Exploring AI-driven Business Opportunities in Customer Service

Masters's Thesis

Spring Term 2024

Advisor:
Stefan Kluge

Chair of Quantitative Marketing and Consumer Analytics
L5, 2 - 2. OG
68161 Mannheim
Internet: www.quantitativemarketing.org

Table of Contents

List of Tables

List of Figures

List of Abbreviations

AI	Artificial intelligence
API	Application programming interface
BERT	Bidirectional encoder representations from transformers
GAN	Generative adversarial network
GPT	Generative pretrained transformer
LLM	Large language model
ML	Machine learning
NLP	Natural language processing
OC	Online community
OLS	Ordinary least squares
QQP	Quora question pairs
STP	Segmentation, targeting, and positioning
VAE	Variational autoencoder
VIF	Variance inflation factor
WLS	Weighted least squares
XML	Extensible markup language

Abstract

Artificial Intelligence (AI) is probably the most frequently used buzzword in every business meeting. Unfortunately, AI often remains a buzzword because companies are not always able to use it to profitably. However, there exist enumerable use cases for AI, especially in the field of marketing. Therefore, this work provides an overview of previous literature on AI-driven applications in marketing, with a focus on customer service, to show how AI can add value in various marketing processes. In addition, a comparative analysis of the similarity between human- and AI-generated (from ChatGPT) answers to questions from Stack Overflow is conducted. This provides insights into how the answers from ChatGPT would be perceived by the users as a substitute to humans. The combined approach led to the conclusion that AI can and should be used profitably while considering safety and ethical use, a customized design for each use case, and an enjoyable user experience for customers.

Keywords: Generative artificial intelligence, ChatGPT, AI-driven marketing, customer service, online communities

1. Introduction

AI is no longer a novelty these days. Many companies are making use of AI, such as Amazon's Alexa voice assistant (Amazon 2024), Allianz employing AI to detect insurance fraud (Allianz 2022), or Netflix' movie recommendation algorithms (Huang and Rust 2021). Furthermore, private investment in AI technology does not appear to be slowing down over the next few years as the amount of investments is forecasted to quadruple from 2020 to 2025 (Goldman Sachs 2023). The relevance of AI is also reflected in scientific research as the publications of academic articles on AI has risen sharply since 2017 (Vlačić et al. 2021). In the area of marketing in particular, AI represents a transformative technology that will have an impact on marketing strategy and customer behavior (Davenport et al. 2020). In response to changing customer behavior, marketers need to strategically employ AI along the customer journey to develop longer-lasting and more value-generating customer relationships (Huang and Rust 2023). Herefore, generative AI, such as chatbots, plays a crucial role in building a personal relationship with each customer (Li, Yao, and Nan 2023). However, there are only few studies on the effects of chatbots on customer engagement or customer satisfaction. Additionally, those studies are quickly outdated as AI models are advancing rapidly with each update extending their capabilities. ChatGPT, for instance, was published at the end of 2022 and there have been several version updates since then (OpenAI 2024a). Therefore, research needs to keep up with the advancements of AI for which this work aims to contribute to. Although there exist many studies on AI in health- and medical-related context (Budler, Gosak, and Stiglic 2023; Musheyev et al. 2024), business-related context (Jarco and Sulkowski 2023), or in research paper writing (Katar et al. 2023), only one study was found that examines the similarity of AI- to human-generated answers concerning technical questions from Stack Overflow (Sarker et al. 2023). The study from Sarker et al. (2023), however, focuses on identifying the semantic differences of words used rather than to find out how the AI-answer would have been perceived

by the users as a substitute of the human answer. The last two aspects are of particular interest, as the insights can help to find out more about how a customer service chatbot would be perceived compared to a human customer service employee without compromising the satisfaction of customers. In light of the collaboration between Stack Overflow and OpenAI, which aims "to create better products that benefit the Stack Exchange community's health, growth, and engagement" (OpenAI 2024b), this investigation becomes even more interesting.

1.1 Objectives and scope

This work follows two main objectives. First, the literature review provides an overview on AI-driven marketing applications, their effects on customers and customer service, and how they improve marketing processes and outcomes. Second, the conduction of a comparative analysis of one large dataset with human- and AI-generated text allows conclusions to be drawn about how well AI is likely to perform as a substitute for humans. This combined approach enables to better understand the results in the context of existing research and to derive more coherent conclusions.

