG

BACHELORARBEIT

Einsatz und Auswirkungen von Chatbots in Online Shops kleiner und mittelständischer Unternehmen

vorgelegt am 18. September 2025

Sören Gooß

Erstprüferin: Prof. Dr. Sabine Schumann
Zweitprüferin: Prof. Dr. Marina Tropmann-Frick

HOCHSCHULE FÜR ANGEWANDTE WISSENSCHAFTEN HAMBURG
Department Medientechnik
Finkenau 35
20081 Hamburg

Zusammenfassung

Diese Bachelorarbeit untersucht den Einsatz und die Auswirkungen von Chatbots in kleinen und mittelständischen Unternehmen. Dazu wurde ein Chatbot zur Kundenberatung und Unterstützung bei Kaufentscheidungen in dem Online Shop eines mittelständischen Unternehmens realisiert und eingesetzt. In dieser Arbeit werden sowohl die Kostenaspekte als auch die Auswirkungen auf die Kunden betrachtet. Zusätzlich wurden zwei Interviews sowie ein fachlicher Austausch mit Experten in dem Bereich Chatbots aus der Industrie und Forschung durchgeführt, um eine umfassendere Perspektive zu liefern.

Diese Arbeit präsentiert ein Vorgehensmodell für die technologische Umsetzung eines Chatbots, der in kleinen und mittelständischen Unternehmen eingesetzt werden kann, das von der Konzeptentwicklung über die Implementation bis hin zur Evaluation reicht. Die Ergebnisse des Projekts zeigen Chancen für die Umsatzsteigerung und eine Transformation der Kundeninteraktion auf, welche aufgrund von fehlender Nutzung des Chatbots allerdings nicht fundiert genug umgesetzt worden sind. Zusätzlich zeigen sie, dass Kunden häufig zusätzlich motiviert werden müssen, um das bereitgestellte Angebot anzunehmen. Sie beleuchten allerdings die Möglichkeiten für den Einsatz eines Chatbots in kleinen und mittelständischen Unternehmen.

Abstract

This bachelor thesis examines the use and impact of chatbots in small and medium-sized companies. To this end, a chatbot was implemented and used in the online shop of a medium-sized company to provide customer advice and support with purchasing decisions. This thesis considers both the cost aspects and the impact on customers. In addition, two interviews and a technical exchange with experts in the field of chatbots from industry and research were conducted to provide a more comprehensive perspective.

This thesis presents a process model for the technological implementation of a chatbot that can be used in small and medium-sized companies, ranging from concept development to implementation and evaluation. The results of the project reveal opportunities for increasing sales and transforming customer interaction, which, however, have not been implemented sufficiently due to a lack of use of the chatbot. In addition, they show that customers often need additional motivation to accept the offer provided. However, they highlight the possibilities for using a chatbot in small and medium-sized companies.

Vorwort

Zur besseren Lesbarkeit dieser Bachelorarbeit wird bei Personenbezeichnungen das generische Maskulinum oder neutrale Begriffe verwendet. Selbstverständlich gelten sämtliche Personenbezeichnungen für alle Geschlechter.

Danksagung

Ich möchte an dieser Stelle meine Dankbarkeit gegenüber allen ausdrücken, die mich in dem Zeitraum meiner Bachelorarbeit unterstützt und begleitet haben.

Ich bedanke mich bei meiner Erstprüferin Prof. Dr. Sabine Schumann für die durchgängige Betreuung und Unterstützung bei allen Fragen und Anliegen, die in diesem Zeitraum aufgekommen sind. Außerdem möchte ich mich bei meiner Zweitprüferin Prof. Dr. Marina Tropmann-Frick bedanken, die ebenfalls ihre Unterstützung und Sichtweisen beigetragen hat, immer für Rückfragen bereitstand und sich bereit erklärt hat, die Rolle der Zweitprüferin zu übernehmen.

Zudem bedanke ich mich bei meinem Unternehmen und meinen Kollegen, die mir dieses Projekt ermöglichten und mich ebenfalls tatkräftig unterstützten.

Zuletzt möchte ich mich bei denjenigen bedanken, die meine Bachelorarbeit korrekturgelesen und somit zu einer verbesserten Ausführung beigetragen haben.

Anmerkung zum Einsatz von KI

Während der Anfertigung dieser Bachelorarbeit wurden generative KI-Tools verwendet, um diverse Aspekte des Arbeitsprozesses zu unterstützen. Besonders wurden Sprachmodelle genutzt, um bei dem Entwickeln und Debuggen von Code, bei dem Heraussuchen von wissenschaftlichen Quellen sowie bei der Formulierung von einzelnen Textpassagen zu unterstützen. Außerdem wurde eine KI-gestützte Rechtschreibprüfung und Übersetzung angewandt. Konkret wurden die Tools Perplexity Pro, Consenus AI, Scribbr Rechtschreibprüfung mit Quillbot sowie DeepL AI angewandt. Alle Konzepte, Implementationen, Evaluationen sowie selbst erstellte Grafiken wurden von dem Autor erstellt.

Inhaltsverzeichnis

Abkürzungsverzeichnis VI

Abbildungsverzeichnis VII

Formelverzeichnis VIII

1 Einleitung 1

1.1 Motivation 1

1.2 Zielsetzung und Fragestellung 4

1.3 Aufbau der Arbeit 5

2 Grundlagen 5

2.1 Entwicklungsgeschichte 5

2.2 Künstliche Neuronale Netze 7

2.2.1 Aktuelle Architekturen 7

2.2.2 Das Transformer-Prinzip 8

2.3 Kontextsensitive LLMs 12

2.4 Restriktionen kleiner und mittelständischer Unternehmen 14

3 Konzept 15

3.1 Hintergrund 15

3.2 Architekturen 16

3.3 Verwendete Modelle 17

4 Implementierung 19

4.1 Entwicklungsumgebung 19

4.2 Prozessabläufe 20

4.3 Herausforderungen bei der Implementation 22

5 Evaluation 24

5.1 Entwicklungs- und laufende Kosten 24

5.2 Vergleich mit Dienstleistern 25

5.3 Return on Investment (ROI) 26

5.3.1 Umsatz des Chatbots 26
5.3.2 Entlastung des Kundenservice 27
5.3.3 Weitere Kennzahlen 27
5.4 Auswirkungen auf die Kundschaft 31
5.5 Datenschutz und Datensicherheit 33
5.5.1 Datenschutz von Kundendaten 33
5.5.2 Datensicherheit 35
5.6 Vergleich mit Experten und anderen Forschungsprojekten 36
6 Fazit 37
6.1 Zusammenfassung 37
6.2 Ausblick 38
Literaturverzeichnis 40
Anhang 49
Interviews 49
Interviewpartner 1: Felix (Anonymisiert) 49
Interviewpartner 2: Maximilian (Anonymisiert) 58
Gedächtnisprotokoll zu fachlichem Austausch: Thomas & Michael (Anonymisiert) 66

Abkürzungsverzeichnis

AIML ..... *Artifical Intelligence Markup Language*
API ..... *Application Programming Interface*
CNN ..... *Convolutional Neural Networks*
CVR ..... *Conversionrate*
DSGVO ..... *Datenschutz Grundverordnung*
FFN ..... *Feed Forward Network*
GPAI ..... *General-Prupose AI*
GPT ..... *Generative Pretrained Transformer*
JSON ..... *JavaScript Object Notation*
KI ..... *Künstliche Intelligenz*
KNN ..... *Künstliche Neuronale Netze*
LLM ..... *Large Language Model*
MCP ..... *Model Context Protocol*
ML ..... *Machine Learning*
QAT ..... *Quantization-Aware Training*
RAG ..... *Retrieval-Augmented Generation*
RNN ..... *Recurrent Neural Network*
ROI ..... *Return on Investment*
SLM ..... Small Language Model
UUID ..... *Unique User Identifier*
VRAM ..... *Virtual Random Access Memory*

Abbildungsverzeichnis

Abbildung 1: Zeitraum, um eine Million Nutzende zu erreichen [Statis 2023a] 1
Abbildung 2: Umfrage zu Arten des Einsatzes von Sprachassistenten im Alltag [Statis 2024] 2
Abbildung 3: Umsatz mit generativer KI weltweit ab 2020 mit Prognose bis 2032 [Statis 2023b] 3
Abbildung 4: Beispiel einer CNN-Architektur [Dharma 2022] 7
Abbildung 5: Der Transformer – Modellarchitektur [VaShPa+ 2017, S. 3]. 9
Abbildung 6: Scaled Dot-Product Attention [VaShPa+ 2017, S. 4]. 10
Abbildung 7: Multi-Head Attention mit mehreren parallelen Layern [VaShPa+ 2017, S. 4]. 11
Abbildung 8: High-Level Overview of Qdrant's Architecture [Qdrant 2025] 13
Abbildung 9: Prozessabläufe in der angewandten Architektur (zweites Konzept) 20
Abbildung 10: Formatting for Gemma IT models [KaFePa+ 2025, S. 4] 23
Abbildung 11: CVR mit und ohne Chatbot-Interaktion für Bestellungen mit vom Chatbot empfohlenen Produkten in Prozent (%) 29
Abbildung 12: Durchschnittlicher Warenkorbwert mit und ohne Chatbot-Interaktion für Bestellungen mit vom Chatbot empfohlenen Produkten in Euro (€) 29
Abbildung 13: CVR mit und ohne Chatbot-Interaktion für alle Bestellungen in Prozent (%) 30
Abbildung 14: Durchschnittlicher Warenkorbwert mit und ohne Chatbot-Interaktion für alle Bestellungen in Euro (€) 30

Formelverzeichnis

Formel 1: Berechnung der Scaled Dot-Product Attention 10
Formel 2: Aktivierungsfunktion des FFN 11
Formel 3: Berechnung des Return on Investment 26

1 Einleitung

Moderne Technologien, wie die der Künstlichen Intelligenz (KI), etablieren sich zunehmend in unterschiedlichen Branchen. Dabei erlangen sie, je nach Anwendungsbereich, verschiedene Akzeptanzniveaus. Einer dieser Bereiche stellen Chatbots dar, die zunehmend im Kundenservice in Online Shops eingesetzt werden. Diese Arbeit befasst sich mit dem Einsatz von Chatbots in kleinen und mittelständischen Unternehmen (KMU). Im Folgenden werden die Motivation und die Relevanz des Themas sowie die Zielsetzung und der Aufbau dieser Bachelorarbeit dargestellt.

1.1 Motivation

Durch die Veröffentlichung des Sprachmodells ChatGPT am 30. November 2022 stieg das öffentliche Interesse an dem Anwendungsbereich der Chatbots stark an. Mit der Erreichung einer Million Nutzender weltweit in den ersten fünf Tagen ist ChatGPT eine der am schnellsten wachsenden Anwendungen jemals (Abb. 1). ChatGPT, das bis zu diesem Zeitpunkt fortschrittlichste öffentlich nutzbare Sprachmodell, wurde durch die Nutzung neuer Technologien entwickelt und anschließend mit einer sehr großen Datenmenge trainiert [OuWuJi 2022, BrMaRy 2020].

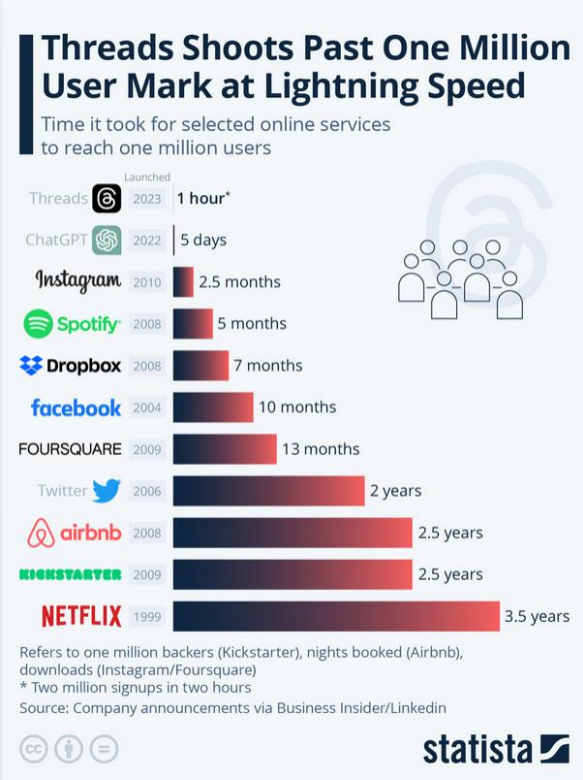

Abbildung 1: Zeitraum, um eine Million Nutzende zu erreichen [Statis 2023a]

Die Geschichte der Chatbots geht bis in das Jahr 1966 zurück, in dem Joseph Weizenbaum den ersten Chatbot namens ELIZA entwickelt hat [Weizen 1966]. Seither wurden in diesem Bereich stetig Fortschritte gemacht und verschiedene Modelle veröffentlicht. 2010 stellte Apple beispielsweise seinen Assistenten Siri vor, gefolgt von Googles Assistant, Microsofts Cortana und Amazons Alexa [BeLóDi 2019]. Aktuell existieren unterschiedliche Sprachmodelle und Sprachassistenten in Form von Large Language Models (LLM), die auf verschiedene Aufgaben ausgerichtet sind. In den vergangenen zwei Jahren ist ein deutlicher Trend zu beobachten, der belegt, dass diese Modelle zunehmend eingesetzt werden und komplexe Aufgaben im Alltag, wie in Abbildung 2 zu erkennen, bewältigen [RaMuFa^{+} 2024] (Abb. 2, Abb. 3).

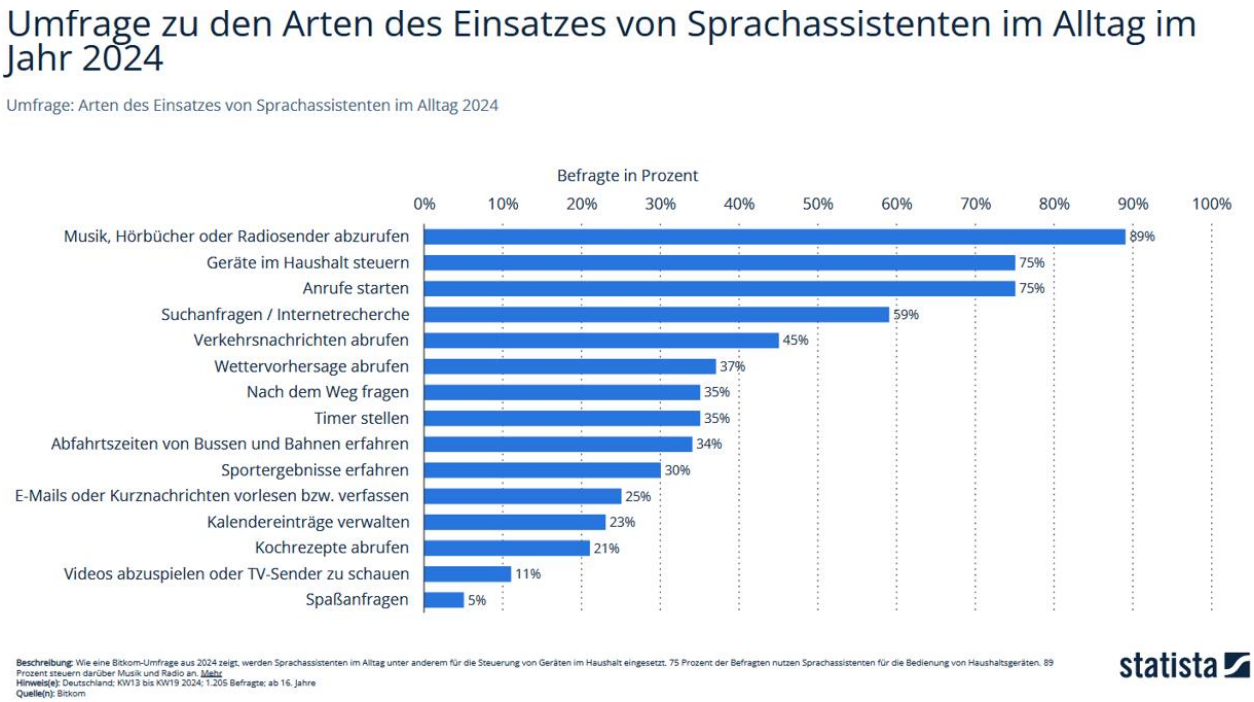

Abbildung 2: Umfrage zu Arten des Einsatzes von Sprachassistenten im Alltag [Statis 2024]

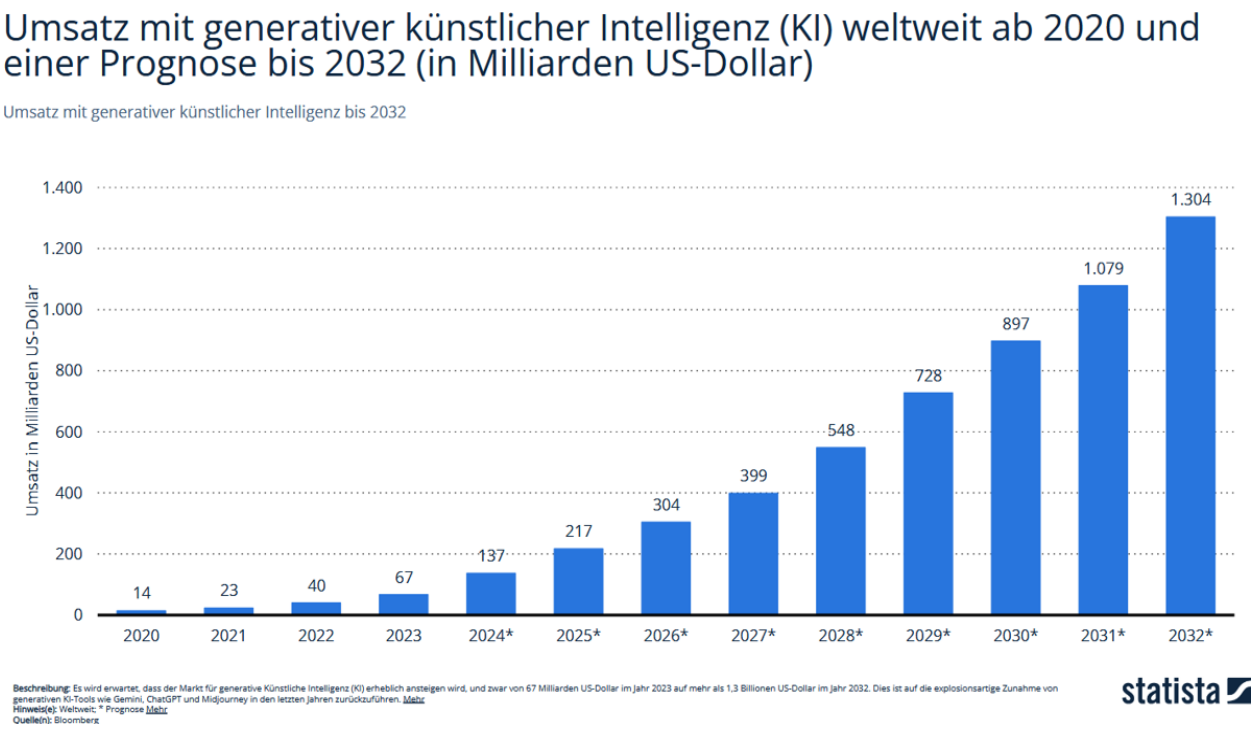

Abbildung 3: Umsatz mit generativer KI weltweit ab 2020 mit Prognose bis 2032 [Statis 2023b]

Angesichts der jüngsten Entwicklungen im Bereich der Chatbots ergibt sich das Interesse, dieses Forschungsfeld mit einem spezifischen Fokus auf den Einsatz in KMU weiter zu untersuchen. Durch die stetigen Fortschritte der aktuellen Modelle werden Chatbots vermehrt im Bereich des Kundenservice eingesetzt. Dabei nutzen Chatbots häufig Modelle oder APIs häufig von existierenden Anwendungen wie ChatGPT von OpenAI, Claude von Anthropic oder Gemini von Google [Stefan 2025, aloa 2025]. Die Anbindung der API besitzt im Gegensatz zu selbst gehosteten Modellen eine leistungsstarke Infrastruktur, die es erlaubt, schnell und präzise Antworten zu liefern. Mittlerweile existieren allerdings unzählige Modelle, die mittels neuester Methoden auch unter der Verwendung weniger Ressourcen zufriedenstellende Ergebnisse in einer angemessenen Zeit liefern.

Die Motivation und Relevanz dieser Bachelorarbeit entsteht aus dem Umsetzungsprojekt eines Chatbot-Assistenten, der aus einer intelligenten Suche mit Unterstützung von KI im Online Shop des Unternehmens entwickelt wurde. Diese Aufgabe wurde zunächst als unabhängiges Projekt entworfen und bestand aus der Basis für den Prototypen des Chatbot-Assistenten. Im Laufe des Umsetzungsprojekts wurde deutlich, dass insbesondere die Auswertung der angestrebten Ergebnisse von besonderem Interesse ist und sich für eine Bachelorarbeit eignet. Für diese Arbeit wurden im Verlauf des Projektes und im Zeitraum der Bachelorarbeit weitere Anpassungen und

Verbesserungen zum Chatbot-Assistenten hinzugefügt und durch Zwischenergebnisse nützliche Daten generiert. Die Endergebnisse sowie erhobene Daten des Projekts werden in dieser Arbeit evaluiert. Das Projekt bietet die Möglichkeit, den Prozess der Entwicklung, Implementierung und Anwendung über den gesamten Zeitraum hinweg zu beobachten und zu analysieren. Besonders für KMU, die versuchen, Unterstützung in Form von KI in ihre Unternehmensprozesse einzugliedern, liefern die Ergebnisse dieser Arbeit wertvolle Einblicke.

1.2 Zielsetzung und Fragestellung

Das Ziel dieser Bachelorarbeit umfasst die Planung, die Umsetzung und den Einsatz eines Chatbots, welcher mittels begrenzter Ressourcen entwickelt wurde. Dieser wurde in dem Online Shop eines mittelständischen Unternehmens betrieben. Außerdem enthält die vorliegende Arbeit eine Aufwands- und Kostenanalyse, die anschließend einer Nutzenanalyse gegenübergestellt wird. Die Arbeit soll aufzeigen, dass KMU ebenfalls in der Lage sind, einen eigenen Chatbot zu entwickeln sowie diesen in ihren Online Shop zu integrieren. Einige Studien wie [ShaMis 2024], [SurCas 2023] und [ErSuLa 2024] zeigen bereits positive Auswirkungen auf die Kundenzufriedenheit und auf den Umsatz des Online Shops, weshalb es von besonderem Interesse ist, diesen Trend aufzugreifen und forschungsrelevante Ergebnisse zu dokumentieren und auszuwerten. Die Hauptaufgabe des Chatbots liegt darin, der Kundschaft einen schnellen Zugang zu Informationen über Produkte und das Unternehmen zu geben, die spezifisch auf die Anfrage des Kunden angepasst sind. Durch die Gegenüberstellung des Entwicklungsaufwands und der erzielten Ergebnisse werden grundlegende und praxisnahe Erkenntnisse über die Wirtschaftlichkeit und Effektivität solcher Anwendungen gewonnen. Zudem werden ausgewählte Grundsätze und Ergebnisse aus verwandten Studien aufgegriffen, um die in dieser Arbeit aufgezeigten Erkenntnisse in einen vergleichbaren Kontext zu stellen. Darüber hinaus liefern drei Interviews mit Experten im Bereich Chatbots aus der Industrie und der Forschung weitere Erkenntnisse. Die Arbeit beschäftigt sich demnach mit der Fragestellung, inwiefern ressourcenschonende Chatbots den Kundenservice von KMU entlasten, ihren Umsatz steigern können, Kosteneinsparungen hervorrufen sowie durch die systematische Auswertung von Konversations- und Bestelldaten verwertbare Einblicke in die Kundenbedürfnisse liefern.

1.3 Aufbau der Arbeit

Diese Arbeit fokussiert sich insbesondere darauf, die gewonnenen Daten aus der praktischen Implementierung eines Chatbots in einem Online Shop zu analysieren und zu bewerten. Gestützt durch einschlägige Studien wird festgestellt, wie die mit der Entwicklung und dem Betrieb verbundenen Kosten durch die erlangten Ergebnisse zu rechtfertigen sind.