1.2 Structure

The work commences with clarifying the technical background of machine learning (ML) and large language models (LLM) as well as describing the dynamics and user behavior of online communities such as Stack Overflow. Then, a broad literature concerning AI-driven marketing applications with a special focus on customer service is conducted which serves as a basis for the data analysis later on. After that, the methodological procedure is explained including the research design, sample collection, and explanation of variables. The results from the data analysis are subsequently presented. The work concludes with a discussion of the results, their limitations, and a future outlook on further research directions.

2. Theoretical Background

This study is strongly concerned with AI models and LLMs, specifically the GPT 3.5 turbo model from OpenAI. For a better understanding of the technical functionalities, it is crucial to introduce the basics of ML and LLMs and clarify relevant terminologies. However, the technical background of ML and LLMs is very complex and has developed over several decades. To stay within the limits of this master thesis, this chapter does not cover every aspect but instead aims to provide a general understanding. Moreover, this study uses data from the online community Stack Overflow. Therefore, it is required to unravel the key aspects of online communities and the dynamics between as well as the behavior of users in online communities. The following paragraphs discuss these areas, beginning with ML, LLMs, and closing with online communities.

2.1 Machine learning.

The question of what AI is and especially of what it is not has not yet been conclusively defined (Bandi, Adapa, and Kuchi 2023; Campbell et al. 2020; Huang and Rust 2021). However, AI is generally used as the overarching notion of various techniques and tools that enable machines to demonstrate and mimick human-like intelligence and behavior (Davenport et al. 2020). One key enabler is machine learning that functions like "An application of AI that […] automatically learns and improves from experience without being explicitly programmed." (Campbell et al. 2020). Generally, machine learning can be categorized into three distinct techniques that vary in learning and processing input: supervised ML, unsupervised ML, and reinforcement learning (Campbell et al. 2020; Overgoor et al. 2019).

For supervised ML, a system is trained on an existing set of data (training data) with labelled responses. This means that the system directly learns to predict new labels correctly based on its knowledge from the training data. Depending on the quality and amount of training data, the capabilities of the model can vary (Campbell et al. 2020). In contrast to that,

unsupervised ML does not attempt to predict outcomes but to identify patterns or clusters within and between datasets. The system is neither familiarized with nor trained on the data. Instead, it learns to recognize structural connections without informing the system about a correct or false conclusion (Campbell et al. 2020; Overgoor et al. 2019). Lastly, reinforcement learning is a process where a system takes different actions and learns continuously based on external feedback by trial and error (Campbell et al. 2020). Analogous to the learning method based on reward and punishment, reinforcement learning rewards a system for desirable behavior or outcomes and sanctions for unwanted actions (Gentsch 2019, p. 38). One simple example in marketing could be a machine learning model that aims to increase the amount of purchases from customers. By taking different actions and showing different advertisements, the model can learn from the customers' purchase decisions and deduct which actions are the most successful (Overgoor et al. 2019).

However, in practice these three different ML techniques are not used separately but combined to improve the overall performance and predictive power of the model. This combination is then called ensemble model (Campbell et al. 2020). Moreover, with the increasing computing power and availability of more data, deep learning has become very popular in the past. Deep learning can be described as a technique that is inspired by the way a human brain works and can learn and improve from experience without explicit intructions for a specific task (Campbell et al. 2020; Overgoor et al. 2019).

2.2 Large Language Models.

Within the context of the taxonomy of AI, LLMs fall in the field of machine learning and deep learning and can be considered generative AI in some cases (Shahab et al. 2024). Advances in natural language processing (NLP) enabled the development of LLMs. The difficulty in reproducing human language is that it is processed by complex cognitive processes in the human brain which is today still unfeasible for machines to replicate (Suri et al. 2024).

However, as previously discussed, machine learning enables machines to train on vast amounts of data and to identify patterns. Thus LLMs are able to observe, for instance, the statistical probability of words and sentences in a specific context without explicit programming (Suri et al. 2024). Significant progress towards LLMs as known today has been achieved in the 1980s with the advances of AI in the direction of deep neural networks and the associated deep learning technique. Deep neural networks are described deep for the depth of various different layers in a neural network through which data passes and from which the model learns progressively as each layer extracts more information (Shahab et al. 2024). Moreover, word embeddings and attention mechanism are two fundamental techniques for the operations of a LLM. The rationale behind word embeddings is to transform text into numbers or vectors "to capture the semantic meaning and contextual relationships between words in a language" (Suri et al. 2024). In order to obtain more context-awareness, attention mechanisms enable LLMs to give more attention to some parts of the input elements and less attention to other parts by assigning different scores (Suri et al. 2024). For example, in a transformer model an attention mechanism can be used to "[…] draw global dependencies between input and output." and allow "[…] for significantly more parallelization […]" (Vaswani et al. 2023). Hereby, the transformer model is only one of several possible architectures for generative AI models. There are also variational autoencoders (VAEs), generative adversarial networks (GANs), diffusion models, language models, normalizing flow model, and various combinations of these models, called hybrid models. These models mainly differ in their architecture components and training method which make them suitable for different tasks, e.g., transformers for NLP or VAEs for image generation and reconstruction (Bandi, Adapa, and Kuchi 2023).

Since this study is mainly concerned with the GPT 3.5 turbo model, only the transformer architecture will be further elaborated on. In simple terms, the transformer architecture consists of an encoder component that processes the input and a decoder component that generates the

outcome (Bandi, Adapa, and Kuchi 2023). As mentioned earlier, transformers typically employ attention mechanisms and can be built on several self-attention layers to be able to adequately process the relationships and dependencies between the input and output sequences (Vaswani et al. 2023). The GPT 3.5 model from OpenAI is a neural network that employs this transformer architecture with multiple layers and parameters. The model is trained on a significant amount of human-generated text and aims to generate human-like text by guessing the next word in a text sequence based on the previous text components (Roumeliotis and Tselikas 2023). The model training follows the approach of a combination of unsupervised and supervised ML practices. First, the model is fed with unlabelled data (text) to identify patterns and characteristics of human text. Then, a smaller dataset with labelled data is used to fine-tune the model on a specific set of tasks like text classification or question answering (Roumeliotis and Tselikas 2023). However, training the model on human-generated text data can also lead to the replication of human behavior, biases, and heuristics like loss aversion and effort reduction (Suri et al. 2024). To conclude, the GPT 3.5 model from OpenAI is a powerful generative AI model capable of processing human text and reproducing human-like text enabling a wide variety of potential applications (Roumeliotis and Tselikas 2023).

2.3 Online communities

This section examines the dynamics of online communities, especially technical online communities, and the behavior of their users. This understanding is beneficial for the analysis and interpretation of the results later on. Stack Overflow falls into the category of an online technical knowledge community (Burtch, Lee, and Chen 2023; Mustafa and Zhang 2022) where users can benefit from each other's computer science knowledge in a question-answering environment (Metzler, Günnemann, and Miettinen 2019). Metzler, Günnemann, and Miettinen (2019) investigated the stability and structure of online question-answering communities by employing a hyperbolic community model. This model is a concept in network science and is

particularly used to detect communities within complex networks by embedding nodes (connections) in a hyperbolic space. This allows for geometric interpretation where the likelihood of a connection between nodes is described as a function of their distance in this hyperbolic space. The shorter the distance the higher the likelihood of a connection and thus the probability of interaction between these nodes (Metzler, Günnemann, and Miettinen 2016). The authors analysed several online question and answering platforms and were able to detect a uniform structure across all observed platforms: “There is a small group of users who is responsible for the majority of the social interactions.” (Metzler, Günnemann, and Miettinen 2019). According to Metzler, Günnemann, and Miettinen (2019), there are approximately 20 % of the users actively contributing to the community and the other 80 % of users are passively participating. However, users who actively contribute to online communities tend to have a sense of reciprocation and thus expect others to repay them (or the community). If that is not fulfilled, the motivation to contribute could potentially diminish and thus decrease participation (Mustafa and Zhang 2022). Nevertheless, other factors may sustain user participation such as community recognition, benefits from social interaction, self-satisfaction, devotion to the community, or demographic features like age, gender or education. In order to reach maximum user participation, there exists no singular optimal configuration of all factors but various effective configurations. For knowledge contribution, it can be observed that the key drivers are online social interaction (sense of connection by supporting each other within the community), community recognition (for example, in form of appreciation or feedback by an upvote) and a sense of reciprocation (Mustafa and Zhang 2022).