Das folgende Kapitel widmet sich zunächst der Geschichte von generativer KI und Chatbot-Technologie, wobei der Fokus insbesondere auf dem aktuellen Transformer-Prinzip liegt, auf dem auch das verwendete Modell grundlegend basiert. Das dritte Kapitel skizziert die Konzeptphase, verwandte Implementationen aus unterschiedlichen Studien und erläutert die Gründe für die verwendeten Methoden. Darauffolgend behandelt das vierte Kapitel die Implementation des Systems. Es beschreibt die Prozessabläufe und schildert Herausforderungen bei der Entwicklung. In dem fünften Kapitel werden die wirtschaftlichen Aspekte und Herausforderungen einer solchen Implementierung in Bezug auf die erlangten Ergebnisse evaluiert. Außerdem stellt das Kapitel den Einfluss eines Shop-Assistenten auf die Kundschaft in Bezug auf die Verlässlichkeit und die Einstellung gegenüber dem Chatbot dar. Anschließend wird das Thema Datenschutz und Datensicherheit beleuchtet, da es vorrangig im Anwendungsbereich Kundenservice eine hohe Relevanz genießt. Da im Rahmen dieser Bachelorarbeit Experteninterviews und ein fachlicher Austausch durchgeführt wurden, werden diese in Begleitung von weiteren Studien in Bezug zu dem Projekt und seinen Ergebnissen gesetzt. Abschließend wird in dem sechsten Kapitel eine konsolidierte Darstellung der Ergebnisse geliefert sowie ein Fazit gezogen und ein Ausblick auf zukünftige Entwicklungen gegeben.

2 Grundlagen

Dieses Kapitel veranschaulicht die Grundprinzipien und die zugrunde liegenden Technologien, auf denen diese Arbeit und das Projekt aufbauen. Zudem werden Restriktionen in Bezug auf Chatbotprojekte, insbesondere bei KMU, erläutert.

2.1 Entwicklungsgeschichte

Mit dem durch Alan Turing geprägten Aufstieg von Computertechnologie nach dem Zweiten Weltkrieg gewann auch die Idee der Künstlichen Intelligenz zunehmend Aufmerksamkeit. Turing definierte ein Imitationsszenario, heute bekannt als Turing-Test, in dem eine Maschine versucht, vorzugeben, ein Mensch zu sein. Dieses legte die Grundidee für intelligente Maschinen dar

[Turing 1950]. 1956 wurde der Begriff „Künstliche Intelligenz“ auf der Dartmouth-Konferenz erstmals gefestigt [McMiRo+ 1955]. Die folgenden Jahre waren geprägt von Optimismus und großen Investments in den Bereich der Künstlichen Intelligenz. Die Anfänge fokussierten sich auf das Lösen logischer und heuristischer Probleme. Nachdem einige frühe Projekte ihre Entwicklungsziele verfehlten und Experten realisierten, dass der damalige Stand der Technologie nicht ausreichte, um signifikante Fortschritte zu machen, stockte die Entwicklung von KI zunächst [ToBoSa+ 2022].

Durch die Verbesserung von Algorithmen und die steigende Rechenleistung erlangten künstliche neuronale Netze (KNN) in den 1980er Jahren, die in den vorherigen Jahren anderen Methoden unterlegen waren, erneut Aufmerksamkeit. Besonders die Implementation der Backpropagation erlaubte es, KNN mit mehreren Ebenen zu trainieren, die komplexere Aufgaben lösen konnten [RuHiWi 1986], [ToBoSa+ 2022]. Mit steigender Menge an Daten und Rechenleistung wurden stetig neue Methoden entwickelt, die sich heute unter dem Begriff „Deep Learning“ einordnen lassen. 2012 zeigte AlexNet, eine spezielle Ausführung einer Art von KNN, erneut die Relevanz von KNN in der Bearbeitung komplexer Aufgaben. 2017 erfolgte durch Google ein wichtiger Schritt in der Sprachanalyse, welcher zum aktuellen Entwicklungsstand beigetragen hat [VaShPa+ 2017].

Auch die Chatbot-Technologie entwickelte sich parallel zu den Methoden der KI weiter. Die Entwicklung von ELIZA durch Joseph Weizenbaum im Jahr 1966 generierte initiales Interesse an Chatbots [Weizen 1966]. Die nächsten größeren Fortschritte sollten allerdings erst circa 30 Jahre später mit der Artificial Intelligence Markup Language (AIML) gemacht werden, welche die Konversationsregeln für den Chatbot A.L.I.C.E. und folgende Chatbots spezifizierte [Wallac 2003]. Darauffolgend kamen zunehmend neuronale Netze zum Einsatz, um Assistenten wie Siri, Google Assistant oder Cortana zu betreiben. Auf Grundlage des Transformer-Prinzips von Google, einer speziellen Ausführung eines KNN [VaShPa+ 2017], wurden zuletzt die heute bekannten Large Language Models (LLM) wie ChatGPT entwickelt [LiSaPo+ 2018].

2.2 Künstliche Neuronale Netze

Künstliche neuronale Netze stellen einen Meilenstein in dem Bereich der KI dar. Durch Nachahmung der Neuronen im menschlichen Gehirn können sie durch spezielle Algorithmen komplexe Aufgaben bearbeiten. Die Idee der Simulierung von Neuronen stammt von McCulloch und Pitts aus dem Jahr 1943 [CulPit 1943] und die Verbindung dieser künstlichen Neuronen wurde erstmals mit dem Perceptron von Rosenblatt gefestigt [Rosenb 1958]. Unterschiedliche Ausführungen von KNN werden im Folgenden erläutert, die in verschiedenen Bereichen wie beispielsweise der Bildanalyse und -erkennung, Sprachanalyse, Musikanalyse, Emotionserkennung und weiteren Bereichen eingesetzt werden [KrSuHi 2012], [ChCeHa 2017], [Apple 2017], [OrDiZe+ 2016].

2.2.1 Aktuelle Architekturen

Convolutional Neural Networks (CNN) sind eine der ersten Ausführungen von KNN. Mit Vorläufern wie dem Neocognitron von Fukushima [Fukush 1979] oder dem TDNN von Waibel [WaHaHi+ 1989] entwickelte LeCun 1989 die Ausführung des CNN [LeBoDe+ 1989]. CNN werden häufig für die Bildanalyse genutzt. Sie bestehen aus einem Input-Layer, mehreren Hidden-Layern und einem Output-Layer (siehe Abb. 4). Besonders das in [KrSuHi 2012] vorgestellte CNN AlexNet erlangte 2012 große Aufmerksamkeit, da es den Fokus auf Deep Learning intensiviert hat.

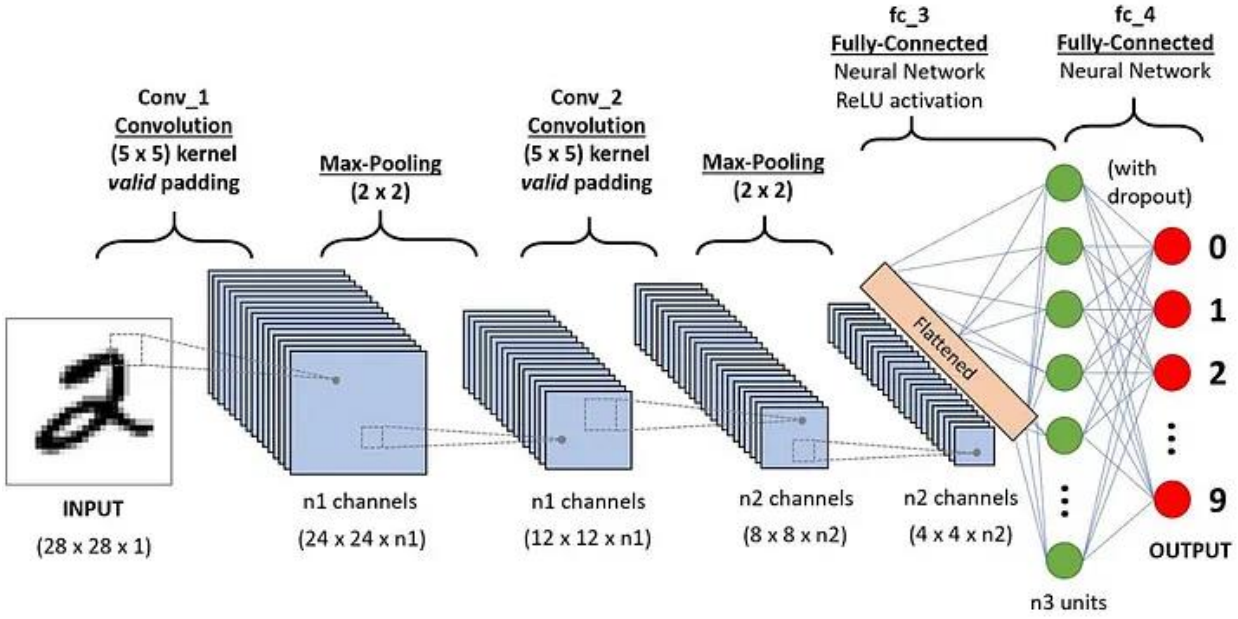

Abbildung 4: Beispiel einer CNN-Architektur [Dharma 2022]

Virtuelle Assistenten wie Apples Siri, Amazons Alexa, Microsofts Cortana oder Google Assistant haben Gebrauch von Dialogsystemen gemacht, die als zugrundeliegende Technologie die Architektur des Recurrent Neural Network (RNN) und dessen Variante Long Short-Term Memory nutzen, um Sprache effektiv zu verarbeiten [ChCeHa 2017]. Diese Architekturen zeichnen sich durch ihre Spezialisierung auf sequenzielle Daten aus. Zusammen mit anderen Architekturen wie beispielsweise Deep Neural Networks, die menschliche Sprache erkennen und umwandeln [Apple 2017], und CNNs mit DeepMinds WaveNet, die Akustik und menschliche Sprache generieren [OrDiZe+ 2016], bilden sie die Basis moderner Assistenten. Aktuelle Ansätze verwenden das 2017 von Google vorgestellte Transformer-Prinzip, das mit Aufmerksamkeitsmechanismen arbeitet, um menschliche Sprache zu decodieren, Kontext zu verstehen und Anweisungen zu interpretieren [VaShPa+ 2017].

2.2.2 Das Transformer-Prinzip

Mit der Forschungsarbeit „Attention Is All You Need" [VaShPa+ 2017] führte Google ein revolutionäres Prinzip der KNN ein. Durch die Implementation von sogenannten Self-Attention-Mechanismen kann das Transformer-Modell von Google menschliche Sprache effektiver decodieren und interpretieren. Attention ist eine Methode des maschinellen Lernens (ML), die die Relevanz von Komponenten in einer Eingabe, auch Token genannt, beschreibt [BaChBe 2014]. Modelle können die Relevanz von Token erkennen, indem sie die sogenannten Attention-Weights jedes Token berechnen. Der Mechanismus ermöglicht also eine selektive Auswahl von Token in der Eingabe. Das bedeutet, dass mit dem Attention-Mechanismus nicht mehr alle Informationen aus einer Eingabe in einen einzigen Vektor komprimiert werden müssen, sondern dass jede Komponente eigene Vektoren bekommt, die eine kontextualisierte Repräsentation der Informationen darstellen [BaChBe 2014]. Schon vor der Transformer-Architektur kamen Attention-Mechanismen zum Einsatz – meist sequenziell in RNNs, seltener in alternativen Varianten (z. B. [PaTäDa+ 2016]). Dadurch sind sie sehr ressourcenintensiv, da sie den Input ausschließlich nacheinander verarbeiten können. Die Neuheit des Transformers ist, dass dieser vollständig mit dem Attention-Mechanismus arbeitet, auf Convolutional- und Recurrent-Layer verzichtet und damit eine Parallelisierung erlaubt. Der Transformer nutzt den abgewandelten Self-Attention-Mechanismus, der es dem Modell erlaubt, die Bedeutung und Relevanz einzelner Elemente aus der Eingabe untereinander abzuwägen. Jeder Token kann nun auf alle anderen Token achten und dadurch komplexe Kontexte abbilden. [VaShPa+ 2017]

Darüber hinaus arbeitet der Transformer mit einer Encoder-Decoder-Architektur, die in Abbildung 5 zu sehen ist. Dabei stellt die linke Seite den Encoder dar, in dem in einem Multi-Head-Sub-Layer einmalig die Eingabe kodiert und daraus kontextuelle Informationen für den Decoder generiert werden. Dieser berechnet zunächst positionelle Kodierungen der Eingabetoken und weist jedem Token drei Vektoren zu: Query (Q), Key (K) und Value (V). Dies ist die Basis des Attention-Mechanismus, der den Kontext zwischen den Token ermöglicht. Die Gewichtung (Weights) dieser Vektoren wird wie bei anderen neuronalen Netzwerken im Laufe des Trainings gelernt. [VaShPa+ 2017]

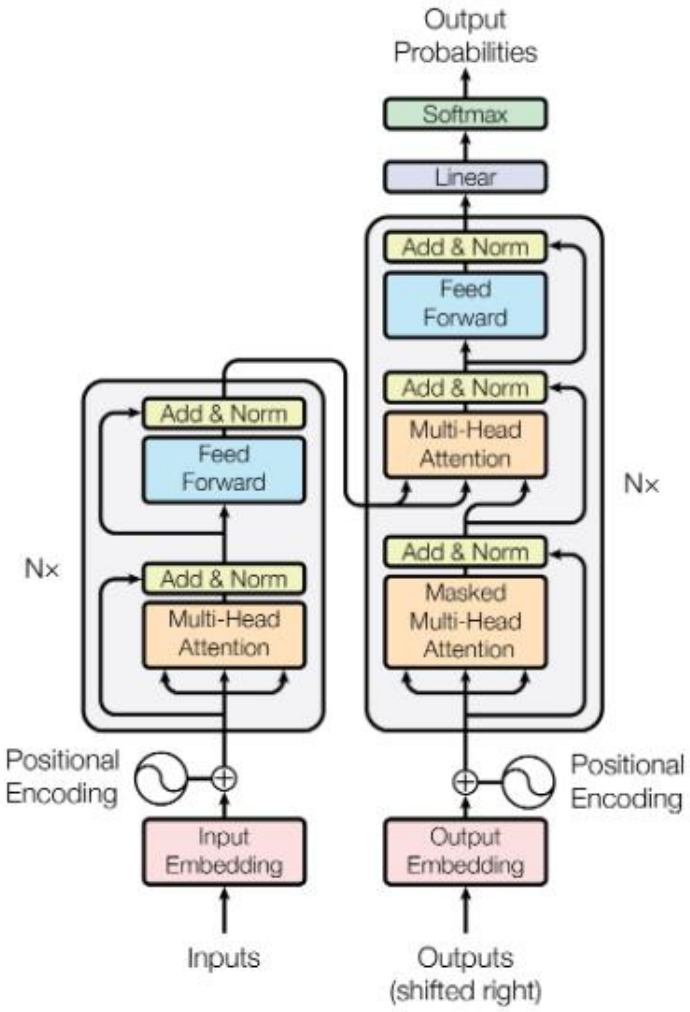

Abbildung 5: Der Transformer – Modellarchitektur [VaShPa+ 2017, S. 3].

Wie in Abbildung 6 illustriert, werden die Q- und K-Vektoren aller Eingabetoken mittels des Skalarproduktes miteinander multipliziert und danach aufgrund der potenziell großen Werte durch die Division mit der Wurzel der Vektordimension herunterskaliert. Daraufhin kommt die aus dem Machine Learning (ML) bekannte Softmax-Funktion zum Einsatz, um die hohen Werte hervorzuheben und die niedrigen Werte zu vernachlässigen. Die Ergebnisse werden mit den V-Vektoren der Eingabetoken multipliziert, um die Beziehungen der Wörter untereinander als Matrix zu erhalten [VaShPa+ 2017].

Dieser Prozess als Formel aus [VaShPa+ 2017] lautet wie folgt:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (1)$$

Formel 1: Berechnung der Scaled Dot-Product Attention

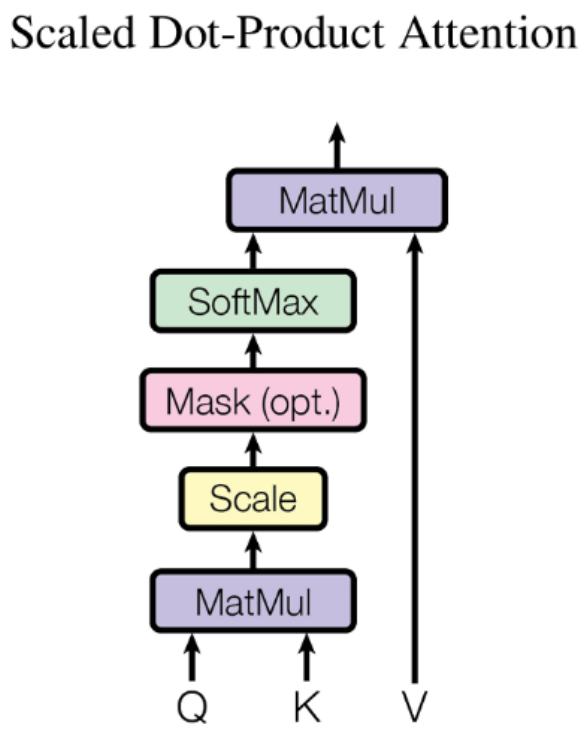

Abbildung 6: Scaled Dot-Product Attention [VaShPa+ 2017, S. 4].

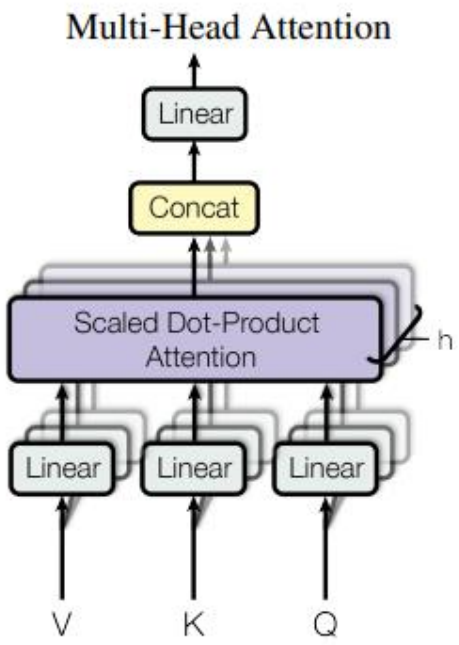

Abbildung 7: Multi-Head Attention mit mehreren parallelen Layern [VaShPa+ 2017, S. 4].

Multi-Head-Attention (siehe Abb. 7) bedeutet, dass diese Operationen h-mal parallel ausgeführt werden, um pro Head verschiedene Beziehungen durch unterschiedliche Weights der Vektoren zu erhalten. Danach wird jeder Vektor in der Ausgabe einzeln durch ein Feed-Forward-Netzwerk (FFN) gegeben und durch dessen Weights mit folgender Formel aus [VaShPa+ 2017] nochmals transformiert und angereichert:

$$FFN(x) = max(0, xW_1 + b_1)W_2 + b_2 \quad (2)$$

Formel 2: Aktivierungsfunktion des FFN

Das FFN fungiert im Transformer als eine Art Gedächtnis gelernter Beziehungen der Token [GeScBe+ 2021].

Der Decoder nutzt im Gegensatz zum Encoder zwei Multi-Head-Sub-Layer. Diese funktionieren wie im Encoder, mit dem Unterschied, dass ein sogenanntes Masking angewandt wird. Dabei werden die Werte aus der Attention-Matrix von Positionen, die größer als die aktuelle Position i sind, auf den Wert $-\infty$ gesetzt, um den Decoder daran zu hindern, auf nachfolgende Token zu achten und sich ausschließlich auf vorherige Token zu beziehen. Das zweite der Multi-Head-Sub-Layer erhält die Werte für die K- und V-Vektoren aus dem FFN des Encoders, um Rückbezug auf das „Gedächtnis“ zu erhalten. Danach wird auch hier wieder ein FFN auf die Ausgabe angewandt.

Zum Abschluss werden die Vektoren durch einen linearen Layer geleitet, der die Vektoren jeweils in einen Logits-Vektor mit der Vokabulargröße des Modells umwandelt. Dieser bildet die rohen Werte aus dem letzten Layer und damit die Einschätzung des Modells für jeden Token ab. Mit einer weiteren Softmax-Funktion wird der Logits-Vektor normalisiert und die Wahrscheinlichkeiten für die Token werden berechnet. Der letzte Vektor korrespondiert dabei mit dem nächsten Token. [VaShPa+ 2017]

2.3 Kontextsensitive LLMs

Damit LLMs spezialisierte Fragen beantworten können, müssen sie darauf trainiert werden oder extern mit Informationen unterstützt werden. Dafür gibt es verschiedene Möglichkeiten, die je nach Anwendungsfall eingesetzt werden können. Eine dieser Möglichkeiten umfasst das Fine-Tuning von Modellen. Dabei wird ein vortrainiertes Modell auf einen spezifischen, auf den Einsatzbereich zugeschnittenen Datensatz trainiert, damit es insbesondere in diesem Bereich Wissen erlangt. Bei dieser Methode werden alle Parameter des Modells aktualisiert, um das neue Wissen zu festigen. Dies erwies sich allerdings in den letzten Jahren als problematisch, da die Modelle deutlich größer wurden und mehr Parameter enthielten. Daher wurde 2021 die Methode Low-Rank Adaptation (LoRA) vorgestellt, die es ermöglicht, Parameter des vortrainierten Modells einzufrieren und lediglich einen kleinen Teil dieser neu zu trainieren. Dafür werden zwei Matrizen in jedem Dense-Layer des Modells eingesetzt und dem Training unterzogen. Damit können die trainierbaren Parameter häufig um den Faktor 10.000 reduziert werden [HuShWa+ 2021]. Diese Methode wird in Bezug auf Chatbots häufig angewandt, um ein spezialisiertes LLM zu trainieren, das zunächst die Nutzereingabe vorverarbeitet, um dem Hauptmodell, das die Anfrage beantwortet, mehr Kontext zu bieten [RanYin 2025].