With the emergence of perfomant AI models like ChatGPT, some researches initiated to investigate their effects on existing online communities. Generally, LLMs have a negative effect on user participation and knowledge creation, partly due to the fact that the abovementioned features, especially community recognition and social interaction, are less

fulfilled (Burtch, Lee, and Chen 2023; Li and Kim 2024). Burtch, Lee, and Chen (2023) have examined the effects of ChatGPT on platform traffic on Stack Overflow and Reddit in a time frame from October 2021 to March 2023. They observed a significant lower traffic and frequency of posted questions, particularly questions related to topics to which ChatGPT is already able to access the information from its training data. On the other side there is an increased website traffic through reduced "costs of content creation" (Li and Kim 2024) by, e.g., posting output from ChatGPT, and thus the amount of low quality contributions and misinformation increases (Li and Kim 2024). Li and Kim (2024) further discovered a decline in activity of highly engaged and qualified users but not for inexperienced users which potentially intensifies the change in content quality on the platform but above all "poses a threat to the expertise they cultivated" (Li and Kim 2024). However, these effects could not be observed for Reddit for the same topics indicating the "importance of social attachment" (Burtch, Lee, and Chen 2023). At this point, it must be considered that this field of research has not yet been sufficiently explored. So far, it cannot ultimately be concluded whether these effects harm the platform, its users or the knowledge generation in online technical knowledge communities (Burtch, Lee, and Chen 2023).

3. Review of Previous Research

As mentioned in the introduction, there has been a significant increase in academic research on AI in marketing since 2017 (Vlačić et al. 2021). Moreover, there is a lack of consistency and sometimes incongruent findings between different research articles, for example, different accuracy and appropriateness outcomes of the same AI model (Hu et al. 2023; Whiles and Terry 2024). Therefore, it is crucial to look deeper into current research and gain a better understanding of the impact and dynamics of AI and AI-driven marketing applications and how they may affect customers. Thus, this chapter provides insights into previous research and highlights the key findings of AI in marketing. The chapter is structured as follows: first, AI applications in marketing are explored, second, potential effects of AI on customers and customer service are examined, and third, the performance of AI in various scenarios is depicted.

3.1 AI-driven Marketing Applications

AI and machine learning algorithms allow for myriad applications in marketing (Ngai and Wu 2022). Thus, to provide a transparent and structured overview of the possible fields of AI applications in marketing and customer service, it is useful to map them into different categories. There exist several different approaches to categorizing, e.g., in a more detailed way using the 7P's (product, price, promotion, place, people, process, and physical evidence) (Ngai and Wu 2022), in a more condensed version into fewer categories (Ljepava 2022; Vlačić et al. 2021), or in the distinction between the type of task (Davenport et al. 2020). For this work, a more condensed categorization adapted from Ljepava (2022) and Vlačić et al. (2021) was chosen to reduce complexity and to focus on the application areas that are in line with the research objective. Moreover, it has to be noted that this is not an exhaustive list of possible applications but instead serves as an overview of the capabilities of AI in marketing. Thus, for this work, the identified applications of AI in marketing are divided into three categories,

namely (1) analysis, (2) customer relations, and (3) marketing strategy. The first category focuses on AI-enhanced techniques to analyze customer demographics, needs, and behavior. The second category covers relevant use cases for direct customer interaction and customer relationship management. The third category encompasses the practices that help marketers in strategic decision making.

Analysis. Large amounts of data are required for marketers to find out more about their customers or customer segments (Hossain et al. 2022). Hereby, AI is the catalyst to collect and process vast amounts of structured and unstructured data (Ngai and Wu 2022; Vlačić et al. 2021). The capabilities of AI in text, voice, image, and video analytics can provide more insights into unstructured data, for example, from online shopping and social media behavior, image recognition or heat mapping, than traditional methods (Campbell et al. 2020; Huang and Rust 2021). Some companies like North Face or Amazon already use enhanced analytics for social media and retailing to improve customer profiling allowing for more precise predictions about consumer choices and purchase behavior (Vlačić et al. 2021). AI-enhanced choice modelling techniques help in making decisions on how to target each individual customer in order to achieve a purchase decision on behalf of the customer (Brei 2020; Davenport et al. 2020).