Oft wird dazu zusätzlich die Architektur der Retrieval-Augmented Generation (RAG) angewandt, die mit der Nutzereingabe externe Informationen abruft und diese in den Kontext des Modells einspeist, um damit präzise Antworten zu generieren. Dazu werden häufig Vektordatenbanken verwendet, die mit Hilfe von semantischer Suche bedeutungsähnliche Informationen basierend auf der Nutzereingabe ausgeben können [GaXiGa+]. Dabei kommt ein sogenanntes Embedding-Modell zum Einsatz, das die Daten für die Vektordatenbank zunächst in Vektoren transformiert, um mit dem Embedding der Nutzereingabe eine Ähnlichkeitssuche durchzuführen. Embedding-Modelle sind darauf trainiert, unterschiedliche Datentypen wie Text, Bilder oder Videos in eine einheitliche Repräsentation zu übersetzen. In der Vektordatenbank können dabei mehrere Kollektionen (Collections) enthalten sein, die aus unterschiedlichen Bereichen mit Informationen

gefüllt werden (siehe Abb. 8). Diese Informationen werden vorverarbeitet, strukturiert aufgeteilt und anschließend als sogenannte Dokumente in eine Collection hinzugefügt. Die Embeddings setzen sich aus den sogenannten Dense- und Sparse-Vektoren zusammen. Die Dense-Vektoren bilden die Sinnhaftigkeit der Informationen ab, während mit den Sparse-Vektoren ein Keyword Matching betrieben werden kann. Mit der Ähnlichkeitssuche werden die Embeddings der Query mit denen der Dokumente verglichen. Für die Embeddings können die verschiedenen Vektoren einzeln oder zusammen genutzt werden. Wenn beide Vektoren verbunden werden, wird das als hybride Suche bezeichnet, da sowohl nach der semantischen als auch nach der direkten Bedeutung oder nach Begriffen gesucht wird. [Qdrant 2025]

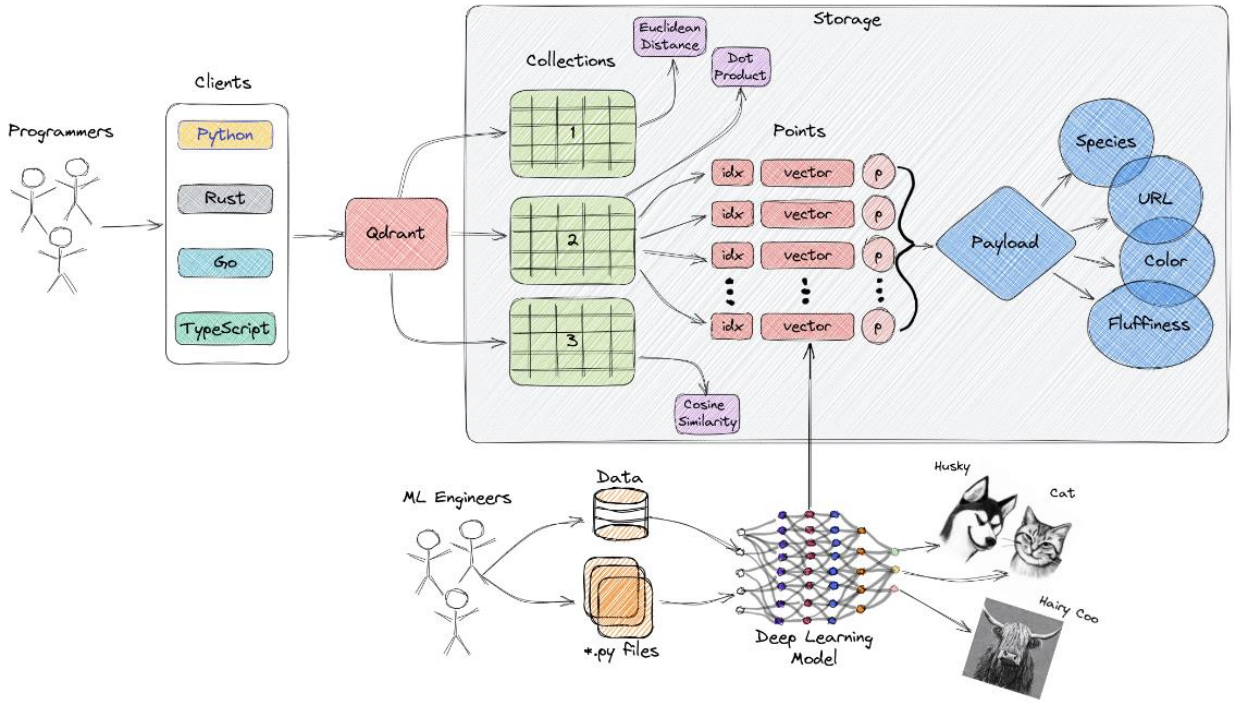

Abbildung 8: High-Level Overview of Qdrant's Architecture [Qdrant 2025]

Mittlerweile existieren verschiedene Stufen von RAG-Architekturen, die unterschiedliche Ergebnisse liefern. Die Naive-RAG-Architektur ist die einfachste Ausführung. Bei dieser wird lediglich die Nutzereingabe für die Vektorsuche verwendet und die gefundenen Dokumente werden dem LLM als Informationen zur Verfügung gestellt. Die Advanced-RAG-Architektur erweitert diese Basis um Methoden wie das Neubewerten (reranking) und Fusionieren (fusion) von Dokumenten. Zuletzt implementiert die Modular-RAG-Architektur komplexe Methoden, um rekursiv die bestmöglichen Ergebnisse für die Nutzereingabe zu finden. [GaXiGa+]

2.4 Restriktionen kleiner und mittelständischer Unternehmen

KMU verfügen häufig über begrenzte Ressourcen, weil sie im Vergleich zu Großunternehmen kleinere Budgets, weniger spezialisiertes IT-Personal und oft auch eingeschränktere Datenbestände haben. Dadurch entstehen hohe Aufwände für Infrastruktur, Systemintegration, Betrieb und Compliance, die sie nur selektiv stemmen können. Dies trifft auch auf die Integration von LLM oder Chatbots zu. Damit fortschrittliche Modelle mit Parametern im Milliardenbereich betrieben werden können, braucht es entweder entsprechende Hardware oder Endpunkte von Anbietern, die sich auf Inferenz von KI-Modellen spezialisiert haben. Bekannte Beispiele sind Huggingface Endpoints [Huggin 2025a], Amazon Sagemaker [Amazon 2025a] oder Cloudflares Workers AI [Cloudf 2025]. Bei diesen Anbietern wird oft nach der Laufzeit der Instanz abgerechnet. Zudem gibt es Anbieter, die pro Anfrage oder pro Token-Anzahl abrechnen, wie OpenAI mit ihren GPT-Modellen [aloa 2025]. Bei gehosteten Anbietern oder APIs muss allerdings die Frage nach dem Datenschutz beantwortet werden, falls persönliche Daten verarbeitet werden sollen. Dieses Thema wird im weiteren Verlauf der Arbeit erneut aufgegriffen.

Auch die Implementierung eines solchen Modells kann eine Herausforderung darstellen. Bei der Nutzung der aufgeführten Dienste gestaltet sich die Einrichtung zunächst unkompliziert, da lediglich ein verfügbares Modell ausgewählt und der Dienst gestartet werden muss. Danach wird der zur Verfügung gestellte Endpunkt angesprochen und das vom Modell vorgegebene Format verwendet, um Informationen korrekt zu übergeben. Bei der Entwicklung eines selbst gehosteten Modells sind allerdings fachkundige Mitarbeitende notwendig, die die Zeit für die Implementierung und Wartung des Modells aufwenden können. Eine andere Möglichkeit ist es, externe Fachkundige oder Dienstleister zu beauftragen. Die Kosten dafür sind allerdings oft signifikant höher und dieser Ansatz bietet weniger Kontrolle über die Rahmenbedingungen des Projekts.

3 Konzept

Diese Bachelorarbeit orientiert sich an bereits existierenden Implementationen und Methoden. Das folgende Kapitel erläutert verwandte Literatur und verdeutlicht, inwiefern diese Implementation von bisherigen Ausführungen abweicht. Zusätzlich werden die Architekturen als Konzepte vorgestellt und verwendete Modelle erläutert.

3.1 Hintergrund

Kontextbasierte LLMs können unter anderem mit den im vorherigen Kapitel beschriebenen Methoden realisiert werden. Diese beiden Methoden werden oft kombiniert, um einen präzisen Kontext für das LLM, das die Nutzereingabe bearbeitet, bereitzustellen. [ViRaDh 2024] und [BaBeCu 2024] nutzen beispielsweise die Vorteile des Fine-Tuning, um das Hauptmodell zunächst auf den Anwendungsbereich zu spezialisieren, und unterstützen anschließend mit einer Vektordatenbank. Diese sorgt dafür, dass eine breite Menge an Nutzeranfragen effektiv verarbeitet werden kann. [RanYin 2025] stellt zunächst ein System dar, in dem die Nutzereingabe in eine Vektordatenbank gegeben und gleichzeitig von einem fine-tuned LLM vorverarbeitet wird. Diese beiden Ergebnisse daraus werden zusammengeführt und in den Kontext des eigentlichen LLMs übergeben. Dadurch bekommt dieses die bestmöglichen Informationen, um die Nutzereingabe präzise zu verarbeiten und zu beantworten. Da dieses spezialisierte, vorgeschaltete LLM zusätzliche Latenz zu einer normalen Anfrage bringt, ist zu beachten, in welchen Situationen diese Architektur verwendet werden soll. Bei der Interaktion mit der Kundschaft sind zu lange Antwortzeiten von mehreren Sekunden oft bereits unvorteilhaft für die Effektivität des Einsatzes von Chatbots [GnMoAd$^+$ 2018]. Dieses Phänomen zeigt sich ebenfalls bei den Ladezeiten von Webseiten, bei denen Nutzende bereits nach zu langer Ladezeit der Webseite diese frühzeitig verlassen [SheNee 2016]. Deshalb bieten Small Language Models (SLMs) eine effektivere Methode, um das vorgeschaltete Modell zu implementieren. Auch [ZhLiPa 2024] betont, dass SLM einen Vorteil in der Genauigkeit von Ausgaben der LLM hervorbringen können. Durch das Vorverarbeiten von der Nutzereingabe oder das Hinzufügen von relevanten Informationen haben die LLMs eine verbesserte Basis, um nicht zu halluzinieren. LLMs neigen zu Halluzinationen, wenn das für eine Anfrage nötige Wissen nicht ausreichend in den Modellparametern verankert ist, wodurch sie plausibel klingende, aber falsche Aussagen erzeugen [HuYuMa$^+$ 2023].

Obwohl SLM gewisse Vorteile bringen können, wird in [RanYin 2025] ein anderes System implementiert, das unter anderem mit Nutzerfeedback arbeitet und damit das Embedding-Modell trainiert, welches die Daten aus der Vektordatenbank ausgibt. Die Verfassenden fokussieren sich aufgrund von Effizienz auf die von ihnen vorgestellte Methode „Quantized Influence Measure“, welche als „KI-Judge“ fungiert, um gefundene Dokumente nochmals zu bewerten. Dadurch kann das Hauptmodell die Fragen mit den passenden Informationen beantworten. Effizienz und Ressourcenschonung sind wichtige Faktoren für kleine und mittelständische Unternehmen. Diese besitzen nicht die nötigen Ressourcen, um LLMs mit vielen Parametern zu betreiben. Daher ist es notwendig, kleine, kompakte Modelle wie in [ViRaDh 2024] zu nutzen. Dadurch wurde sich für ressourcenschonende Architekturen entschieden, die im Folgenden erläutert werden.

3.2 Architekturen

Da die Architektur so effizient wie möglich arbeiten soll, um der Kundschaft in Echtzeit eine Antwort liefern zu können, muss untersucht werden, wie hoch die Latenzen bei bestimmten Methoden sind. Aus diesem Grund werden im Folgenden zwei unterschiedliche Architekturen verfolgt und in der Implementation gegenübergestellt. Der grundlegende Aufbau bleibt bei beiden Ansätzen gleich und besteht aus einem Proxy, der Anfragen annimmt, die Vektorsuche durchführt und Anfragen an das LLM schickt. Das LLM liegt auf einem Server des Inferenzanbieters Huggingface. Huggingface besitzt den Dienst Endpoints, der speziell für die Inferenz von LLMs gedacht ist. Die Endpoints sind APIs, welche von Unternehmen oder Individuen genutzt werden können, um Modelle zu deployen, ohne eine eigene Serverinfrastruktur aufzubauen [Huggin 2025a]. Es wird bewusst auf das Fine-Tuning eines Modells verzichtet, da es ressourcenintensiv und ineffektiv bei modularen Daten, die sich stetig ändern, ist. Mit solchen modularen Daten wird in diesem Unternehmen gearbeitet.

Der erste Ansatz arbeitet lediglich mit einer Vektordatenbank und deren Reranking-Funktionen, um den bestmöglichen Kontext darzustellen. Zudem wird ein Influence Scoring eingebaut, bei dem die jeweiligen Collections der Vektordatenbank bestimmte Keywords zugewiesen bekommen. Diese Keywords können beliebig erweitert werden und sorgen dafür, dass manche Collections höher als andere bewertet werden, wenn eine Nutzeranfrage die vorher definierten Keywords enthält. So kann der Vektorkontext, basierend auf Datenauswertung der laufenden Anfragen von Kunden, in eine bestimmte Richtung gelenkt werden. Dadurch, dass alle Collections im Kontext des LLM vorkommen können, besitzt es die wichtigsten Informationen, um die Anfrage kollektionsübergreifend zu beantworten.

Der zweite Ansatz baut auf dem ersten auf, verfolgt zusätzlich aber das Prinzip eines vorgeschalteten Modells, das die Historie der Nutzeranfragen verarbeitet, um auch der Vektordatenbank einen gewissen Kontext zu geben. Dadurch können auch Nachfragen von Nutzern effektiv verarbeitet werden, da sich die Suche nicht mehr ausschließlich auf die aktuelle Anfrage, sondern auch auf vorherige Anfragen bezieht. Das Modell soll entscheiden, ob sich die Anfragen aufeinander beziehen, und im Falle dessen eine zusammengefasste Aussage ausgeben. Außerdem wird ein Modell parallelgeschaltet, das die Anfrage auf schädliche Inhalte überprüft und die Hauptanfrage zur Not unterbricht und den Nutzenden auf die Nutzungsrichtlinien hinweist. Schädliche Anfragen umfassen zum Beispiel Rollenspiele, bei denen dem Modell die Anweisung gegeben wird, sich unabhängig seiner Richtlinien zu verhalten oder Informationen zu generieren, die für schädliche Zwecke genutzt werden. Dies nennt sich auch Jailbreaking [ShChBa$^+$ 2023]. Damit einher geht auch das Risiko, dass schädliche Antworten des Chatbots gegen das Unternehmen genutzt werden können. Dieser Ansatz stellt sicher, dass das Modell sowohl kontextsensitive als auch unschädliche Antworten liefert.

Insgesamt ist der zweite Ansatz als robuster und einsatzfähiger zu bewerten, da er weniger anfällig für Kontextverlust, Jailbreaking oder Verbreiten von falschen Informationen ist. Die geringfügig höhere Latenz kann damit ausgeglichen werden, dass Token sofort dem Nutzenden weitergeleitet werden, während das LLM weitere Token generiert, anstatt die gesamte Nachricht nach einigen Sekunden auf einmal anzuzeigen. Das vermittelt außerdem das Gefühl, eine echte Konversation zu führen.

3.3 Verwendete Modelle

In dem Projekt werden zwei LLMs verwendet. Einerseits das Hauptmodell, das die Aufgabe hat, die Anfragen der Nutzenden zu beantworten. Andererseits das Embedding-Modell, welches die Vektor-Embeddings für die Dokumente und die Nutzeranfragen generiert. Als Hauptmodell wurden zunächst einige Modelle betrachtet. Das erste Modell war das LeoLM mit 13 Milliarden Parametern, was speziell auf deutsche Datensätze trainiert wurde [Plüste 2023]. Dieses erwies sich jedoch als ungeeignet, da es auf Llama 2 basierte, somit bereits veraltet war und im Vergleich mit neueren Modellen schlechtere Antworten gab. Auch Qwen3 [Qwen 2025] oder Mistral NeMo [Mistra 2025] wurden betrachtet, da beides kleinere Modelle sind, die trotzdem hohe Qualitäten aufweisen.

Letztendlich wurde sich aber für das in diesem Jahr veröffentlichte Gemma 3 mit zwölf Milliarden Parametern von Google entschieden, da es in Befolgen von Anweisungen punktet und sehr anpassungsfähig ist. Das Modell ist zusätzlich „instruction-tuned“, um effektiv Anweisungen zu folgen. Darüber hinaus ist es auf eine 4-Bit-Präzision quantisiert worden [Google 2025]. Bei der Quantisierung wird die Repräsentation der Gewichte des Modells auf eine kleinere Bitanzahl beschränkt und so die Speicher- und Rechenanforderungen reduziert. Google nutzt bei diesem Modell das sogenannte Quantization-Aware-Training (QAT), was dabei hilft, die Genauigkeit und Qualität des Modells zu behalten, obwohl die Präzision der Gewichte verringert wird. Normalerweise werden Modelle erst nach dem Training quantisiert, bei QAT wird bereits im Training mit quantisierten Daten gearbeitet [JaKlCh+ 2017]. Das Modell basiert auf der in [VaShPa+ 2017] vorgestellten Transformer-Architektur und unterstützt einen Kontext von 128.000 Token [KaFePa+ 2025].

Allerdings arbeitet auch dieses Modell, wie fast alle Conversational-LLMs wie GPT, LLama oder Mistral, mit einer sogenannten „Decoder-only“-Architektur. Diese Architektur bringt den Vorteil, dass sie effektiver skaliert werden kann, da kein Encoder benötigt wird und damit fast die Hälfte der Parameter wegfällt [LiSaPo+ 2018]. Sie besteht lediglich aus Masked-Self-Attention-Mechanismen in dem Decoder-Block (Abb. 4), die, wie in Kapitel 2.2.2 erläutert, sicherstellen, dass sich das Modell für den nächsten Token ausschließlich auf vorangegangene Token bezieht. Außerdem nutzt es Grouped-Query-Attention, was für effizientere Inferenzen sorgt, indem es die Query-Heads gruppiert und weniger Key-Value-Heads zuordnet, und somit weniger Overhead produziert [AiLeJo+ 2023]. Für dieses Modell wurde sich entschieden, weil es eine sehr gute Performance bei fast gleichbleibender Qualität auch bei niedriger Präzision und Parameteranzahl bietet. Mit zwölf Milliarden Parametern lässt es sich erfahrungsgemäß mit akzeptabler Latenz auf einer Nvidia A10G GPU mit 24GB VRAM betreiben.

Das Embedding-Modell ist das BGE-M3 von der Beijing Academy of Artificial Intelligence. Dieses Modell wurde gewählt, weil es sowohl mehrere Sprachen unterstützt als auch Dense- und Sparse-Vektoren nativ generieren kann. Zusätzlich unterstützt es einen Kontext von 8129 Token, was dabei hilft, auch lange Produktdaten als gesamtes Dokument in die Vektordatenbank aufzunehmen [BGE-M3 2024]. Andere Modelle wie das Jina-Embeddings-V2-Base-De sind spezifisch auf Deutsch und Englisch trainiert, können aber nativ keine Sparse-Vektoren generieren [JinaV2 2024].

4 Implementierung

Die Implementierung stellt die Grundlage dieser wissenschaftlichen Arbeit dar, da in diesem Kapitel entschieden wird, welche Programmiersprache, welches Framework und welches LLM-Modell angewandt werden. Alle Bestandteile der Implementierung werden im Folgenden dargestellt und es wird erläutert, warum diese verwendet wurden.

4.1 Entwicklungsumgebung

Die Entwicklungsumgebung teilt sich in zwei Bereiche auf: die Testumgebung und die Produktionsumgebung. Die Testumgebung setzt sich aus einer einfachen Python-Umgebung mit Jupyter Notebooks zusammen. Dies diente dazu, sich mit den Methoden vertraut zu machen. In dieser Phase wurde außerdem mit verschiedenen LLM-Modellen experimentiert, um das geeignetste zu identifizieren. Diese wurden dabei auf einem lokalen Rechner mittels PyTorch aufgesetzt und über das lokale Netzwerk angesprochen.

Die Basis der Produktionsumgebung ist anders als die Testumgebung aufgebaut. Der Proxy wird in der im Unternehmen bevorzugten Programmiersprache Rust aufgesetzt. Obwohl Python die präferierte Sprache für den Bereich des Machine Learning ist, sind zunehmend auch Alternativen verfügbar. Eine davon ist die Nutzung von sogenannten Open Neural Network Exchange (ONNX)-Modellen in Rust. Zudem zeigen Benchmarks wie [Šuboni 2024], dass Rust teilweise signifikante Leistungs- und Geschwindigkeitsvorteile gegenüber Python bietet. Aufgrund genannter Gründe und der grundlegenden Architektur von Rust wird Rust als effiziente Alternative bewertet. Der Nachteil von Rust ist die fehlende Adaption von bekannten Python-Frameworks, die für Anwendungsfälle wie Chatbots verwendet werden. Aufgrund dessen wurde sich zunächst nicht intensiv mit Alternativen im Ökosystem von Rust beschäftigt und die Architektur eigenständig entwickelt. Letztendlich wurden Frameworks wie LangChain oder LlamaIndex auf ihren Nutzen untersucht, aber aufgrund ihrer hohen Abstraktion von Vorgängen und der nötigen Anpassung der bereits existierenden Codebase nicht verwendet. Stattdessen wurden die Vorgänge wie das Laden des Embedding-Modells, das Generieren der Embeddings, die Vektorsuche und die Abwicklung der Anfrage über den Inferenz-Endpunkt selbst implementiert.

Ein weiterer Hauptbestandteil stellt die Vektordatenbank dar, die mittels des Vektordatenbank-Tools Qdrant realisiert wurde. Qdrant ist auf Performance optimiert und in Rust geschrieben. Aus diesem Grund wurde das Tool für die Vektordatenbank ausgewählt. Qdrant unterstützt außerdem die Hybrid-Search, bei der sowohl Dense- als auch Sparse-Vektoren ausgegeben werden, um das bestmögliche Ergebnis zu erhalten. Zusätzlich lassen sich in Qdrant Filter einbauen, die ein Suchergebnis weiter präzisieren können [Qdrant 2025].

4.2 Prozessabläufe

Damit der Chatbot als Ganzes funktionieren kann, müssen die verschiedenen Prozesse miteinander verbunden werden. Dabei teilt sich die Architektur in drei Grundkomponenten auf. Das Frontend beinhaltet eine mit Alpine.js realisierte Chatfensterlösung, die durch das Interface Anfragen der Nutzenden an das Backend leitet und entsprechend der Antworten des LLM anzeigt. Außerdem werden die Aktivitäten der Nutzenden nach Akzeptieren der Cookies mittels des lokalen Speichers des Browsers aufgezeichnet, um so an verwertbare Daten zu kommen. Der Chat-Button wird auf jeder Seite unten rechts in der Ecke angezeigt und begrüßt den Kunden bei den ersten zwei Seitenaufrufen. Darüber hinaus ist der Button animiert, um das Engagement zu steigern.

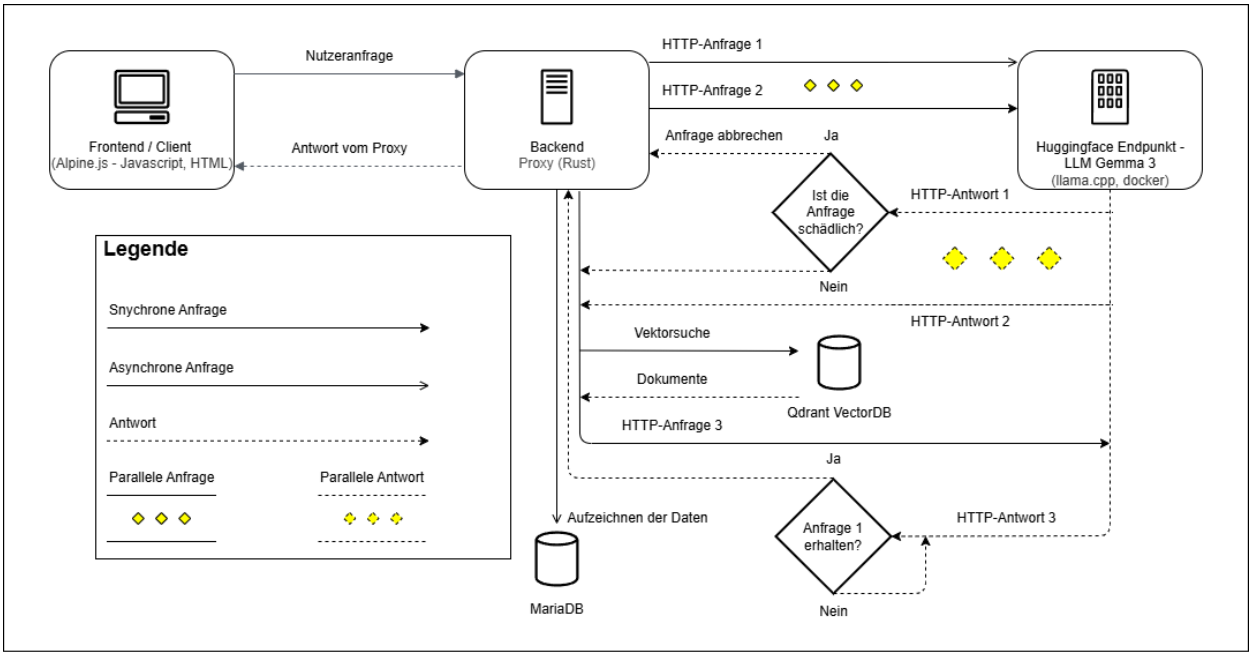

Abbildung 9: Prozessabläufe in der angewandten Architektur (zweites Konzept)

Wie in Abbildung 9 zu erkennen, beinhaltet das Backend den in Rust geschriebenen Proxy, der die Anfragen des Frontends annimmt, sie verarbeitet und an den LLM-Endpunkt weiterleitet. Zusätzlich ist der Proxy für das Aufzeichnen von Nutzerinteraktionen zuständig. Wenn Nutzende den Chat-Button im Frontend oder einen Link, der vom Chatbot bereitgestellt wurde, öffnen, wird dies in einer MariaDB-Datenbank gespeichert. Außerdem werden alle Konversationen sowie die Ergebnisse aus der Vektorsuche mit dem korrespondierenden Unique User Identifier (UUID), der nach der ersten Konversation für jeden Nutzenden generiert wird, in der Datenbank festgehalten.

In dem ersten Konzept der Architektur wird die Nutzeranfrage direkt für die Vektorsuche verwendet und der entstandene Kontext mit dem vorgefertigten Prompt an den LLM-Endpunkt gesendet. Diese Antwort wird anschließend im Proxy formatiert, indem zum Beispiel Links in die richtige HTML-Formatierung gebracht, danach wieder an das Frontend gesendet und dort abgebildet werden. Das in Abbildung 9 abgebildete zweite Konzept schickt zwei parallele Anfragen an den Endpunkt. Die erste beinhaltet einen System-Prompt, der dem Modell die Anweisung gibt, die aktuelle Nutzeranfrage auf schädliche Inhalte zu untersuchen. Wenn das Modell die Anfrage als schädlich einordnet, wird der gesamte Prozessablauf unterbrochen und dem Nutzenden wird eine Fehlermeldung ausgegeben. Diese Anfrage ist asynchron und muss erst am Ende ausgewertet werden. Die zweite Anfrage enthält die letzten fünf Nutzeranfragen sowie einen System-Prompt, welcher das Modell anweist, zu untersuchen, ob diese miteinander zusammenhängen. Wenn dies der Fall ist, soll das Modell eine aufbereitete Nutzeranfrage zurücksenden, die die Beziehungen zwischen den vorherigen Nutzeranfragen abbildet und zusammenfasst. Diese wird dann statt der originalen Nutzeranfrage für die Vektorsuche verwendet, sodass der Vektorsuche ein Kontext bereitsteht, der dafür sorgt, dass passende Dokumente abgerufen werden können. Danach findet derselbe Prozess wie bei dem ersten Konzept statt, nur ist die Hauptanfrage in dieser Ausführung Anfrage 3, die an den Huggingface Ednpunkt gesendet wird. Sollte die Antwort auf Anfrage 1, anch Antwort auf Anfrage 3 noch nicht angekommen sein, wird so lange gewartet, bis diese angekommen ist (siehe Abb. 9).

Der Endpunkt für das LLM wird über den Anbieter Huggingface gehostet. Dabei wird, sofern verfügbar, die Nvidia A10G GPU mit 24 GB VRAM für einen US-Dollar die Stunde verwendet. Der Endpunkt ist so gestaltet, dass lediglich ein auf Huggingface existierendes Modell ausgewählt, Grundkonfigurationen eingestellt und der Endpunkt gestartet werden muss. Danach können Anfragen direkt in einem bestimmten JSON-Format [W3Scho 2025] durchgeführt werden. Im Hintergrund läuft der Standard-Container mit der Bibliothek und Runtime llama.cpp. Die Runtime

hat die Aufgabe, effiziente Inferenz mit verschiedenen Modellen auszuführen. Aufgrund dessen, dass diese Implementation ein Prototyp ist und er in einem mittelständischen Unternehmen eingesetzt wird, ist der Endpunkt lediglich täglich von 14 bis 20 Uhr aktiv. Während dieser Uhrzeiten sind die meisten Nutzenden im Online Shop. Huggingface bietet die Möglichkeit, den Endpunkt nach einer bestimmten Zeit der Inaktivität, zum Beispiel 15 Minuten, herunterzuskalieren, sodass nicht unnötig Ressourcen verwendet werden. Durch Blockieren von Anfragen im Proxy und Ausblenden des Chatfensters im Frontend wird sichergestellt, dass außerhalb dieser Zeiten keine Anfragen an den Endpunkt gestellt werden [Huggin 2025b].

4.3 Herausforderungen bei der Implementation

Mit der Implementierung eines Chatbots gehen einige Herausforderungen einher. Dazu zählen auch spezifische Szenarien, die vor dem Betrieb nicht berücksichtigt werden können. Eine wesentliche Herausforderung stellte die Kontextbeibehaltung der Konversation dar, da der Chatbot bei Anschlussfragen häufig fehlerhaft reagierte. Die Gründe dafür umfassen zum einen einen falschen Kontext aus der Vektorsuche, der dem Modell zur Verfügung gestellt wird, und zum anderen eine Überschreibung alter Nachrichten durch einen zu großen Kontext.

Die ersten Versionen des Chatbots, die die Architektur nutzen, die im ersten Konzept vorgestellt wurde, verwenden lediglich die aktuelle Anfrage, um die Vektorsuche durchzuführen. Das resultierte in Ergebnissen mit unpassenden Informationen, wodurch der Chatbot falsche Antworten gegeben hat. Fragt ein Nutzender beispielsweise etwas zu einem Produkt wie „Ist das Produkt XY vegan?“, wird zunächst der richtige Kontext für Produkt XY ausgegeben. Sollte der Nutzende aber eine Nachfrage wie „Ist es auch glutenfrei?“ stellen, wird der Kontext mit hoher Wahrscheinlichkeit nicht erneut das Produkt aus der ersten Frage enthalten. Dieses Problem wurde in späteren Versionen mittels der zweiten Architektur aus Kapitel 3.2 gelöst, indem vor der Vektorsuche eine weitere Anfrage an den Endpunkt gesendet wird, um den Kontext der Nutzeranfrage zu erfassen und die umformulierte Nutzeranfrage für die Vektorsuche zu nutzen.

Da der Endpunkt als Runtime llama.cpp benutzt, formatiert dieser die Anfragen automatisch in das modellspezifische Prompt-Format (vgl. Abbildung 10 für das verwendete Gemma 3-Format). Dabei ist das sogenannte „Lost in the middle“-Problem von Bedeutung, bei dem relevante Informationen, die im mittleren Teil des Kontextfensters liegen, vom Modell seltener berücksichtigt werden [LiLiHe+ 2023]. Obwohl mittlerweile Methoden entwickelt wurden, die dieses Problem umgehen, ist es in heutigen Modellen immer noch präsent und lässt sich durch eine effektive

RAG-Architektur ausgleichen [GaXiWu+ 2025]. Die Erkenntnis, dass mithilfe der zusammengefassten Anfrage des vorgeschalteten Modells das Hauptmodell vorherige Nachrichten besser bearbeitet, weist auf dieses Problem ebenfalls hin. Dadurch, dass der Prompt so formatiert wird, dass zuerst die Systemanweisungen, danach die Nachrichten und dann der Kontext zur aktuellen Anfrage stehen, sind die Nachrichten im mittleren Teil des Kontextfensters.

Context	Formatting
User turn	<start_of_turn>user
Model turn	<start_of_turn>model
End of turn	<end_of_turn>
Example of discussion:	
User: Who are you? Model: My name is Gemma! User: What is 2+2? Model: 2+2=4.	
Model input:	
[BOS]<start_of_turn>user Who are you?<end_of_turn> <start_of_turn>model My name is Gemma!<end_of_turn> <start_of_turn>user What is 2+2?<end_of_turn> <start_of_turn>model	
Model output:	
2+2=4.<end_of_turn>	

Abbildung 10: Formatting for Gemma IT models [KaFePa+ 2025, S. 4]

Zudem stellte es sich als herausfordernd heraus, Produkte bereitzustellen, die gewisse Inhaltsstoffe nicht enthalten. Letztendlich wurde durch die Nutzung der Filterfunktion von Qdrant, um ein bestimmtes Feld in dem JSON-Payload zu filtern, eine Verbesserung erreicht. Dazu müssen die Produkte allerdings konsistent Felder wie „Tags“ enthalten, die die Eigenschaften und Inhaltsstoffe des Produktes beschreiben. Alternativ müssen eigene Filterfunktionen implementiert werden, die nachträglich die Ergebnisse filtern, was jedoch zu einem qualitativ schlechteren Kontext führt, da Qdrant eine Vektorsuche durchführt, die Filter direkt mitberücksichtigt, um die passenden Dokumente zu finden [AquMyr 2024].

Auch das sogenannte Jailbreaking eines Modells ist ein präsentes Problem. Im ersten Konzept reagierte das Modell meist korrekt auf direkt und indirekt schädliche Anfragen, indem es diese ablehnte. Mit dem Hinweisen auf vorherige Nachrichten wurde, reagierte das Modell anfälliger auf Anfragen, die beispielsweise ein Rollenspiel enthielten. Dabei geht es darum, dass das Modell vorgeben soll, andere Rollen einzunehmen und keine Anweisungen zu befolgen. Als zusätzlicher Sicherheitsmechanismus fungiert die erste Anfrage aus dem zweiten Konzept wie in Kapitel 4.2 beschreiben. Bei Erkennung von schädliche Inhalten in der Anfrage wird die Nutzeranfrage sofort abgebrochen und der Nutzende auf Richtlinien hingewiesen.

5 Evaluation

In diesem Kapitel wird die Wirtschaftlichkeit der entwickelten Chatbot-Implementation betrachtet. Der Schwerpunkt liegt auf den Kosten für die Entwicklung, Wartung und das Hosting sowie einem Vergleich mit aktuellen Angeboten von externen Dienstleistern. Außerdem wird auf die Auswirkungen auf die Kundschaft und auf den Schutz sensibler Daten eingegangen. Als Abschluss wird ein weiterer Einblick und Vergleich zur Industrie durch Experteninterviews und ausgewählte Studien präsentiert.

5.1 Entwicklungs- und laufende Kosten

Die Entwicklung erfolgte über einen Zeitraum von fünf Monaten, wobei ein Monat Vollzeit (circa 106 Stunden), drei Monate mit je 48 Stunden Arbeit und ein Monat mit 42 Stunden Arbeit für das Chatbot-Projekt geleistet wurden. Das entspricht einem Gesamtaufwand von 292 Stunden. Dieser bezieht die Arbeitszeit mit ein, die entrichtet wurde, bevor das Projekt als Grundlage für diese Bachelorarbeit diente. Die erste funktionsfähige Version, die bereits in der Live-Umgebung getestet wurde, konnte nach einem Arbeitsaufwand von schätzungsweise 200 Arbeitsstunden bereitgestellt werden. Danach wurde der Chatbot durch die Analyse von Nutzeranfragen und auftretenden Problemen verbessert. Insgesamt ergeben sich daraus bei einem Stundenlohn von 15 € pro Stunde geschätzte Entwicklungskosten von 4 350 €.

Die Wartungskosten werden auf eine Stunde Arbeitsaufwand pro Tag geschätzt, was bei dem genannten Stundenlohn etwa 300 € pro Monat entspricht. Mit entwickelten Automatisierungen, wie der Aktualisierung der Produktdaten oder dem Abgleichen und Analysieren von Anfragen und getätigten Bestellungen, wird sich dieser Aufwand mittelfristig reduzieren. Allerdings sollten die Daten aus den Anfragen regelmäßig manuell ausgewertet und Anpassungen an dem Kontext durchgeführt werden.

Zusätzlich fallen monatliche Kosten für das Hosting des Proxy-Servers und des verwendeten Modells an. Der Proxy liegt auf einer Amazon-EC2-T3-Instanz mit der Spezifizierung t3a.xlarge, welche circa 0,17 € pro Stunde kostet [Amazon 2025b]. Im laufenden Betrieb beanspruchte der Proxy ungefähr 30 % der Instanzkapazitäten, womit sich die monatlichen Kosten je nach Betriebszeit von 6–24 Stunden täglich auf 9–37 € belaufen. Der derzeitige Betrieb des LLM erfolgt über den Anbieter Huggingface, bei dem der Endpunkt mit einer NVIDIA A10G GPU (24GB VRAM) betrieben wird. Die Kosten hierfür betragen 1 $ pro Stunde [Huggin 2025c]. Bei der

aktuellen Nutzung von sechs Stunden täglich ergibt das monatliche Kosten von circa 180 $. Damit wäre der Chatbot allerdings nicht dauerhaft erreichbar, was den Kernpunkt dieser Technologie ausmacht. Bei einem dauerhaften Betrieb steigen die Kosten auf 720 $ im Monat. Umgerechnet zum Kurs vom 08.08.2025 ergeben sich somit monatliche Gesamtkosten von 489 € für eine Nutzung von sechs Stunden täglich beziehungsweise etwa 1057 € bei dauerhaftem Betrieb. Bei allen genannten Beträgen sind weder Kosten für Softwarelizenzen noch für externe Unterstützung enthalten.

5.2 Vergleich mit Dienstleistern

Ein direkter Kostenvergleich mit etablierten Dienstleistern wie assono GmbH, moinAI oder AI Works München konnte im Rahmen dieser Arbeit nicht aufgestellt werden, da die Lösungen der Dienstleister nicht praktisch getestet wurden und die Preisgestaltung teilweise undurchsichtig ist und erst auf Anfrage ersichtlich wird. Nach verfügbaren Informationen liegen die monatlichen Kosten je nach Paket und Anfragevolumen bei 475 € bis 2000 €, zuzüglich unbekannter Einrichtungskosten [assono 2025, moinAI 2025].

Die durchschnittlichen Entwicklungskosten von einer Chatbot-Implementation bewegen sich schätzungsweise im Bereich zwischen 10.000 € und 100.000 € [Verma 2025], [AltKho 2025], [vizolo 2024]. Dies sind lediglich allgemeine Schätzungen und auf die Anfragen an die verschiedenen Dienstleister wurde keine eindeutige Antwort geliefert. Die monatlichen Kosten bewegen sich nach obiger Angabe etwa in derselben Größenordnung wie die Kosten der selbst entwickelten Lösung. Die deutlich geringeren Entwicklungskosten sind unter anderem darauf zurückzuführen, dass das Unternehmen einen niedrigeren Stundenlohn entrichtet, als ein professioneller Dienstleister in Rechnung stellt. Zusätzlich wurde dieses Projekt ohne umfassende Vorerfahrung gestartet, wodurch sich der Arbeitsaufwand des eigentlichen Projektes nochmals verlängert und damit auch die direkten Entwicklungskosten.

5.3 Return on Investment (ROI)

Die Evaluation des Projektes stützt sich außerdem auf den Return on Investment (ROI) als Kennzahl. Der ROI bietet eine effiziente Möglichkeit, zu bewerten, ob ein Projekt auf Dauer rentabel ist [Stobie 2020]. Dieser berechnet sich wie folgt:

$$ROI = \frac{Nettogewinn - Investitionskosten}{Investitionskosten} \times 100 \quad (3)$$

Formel 3: Berechnung des Return on Investment [Stobie 2020]

Im Rahmen dieses Projektes werden der Umsatz, der unmittelbar durch den Chatbot entstanden ist, sowie die Entlastung des Kundenservice betrachtet, indem häufig gestellte Anfragen abgefangen wurden. Die Auswirkungen auf die Kundschaft können im ROI nicht direkt berücksichtigt werden. Dazu zählen zum Beispiel die schnellere Navigation auf der Shop-Seite oder persönliche Empfehlungen zu Produkten.

5.3.1 Umsatz des Chatbots

Der entstandene Umsatz wird durch den Abgleich der vergebenen UUID in der Chatbot-Konversation und das Hinterlegen dieser in den Bestellattributen errechnet. Dabei wird zunächst untersucht, welche Produkte in der Bestellung der Kundschaft vorhanden sind, die der Chatbot in seinen Antworten erwähnt hat. Danach wird manuell abgeglichen, ob bei der automatisierten Auswertung Fehler beim Produktabgleich vorgekommen sind. Über einen Zeitraum von zwei Monaten wurde der Chatbot im Online Shop des Unternehmens getestet und Daten von insgesamt 270 Konversationen gesammelt. Rund zwei Drittel (174) der Anfragen fallen unter die Kategorie „Produktanfragen“. Dabei werden Fragen zu Produkten gestellt oder nach Produktempfehlungen gefragt. Das andere Drittel wird hauptsächlich von Kundenservice-Anfragen belegt (84) und lediglich 12 Anfragen sind unter der Kategorie „Andere“ einzuordnen. Diese beinhaltet sowohl Beschwerden über den Chatbot als auch Begrüßungen. Von den 174 Produktanfragen sind 51 einer anschließenden Bestellung zuzuordnen. Davon enthielten 20 Bestellungen Produkte, die der Chatbot in seinen Antworten erwähnt und empfohlen hat. Diese Bestellungen können direkt auf die Empfehlungen des Chatbots zurückgeführt werden. Dabei muss allerdings beachtet werden, dass keine Ergebnisse vorliegen, dass die Kundschaft ohne eine Interaktion mit dem Chatbot keine Bestellung getätigt hätte.

Zusammen ergibt das mit den zugehörigen Bestellwerten einen Umsatz von circa 232 € und bei einer Marge von 30 % einen Gewinn von 69,60 €. Dieser Umsatz wurde insgesamt ausgeglichen über den gesamten Testzeitraum generiert. Getätigte Verbesserungen in dem Zeitraum hatten keine Auswirkungen auf die Anzahl der Interaktionen oder den Umsatz. Da sich die Betriebskosten des Hauptmodells bis zu dem Zeitpunkt des 08.08.2025 auf 289,49 € belaufen, hat der Chatbot 24,04 % seiner Betriebskosten mit direkten Produktvermittlungen gedeckt.

5.3.2 Entlastung des Kundenservice

Die Entlastung des Kundenservice wird durch den Vergleich des Aufkommens von häufig gestellten Fragen berechnet. Dabei werden der durchschnittliche Stundenlohn eines Kundenservicemitarbeiters und die durchschnittliche Antwortzeit solcher Anfragen berücksichtigt. Dies ergibt die Entlastung des Kundenservice als numerischen Wert. In dieser Testphase sind zwar einige Anfragen aufgekommen, die in den Bereich der häufig gestellten Fragen zählen würden, allerdings hat der Kundenservice keine spürbaren Effekte vermerkt. Dies lässt sich unter anderem mit dem geringen Anfragevolumen erklären. Der Kundenservice bekommt täglich durchschnittlich 120 E-Mails und 100 Anrufe. Der Chatbot vermerkte circa 5 Anfragen pro Tag. Da nur ein Drittel aller Anfragen dem Bereich des Kundenservice zuzuordnen ist, übernimmt der Chatbot durchschnittlich 1,66 Anfragen pro Tag für den Kundenservice. Darüber hinaus wurde der Chatbot hauptsächlich als Produktberater eingesetzt und konnte nicht alle Kundenserviceanfragen beantworten. Deshalb verweist dieser bei einer Kundenserviceanfrage häufig direkt auf den Kundenservice. Mit weiteren Anpassungen und Hinterlegungen von Antworten in Absprache mit dem Kundenservice sowie steigendem Engagement wird erwartet, dass auch in diesem Bereich positive Ergebnisse erzielt werden.

5.3.3 Weitere Kennzahlen

Zusätzlich zu dem direkten Umsatz lassen sich noch andere Kennzahlen miteinander vergleichen. Dazu zählen die Nutzungsrate des Chatbots im Vergleich zu den gesamten Sitzungen im Online Shop, die Conversionrate (CVR) im Vergleich zum normalen Betrieb und der durchschnittliche Warenkorbwert eines Chatbot-Nutzenden im Vergleich zu einem Nutzenden, der diesen nicht nutzt. In diesem Kontext misst die CVR zum einen die Kaufquote von Kundinnen und Kundschaft nach einer Chatbot-Interaktion und zum anderen die allgemeine Bestellquote aller Website-Besucher.

Insgesamt wurde der Chatbot im Vergleich zu den gesamten Sitzungen wenig genutzt. Etwa 0,028 % der Kundschaft haben eine Interaktion mit dem Chatbot im Einsatzzeitraum durchgeführt. Das Chatfenster wurde insgesamt 2318 Mal aufgerufen, was mit der Anzahl der Interaktionen eine Interaktionsrate von 11,65 % ergibt. Die Click-Through-Rate (CTR) von bereitgestellten Links beträgt etwa 49,63 %, da 134 von 270 Personen (mehrfaches Tracking ausgeschlossen) auf die Links des Chatbots klickten. Die Unterschiede der CVR und des durchschnittlichen Warenkorbwertes sind in den Abbildungen 11 bis 14 zu sehen. Für den Vergleich wurden die Personen berücksichtigt, die mit dem Chatbot interagiert und anschließend eine Bestellung abgeschlossen haben. Der durchschnittliche Warenkorbwert errechnet sich aus allen Waren, die in der Bestellung vorhanden sind, und nicht ausschließlich aus den Produkten, die der Chatbot vorgeschlagen hat.

Abbildung 11 und 12 zeigen die CVR und den Warenkorbwert für Bestellungen mit Produkten, die direkt auf den Chatbot zurückzuführen sind. Abbildung 13 und 14 stellen alle Bestellungen dar, die getätigt wurden, nachdem eine Interaktion mit dem Chatbot stattgefunden hat. Die Abbildungen zeigen drei unterschiedliche Szenarien, um einen tieferen Einblick in die verschiedenen Abläufe zu präsentieren. Zunächst werden Bestellungen dargestellt, die noch am selben Tag getätigt wurden, danach Bestellungen, die ein bis zwei Tage nach der Interaktion durchgeführt wurden, und zuletzt alle anderen Bestellungen, die nach zwei oder mehr Tagen getätigt wurden. In den Abbildungen ist ein deutlicher Unterschied zwischen den Bestellungen zu erkennen, welche die vom Chatbot empfohlenen Produkte enthielten, und denen, die lediglich eine vorherige Interaktion aufwiesen. Dabei ist zu erkennen, dass sowohl die CVR als auch der Warenkorbwert durch das Vergleichen aller Bestellungen mit Interaktionen deutlich zunehmen. Der Warenkorbwert bei Bestellungen mit Produktempfehlungen des Chatbots liegt unter dem durchschnittlichen Warenkorbwert aller Bestellungen ohne Chatbot-Interaktion. Darüber hinaus ist ein Unterschied zwischen den Bestellungen, die am selben Tag der Interaktion getätigt wurden, und allen anderen, die später durchgeführt wurden, zu erkennen. Bei den später durchgeführten Bestellungen kann der Einfluss des Chatbots nicht gemessen werden, da der Zeitraum zwischen Interaktion und Bestellung in den Ergebnissen von 1–51 Tagen reicht. Dieser Beitrag müsste durch eine manuelle oder eine mit einem LLM durchgeführte Analyse der Nutzeranfrage und des Warenkorbs bestimmt werden. Wenn Nutzende Produktempfehlungen des Chatbots in den Warenkorb legen und später bestellen, ist das jedoch ein Indikator für den Einfluss des Chatbots. Im Gesamtkontext dienen die folgenden Abbildungen als trendhafte Hinweise; robuste Schlussfolgerungen erfordern größere Stichproben und längere Beobachtungszeiträume.