Customer relations. Based on the aforementioned big data analysis of customer data, several opportunities arise to promote customer relations and to improve the customer journey. This can be achieved, for instance, by automating parts of the communication channels with customers using LLM-powered chatbots (Kim, Kim, and Baek 2024; Rivas and Zhao 2023). AI models like ChatGPT may increase customer engagement and satisfaction as well as reduce waiting times for customers by simultaneously handling several inquiries (Raj et al. 2023). Moreover, AI allows for a new level of personalization not only for customer segments but also for individuals. This personalization covers topics such as dynamic content creation or

individual targeting and personalized offers (Campbell et al. 2020). In addition to that, many companies already use AI-driven recommendation systems to better respond to the needs of customers like Netflix' movie recommendation algorithms or Amazon's cross selling suggestions (Huang and Rust 2021).

Marketing strategy. At the marketing strategy stage, AI is significantly useful in key strategic decision-making in segmentation, targeting, and positioning (STP). While AI helps with segmentation to recognize differences within target groups and discover micro-segments, it enables the implementation of sophisticated dynamic pricing and churn management techniques when addressing target groups (Huang and Rust 2021). In addition, the most frequently mentioned applications of AI in marketing strategy relate to demand and sales forecasting as well as estimating market growth (Brei 2020; Ljepava 2022). Moreover, there is also potential to improve brand image and brand development in terms of positioning. The brand Under Armour, for example, uses sentiment analysis to find out how its customers perceive their brand and to identify opportunities to further develop the brand (Campbell et al. 2020). Generally, AI can help to improve many areas in marketing strategy such as analyzing competitors (Huang and Rust 2021), identifying market or product gaps and developing new products (Campbell et al. 2020), implementing human-AI collaboration (Xueming Luo et al. 2021), and many more (Ljepava 2022; Ngai and Wu 2022).

3.2 Customer Perception and Acceptance of AI

In general, individuals differ in their readiness for technology, which affects whether customers accept and even endorse the use of technology in the customer journey or neglect it. Consequently, regardless the configuration or technical capabilities of a chatbot, for example, this individual preposition can influence a customer's engagement behavior (Yin, Li, and Qiu 2023). However, there are two factors that independently influence people's technology acceptance, namely perceived usefulness and perceived ease of use. Perceived usefulness refers

to the extent to which the individual believes that using the technology would be beneficial. Perceived ease of use refers to the effort that is considered necessary to use the technology (Davis 1985). From a causality perspective it is assumed that ease of use has an indirect effect on the use of technology, as it is mediated through usefulness. The reasoning behind this is that the ease of use positively affects usefulness because the less effort is required to use a system the more time is available for other tasks (Davis 1989). However, these two factors only explain technology acceptance if the use of the technology is equated with the acceptance of the technology. Additionally, this work focuses on the effects of the usage of AI in customer service. For AI, being the technology, more recent studies indicate that there are more factors that influence customers' perception and acceptance of AI. First, individuals are seen to consider ethical factors and the safety of using the technology (Arango, Singaraju, and Niininen 2023; Hui et al. 2023; Li, Yao, and Nan 2023). Individuals are sensitive to the type of information that is being processed by AI, i.e. individuals are less sensitive to the disclosure of information such as name or date of birth than to the disclosure of their ID number (Li, Yao, and Nan 2023). Moreover, if AI is used for ethical reasons, e.g., by generating artificial instead of real images of children to protect their privacy rights, individuals are more likely to support the usage of technology (Arango, Singaraju, and Niininen 2023). Second, the setting in which AI is used and the emotional state of individuals matter (Budler, Gosak, and Stiglic 2023; Crolic et al. 2022; Prentice and Nguyen 2020). Generally, conversational AI is perceived user-friendly and useful (Budler, Gosak, and Stiglic 2023). This also applies to retail, as AI-supported shopping in stores is positively correlated with the service experience (Farooqui 2022). Yet in "people-intensive industries such as hotels" (Prentice and Nguyen 2020) customers tend to prefer to be served by people rather than AI (Prentice and Nguyen 2020). Moreover, for customers that are in an angry emotional state, anthropomorphic AI chatbots, for example, can reduce customer satisfaction and therefore the likelihood of purchase (Crolic et al. 2022). As a