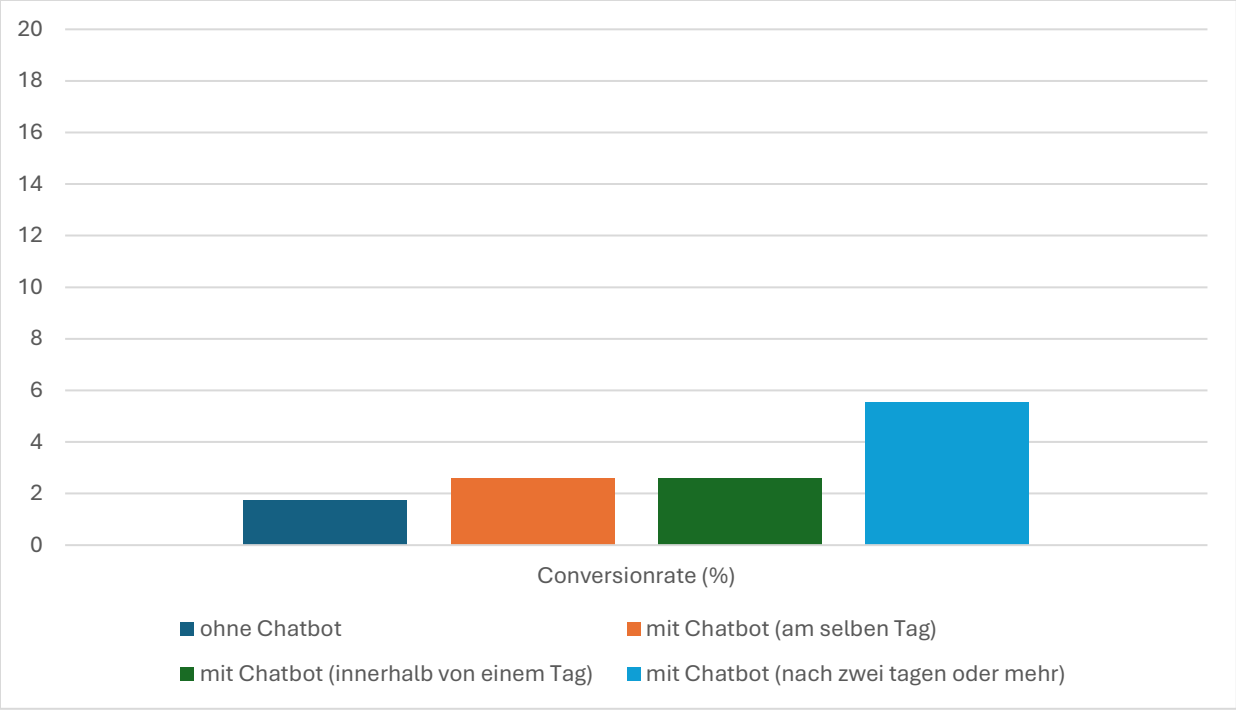

Abbildung 11: CVR mit und ohne Chatbot-Interaktion für Bestellungen mit vom Chatbot empfohlenen Produkten in Prozent (%)

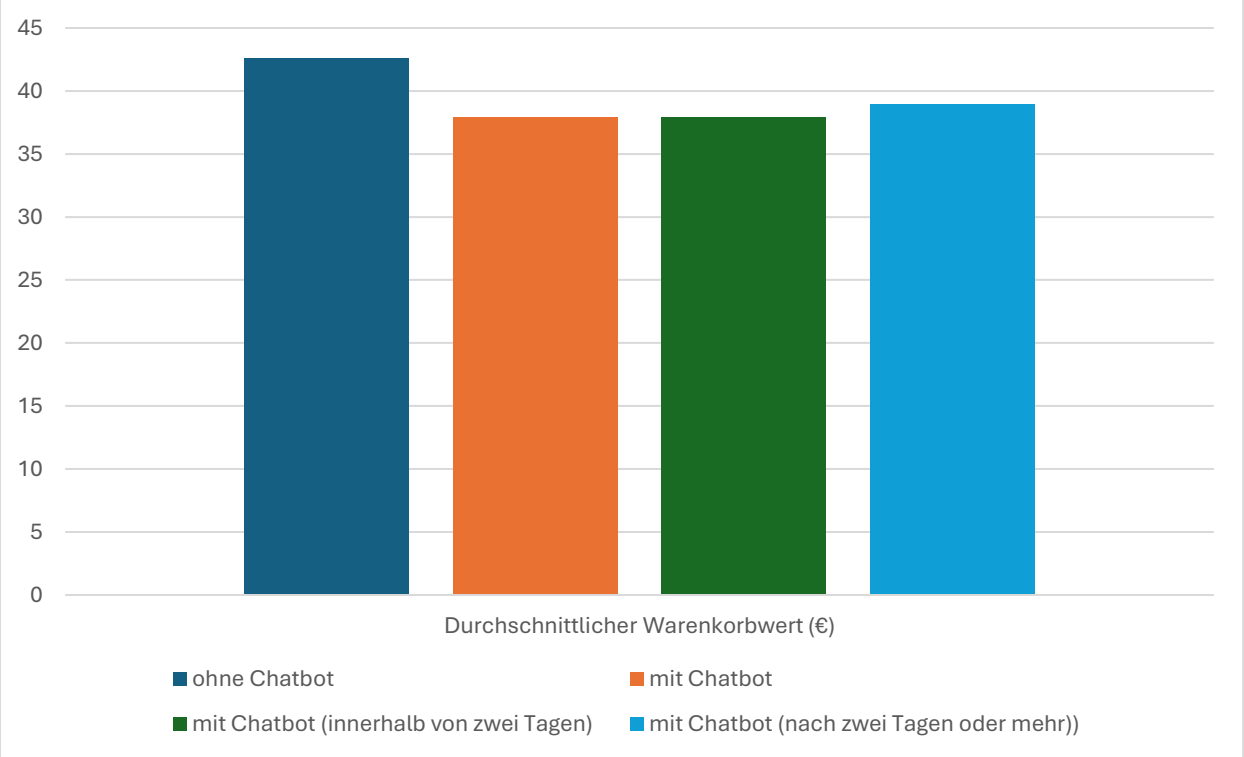

Abbildung 12: Durchschnittlicher Warenkorbwert mit und ohne Chatbot-Interaktion für Bestellungen mit vom Chatbot empfohlenen Produkten in Euro (€)

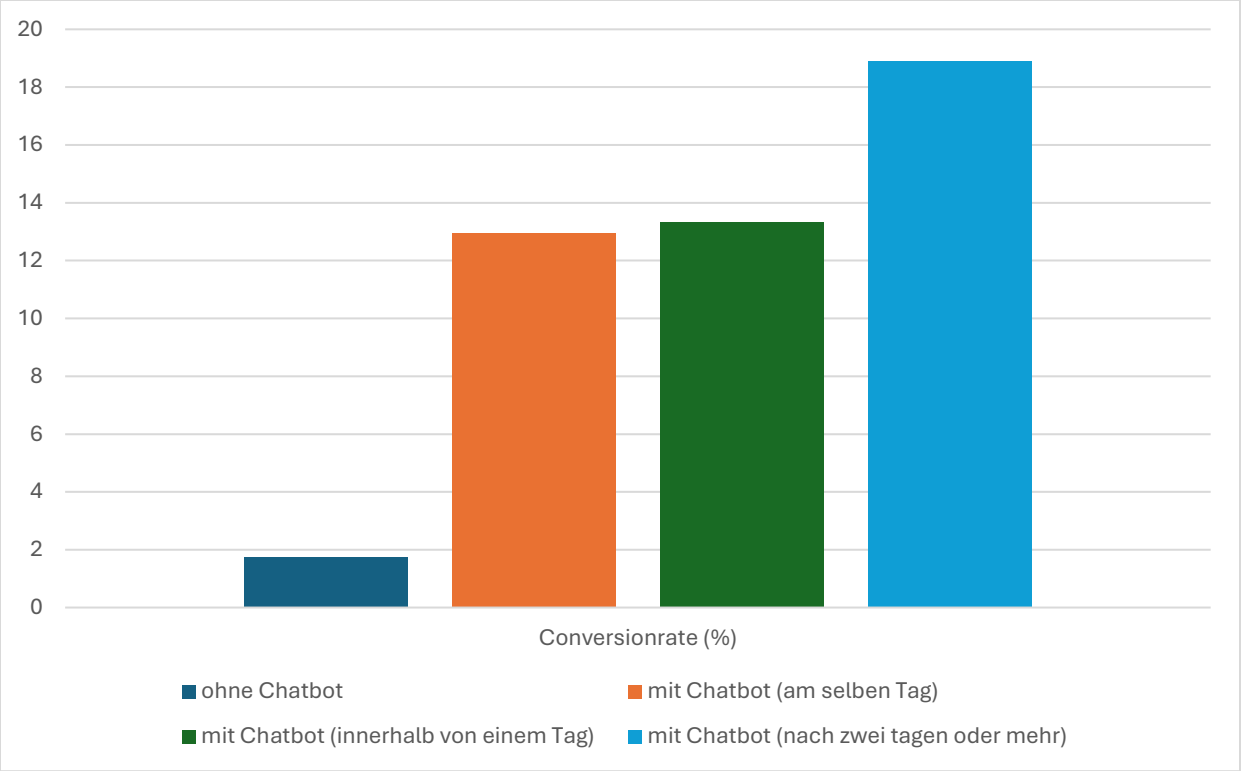

Abbildung 13: CVR mit und ohne Chatbot-Interaktion für alle Bestellungen in Prozent (%)

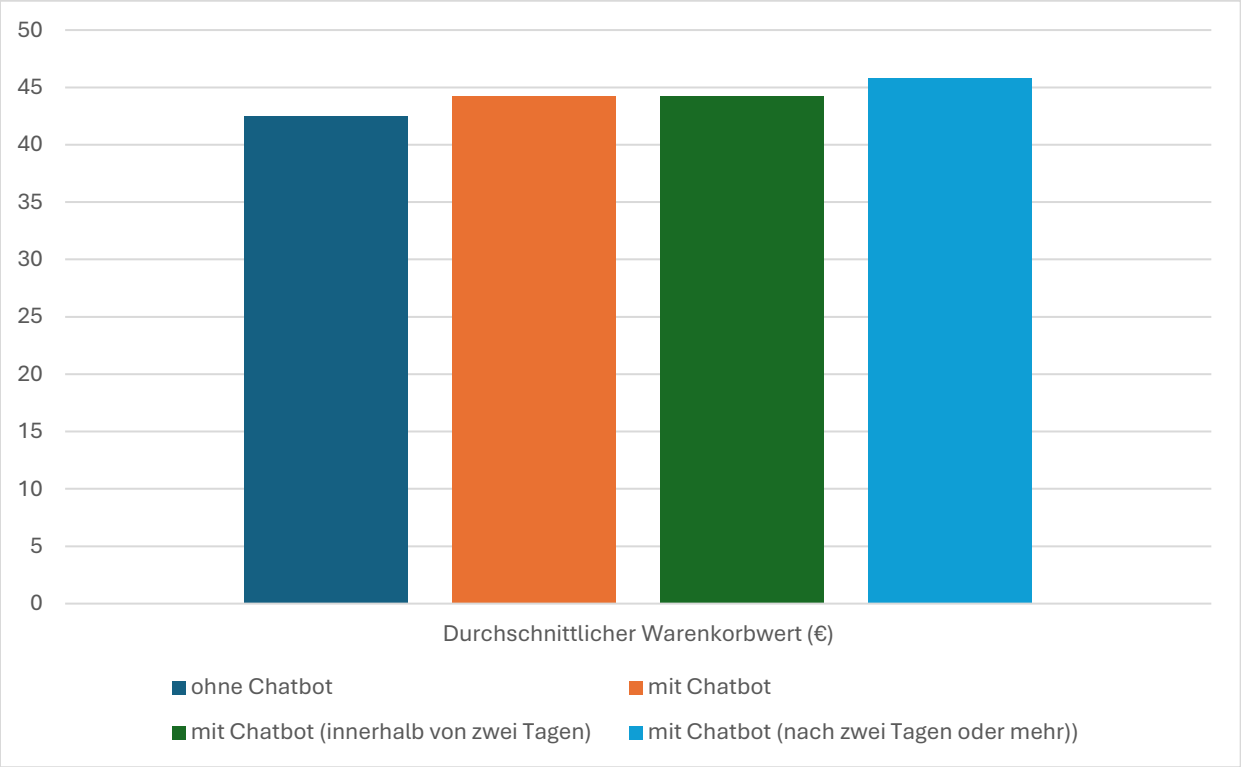

Abbildung 14: Durchschnittlicher Warenkorbwert mit und ohne Chatbot-Interaktion für alle Bestellungen in Euro (€)

Im Wesentlichen hat der Chatbot im betrachteten Zeitraum seine Kosten noch nicht wieder ausgeglichen. Allerdings hat die Evaluation ergeben, dass die Kostendeckung des Chatbots durch die Steigerung der Interaktionsrate und die Anpassung an den Kundenservice erreicht werden kann. Da sich der Nettogewinn effektiv auf rund 70 € beläuft, die der Chatbot mit direkten Bestellungen erwirtschaftet hat, liegt der ROI für dieses Projekt bei circa −98,5 %. Die Erwirtschaftung der Entwicklungs- und Betriebskosten durch den Chatbot in diesem Zeitraum lag allerdings nicht in der Erwartungshaltung. Aufgrund dessen, dass lediglich 0,028 % der Kundschaft mit dem Chatbot über zwei Monate hinweg interagiert haben, lässt sich bereits mit einer Steigerung auf 3 % der mögliche Umsatz laut aktuellen Statistiken um ein Hundertfaches erhöhen. Da sich die Implementierung des Chatbots in der Anfangsphase befindet, der technische Stand aber fundiert ist, wird der Fokus bei einer Weiterentwicklung darauf liegen, wie die Kundeninteraktion mit dem Chatbot ansprechender gestaltet werden kann, um das Engagement zu erhöhen. Dies ist erstrebenswert, da die obigen Abbildungen zeigen, dass Bestellungen, die eine Interaktion mit dem Chatbot aufweisen, sowohl eine höhere CVR als auch einen höheren Warenkorbwert besitzen.

5.4 Auswirkungen auf die Kundschaft

Ebenfalls von zentraler Bedeutung sind die Auswirkungen auf die Kundschaft und deren Kaufverhalten. Dabei stehen das Kundenvertrauen und die Benutzerfreundlichkeit sowie die langfristige Kundenzufriedenheit im Fokus. Sollte der Kunde negative Erfahrungen mit dem Chatbot machen – sei es durch technische Probleme, unpassende Antworten oder eine zu aufdringliche Präsenz –, kann dies zu erheblichen Konsequenzen führen. Der Kunde fühlt sich in seinem Einkaufserlebnis gestört, entwickelt Misstrauen gegenüber dem Chatbot oder der Marke und reduziert seine Kaufaktivitäten [RaBeMe 2024].

Da keine Bewertungen der Kundschaft nach dem Chatverlauf erfasst wurden, kann in dieser Arbeit lediglich auf die Chatverläufe eingegangen werden. Allerdings generiert auch die Erhebung und Auswertung der Daten ohne direkte Bewertung durch die Kundschaft wichtige Einblicke in die Bedürfnisse dieser, da sie ihre Probleme direkt und anonym schildern können [RaBeMe 2024]. Um zusätzliche Daten zu generieren, sollte in zukünftigen Implementationen trotzdem ein Bewertungssystem implementiert werden. Als Beispiel erfragt ein kurzes Feedback, wie die Konversation wahrgenommen wurde und ob der Kauf auch ohne Chatbot erfolgt wäre. Zusätzlich kann zukünftig die Demografie der Kundschaft analysiert werden, um eine umfassendere Perspektive zu erhalten.

Insgesamt lassen sich die Interaktionen als neutral bewerten. Lediglich zwei der 270 Interaktionen waren negativ konnotiert, da geschildert wurde, dass das Chatfenster nervig sei. Es gab einen Zeitraum von mehreren Tagen, in dem das automatische Update der Vektordatenbank fehlschlug und dies erst am Anfang der darauffolgenden Woche bemerkt wurde, wobei der Chatbot seine korrekten Informationen verlor. Dabei hat der Chatbot Links ausgegeben, die nicht existierten. In diesem Zeitraum konnten Frustrationen bei den Anfragen erkannt werden, da die Kundschaft sich auf korrekte Informationen des Chatbots verlässt. Solche Probleme sorgen für Irritationen bei der Kundschaft und sollten zukünftig vermieden werden. Das anfängliche Vergessen von Konversationsnachrichten sorgte für Frustrationen, da die Kundschaft ihr Anliegen mehrfach wiederholen musste, wenn der Chatbot den Sachverhalt in seinem Kontext nicht finden konnte. Das lässt sich auf die fehlende Kontextbehandlung in der Vektorsuche und das in Kapitel 4.3 genannte „Lost in the middle“-Problem zurückführen. Im Gegensatz dazu bedankten sich jedoch auch 14 Personen für die Hilfsbereitschaft des Chatbots.

5.5 Datenschutz und Datensicherheit

Der besondere Schutz von kundschaftbezogenen Daten spielt auch im Bereich der Chatbot-Anwendung eine wichtige Rolle. Hierbei gilt es, sowohl auf die DSGVO- als auch auf die EU-DSGVO-Konformität der Prozesse, über die die Daten erfasst werden, zu achten, als auch einen sicheren Speicher für sensible Kundendaten zu schaffen. Im Folgenden wird erläutert, wie in diesem Projekt mit Datenschutz und Datensicherheit umgegangen wurde und was bei zukünftigen Erweiterungen beachtet werden muss. Im Folgenden wird allgemein von der DSGVO gesprochen, gemeint sind die DSGVO von Deutschland sowie die DSGVO der EU.

5.5.1 Datenschutz von Kundendaten

Bei der Verarbeitung von sensiblen Kundendaten spielt der Datenschutz eine wichtige Rolle. Die Verarbeitung von diesen Daten erfordert stets eine DSGVO-konforme Implementation und Überprüfung.

Wesentliche Grundsätze nach [DSGVO 2025a] umfassen:

- **Rechtmäßigkeit**: Daten dürfen nur auf einer gültigen Rechtsgrundlage verarbeitet werden.
- **Zweckbindung**: Die Daten dürfen nur für festgelegte, eindeutige und legitime Zwecke verarbeitet werden.
- **Datenminimierung**: Nur die nötigsten Daten dürfen verarbeitet und gespeichert werden.
- **Richtigkeit**: Die Daten müssen sachlich korrekt und aktuell gehalten werden.
- **Speicherbegrenzung**: Daten müssen gelöscht werden, sobald sie nicht mehr benötigt werden.
- **Integrität und Vertraulichkeit**: Die Daten müssen technisch und organisatorisch sicher gehandhabt werden.

Ergänzend zum Datenschutz gilt für Chatbots auch die Transparenzpflicht des EU-AI-Act: Nutzende sind darauf hinzuweisen, dass sie mit einem KI-System interagieren, und synthetische Inhalte sind entsprechend zu kennzeichnen. Der EU-AI-Act ist seit 2024 in Kraft und ergänzt den Datenschutzrahmen um spezifische Vorgaben für den Einsatz von KI-Systemen. Dieser richtet sich an Unternehmen als Anbieter und Verwender solcher Systeme [EUAI 2024 Art. 50].

Um diese Grundsätze und die Transparenzpflicht einzuhalten, ist besonders bei LLMs von Bedeutung, nur die nötigsten Informationen in dem Kontext bereitzustellen und die Kundschaft auf die Verwendung eines KI-Modells aufmerksam zu machen. Da LLMs zu Halluzinationen neigen und unabsichtlich sensible Daten von Kundschaft rausgeben könnten, müssen sichere Workflows für das Abrufen dieser Daten implementiert werden. Im Rahmen dieser Bachelorarbeit wurde lediglich ein Prototyp eines Chatbots implementiert und getestet, um dessen Potenzial zu untersuchen. Dabei ist eine Verarbeitung von sensiblen Daten nicht implementiert worden. Die Implementation hätte weiteren Aufwand erfordert, der den zeitlichen Rahmen überschritten hätte. Daher ist der Chatbot lediglich als Produktberater und Vorstufe zum Kundenservice tätig. Allerdings liegt bereits ein Konzept vor, das bei weiterer Verfolgung des Projektes umgesetzt werden könnte. Dies wird im Folgenden erläutert.

Das Hauptziel der Implementierung des Kundensupports besteht darin, dem Chatbot nach Authentifizierung der Kundschaft relevante Daten zu Bestellungen bereitzustellen. Dazu zählt beispielsweise das Bereitstellen der Trackingnummer oder des direkten Versandverlaufs, die Änderung von Liefer- und Rechnungsadressen oder das Stornieren einer Bestellung. Plattformen wie Shopify bieten Möglichkeiten, um mit spezifischen Berechtigungsbereichen und Authentifizierung durch einen Login der Kundschaft sensible Daten präzise abzurufen und zu verarbeiten. Mit bestimmten APIs oder dem Early Access eines Model Context Protocol (MCP)-Servers, welcher speziell dafür entwickelt wird, KI-Modelle mit Echtzeitdaten zu versorgen, kann die Kundschaft sicher authentifiziert werden [Shopif 2025]. Durch den Zugriff auf die Kundendaten können nicht nur hilfreiche Informationen bereitgestellt, sondern auch Daten wie die Bestellhistorie genutzt werden, um der Kundschaft personalisierte Produktempfehlungen zu geben. Allerdings sind auch für diese Form der Implementation Sicherheitsregeln zu beachten, da hier ebenfalls die Konversationen der Kundschaft gespeichert werden.

5.5.2 Datensicherheit

Die Datensicherheit stellt einen weiteren wichtigen Faktor dar und umfasst die technische Implementierung der DSGVO-Richtlinien bei der Verarbeitung von Kundendaten. Datensicherung bedeutet vor allem, wie und wo die sensiblen Daten gespeichert werden, ob regelmäßige Backups und Redundanzen vorhanden sind, ob die Datenübertragung verschlüsselt abläuft und ob Schutzmaßnahmen gegen Sicherheitslücken oder Angriffe bestehen. Dabei ist insbesondere bei Drittanbietern darauf zu achten, dass verwendete Daten ausschließlich innerhalb der EU gespeichert werden und nur unter Umständen mit entsprechenden Sicherheitsprotokollen verschlüsselt versendet werden. [DSGVO 2025b]

Trotz des Fokus auf Funktionalität und Effektivität des Systems mussten auch hier Vorkehrungen getroffen werden. Ein Beispiel dafür ist ein Kunde, der nicht weiß, dass der Chatbot noch keine personalisierten Anfragen verarbeitet, aber bereits in der ersten Nachricht persönliche Daten von sich preisgibt, in der Hoffnung, Auskunft über seine Bestellung zu erhalten. Da in dem Unternehmen die Strukturen für einen DSGVO-konformen Umgang mit sensiblen Daten notwendigerweise vorhanden sind, wurden diese auch genutzt. Dazu zählen das Hosting der Datenbank und des Proxys zum Speichern und Verarbeiten der Anfragen auf einem in der EU gehosteten Server und das Weiterleiten der Anfragen an einen Huggingface-Endpunkt mit einem unterliegenden Server, der sich ebenfalls in der EU befindet. Huggingface versichert beispielsweise, dass die Endpunkte keine sensiblen Daten wie das Payload oder die Token, welche an den Endpunkt gesendet werden, speichern und dass Log-Dateien nach 30 Tagen gelöscht werden [Huggin 2025a]. Zusätzlich anzumerken ist, dass für General-Purpose AI (GPAI)-Modelle EU-weit seit dem 2. August 2025 spezifische Vorgaben zu Transparenz, Sicherheit und einer öffentlichen Zusammenfassung wesentlicher Trainingsdatenquellen gelten, die ebenfalls einzuhalten sind. Die EU definiert GPAI-Modelle als KI-Modelle mit allgemeiner Verwendbarkeit, die für verschiedene direkte als auch indirekte (zum Beispiel eingebettet in andere KI-Systeme) Zwecke dienen können [EUAI 2024 Art. 3 Nr. 63].

Allerdings ist bei zukünftiger Implementation von Kundendaten zu beachten, dass Huggingface ein US-basiertes Unternehmen ist und diese auf Anfrage der US-Regierung aufgrund des 2018 verabschiedeten CLOUD-Acts geforderte Daten herausgeben müssen [CLOUD 2018]. Auch wenn diese aus eigener Angabe nicht gespeichert werden, sollte in diesem Fall ein anderer Dienstleister genutzt werden, um das Modell zu hosten. Außerdem ist das Verwenden des Shopify-MCP-Servers notwendig, um die Kundendaten abzurufen und die DSGVO-Konformität

beizubehalten. Über den MCP-Server können die benötigten Daten gezielt abgerufen werden, sodass Daten der übrigen Kundschaft nicht gefährdet werden. So sind diese selbst bei einem gelungenen Jailbreak oder einer Halluzination des LLM sicher und werden dem Modell nicht zur Verfügung gestellt. Jedoch ist darauf zu achten, dass Shopify ebenfalls ein US-basiertes Unternehmen ist und Dienste wie Cloudflare CDN nutzt, um Webseiten schnell zu laden, sowie zwischenzeitlich sensible Daten wie IP-Adressen zu speichern. Auch wenn 2023 der Angemessenheitsabschluss „Data Privacy Framework“ zwischen EU und USA angenommen wurde, ändert das nicht die Gesetzeslage der USA und bestehende Vorgaben müssen weiterhin beachtet werden. [EUDPF, 2023]

5.6 Vergleich mit Experten und anderen Forschungsprojekten

Im Rahmen dieser Bachelorarbeit wurden zwei Interviews sowie ein fachlicher Austausch mit Experten auf dem Fachgebiet der Chatbots geführt. Die Auswahl der Gesprächspartner erfolgte opportunitätsbasiert; im Erhebungszeitraum konnten keine fachlich passenden Expertinnen gewonnen werden. Für weiterführende Untersuchungen wird eine ausgewogene Geschlechterverteilung empfohlen, um zusätzliche Perspektiven einzubeziehen. Die befragten Experten bringen sowohl viel Erfahrung als auch Wissen mit ein, da sie schon in verschiedenen Bereichen Chatbot-Projekte umgesetzt haben. Der generelle Konsens und die klare Empfehlung sind, dass Unternehmen sich durchaus und so früh wie möglich mit dieser Technologie befassen sollten [Felix Frage 15], [Maximilian Frage 15], [Thomas & Micheal Frage 4]. Die Experten haben die Hürden und Chancen ebenfalls festgestellt, die bereits in dieser Arbeit erläutert wurden. Darunter zählen beispielsweise das Kundenvertrauen, die falsche Beantwortung von Fragen, das Verlieren des Kontextes, sodass die Kundschaft sich wiederholen muss, aber auch die dauerhafte Verfügbarkeit, die sofortige Reaktionszeit und das effiziente Bearbeiten von Kundenanfragen sowie die Einblicke in Kundenbedürfnisse als Marketing-Mehrwert [Felix Frage 10], [Thomas & Micheal Frage 3]. Die Hürden lassen sich bereits heute mit einer robusten Implementation effektiv überwinden und die Chancen, die dadurch entstehen, überwiegen deutlich.

Ein Experte vergleicht das Voranschreiten der KI-Technologie mit der Digitalisierung [Felix Frage 15] und empfiehlt besonders auch KMU, sich untereinander sowie mit Forschungsinstituten auszutauschen und sich mit der Technologie auseinanderzusetzen [Felix Frage 16]. Besonders wird aber auch die Implementation mit der Webseite des Unternehmens einen wichtigen Faktor darstellen [Maximilian Frage 3]. Die Oberflächen sollten so nutzerfreundlich wie möglich gestaltet werden [Felix Frage 4]. Diverse Funktionen garantieren das initiale und auch beständige

Engagement der Kundschaft. Die Einbindung des Chatbots in das User Interface (UI) ist damit einer der wichtigsten Bestandteile, der maßgeblich entscheidet, wie sich die Kundschaft gegenüber der Implementation verhält. Das wurde ebenfalls im Rahmen dieses Projekts festgestellt, da das Interface des Chatbots lediglich durch einen Button und eine Chatbox abgebildet wurde und es noch keine weitere Einbindung in andere Funktionen des Online Shops wie beispielsweise in die Suche oder die Produktseiten gab.

Auch Arbeiten wie [KeEjId+ 2024], [ShaMis 2024], [SurCas 2023] und [ErSuLa 2024] unterstreichen die Relevanz von Chatbots, insbesondere in KMU, da diese wertvolle Einblicke in Kundenintentionen gewonnen werden können. Gleichzeitig betonen sie die nötigen Voraussetzungen für den effektiven Einsatz von Chatbots. Diese Studien identifizieren positive Auswirkungen auf diverse Aspekte wie die Umsatzsteigerung, die Kundenzufriedenheit, die Kundenbindung sowie den Marketingwert, der durch das Aufzeichnen der Konversationen generiert wird. Auch die Experten erkennen diesen Marketingwert und gehen über die rein funktionalen Anwendungen von Chatbots hinaus. Sie empfehlen Unternehmen, ihre Prozesse kritisch zu analysieren und zu erwägen, wie sich unter anderem durch intern ausgerichtete Chatbots diese mittels KI transformieren lassen [Felix Frage 16], [Maximilia Frage 15]. Dabei wird deutlich, dass viele Unternehmen und Mitarbeitende das volle Potenzial des KI-Wandels bislang nicht erkannt haben, was dazu führen kann, dass Chancen ungenutzt bleiben [Felix Frage 15].

6 Fazit

Im Fazit werden die wesentlichen Erkenntnisse dieser Bachelorarbeit zusammengefasst und kritisch reflektiert; zudem werden Perspektiven für weitere Forschung und konkrete Verbesserungsmöglichkeiten aufgezeigt.

6.1 Zusammenfassung

In dieser Bachelorarbeit wurde ein Chatbot zwei Monate in einem Online Shop eines mittelständischen Unternehmens erprobt. Sie liefert Einblicke in die Entwicklung und den Einsatz von Chatbots in KMU und zeigt die Machbarkeit und Sinnhaftigkeit einer Implementierung. Die LLM-Modelle, die aktuell zur freien Verfügung stehen, reichen bereits aus, um ein effizientes System ohne spezielle Hardwareanforderungen und mit begrenzten Betriebskosten aufzubauen. Zudem stehen ausreichend Open-Source-Werkzeuge zur Verfügung sowie Hosting-Dienstleister, die die Umsetzung und den Betrieb von solchen Modellen in einem KMU erleichtern.

Eine Vollzeitkraft realisiert den Prototyp in circa zwei Monaten. Trotz des aktuell negativen ROI sowie Limitationen wie begrenzten Ressourcen, Kontextverlusten des LLMs oder Schwierigkeiten beim Filtern der richtigen Produkte besteht Wirtschaftlichkeitspotenzial bei höherem User-Engagement. Es liegt eine tragfähige technische Grundlage vor, die sich beliebig erweitern lässt. Der Chatbot liefert Einblicke in die Kundenbedürfnisse und in die Verbesserung des Kundenerlebnisses.

Die Ergebnisse zeigen auch bei geringem Engagement positive Entwicklungen, die insgesamt als Erfolg gewertet werden können und wertvolle Erkenntnisse liefern. Es genügt jedoch nicht, der Kundschaft lediglich die Möglichkeit zu geben, sich mit einem Chatbot unterhalten zu können und diesem ihre Anliegen anzuvertrauen. Um die Interaktionsrate zu fördern, müssen weitere, bereits geschilderte Maßnahmen unternommen werden, um den Chatbot passend in die Umgebung des Online Shops einzubinden und die Kundschaft von dessen Nutzen zu überzeugen. Das stellt sicher, dass die Kundschaft kein schlechteres, sondern ein besseres Kauferlebnis im Online Shop hat. Insgesamt lässt sich mit Rückbezug auf die Fragestellung schlussfolgern, dass ressourcenschonende Chatbots Potenziale zur Umsatzsteigerung bieten – sowohl durch direkte Effekte wie Produktempfehlungen als auch durch einen Marketing-Mehrwert aus der systematischen Analyse von Konversations- und Bestelldaten, der zur Anpassung und Entwicklung von Strategien genutzt werden kann.

6.2 Ausblick

Aufbauend auf dieser Arbeit können über einen längeren Zeitraum hinweg mehr und aussagekräftigere Daten gesammelt werden, die umfangreichere Erkenntnisse liefern. Da der Rahmen dieser Bachelorarbeit zeitlich begrenzt ist und durch fehlendes User-Engagement nur bedingt Verbesserungen durchgeführt werden konnten, ist es von Bedeutung, die vorgestellten Ansätze weiter zu untersuchen. Dazu sollte ein Bewertungssystem eingebaut werden, das es der Kundschaft ermöglicht, die Konversationen mit dem Chatbot zu bewerten. Außerdem ist es von Interesse, die Demografie der Kundschaft zu untersuchen, um eine Verbesserung der Antworten des Chatbots durchführen zu können.

Diese Arbeit bietet den Grundbaustein und eine Anleitung für ein Chatbot-System, welches in KMU selbst ohne fundiertes Vorwissen realisiert und eingesetzt werden kann. Ein besonders wichtiger Punkt, der sich in dieser Arbeit allerdings nicht endgültig klären ließ und weiteren

Forschungsbedarf aufweist, ist der des User-Engagements. Das Projekt wird in dem Unternehmen im Anschluss weitergeführt und mögliche Lösungen für aufgeführte Hürden werden erörtert. Mit bereits vorgestellten Ansätzen wird versucht, dieses Projekt im Online Shop zu etablieren und rentabel zu gestalten. Die Chancen für die Einblicke in Kundenintentionen, das Entlasten des Kundenservice, die Steigerung des Umsatzes und die Kundenbindung überwiegen deutlich den genannten Limitationen.

Ausgehend von dieser Bachelorarbeit wird Unternehmen empfohlen, sich frühestmöglich mit dieser Technologie auseinanderzusetzen. Dabei hilft der Austausch mit anderen Unternehmen oder auch Forschungsinstituten wie Universitäten oder Hochschulen. Die rasante Entwicklung dieser Technologie in den letzten Jahren ermöglicht es, deutlich einfacher einen Nutzen daraus zu ziehen, birgt aber auch die große Gefahr, ebenso schnell den Anschluss zu verlieren.

Literaturverzeichnis

[assono 2025] Chatbot mit Künstlicher Intelligenz und LLM für Ihr Unternehmen. Prozesse automatisieren. Fragen beantworten. Kunden begeistern.
https://www.assono.de/chatbot, letzter Zugriff am 08.08.2025

[AiLeJo+ 2023] Ainslie, Joshua; Lee-Thorp, James & de Jong, Michiel et al.:
GQA: Training Generalized Multi-Query Transformer Models from Multi-Head Checkpoints.
https://arxiv.org/abs/2305.13245, letzter Zugriff am 17.08.2025

[aloa 2025] ChatGPT vs Claude vs Gemini.
https://aloa.co/ai/comparisons/llm-comparison/chatgpt-vs-claude-vs-gemini, letzter Zugriff am 13.09.2025

[AltKho 2025] Altynpara, Evgeniy & Khodukina, Kateryna:
How Much Does It Cost to Build a Chatbot: A Complete Budgeting Guide.
https://www.cleveroad.com/blog/chatbot-development-cost/, letzter Zugriff am 17.09.2025

[Amazon 2025a] Amazon Sagemaker.
https://aws.amazon.com/de/sagemaker/, letzter Zugriff am 13.09.2025

[Amazon 2025b] Amazon-EC2-T3-Instances.
https://aws.amazon.com/de/ec2/instance-types/t3/, letzter Zugriff am 17.09.2025

[Apple 2017] Siri Team:
Hey Siri: An On-device DNN-powered Voice Trigger for Apple's Personal Assistant.
https://machinelearning.apple.com/research/hey-siri, letzter Zugriff am 16.07.2025

[AquMyr 2024] Aquino, Sabrina & Myrial David:
A Complete Guide to Filtering in Vector Search.
https://qdrant.tech/articles/vector-search-filtering/, letzter Zugriff am 17.08.2025

[BaBeCu+ 2024] Balaguar, Angels; Benara, Vinamra & de Freitas Cunha et al.:
RAG vs Fine-Tuning: Pipelines, Tradeoffs, and a Case Study on Agriculture.
https://arxiv.org/abs/2401.08406, letzter Zugriff am 16.07.2025

[BaChBe 2014] Bahdanau, Dzmitry; Cho, Kyunghyun; Bengio, Yoshua:
Neural Machine Translation by Jointly Learning to Align and Translate.
https://arxiv.org/abs/1409.0473, letzter Zugriff am 17.09.2025

[BeLóDi+ 2019] Berdaso, Ana; López, Gustavo & Diaz, Ignacio et al.:
User Experience Comparison of Intelligent Personal Assistants: Alexa, Google Assistant, Siri and Cortana.
https://www.mdpi.com/2504-3900/31/1/51, letzter Zugriff am 16.07.2025

[BGE-M3 2024] BGE-M3 Embedding Model.
https://huggingface.co/BAAI/bge-m3, letzter Zugriff am 28.07.2025

[Cloudf 2025] Cloudflare Workers AI.
https://developers.cloudflare.com/workers-ai/, letzter Zugriff am 13.09.2025

[CLOUD 2018] DIVISION V – CLOUD ACT.
https://www.justice.gov/criminal/media/999391/dl?inline, letzter Zugriff am 17.09.2025

[CulPit 1943] McCulloch, Warren S. & Pitts, Walter:
A logical calculus of the ideas immament in nervous activity.
https://link.springer.com/article/10.1007/BF02478259, letzter Zugriff am 13.09.2025

[Dharma 2022] Dharmaraj:
Convolutional Neural Networks (CNN) – Architecture Explained.
https://medium.com/@draj0718/convolutional-neural-networks-cnn-architectures-explained-716fb197b243, letzter Zugriff am 17.09.2025

[DSGVO 2025a] Grundsätze für die Verarbeitung personenbezogener Daten.
https://dsgvo-gesetz.de/art-5-dsgvo/, letzter Zugriff am 22.08.2025

[DSGVO 2025b] Sicherheit der Vearbeitung.
https://dsgvo-gesetz.de/art-32-dsgvo/, letzter Zugriff am 17.09.2025

[ErSuLa 2024] Ernestivita, Gesty; Subagyo; Laras, Titi:
Implementation of a Chatbot and Auto Promote in Omproving Customer Experience on E-Commerce Sites.
https://www.ijrrjournal.com/IJRR_Vol.11_Issue.1_Jan2024/IJRR40.pdf, letzter Zugriff am 17.09.2025

[EUAI 2024] REGULATION (EU) 2024/1689 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL.
https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng, letzter Zugriff am 17.09.2025

[EUDPF 2023] COMMISSION IMPLEMENTING DECISION EU 2023/1795.
https://eur-lex.europa.eu/eli/dec_impl/2023/1795/oj, letzter Zugriff am 17.09.2025

[Fukush 1979] Fukushima, Kunihikio:
Self-organization of a neural network which gives position-invariant response.
https://dl.acm.org/doi/abs/10.5555/1624861.1624928, letzter Zugriff am 13.09.2025

[GaXiGa+ 2023] Gao, Yunfan; Xiong, Yun & Gao, Xinyu et al.:
Retrieval-Augmented Generation for Large Language Models: A Survey.
https://arxiv.org/abs/2312.10997, letzter Zugriff am 16.07.2025

[GaXiWu+ 2025] Gao, Yunfan, Xiong, Yun & Wu, Wenlong et al.:
U-NIAH: Unified RAG and LLM Evaluation for Long Context Needle-In-A-Haystack.
https://arxiv.org/abs/2503.00353, letzter Zugriff am 09.08.2025

[GeScBe+ 2021] Geva, Mor; Schuster, Roei & Berant, Jonathan et al.:
Transformer Feed-Forward Layers Are Key-Value Memories.
https://arxiv.org/abs/2012.14913, letzter Zugriff am 17.08.2025

[GnMoAd+ 2018] Gnewuch, Ulrich; Morana, Stefan & Adam, Marc T. P. et al.:
Faster Is Not Always Better: Understanding the Effect of Dynamic Response Delays in Human-Chatbot Interaction.
https://www.researchgate.net/publication/324949980_Faster_Is_Not_Always_Better_Understanding_the_Effect_of_Dynamic_Response_Delays_in_Human-Chatbot_Interaction, letzter Zugriff am 17.09.2025

[Google 2025] Google Team:
Gemma 3.
https://huggingface.co/google/gemma-3-12b-it-qat-q4_0-gguf, letzter Zugriff am 28.07.2025

[Huggin 2025a] Huggingface Endpoints.
https://huggingface.co/inference-endpoints/index, letzter Zugriff am 28.07.2025

[Huggin 2025b] Autoscaling.
https://huggingface.co/docs/inference-endpoints/autoscaling, letzter Zugriff am 15.09.2025

[Huggin 2025c] Pricing.
https://huggingface.co/docs/inference-endpoints/support/pricing, letzter Zugriff am 17.09.2025

[HuShWa+ 2021] Hu, Edward J.; Shen, Yelong; Wallis, Phillip et al.:
LoRA: Low-Rank Adaptation of Large Language Models.
https://arxiv.org/abs/2106.09685, letzter Zugriff am 16.07.2025

[HuYuMa+ 2023] Huand, Lei; Yu, Weijiang & Ma, Weitao et al.:
A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions.
https://arxiv.org/abs/2311.05232, letzter Zugriff am 17.09.2025

[JaKlCh+ 2017] Jacob, Benoit; Kligys, Skirmantas & Bo, Chen et al.:
Quantization and Training of Neural Networks for Efficient Integer-Arithmetic-Only Inference.
https://arxiv.org/abs/1712.05877, letzter Zugriff am 17.09.2025

[JinaV2 2024] Jina-Embeddings-V2-Base-De Model.
https://huggingface.co/jinaai/jina-embeddings-v2-base-de, letzter Zugriff am 28.07.2025

[KaFePa+ 2025] Kamath, Aishwarya; Ferret, Johan & Pathak, Shreya et al.:
Gemma 3 Technical Report.
https://arxiv.org/abs/2503.19786, letzter Zugriff am 29.07.2025

[KeEjId+ 2024] Kedi, Wagobera Edga; Ejimunda, Chibundoom & Idemudia, Courage et al.:
AI Chatbot integration in SME marketing platforms: Improving customer interaction and service efficiency.
https://fepbl.com/index.php/ijmer/article/view/1327, letzter Zugriff am 29.08.2025

[KrSuHi 2012] Krizhevsky, Alex, Sutskever, Ilya; Hinton, Geoffrey E.:
ImageNet Classification with Deep Convolutional Neural Networks.
https://papers.nips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf, letzter Zugriff am 17.09.2025

[LeBoDe+ 1989] LeCun, Y.; Boser, B. & Denker, J. S. et al.:
Backpropagation Applied to Handwritten Zip Code Recognition.
https://ieeexplore.ieee.org/document/6795724, letzter Zugriff am 05.09.2025

[LiLiHe+ 2023] Liu, Nelson F.; Lin, Kevin & Hewitt, John et al.:
Lost in the Middle: How Language Models Use Long Contexts.
https://arxiv.org/abs/2307.03172, letzter Zugriff am 09.08.2025

[LiSaPo+ 2018] Liu, Peter J.; Saleh, Mohammad & Pot, Etienne et al.:
Generating Wikipedia by summarizing long sequences.
https://arxiv.org/abs/1801.10198, letzter Zugriff am 17.08.2025

[McMiRo+ 1955] McCarthy, J; Minsky, M. L. & Rochester, N. et al.:
A Proposal for the Darthmouth Summer Research Project on Artificial Intelligence.
http://jmc.stanford.edu/articles/dartmouth/dartmouth.pdf, letzter Zugriff am 17.09.2025

[Mistra 2025] Mistral AI:
Mistral Small NeMo.
https://mistral.ai/news/mistral-nemo, letzter Zugriff am 18.09.2025

[moinAI 2025] Preisübersicht moinAI.
https://www.moin.ai/preise, letzter Zugriff am 08.08.2025

[OrDiZe+ 2016] van den Oord, Aaron; Dieleman, Sander & Zen, Heiga et al.:
WaveNet: A Generative Model for Raw Audio.
https://arxiv.org/abs/1609.03499, letzter Zugriff am 16.07.2025

[PaTäDa+ 2016] Parikh, Ankur; Täckström, Oscar & Das, Dipanja et al.:
A decomposable attention Model. In Empirical Methods in Natural Language Processing.
https://arxiv.org/pdf/1606.01933, letzter Zugriff am 16.07.2025

[Plüste 2023] Plüster, Björn:
LEOLM: Ein Impuls für deutschsprachige LLM-Forschung.
https://laion.ai/blog-de/leo-lm/, letzter Zugriff am 15.09.2025

[Qdrant 2025] Qdrant Documentation.
https://qdrant.tech/documentation/, letzter Zugriff am 28.07.2025

[Qwen 2025] Qwen Team:
Qwen3: Think Deeper, Act Faster.
https://qwenlm.github.io/blog/qwen3/, letzter Zugriff am 15.09.2025

[RaBeMe 2024] Ranieri, Angelo; Di Bernardo, Irene & Mele, Christina:
Serving customers through chatbots: positive and negative effects on customer experience.
https://www.emerald.com/jstp/article/34/2/191/1215401/Serving-customers-through-chatbots-positive-and, letzter Zugriff am 23.08.2025

[RanYin 2025] Rangan, Keshav & Yin, Yiqiao:
A fine-tunin enhanced RAG system with quantized influence measure as AI judge.
https://www.nature.com/articles/s41598-024-79110-x, letzter Zugriff am 16.07.2025

[RaMuFa+ 2024] Raiaan, Mohaimenul Azam Khan; Mukta, Saddam & Fatema, Kaniz et al.:
A Review on Large Language Models: Architectures, Applications, Taxonomies, Open Issues and Challenges.
https://ieeexplore.ieee.org/document/10433480, letzter Zugriff am 16.07.2025

[Rosenb 1958] Rosenblatt, F.:
The Perceptron: A probabilistic model for information storage and organization in the brain.
https://www.ling.upenn.edu/courses/cogs501/Rosenblatt1958.pdf, letzter Zugriff am 13.09.2025

[RuHiWi 1986] Rumelhart, David, E.; Hinton, Geoffrey E. & Williams, Ronald J.:
Learning representations by back-propagating errors.
https://www.nature.com/articles/323533a0, letzter Zugriff am 30.08.2025

[ShChBa+ 2023] Shen, Xinyue; Chen, Zeyuan & Backes, Micheal et al.:
„Do Anything Now“: Characterizing and Evaluating In-The-Wild Jailbreak Prompts on Large Language Models.
https://arxiv.org/abs/2308.03825, letzter Zugriff am 17.09.2025

[ShaMis 2024] Sharma, Rishabh & Mishra, Abhinav:
Advanced NLP and ML Techniques in E-Commerce: Enhancing Customer Experience with AI Chatbots.
https://ieeexplore.ieee.org/document/10692186/authors, letzter Zugriff am 13.09. 2025

[SheNee 2016] Shellhammer, Alex & Neel, Juliette:
The need for mobile speed.
https://blog.google/products/admanager/the-need-for-mobile-speed/, letzter Zugriff am 13.09.2025

[Shopif 2025] Customer Accounts MCP server.
https://shopify.dev/docs/apps/build/storefront-mcp/servers/customer-account, letzter Zugriff am 22.08.2025

[Stobie 2020] Stobierski, Tim:
How to Calculate ROI to Justify a Project.
https://online.hbs.edu/blog/post/how-to-calculate-roi-for-a-project, letzter Zugriff am 08.08.2025

[Stefan 2025] Graziano Stefanelli:
ChatGPT vs Microsoft Copilot vs Claude vs Gemini: Full Report and Comparison. Accuracy, Speed, Interface, Integrations, Pricing, Enterprise Tools, Multimodal Support, Availability, and Unique Features.
https://www.datastudios.org/post/chatgpt-vs-microsoft-copilot-vs-claude-vs-gemini-full-report-and-comparison-accuracy-speed-inter, letzter Zugriff am 13.09.2025

[SurCas 2023] Surjandy & Cassandra, Cadelina:
The Effect of Using Chatbots at e-Commerce Services of Customer Satisfaction, Trust and Loyalty.
https://ieeexplore.ieee.org/document/10285799, letzter Zugriff am 13.09.2025

[Šuboni 2024] Šubonis, Martynas:
9x Model Serving Performance Without Changing Hardware.
https://martynassubonis.substack.com/p/optimize-for-speed-and-savings-high, letzter Zugriff am 09.08.2025

[ToBoSa+ 2022] Toosi, Amirhosein; Bottino, Andrea & Saboury, Babak et al.:
A brief history of AI: How to prevent another winter (a critical review).
https://arxiv.org/pdf/2109.01517, letzter Zugriff am 30.08.2025

[Turing 1950] Turing, A. M.:
COMPUTING MACHINERY AND INTELLIGENCE.
https://courses.cs.umbc.edu/471/papers/turing.pdf, letzter Zugriff am 16.07.2025

[VaShPa+ 2017] Vaswani, Ashish; Shazeer, Noam & Parmar, Niki et al.:
Attention Is All You Need.
https://arxiv.org/abs/1706.03762, letzter Zugriff am 16.07.2025

[Verma 2025] Verma, Sanjeev:
Enterprise AI Chatbot Development Cost in 2025– Let's Explore in Detail.
https://www.biz4group.com/blog/enterprise-ai-chatbot-development-cost, letzter Zugriff am 17.09.2025

[ViRaDh 2024] Vidivelli, S.; Ramachandran, Manikandan & Dharunbalaji, A.:
Efficiency-Driven Custom Chatbot Development: Unleashing LangChain,RAG, and Performance-Optimized LLM Fusion.
https://www.sciencedirect.com/org/science/article/pii/S1546221824005642, letzter Zugriff am 16.07.2025

[vizolo 2024] Chatbot-Betrieb: Preise erklärt.
https://vizologi.com/de/Erl%C3%A4uterung-der-Preise-f%C3%BCr-den-Betrieb-von-Chatbots/, letzter Zugriff am 17.09.2025

[WaHaHi+ 1989] Waibel, Alexander; Hanazawa, Toshiyuki & Hinton, Geoffrey et al.:
Phoneme Recognition Using Time-Delay Neural Networks.
https://www.cs.toronto.edu/~fritz/absps/waibelTDNN.pdf, letzter Zugriff am 13.09.2025

[Wallac 2003] Wallace, Dr. Richard S.:
The Elements of AIML Style.
https://files.ifi.uzh.ch/cl/hess/classes/seminare/chatbots/style.pdf, letzter Zugriff am 30.08.2025

[Weizen 1966] Weizenbaum, Joseph:
ELIZA—a computer program for the study of natural language communication between man and machine.
https://dl.acm.org/doi/10.1145/365153.365168, letzter Zugriff am 16.07.2025

[W3Scho 2025] JSON Syntax.
https://www.w3schools.com/js/js_json_syntax.asp, letzter Zugriff am 17.09.2025

[ZhLiPa 2024] Zhang, Qin; Liu, Ziqi & Pan, Shirui:
The Rise of Small Language Models.
https://ieeexplore.ieee.org/abstract/document/10897262, letzter Zugriff am 16.07.2025

Anhang

Interviews

Hinweis: Eckige Klammern […] sind dort gesetzt, wo Aussagen die Anonymität der Person gefährden könnten, das Thema zu weit gefasst ist oder sich der Kontext gedacht werden kann. Das Einverständnis, die folgenden Aussagen zu verwenden, wurde von jedem Gesprächspartner eingeholt.