countermeasure, Crolic et al. propose that AI should be used to recognize whether a customer is in an angry or non-angry emotional state, although there is a considerable probability that the customer's emotional state will be misidentified, especially if the customer is dishonest or expresses vaguely (Huang and Rust 2023). Last, the configuration and empathy of conversational AI influences the way individuals perceive it (Elmashhara et al. 2024; Hui et al. 2023; Kim, Kim, and Baek 2024; Li, Yao, and Nan 2023; Puntoni et al. 2021). To increase customer engagement, chatbots that are able to create an emotional connection tend to increase positive customer engagement behavior (Li, Yao, and Nan 2023). Further, Kim, Kim, and Baek (2024) propose that "AI chatbots should be crafted with a focus on enjoyable, user-centered interfaces that foster long-term user satisfaction and engagement". In this context, Elmashhara et al. (2024) found that gamification of chatbots with simple games of chance, for example flipping a coin to win a discount, can increase both customer engagement and purchase outcomes.

To conlcude, AI can be both beneficial and harmful in customer service. Customers naturally have different levels of technology readiness which influences the acceptance for the use of AI along the customer journey (Yin, Li, and Qiu 2023). Nevertheless, conversational AI, for example, needs to be configurated individually for different use cases or target groups and requires a certain degree of adaptability (Garvey, Kim, and Duhachek 2023). Additionally, design decisions for AI applications in marketing must be made carefully, with the customer at the center and ethical guidelines in mind (Puntoni et al. 2021).

3.3 Performance of AI

To assess the performance of AI in marketing, researchers usually separate between two perspectives, with the first focusing on the comparison of AI to existing systems or other benchmarks and the second focusing on the outcome and thus the contribution to overall

(marketing) performance by using AI (Vlačić et al. 2021). Here, one more category is added examining the accuracy of AI in various scenarios.

Comparison. Most studies compare an AI model (like ChatGPT) to other AI models, to a human benchmark, or both. Regarding the comparison between various AI models, ChatGPT – and especially ChatGPT 4.0 – tends to perform best (Leong 2023; J. Liu et al. 2023; Lozić and Štular 2023). In a study that compared academic abstracts generated from Bard, ChatGPT, and Poe Assistant, ChatGPT is observed to generate most human like results across categories and differs from the other models mainly "in the use of subordinate, elaborating, and finite adverbial clauses" (Leong 2023, p. 127). The superiority of ChatGPT 3.5 and 4 has also been demonstrated by Lozic and Štular (2023). By comparing the correctness and scientific contribution of six academic articles generated by different AI models, Bing, Bard, Claude 2, Aria, Chat-GPT 3.5, and Chat-GPT 4, both chat GPT models showed more advanced results, especially due to their ability to reuse existing knowledge (Lozić and Štular 2023). Additionally, ChatGPT is considered a superior assistant in solving programming tasks more efficiently and with higher quality (Kosar et al. 2024; J. Liu et al. 2023). However, regarding the comparison between AI models and humans, human-generated text is still significantly different from AI-generated text (Hulman et al. 2023; Leong 2023; Sarker et al. 2023; Zaitsu and Jin 2023).

Outcome. AI can improve operational outcomes in terms of higher customer satisfaction or sales volume when it is used to complement human intelligence (Huang and Rust 2022). This can be twofold: time-savings, for example from delegating routine tasks to AI tools, or enhancing humans with the generative and analytical strengths of AI (Raj et al. 2023). Time-savings mainly stem from a reduced workload which allows employees to focus more on strategic responsibilities and decision making (Davis 1989; Raj et al. 2023). For example, ChatGPT can serve employees as a quick and informative source that helps to create a basic

knowledge of a particular topic within seconds (Karakose et al. 2023). Additionally, Katar et al. (2023) found that ChatGPT helps researchers to write academic articles by accelerating the acquisition of knowledge or processing more information in less time. While ChatGPT does not have the capability to write an entire article by itself, the authors state that it "can be used as an adjunct tool by the researchers while preparing their research papers" (Katar et al. 2023). In addition, ChatGPT is a useful tool to improve the decision-making process, e.g., for business consultants, by providing them with more information that enables more informed decision-making (Jarco and Sulkowski 2023). Moreover, employees need constant on the job training with researchers finding that, at least for sales agents, a combination of human- and AI coach boosts their sales performance more than just a human or AI coach (Xueming Luo et al. 2021). From a customer service perspective, AI has shown to produce high quality content tailored to customer preferences and needs by considering more factors and insights into customer data from big data analytics, leading to higher customer engagement (Raj et al. 2023). Additionally, chatbots in customer service can both increase customer satisfaction, by "Providing quick, informative, and more natural responses" (Raj et al. 2023) in less time, as well as increase efficiency of service employees by reducing the time and effort required to process customer inquiries (Andrade and Tumelero 2022; Raj et al. 2023).