Interviewpartner 1: Felix (Anonymisiert)

Arbeitsbereich/Fachbereich: Doktorand im Bereich KI-Strategie, Selbstständig im Bereich der Informationsvermittlung zum Umgang mit KI sowie Transformation von Unternehmens Prozessen mithilfe von KI

Frage 1:

Sören: In welchem Gebiet der Künstlichen Intelligenz bist du aktuell tätig?

Felix: Übergeordnet ist es das Thema KI-Transformation. Ich beschäftige mich damit, wie man Organisationen dabei helfen kann, das Thema ganzheitlich zu betrachten. Konkret gesprochen handelt es sich um das Thema KI-Strategieentwicklung, also darum, gemeinsam mit verschiedenen Organisationen, von kleinen über größere Unternehmen bis hin zu Konzernen, zu untersuchen, welche strategischen Fragestellungen relevant sind, um den KI-Wandel anzugehen. Implementierung mache ich zum Beispiel nicht. Die Schlagwörter sind also KI-Strategie und KI-Governance.

Frage 2:

Sören: Arbeiten Sie in einem Unternehmen oder eher in der Forschung?

Felix: Ich arbeite eher in der Forschung. Genauer gesagt habe ich in den ersten Jahren meiner Promotionstätigkeit einiges mit Conversational Agents (CA) gemacht und mir diese aus unterschiedlichen Gesichtspunkten angeschaut. Wie kann man eine Art Monitoring schaffen? Welche Qualifikationen gibt es für einen gut funktionierenden Chatbot, sprich: Worauf müssen Entscheider achten, wenn sie solche Systeme aufbauen, designen und auch implementieren und monitoren wollen? Ich habe viel mit Prototypen von CA beschäftigt und mit ihnen gearbeitet, sei es mit LLM-Anbindung oder hardcoded.

Frage 3:
Sören: Welche Projekte im Bereich KI haben Sie betreut beziehungsweise verwirklicht?

Felix: Ja, ich war involviert bei einem Unternehmen, das Online-Produkte verkauft, welches einen Chatbot implementiert hat, der Bestellprozesse sozusagen abbildet. Das heißt, es war ein nach intern gerichteter Chatbot, den man nutzen konnte, um Hardware, Peripherie etc. selbst zu bestellen. Das ist häufig in Organisationen immer mal eine kleine Herausforderung, wenn man jetzt in Konzernstrukturen denkt, tausende Mitarbeitende hat und diese nicht wissen, wo sie den Link oder das Formular zur Bestellung eines Laptops zum Beispiel finden. Da war ich involviert, hab mir das angeschaut und hatte Beratungseinfluss. Darüber hinaus war ich involviert bei einem Projekt, bei dem es um eine Art Chatbot-Listener ging. Das fand ich damals sehr spannend, weil man im Grunde in Konversationen zwischen Mitarbeitenden einen Listener eingebettet hat, der dann einfach mitgelesen und gehört hat. Datenschutzrechtlich muss man das so betrachten, dass die Organisation damit einverstanden ist und dass es eher so ein Experiment war. Der Chatbot konnte dann unterschiedliche Dinge tun, also Auswertungen zu bestimmten Themen, Zusammenfassungen oder Analysen durchführen. Dazu muss man aber sagen, dass das noch vor den ganzen LLM war. Jetzt geht es halt weiter in Projekten, wo Chatbots mit LLM zusammen gedacht werden, um ein gutes Interface abzubilden.

Frage 4:
Sören: Gab es besondere Herausforderungen oder Erfolgsfaktoren bei diesen Projekten?

Felix: Ja, das, was ich immer wieder sehe in der Forschung, aber auch vor allem in der Praxis, ist, dass die LLM-Modelle viel schneller weiterentwickelt werden als die Oberflächen beziehungsweise die Chatbots. Wenn ich von Chatbots rede, dann meine ich im Grunde damit häufig eine visuelle Oberfläche, hinter der dann Funktionalität steckt im Sinne von zum Beispiel LLM. Für mich sind Chatbots die visuelle Repräsentation und Verknüpfung mit diesen LLM. Auch wenn es so scheint, dass die Magie in den LLM steckt, ist es wichtig, dass die Chatbots eben mitgedacht werden, und das ist häufig nicht der Fall. Als konkretes Beispiel: Ich war in 2, 3 Organisationen, wo erstmals ein LLM angebunden und für die Organisation freigeschaltet wurde. Dann kam aber recht schnell das Tal der Ernüchterung, also dass die Mitarbeitenden sehr schnell unzufrieden waren, weil sie es anders gewohnt waren von beispielsweise ihrem eigenen ChatGPT oder Unternehmens-GPT. Das lag häufig auch nicht an den LLM, da man dort ja dieselben Modelle, zum Beispiel GPT, nutzen kann, aber die Oberflächen haben einfach nicht

mehr hergegeben. Es gab keine Kommentar- oder Referenzfunktion, die Multimodalität konnte nicht abgebildet werden und die Kontextualisierung hat nicht gut funktioniert. Eines der Erfolgskriterien ist, denke ich, die Modularität, dass der Chatbot möglichst anpassbar ist. Ich glaube, es ist sehr wichtig, dass das Ganze vor allem nutzerfreundlich ist. Und nicht nur das, sondern angelehnt an dem, was die breite Masse nutzt. Wenn man eine UI entwickelt, die der von ChatGPT komplett entweicht, hat man vielleicht einfach geringere Akzeptanzraten und das Thema Akzeptanz ist bei KI ja enorm.

Frage 5:

Sören: Wie bewerten Sie die aktuelle Entwicklung und Anwendung von Chatbots in verschiedenen Branchen? Welche Aspekte sehen Sie als besonders vorteilhaft oder kritisch?

Felix: Mhm, also ich glaube, dass Chatbots nach wie vor eine hohe Relevanz haben, auch im Kundenbereich. Vor allem, wenn es um den Servicebereich geht, weil es einfach das Interface derzeit ist, das immer mehr Raum einnimmt. Früher war es Telefonie, aber Chatbots bieten viel mehr Vorteile, wie zum Beispiel dauerhaften Service. Man hat trotzdem die Möglichkeit, an einen Mitarbeiter vermittelt zu werden. Mittelfristig wird das schon der Weg sein, mit Kunden zu kommunizieren. Ich sehe aber auch große Herausforderungen, weil Menschen viel zu hohe Erwartungen an solche Technologien haben. Wenn sie KI lesen, es nicht verstehen, wie es funktioniert, und denken, es sei ein Alleslöser. Und ähnlich ist es auch jetzt mit Chatbots und LLM-Integration. Man muss wissen, wie diese Systeme funktionieren, man muss lernen, zu prompten, man muss die Kniffe kennen, damit man die Systeme möglichst gut nutzt. Und man hatte das früher stärker, also wenn Chatbots nicht die richtige Antwort gegeben haben, hat man sie nicht mehr genutzt. Und auch heute ist es so: Wenn die Systeme nicht tun, was du willst, kommt die große Enttäuschung. Solche Systeme müssen kontinuierlich weiterentwickelt werden. Und langfristig bin ich mir gar nicht sicher, ob Chatbots wirklich die Lösung sind. Ich denke, dass sie noch einige Jahre Relevanz haben werden, aber mein Blick auf die ganze Geschichte ist, dass der Mensch faul ist. Er will Bequemlichkeit. Wenn etwas sehr gut funktioniert, dann wird der Mensch es irgendwann akzeptieren. Auch heute schreiben Millionen Nutzende über Dinge mit ChatGPT, ohne zu verstehen, wie es funktioniert, und trotzdem ist es ihnen egal, sie akzeptieren die Ergebnisse. Ich glaube, der Trend geht dahin, dass wir später alles per Voice machen und es später eine Art Assistent gibt, der mit unterschiedlichen Modulen verknüpft ist. Das gibt es heute bereits in Agentensystemen.

Frage 6:

Sören: Wie stehen die Unternehmen, mit denen du gearbeitet hast, zu der Entwicklung? Wissen sie auch eher nicht, wie Chatbots funktionieren und was deren Potenzial oder Herausforderungen sind?

Felix: Das ist eine gute Frage. Ich denke, die Kern-IT-Abteilung weiß es ganz genau, das heißt diejenige, die wirklich am Chatbot arbeitet. Aber ich glaube, eine große Herausforderung bei den Unternehmen, in die ich hineinblicken konnte, ist, über diese IT-Abteilung hinauszugehen. Also, wie erklärt man einem einfachen Mitarbeitenden, dass dieser Chatbot jetzt wirklich einen Mehrwert hat? Nehmen wir an, man hat einen super Chatbot, eine gute Kontextualisierung, super Antworten und man setzt da einen unerfahrenen Verkaufsmitarbeiter davor. Ob das so gut funktioniert, wage ich zu bezweifeln. Vielleicht beginnt man erst mit einer Sensibilisierung von Mitarbeitenden: Was ist KI, was kann sie? Und man zeigt ihnen Anwendungsbeispiele, bevor sie selbst damit arbeiten. Dazu haben Organisationen unterschiedliche Rollen wie Chief-AI-Officer, der das Ganze strategisch betrachtet.

Frage 7:

Sören: Denken Sie, die Entwicklung wird so schnell weitergehen oder erreichen wir eine Grenze?

Felix: Ich glaube, da wird sich in den nächsten Jahren einiges tun und die Entwicklung wird auch weiter voranschreiten. Ich glaube, diese Voice-Komponenten werden einfach immer präsenter, die hat man ja jetzt auch. Das hängt auch davon ab, wem man einen Chatbot vorsetzt. Ich habe Workshops gehabt mit älteren Menschen und die haben keine Lust auf eine Tastatur. Die wollen alles per Sprache machen. Ich denke, dass es zukünftig in den nächsten 5 bis 10 Jahren vielleicht alles über Personal Assistance laufen wird. Jetzt schon hat jeder ein Handy in der Hosentasche und mit den Assistenten hat man das jetzt schon mit Alexa, das sind erst die Anfänge. Man kann jetzt schon Echtzeitgespräche mit GPT-4 oder -5 führen, was eine unglaubliche Entwicklung ist, und ich glaube, mit der derzeitigen Entwicklungsgeschwindigkeit wird sich da einiges tun in den nächsten 3, 4, 5 Jahren. Wenn Sie eine weite Perspektive haben wollen, denke ich, dass wir vielleicht auch in 10 Jahren keine Handys mehr haben, sondern zum Beispiel durch Brillen oder Chips, wie Elon Musk es vorhat. Da bin ich vielleicht nicht der richtige Ansprechpartner, aber ich denke, das Handy wird einfach überflüssig werden und dann läuft sehr viel über Personal Assistance.

Frage 8:

Sören: Gibt es Anwendungsbereiche, in denen Sie den Einsatz von Chatbots als eher hinderlich ansehen?

Felix: Ja, die Frage finde ich schwierig. Ich sehe häufig die Potenziale. Also im HR-Bereich kannst du Chatbots einsetzen, da sehe ich auch überall Potenziale. Es gibt aber auch, dass diese Chatbots dann Bewerbende anderen bevorzugen oder diskriminieren. […] Es gibt aber auch Chatbots, die beispielsweise LGBTQ-Angehörigen vermitteln, so eine Art unbiased Instanz, die den Leuten dann versucht, Termine zu verschaffen, und das ist supergut. Allerdings gibt es auch Chatbots, die irgendwie einen Partner ersetzen sollen, das ist, glaub ich, ein Anwendungsbeispiel. Also wenn man einen Chatbot nutzt, um fehlende zwischenmenschliche Beziehungen zu simulieren. Das ist eine große Herausforderung und ein großer Trend, vor allem im asiatischen Raum. Die Leute nutzen und verlieben sich teilweise in solche Bots und meine Meinung ist, dass man halt von der Realität wegläuft. Allerdings ist das auch nochmal eine Diskussion, die den Rahmen sprengen würde.

Frage 9:

Sören: Welche konkreten Vorteile konnten Sie bei dem Einsatz von Chatbots im Bereich des Kundenservice im Vergleich zu herkömmlichen Methoden wie FAQ, Tickets, Formularen, E-Mail oder Telefon bereits erkennen.

Felix: Ja, also man hat mehr Möglichkeiten. Man hat die Wahl, mit dem Chatbot zu chatten, einfach zu sprechen oder weiterverbunden zu werden. Es ist deutlich schneller, nicht für alle, aber es kann deutlich schneller werden. Man hat dauerhaft Zugriff, man hat einen relativ neutralen, freundlichen Bot, man hat immer jemanden, den man versteht. Letzteres ist auch noch etwas, weil man die Mitarbeitenden im Kundenservice manchmal aus unterschiedlichen Gründen nicht versteht. Schnellere Bearbeitung, Verfügbarkeit, unterschiedliche Fragen beantworten, das sind alles so Vorteile. […] Je nachdem, wie der aufgebaut ist, kann man den Chatbot X Sachen fragen.

Frage 10:
Sören: Gibt es auch Nachteile oder Probleme, die Sie feststellen konnten?

Felix: Ja, die Kontext-Awareness ist manchmal so ein Aspekt, also dass der Kontext manchmal dem Chatbot schwieriger zu vermitteln ist als einem Mitarbeitenden, dem du dann in einem zweiten oder dritten Nebensatz nochmal mehr Kontext geben kannst. Das ist beim Chatbot natürlich anders. Wenn der von Beginn an nicht den richtigen Kontext hat, dann arbeitet er halt nur mit dem. Klar kann er sich seinen Teil denken, aber das ist nochmal ein bisschen schwieriger. Vermenschlichung ist noch ein Aspekt, ich finde das oft nicht gut gelungen. Man versucht ja immer, anthropomorphe Systeme zu bauen, und ich finde, das ist Nonsens. Man erreicht das einmal natürlich durch Sprachdesign, wo Linguisten häufig involviert sind, aber auch durch Chats, also wo beispielsweise diese drei Punkte angezeigt werden, und im Grunde ist das Zeitverschwendung. Man könnte es auch einfach ausgeben. Das sehen viele aber wahrscheinlich auch ganz anders.

Frage 11:
Sören: Sind Sie überzeugt von Chatbots als Aushilfe für den Kundenservice und die Kundenzufriedenheit?

Felix: Ja, absolut. Also wie gesagt, ich denke, wenn man einen Chatbot mit einer KI zusammendenkt, dann kann man im Kundenservice eine Menge erreichen und ich würde sogar so weit gehen, dass es mittelfristig das Ding sein wird. Und es gibt einige Organisationen, insbesondere mit der ich auch gesprochen habe, die jetzt im Kundensupport einen KI-Chatbot implementiert haben, der die Mitarbeitenden erst unterstützen soll und langfristig – wurde so nie gesagt, aber ist meine Annahme – auch ein Stück weit ersetzen soll, weil man feststellt, dass der Chatbot die Anliegen besser und schneller lösen kann. Kundensupport ist, denke ich, das erste Jobprofil und der erste Anwendungsfall, der zuerst angegangen wird, und alle, die in dem Bereich tätig sind, füttern jetzt diese Bots und irgendwann kippt der Punkt. Viele Unternehmen sprechen auch von kultureller Weiterbildung oder interner Weiterbildung, dass sie ihre Mitarbeitenden nicht kündigen, sondern einfach weiterschulen, aber das ist tatsächlich eine große Herausforderung.

Frage 12:

Sören: Da haben Sie mir meine nächste Frage schon ein bisschen vorweggenommen. Ich mach es ein bisschen konkreter: Denken Sie, dass Chatbots aktuell oder in den nächsten zwei Jahren den Kundenservice komplett ersetzen?

Felix: Komplett glaube ich nicht, aber was heißt komplett? Also, dass alle Organisationen jetzt KI-Chatbots implementieren, das glaub ich nicht. Wenn Sie das mit „komplett„ meinen, glaube ich, dass viele Organisationen einen KI-Chatbot implementieren würden, der ein gewisses Level an Standardisierung und Performance und auch die Möglichkeit liefert, Einsparungen im Personal vorzunehmen. Das glaube ich, wird in den nächsten zwei Jahren der Fall sein, absolut. Ich denke, es wird weiterhin eine Handover-Funktion geben, wenn der Chatbot nicht weiterkommt, dass an einen Mitarbeitenden übergeben wird, aber bis der Mensch komplett ersetzt ist, wird es noch lange dauern.

Frage 13:

Sören: Sehen Sie realistische Chancen für kleine und mittelständische Unternehmen, Chatbots aktuell erfolgreich einzusetzen? Wenn ja, unter welchen Voraussetzungen?

Felix: Ich glaube schon, dass sie Möglichkeiten haben. Großunternehmen haben natürlich andere personelle und finanzielle Mittel, das ist was anderes. Aber es gibt ja einige Open-Source-Frameworks für Chatbots, die man nutzen kann, wie zum Beispiel Rasa. Wenn nicht eigenständig, haben sie die Möglichkeit, auf Softwaredienstleister zurückzugreifen.

Frage 14:

Sören: Genau, da habe ich auch direkt etwas zu: Würden Sie KMU eher dazu raten, interne Lösungen zu entwickeln, oder auf Dienstleister zurückzugreifen, da beispielsweise nicht genügend oder die nötigen Ressourcen vorhanden sind?

Felix: Ja, also Eigenentwicklung ist dann möglich, wenn man wirklich Kapazitäten, also personell und finanziell, hat, aber eine pauschale Aussage zu treffen, ist schwierig. Oft braucht man IT-Entwickler für andere Dinge. Dann sollte man Frameworks oder bestehende Standardlösungen nutzen und häufig ist es doch besser, auch Experten von Dritten mit einzubeziehen. In KMU sind Mittel begrenzt und wenn sie investieren und es nicht funktioniert, wird das nicht noch einmal ausgerollt und eine Fehleranalyse durchgeführt wie in Konzernen beispielsweise. Das heißt,

diese One-Shots in KMU müssen sitzen. Also kurz und knapp: Dritte einbeziehen, ja, interne Mittel freimachen, ja, und dann aber eher auf Standardlösungen eingehen und weniger individualisieren.

Frage 15:

Sören: Dann sind wir so weit durch. Möchten Sie noch etwas im Zusammenhang mit dem aktuellen und zukünftigen Einsatz von Chatbots im Unternehmenskontext, auch im Hinblick auf KMU, ergänzen, was Sie als besonders wichtig erachten und was noch nicht angesprochen wurde?

Felix: Mhm, ich glaube, im Hinblick auf KMU müssen diese einfach den Schritt wagen und experimentieren, solche Systeme zu denken. Auch Ideen von unten, sag ich mal, in der Hierarchie zulassen und einfach mal ausprobieren und vor allem diesen KI-Wandel nicht vergessen und einfach schon mal einen Fuß in der Tür haben, sodass man nicht bei Null steht. Das ist nämlich auch noch so ein Aspekt, den man bei der Digitalisierung in Deutschland gemerkt hat, und ich würde behaupten, das geht bei KI sogar noch schneller. […] Das heißt, man sollte einige Personen im Haus haben, die sich ständig mit dem Thema beschäftigen und nach Potenzialen Ausschau halten, wie man das für diesen Vertrieb von Produkten nutzen kann, Verbesserungen von internen Prozessen etc. Das wird häufig in KMU nicht mitgedacht, weil es nicht unmittelbar Value hat, und Value heißt in dem Sinne Geld. Wenn die Dinge dann einfach verfügbar sind und die Konkurrenz es auch macht, ist es meist schon zu spät. Es gibt auch genug große Konzerne, die sich nicht genug damit beschäftigen. In den meisten Organisationen, in denen ich war, haben sehr wenige Menschen einen guten Plan davon, was KI eigentlich ist, was sie kann und wie sie eingesetzt werden kann.

Frage 16:

Sören: Daran anknüpfend: Haben Sie konkrete Empfehlungen für Unternehmen, die planen, Chatbots einzusetzen?

Felix: Austausch mit anderen Unternehmen und Experten ist ganz wichtig. Lernen von deren Erfolgen und Misserfolgen. In Communities hineingehen, eigenständig experimentieren, ich würde sogar sagen, die Kooperation mit Universitäten kann extrem lukrativ sein, weil Universitäten Netzwerk-Enabler sind. Man bekommt sehr gute Forschende und vielleicht auch Kontakt zu anderen Organisationen. Man sollte Messen und Konferenzen besuchen, die das

Thema KI auch strategischer betrachten. Man sollte KI nicht nur kurzfristig machen, also nicht nur den Chatbot implementieren, sondern weiterdenken. Wo könnte KI interne Prozesse besser machen? Man sollte im Backoffice-Kellerchen daran arbeiten und wenn es so weit ist und gute Practices gibt, ist man schnell am Zahn. Wie war es denn mit ChatGPT und Co.? Das habe ich mich persönlich gefragt. Sobald OpenAI ChatGPT released hat, kamen innerhalb von wenigen Wochen auf einmal ganz viele mit solchen Modellen an, aber keine wagte, den ersten Schritt zu tätigen, aber alle öffneten nach ChatGPT trotzdem ihre Schubladen. Organisationen, die mit dem Fluss der Zeit gegangen sind, hatten eher die Möglichkeit, auf den Zug aufzuspringen.

Sören: Ja, das ist eine interessante Ansicht. Damit wären wir fertig mit dem Interview. Ich danke Ihnen vielmals für diese diversen Einsichten!

Felix: Sehr gerne, hat Spaß gemacht!

Interviewpartner 2: Maximilian (Anonymisiert)

Arbeitsbereich/Fachbereich: Doktor im Bereich Chatbots, Projektleiter im Bereich KI-Strategie sowie Chatbot-Entwicklung

Frage 1:

Sören: In welchem Gebiet der Künstlichen Intelligenz bist du aktuell tätig?

Maximilian: Also ich bin generell Teamleiter, gleichzeitig Berater, Projektleiter und Projektmanager. Mein Team befasst sich generell mit KI-Fragestellungen und mit IT-Management-Fragestellungen. Im KI-Bereich machen oder begleiten wir Unternehmen vor allem auch zu Beginn, wenn sie noch nichts mit KI gemacht haben, so in die Richtung KI-Roadmapping. Also, wie geht man eigentlich an das Thema KI ran? Wie etabliert man verschiedene KI-Services? Das muss man organisatorisch auch abklären und das Ganze sozusagen etablieren, also in Richtung Datenschutz, Betriebsrat, ethische Fragestellungen auch teilweise. Der zweite Schritt ist dann sozusagen, dass wir wirklich Use Cases identifizieren im SAP-Bereich […]. Und da hat man am Ende auch einen Kundenservice und da stößt man relativ schnell auf die Möglichkeit, Chatbots oder Voicebots einzusetzen. Letzteres gewinnt im Moment so ein bisschen an Fahrt. Ich betreue Organisationen, die ganz klassisch ihre Kundenportale, ihre Kundenwebsites haben, wo Chatbots eingesetzt werden. Die Chatbots beantworten dann Fragen zu unterschiedlichen Dingen wie […]. Manchmal beantworten diese auch Fragen zu Karriere und HR-Services. Aktuell haben wir mehrere Chat- und Voicebots bei unseren Endkunden laufen.

Frage 2:

Sören: Welche Projekte im Bereich KI haben Sie konkret betreut beziehungsweise verwirklicht?