Accuracy. Although few studies on the accuracy of AI, in particular ChatGPT, in question answering could be found that are comparable to this work, this paragraph provides a brief insight into the results of other or similar experiments. Generally, ChatGPT, in terms of accuracy, has not proven to be the best model for each and every NLP task. Kocoń et al. (2023) tested ChatGPT 3.5 and ChatGPT 4 in 25 different tasks whereby both models showed solid results. However, they lose in performance compared "to the best models currently available […], from 4 to over 70%" (Kocoń et al. 2023). This is particularly evident for more difficult and pragmatic tasks or tasks involving the evaluation of emotions. The authors conclude that

ChatGPT, regardless of the model version, is a model that covers a wide range of NLP tasks but does not perform best compared to other models. However, as ChatGPT is not fine-tuned to specific tasks or topics, Kocoń et al. (2023) expect better results from a fine-tuned ChatGPT model. Furthermore, in answering medical question, ChatGPT showed solid results, especially for very popular questions (Musheyev et al. 2024). In comparing medical approaches suggested from doctors versus ChatGPT for example, ChatGPT provided nine of the top 20 ranked responses (S. Liu et al. 2023). On the other hand, Whiles and Terry (2024) found that answers from ChatGPT were generally appropriate but "they frequently provide information which is either not factual or not comprehensive" (Whiles and Terry 2024). Hu et al. (2023) came to a similar conclusion as their analysis "suggests that ChatGPT may generate answers with incorrect facts and still lack knowledge" (Hu et al. 2023). The current state of research does not allow a final conclusion to be drawn regarding the accuracy of ChatGPT. It can be said that it generally performs well on most tasks but also has strengths and weaknesses that need to be considered.

4. Methodology

After discussing the theoretical background and reviewing previous literature, this section elaborates on the methodological procedure of the empirical study. The section starts with the research design and assumptions, then illustrates the procedure of the data collection, preparation, and analysis, and concludes with the explanation of the variables.

4.1 Assumptions and Design

The objective of this work is to investigate the similarity between human-generated and AI-generated answers to questions from Stack Overflow, identify significant predictors, and quantify their respective effects to subsequently draw inferences on how the responses of AI (ChatGPT) would possibly have been perceived by humans. The results together with the findings from previous research will then be used to assess AI-driven business opportunities in customer service. The study design incorporated a multivariate regression model to comprehensively assess the impact of multiple variables on the similarity between human-generated answers and AI-generated answers, in a question and answering scenario.

The chosen method is considered appropriate for this research objective for several reasons. Firstly, a multivariate regression model allows to examine several predictor variables simultaneously (Hair et al. 2013, p. 165). Given the complexity of the human- and AI interaction dynamics this method facilitates to identify the key variables. Secondly, the model provides insights into (1) how well each predictor variable explains the variability of the outcome variable and (2) the predictive power of the model including the strength of the relationship between the variables (Hair et al. 2013, p. 165). Furthermore, a multivariate regression model allows to assess not only the individual but also the combined effects of the predictor variables on the outcome variable and thus enabling a more nuanced understanding of the analysis (Hair et al. 2013, p. 166). Lastly, it allows to control for confounding variables

by including them as covariates to enhance the validity of the internal validity of the model (Hair et al. 2013, p. 166).

The hypotheses formulated in this study originate from theoretical and empirical considerations and underlie several assumptions. These considerations and assumptions will be critically discussed in chapter 6.3 Limitations and Future work.

H1 is based on the premise that the popularity of a human-generated answer refers to a high value in the variable score relative. The score relative represents the usefulness of an answer and reflects a common consensus on a topic within the community or a preferred response by the majority of the users. Corresponding AI-generated answers to the same question with a high similarity value are then considered to reflect a similar popularity as the human-generated answer. Since ChatGPT has been found significantly useful in supporting in programming tasks – even more useful than Stack Overflow – the similarity of AI-generated answers to human-generated answers is considered to be higher for more popular human-generated answers (J. Liu et al. 2023):

> H1: The more popular human-generated answers are, the more similar they are to AI-generated answers for the same question, compared to the less popular human answers.