Maximilian: Ich bin seit einigen Jahren dabei, Chatbot-Entwicklung zu begleiten. Also ich habe in diesem Bereich promoviert und ein Forschungsprojekt geleitet, was sich um […] handelt. Wir haben bei Unternehmen interne Kundenservice-Chatbots eingeführt, auch da gab es ähnliche Herausforderungen, die ich gerade meinte. Bei einem Unternehmen beispielsweise hat sich das ganze Unternehmen gegen die Einführung eines Chatbots gesträubt und das Projekt wurde nach 2 Jahren gestoppt und das war eine große Fehlinvestition. Ich habe viel dazu geforscht, wie man Chatbots effektiv einführt und was die Designkriterien sind.

Frage 3:
Sören: Gab es besondere Herausforderungen bei diesen Projekten?

Maximilian: Ja, also ich glaube, es gibt verschiedene Ebenen, also zum Beispiel eine organisatorische, wo man generell so ein gewisses KI-Verständnis in der Organisation bereits besitzen muss, um zu verstehen, was können eigentlich KI-Technologien lösen und was nicht. Bei Endkunden ist es natürlich nochmal schwieriger, aber wenn man jetzt zumindest Chatbots intern einführt, dann muss natürlich die Belegschaft ein Stück weit darauf vorbereitet sein. Das ist bei Kunden schwieriger als bei der Belegschaft, die man darauf auch schulen kann. Bei Kunden ist das eine Art Blackbox und man muss das Ganze irgendwie datenschutzrechtlich, betriebsrattechnisch und ethisch vorbereiten. Viele brauchen dann erstmal einen Ethik-Code of Conduct, zum Beispiel. Meistens müssen die Themen bürokratisch und strategisch aufbereitet sein. Erst dann kann ein Entwicklerteam anfangen, den Chatbot zu entwickeln, sonst kann so ein Projekt auch schnell gestoppt werden. Eine Hauptherausforderung ist aus meiner Sicht, dass Chatbots zu Anfang recht wenig können. Also, es ist immer die Frage nach den Modellen und den Anwendungsfällen, aber so einen generischen Wissensdatenbank-Chatbot muss man nach und nach pflegen. Und gerade zu Anfang denken viele, dass der neue Chatbot ja so viel kann wie ChatGPT und Co. Die Herausforderung ist, das Unternehmenswissen da anzubinden, aufzubereiten und zu strukturieren. Auch, dass die Kunden oder Mitarbeitenden verstehen, dass der Chatbot eine Art Probephase ist. Man weiß zu Beginn häufig gar nicht, was die Kunden für Fragen stellen. Bei uns haben viele Kunden zum Beispiel auch Fragen zur UX und UI der Webseite gestellt. Dabei muss das Ziel dann auch manchmal sein, die Webseite zu verbessern oder eine Art Omnichannel zu schaffen. Und dann kommen noch die ganzen Gestaltungsthemen dazu, wie man mit Buttons, Freitextfeldern und Avataren arbeitet. Antwortzeiten sind wichtig, vielleicht auch sowas wie Handover-Funktionen, zum Beispiel Servicezeiten von 8 und 12 Uhr.

Frage 4:
Sören: Sehen Sie auch bestimmte Erfolgsfaktoren, zum Beispiel bei dem Design der Webseite, wie Sie erwähnten?

Maximilian: Ja, ich glaube, wenn ich mir eine perfekte Webseite vorstellen könnte, dann wäre das in der Zukunft so, dass der Chatbot einen vielleicht sogar viel mehr durch den Prozess begleitet. Also, wenn ich eine Produktinfo haben möchte, öffne ich eine Webseite und das Produkt und dann kann ich mit dem Chatbot chatten und das Produkt oder die Visualisierung passt sich

an. Wenn ich jetzt frage: „Hey, gibt es die Kopfhörer auch in Rot?“ Dann würde die Webseite direkt zu der roten Option des Produktes wechseln, sowas. Oder beispielsweise bei einem Handy die verschiedenen Speichergrößen. Also ich kann mir gut vorstellen, dass die Oberflächen immer mehr ineinandergreifen, werden in Zukunft. Aktuell ist es halt meistens so, dass die Chat-Bubble gut versteckt und unten rechts in der Ecke ist. So zumindest Direktverlinkungen zu gewissen Webseitenbereichen, zu Portalen oder so. Das wäre das erste Wünschenswerte. Also eine Verzahnung zwischen Chatbots und den Direktlinks, Webseiten, Plattformen oder Portalseiten.

Frage 5:

Sören: Wie bewerten Sie die aktuelle Entwicklung und Anwendung von Chatbots in verschiedenen Branchen? Welche Aspekte sehen Sie als besonders vorteilhaft oder kritisch?

Maximilian: Ja, also was ich zum einen merke, ist natürlich, dass die Sprachinterpretation sowohl von den Modellen her extrem gut inzwischen geworden ist. Vor ein paar Jahren war alles noch sehr regelbasiert, intentbasiert und richtig schlecht. Auch die großen Provider haben nichts Gutes geliefert. Jetzt, durch den Trend um ChatGPT und mit anderen KI-Modellen, merkt man halt, dass die Sprachmodelle an sich halt viel besser sind. Das wird in den nächsten Jahren, glaub ich, noch viel stärker werden, sodass wir uns wahrscheinlich um Spracherkennung, -interpretation und Antwortgenerierung keine Gedanken mehr machen müssen. Das wird immer besser werden, aber Wissen aufbereiten, interpretieren und Co. ist halt immer noch schwierig und teilweise wird schlechtes Wissen ausgegeben und halluziniert. Das wird ein Stück weit besser werden, was man durch die Agententrends merkt, wo es einen Hauptagenten gibt und dieser dann als Verteiler zwischen Unteragenten agiert, welche jeweils ihren eigenen Bereich haben. Wir haben jetzt bei einem Kunden so 10 Agenten eingeführt, wo ein Hauptagent dann immer nur aufruft, und man merkt, dass die Interpretation immer besser wird. Ich kann mir vorstellen, dass dieser Trend viel mehr in die Richtung Modularität geht, dass sozusagen Wissenscluster aufgebaut werden und die Agenten dann miteinander interagieren. Das wird der nächste große Trend, um das Wissensproblem zu lösen.

Frage 6:
Sören: Wie steht Ihr Unternehmen dazu?

Maximilian: Uns fällt eigentlich nur der Markt auf, beziehungsweise im Zweifel der Gesetzgeber. Das heißt, wir als Berater würden jetzt erstmal, glaube ich, alles umsetzen und automatisieren, was geht. Also wenn es ein Projekt gibt und es heißt, es sollen 5 % der Anliegen automatisiert werden, bitte identifiziert die besten 5 % oder die am leichtesten umsetzbaren 5 %, dann würden wir das tun. Man muss ehrlicherweise sagen, dass bei dem Kundenservice auch das Problem besteht, dass viel zu wenig Mitarbeitende bestehen und zu viel Druck auf dem Kundenservice ist. Das heißt, wir tun den Mitarbeitenden sogar einen Gefallen, aber auch hier müssen Mitarbeitende zur Verfügung stehen, um den Chatbot mitzutrainieren, und das ist meistens ein Engpass auf der Seite des Kunden, dass keine Ressourcen zur Verfügung stehen, um zu helfen.

Frage 7:
Sören: Denken Sie, die Entwicklung wird so schnell weitergehen oder erreichen wir eine Grenze?

Maximilian: Also, bisher ist es immer noch so, dass viele Serviceanliegen, zumindest bei uns in der Branche, noch manuell erledigt werden. Ich würde behaupten, noch 80 bis 90 % der Anliegen. Ich glaube aber, dass sehr viel mehr automatisiert werden kann. Also ich kann mir gut vorstellen, dass wir in 2, 3 Jahren 20, 30, 40 % mehr automatisieren werden. Ob es irgendwann eine Art Vollautomatisierung gibt, wage ich mal zu bezweifeln, weil ich das Gefühl habe, dass da die menschliche Ebene eine wichtige Rolle spielt. Das merken wir in Chatverläufen. Wenn man einen Vollakademiker hat, der genau sein Anliegen ausdrücken kann, was von einer generativen KI interpretiert werden kann, und die Prozesse alle perfekt sind, kann ich mir vorstellen, dass das in 5 bis 6 Jahren möglich sein wird. Aber wir stellen halt immer noch fest, dass auch viele Anliegen nicht richtig beschrieben werden können oder es geht um Sonderfälle, […], die teilweise einer sehr hohen Spezialisierung bedürfen. Aber ich glaube, wenn es um generelle Anliegen wie Produkte, Tarife, […] geht, da wird bestimmt ein Großteil automatisiert.

Frage 8:

Sören: Gibt es Anwendungsbereiche, in denen Sie den Einsatz von Chatbots als eher hinderlich ansehen?

Maximilian: Ja, ich glaube halt, alles, was komplexere Anwendungen angeht, die grafisch und tabellarisch arbeiten, wird mit Chatbots kompliziert werden. Wenn ich so nachdenke: Excel benutze ich den Co-Piloten gar nicht, bei Word ein bisschen und bei Outlook gefühlt jeden Tag. Das zeigt auch, wie gut Chatbots unterstützen können. Aber gerade im Bereich Marketing, Grafikbearbeitung und Co. kann ich mir gut vorstellen, dass generative KI zum Inspirieren aktuell gut ist, aber nicht so richtig, um da zu helfen. Im Bereich tabellarische Kalkulation oder Aufbau und Spezialsysteme glaube ich ebenfalls, dass es schwierig ist. Aber gerade was Wissensthemen, Wissensaufbereitung und Co. angeht, werden die Modelle sehr viel übernehmen. Ich kann mir vorstellen, dass Software immer enger verzahnt wird, wie ich bereits erwähnt habe. […] Ich glaube, gerade im SAP wird es Ewigkeiten dauern, bis bei komplexen Unternehmensdaten etwas Gutes herauskommt.

Frage 9:

Sören: Da sind RAG-Systeme oder Agentensysteme, die Sie erwähnt haben, vermutlich der richtige Ansatz. Ist es auch dort noch schwierig, modulare Daten zusammenzuführen?

Maximilian: Ja. Aber vor allem dieses automatisierte Verzahnen am Ende, was dann so die SAP und Co., wer auch immer das verspricht, wo man auf seinen Daten mal eben kurz einen Report generieren lassen kann, im Alltag, stell ich mir sehr schwierig vor. Die Daten sind teilweise so komplex und vielseitig, was es ohne Datenaufbereitung extrem schwer macht, sich dort etwas Sinnvolles generieren zu lassen, selbst wenn es technologisch eigentlich möglich ist.

Frage 10:

Sören: Dann kommen wir direkt weiter zum Kundenservice. Welche konkreten Vorteile konnten Sie bei dem Einsatz von Chatbots im Bereich des Kundenservice im Vergleich zu herkömmlichen Methoden wie FAQ, Tickets, Formularen, E-Mail oder Telefon bereits erkennen?

Maximilian: Ja, ich glaube, ein Riesenvorteil vom Einsatz von Chatbots im Kundenservice oder generell von Sprachassistenten ist, dass man diese dauerhaft und auch global ausrollen kann und diese immer verfügbar sind im Vergleich. Sonst hat man oft ein Service-Gap, wenn man nach

17 oder 18 Uhr und vor 8 Uhr niemanden oder nur Randpersonal, das zugeschaltet wird, im Kundenservice erreicht. Das ist vielleicht nicht so qualifiziert und eventuell auch sprachlich nicht. Wir haben zum Beispiel oft auch die Anforderung von mehreren Sprachen und ein Servicemitarbeiter spricht meistens zwei und ein Chatbot kann, wenn man zum Beispiel auch DeepL vorschaltet, schnell 4, 5, 6 Sprachen einwandfrei. Man kann also superschnell sowohl global als auch sprachlich skalieren und den Chatbot dauerhaft verfügbar machen. Man kann auch dahingehend Anliegen viel umfassender beantworten und es gibt keinen Qualitätsverlust, denn wenn ein Anliegen einmal gut dargelegt wurde, ist es ein standardisierter Ablauf. Das Wissen wird im Modell am Ende abgebildet. Neue Kundenservice-Mitarbeitende muss man immer wieder anlernen.

Frage 11:
Sören: Gibt es auch Nachteile oder Probleme, die Sie direkt im Vergleich feststellen konnten?

Maximilian: Ja, wie gesagt, die soziale Interaktion, gerade wenn es komplexere Probleme sind, wie ich meinte, zum Beispiel kritische persönliche Themen. […] Man muss mit den Menschen interagieren und ein bisschen heraushören: Wie tickt der Mensch, wie kann man ihm helfen? Und man muss auch ein bisschen Empathie an der Stelle zeigen. Ich denke, so 50, 60, 70 % kann man bestimmt automatisiert abbilden, aber so 20–30 % bleiben soziale Interaktion.

Frage 12:
Sören: Sind Sie überzeugt von Chatbots als Aushilfe für den Kundenservice und die Kundenzufriedenheit?

Maximilian: Auf jeden Fall, also gerade, wenn ich mir jetzt vorstelle, wie oft ich irgendwo beim Kundenservice anrufe und nur gebrochenes Deutsch geredet wird, weil der Service outgesourced wurde, da bin ich manchmal mit einem gut funktionierenden Chatbot oder ChatGPT-ähnlichem Chatbot zufriedener. Also wenn der Chatbot mein Anliegen am Ende löst, bin ich zufriedener, das per Chat kurz geklärt zu haben. […] Also ich glaube, auch gerade das Thema Warteschleife, Wartezeiten und schlechte Interaktion ist halt wieder das Gegenteil des Kundenservice, wo Sprachassistenten auf jeden Fall jetzt schon deutlich besser sind, da bin ich mir sicher.

Frage 13:

Sören: Ja das stimmt. Denken Sie, dass Chatbots aktuell oder in den nächsten zwei Jahren den Kundenservice sehr weit bis komplett ersetzen könnten, oder dauert das noch?

Maximilian: Ja, bei uns wird es auf jeden Fall, denke ich, noch länger dauern. Ich kann mir gut vorstellen, dass man vielleicht in zwei Jahren so 30, 35 % der Anliegen abgebildet hat. Großkonzerne können da vielleicht schon wirklich diese 30, 40 % erreicht haben. Ich glaube, dass es für KMU schwierig ist, und ich sag mal so: Kleinere Stadtwerke mit 1000 Mitarbeitenden haben vielleicht eher die Herausforderung, auch mit aufzuspringen, weil das natürlich einen hohen Invest bedeutet, und der Return ist bei Großkonzernen und der Menge der Anliegen viel höher. Bei Millionen von Anliegen 30 % zu automatisieren, birgt natürlich einen riesigen betriebswirtschaftlichen Vorteil.

Frage 14:

Sören: Sehen Sie generell realistische Chancen für kleine und mittelständische Unternehmen, Chatbots aktuell erfolgreich einzusetzen? Wenn ja, unter welchen Voraussetzungen?

Maximilian: Mhm, schwierige Frage. Ich glaube, zum einen fehlt es halt an technischem Personal. Dann bleibt halt noch eine Beratung oder eine Plattform über und wenn man in diese Richtung geht, dann gibt es viele Pflege- und Aufsetzungsaufwände, wo wiederum Mitwirkung bestehen muss. […] Und ich glaube, da besteht nicht viel Budget und keine Mitwirkungsmöglichkeiten. Im Zweifel ist es deutlich schwieriger für KMU im Vergleich.

Frage 15:

Sören: Würden Sie KMU eher dazu raten, interne Lösungen zu entwickeln oder auf Dienstleister zurückzugreifen, da beispielsweise nicht genügend oder die nötigen Ressourcen vorhanden sind?

Maximilian: Ich glaube, ich würde ihnen eher empfehlen, auf interne Ressourcen zuzugreifen, sodass man lieber jemanden hat, der dann im KI-Thema technische Expertise besitzt, und lieber da rein investieren, bevor man wahllos auf eine Plattform und einen Dienstleister setzt, der dann am Ende vielleicht nicht das halten kann, was er verspricht. Es besteht einfach diese Komplexität bei der Entwicklung von Sprachassistenten, wo auch viel, wie gesagt, mitgewirkt werden muss. Deswegen würde ich gerade mittelständischen Unternehmen eher raten, ein, zwei KI-Experten

einzusetzen, die das Thema auch ein bisschen ganzheitlicher aufbauen und sich auch Prozesse angucken. Nicht nur Sprachassistenten, sondern generell Prozesse und Abläufe im Unternehmen. Und Chatbots können da ein Baustein sein.

Frage 16:

Sören: Dann sind wir so weit durch. Möchten Sie noch etwas im Zusammenhang mit dem aktuellen und zukünftigen Einsatz von Chatbots im Unternehmenskontext, auch im Hinblick auf KMU, ergänzen, was Sie als besonders wichtig erachten und was noch nicht angesprochen wurde?

Maximilian: Mhm, ich stelle immerhin fest, dass dieses Agententhema immer interessanter wird. Das könnte, glaube ich, noch interessant sein, das immer mehr in Betracht zu ziehen. Ich glaube, sonst ist auch diese Verzahnung zwischen verschiedenen Kanälen interessant. Das halte ich für besonders herausfordernd. Ich denke, da gibt es noch viel zu tun und auch zu forschen.

Frage 17:

Sören: Dann noch eine kleine Frage zum Schluss. Welche konkreten Empfehlungen haben Sie für Unternehmen, die planen, Chatbots einzusetzen?

Maximilian: Ja, wo soll man da anfangen? Ich glaube, es ist erstmal sinnvoll, sich seine Prozesse und seine Daten zu Gemüte zu führen und erstmal zu wissen, was man eigentlich mit dem Chatbot erreichen will, und vielleicht auch eine Art Use-Case-Bewertung zu machen. Also, wo sind im Kundenservice die meisten Anliegen und welche davon können mit wenig Aufwand am ehesten automatisiert werden? […] Auch das ist generell ein Thema: Wo kann der Chatbot überhaupt das Anliegen komplett übernehmen? Dann sollte man eine Art Auswertung machen und die Themen priorisieren, bevor man so ein Chatbot-Projekt angeht. Man muss so ein Projekt vorbereiten und sich auch für die Initialisierung Zeit nehmen. Das ist oftmals leider nicht der Fall gewesen in den letzten Jahren und dadurch scheitern auch viele Chatbot-Projekte.

Sören: Ja, dann danke ich Ihnen für Ihre Einschätzungen und das Interview.

Maximilian: Danke auch Ihnen.

Gedächtnisprotokoll zu fachlichem Austausch: Thomas & Michael (Anonymisiert)

Arbeitsbereich/Fachbereich: Chatbot-Entwicklung im Bereich der KMU

Frage 1:

Sören: In welchem Gebiet der Künstlichen Intelligenz bist du aktuell tätig?

Thomas: Wir arbeiten aktuell mit KMU zusammen, um verschiedene Chatbot-Projekte zu entwickeln und einzusetzen. Wir haben diese Projekte allerdings häufig nur bis zur Übergabe nach der Entwicklung begleitet. Es gab ein paar Projekte im Bereich Support, wo ein Modell beispielsweise intelligent nach Tickets sucht, und diese dem Kundenservice verlinkt.

Frage 2:

Sören: Gab es besondere Erfolgsfaktoren oder Herausforderungen bei den Projekten?

Thomas: Man sollte bei kleinen oder generell Chatbots das Kontextfenster eher klein halten. Viele denken, man könne dem Modell einfach einen riesigen Kontext überreichen und es würde sich die relevanten Informationen heraussuchen, um die Anfrage perfekt zu beantworten. Allerdings existiert da das sogenannte „Lost in the Middle“-Problem, bei dem Modelle mit viel Kontext Informationen aus dem mittleren Bereich vergessen und oft nicht verwenden. Weiter ist es sinnvoll, überhaupt RAG-Methoden zu verwenden, da die Modelle je nach Anwendungsfall häufig nur eine dumme, intelligente Suche sind. Mit RAG kann man Daten- und Kontextaufbereitung durchführen und den Modellen nur die wichtigsten Informationen bereitstellen. Weiter bringt Payload-Indexing ebenfalls etwas, wenn man beispielsweise mit Filtern oder Metatags arbeitet und bei der Vektorsuche in dem RAG-Prozess bereits Dokumente herausfiltert und somit die Suche verfeinert.

Frage 3:

Sören: Welche konkreten Vorteile oder Nachteile konnten Sie bei dem Einsatz von Chatbots im Bereich des Kundenservice im Vergleich zu herkömmlichen Methoden wie FAQ, Tickets, Formularen, E-Mail oder Telefon bereits erkennen?

Thomas: Es entsteht viel Frustration bei den Kunden, wenn der Chatbot ihnen nicht helfen kann. Wenn bei einem Chatbot, der für Kundenservice entwickelt wurde, nach jeder zweiten Frage auf den Kundenservice verwiesen wird, wird die Kundschaft schnell aufhören, den Chatbot zu nutzen.

Außerdem ist der Datenschutz da eine besondere Herausforderung, besonders wenn man auf den Aspekt schaut, dass heutzutage die meisten Daten über Dienste von US-amerikanischen Unternehmen wie Amazon AWS oder Microsoft Azure gespeichert werden. Zudem gibt es auch Beispiele, wo Chatbots falsche Informationen verbreitet, falsche Verkäufe vermittelt oder Datenschutz gebrochen haben. Vorteile wären der dauerhafte Service und dass man sich dabei selbst durchklicken kann. Außerdem ist es eine gute Einsicht für Unternehmen, wenn diese sich anschauen, was Kunden angefragt haben. Dadurch kann ein weiterführendes Marketing entstehen.

Frage 4:
Sören: Sehen Sie generell realistische Chancen für kleine und mittelständische Unternehmen, Chatbots aktuell erfolgreich einzusetzen? Wenn ja, unter welchen Voraussetzungen?

Thomas: Ja, auf jeden Fall, wenn genügend Ressourcen vorhanden sind, sollten, KMU sich definitiv damit beschäftigen. Digitale Souveränität ist wichtig, damit man sich nicht dauerhaft von ausländischen beziehungsweise externen Systemen, wie zum Beispiel AWS und so weiter, abhängig macht.

Micheal: Die Chancen bestehen auch in verschiedenen Bereichen. Es muss nicht immer ein Chatbot nach außen gerichtet sein. Ein interner Bot kann auch bereits wertvolle Insights liefern, wie man Prozesse verbessern kann.

Frage 5:
Sören: Würden Sie KMU konkret eher dazu raten, interne Lösungen zu entwickeln oder auf Dienstleister zurückzugreifen, da beispielsweise nicht genügend oder die nötigen Ressourcen vorhanden sind?

Thomas: Jeder sollte entwickeln, wie er kann. Wenn es fix gehen soll, kann man auf Dienstleister zurückgreifen, aber wie gesagt, es ist langfristig besser, sich nicht abhängig zu machen.

Micheal: Wenn die Ressourcen nicht ausreichen für einen vollumfänglichen Chatbot, kann man auch bereits anfangen, eine intelligente Suche oder Website anzustreben. Dadurch kann die Navigation ebenfalls besser werden und bereits Erfolge bringen. Wenn die IT aber genug

Kapazität hat, dann kann man sich direkt selbst an die Entwicklung setzen und sich Kosten sparen.

Frage 6:

Sören: Welche konkreten Empfehlungen haben Sie für Unternehmen, die planen, Chatbots einzusetzen?

Thomas: Generell mit dem Thema Chatbots beschäftigen, das wird in den nächsten Jahren immer präsenter und all werden Chatbots nutzen. Da wird es wichtig sein, sich bereits damit auseinandergesetzt zu haben, damit nicht abgehängt wird.

Micheal: Man sollte sich auch mit anderen Unternehmen austauschen und so ein Projekt kann auch ein Leuchtturmprojekt für andere Unternehmen sein. Es gibt Messen, auf denen Unternehmen genau solche Projekte auch vorstellen und damit in den Austausch mit anderen geraten.

Sören: Ich danke Ihnen für die diversen Einsichten in diesem Gespräch.

Thomas: Ich danke auch Ihnen für das interessante Gespräch.

Micheal: Danke auch an Sie.

Eigenständigkeitserklärung

Hiermit versichere ich, dass ich die vorliegende Bachelorarbeit mit dem Titel:

Einsatz und Auswirkungen von Chatbots in Online Shops kleiner und mittelständischer Unternehmen

selbständig und nur mit den angegebenen Hilfsmitteln verfasst habe. Alle Passagen, die ich wörtlich aus der Literatur oder aus anderen Quellen wie z. B. Internetseiten übernommen habe, habe ich deutlich als Zitat mit Angabe der Quelle kenntlich gemacht.

Hamburg, den 18.09.2025
Ort, Datum

Sören Gooß