H2 is based on the idea that the amount of time between when a question has been raised and when it was answered by humans reflects the complexity of the question and thus potentially leads to different similarity values. Questions that are answered after a longer time are considered to lack a common consensus within the community at the point when the question was raised and thus is more challenging to answer directly. For the questions that are answered immediately or after a short amount of time it is assumed that the community has developed a common consensus already. Since the strength of AI typically is to reuse existing knowledge instead of creating new, the similarity from AI-generated answers to human answers

is expected to be higher for answers with a shorter time passage (Hu et al. 2023; Roumeliotis and Tselikas 2023):

H2: The faster human answers are generated to a proposed question, the more similar they are to AI-generated answers for the same question.

Lastly, H3 is built on the idea that the similarity of a human-generated answer to its corresponding question refers to how closely the answer is related to the defined problem statement in that question. Thus, a high similarity to the question is considered to answer the question more intelligibly and accurately than answers with lower similarity. Consequently, AI-generated answers that are similar to human-generated answers with high similarity to their question reflect their intelligibility and accuracy. Generally, AI (especially ChatGPT) is considered to provide comprehensive answers with high understandability and relevance (Hu et al. 2023; S. Liu et al. 2023; Lozić and Štular 2023). Thus, it is expected to observe higher similarity between human- and AI-generated answers for high similarity to the initial question:

H3: The more intelligible and accurate human answers are, the more similar they are to AI-generated answers for the same question.

The idea behind the abovementioned hypotheses is to provide sound empirical evidence regarding the factors that shape the dynamics between human- and AI-generated text. This approach allows to quantify the effects of the respective variables and thus enhances the interpretability of the findings. Overall, the methodology chosen aligns with the study's goal and can provide actionable insights for practitioners and marketers.

4.2 Procedure

To investigate the similarity between human-generated and AI-generated answers, a preprocessed dataset from Gleasure et al. (2024) was utilized. The authors used an anonymized

dataset from Stack Overflow. The dataset was initially compiled from the quarterly official public database from the Stack Overflow Archive between 30 November 2020 and 1 September 2023. During that time Stack Overflow accumulated 3,923,852 questions in total. However, the dataset was then reduced to 536,286 questions and 1,290,998 answers by excluding entries with only one answer. Another 2,141 questions were removed by the authors because due to their length which exceeded the capabilities of the AI model that the researchers used for generating the answers. Hereby, the authors used the gpt-3.5-turbo-4k-0613 model, through the OpenAI API, to generate answers for the remaining 534,145 questions. The model is a version of the ChatGPT 3.5 turbo model (OpenAI 2024a) and will further be referred to ChatGPT. This led to a dataset of 534,145 questions from users from Stack Overflow with 1,290,998 corresponding human answers and additionally 534,145 AI-generated answers from ChatGPT to the same questions. To enable a comparison between human- and AI-generated text, the answers were then further processed, and the authors computed the semantic similarity with a QQP (Quora Question Pairs) fine-tuned BERT (Bidirectional Encoder Representations from Transformers) model. BERT is a state-of-the-art natural language processing (NLP) model (Devlin et al. 2019). QQP consist of a dataset derived from Quora and is widely used for tasks determining the semantic similarity of text (Devlin et al. 2019). The computed similarity value represents the predicted probability of the human- and AI-generated answers being duplicates. This measure enables inferences of how the AI-generated answer would have been perceived by the users of Stack Overflow and allows for comparability between answers.

Furthermore, a part of this preprocessed dataset was then obtained from one of the authors in a comma-separated value file. This file was uploaded to Excel via PowerQuery to obtain a first impression of the structure, size, and extent of the data. After that, each human-generated answer was mapped to its corresponding question, in a new Excel spreadsheet, using the answer and question parent ID so that each row in Excel represents one human-generated

answer and is unequivocally distinguishable from the other human-generated answers. In the following columns the respective values for each variable were entered. The variables are similarity to LLM answer, score relative, time since question, and similarity to question, and will be further elaborated on in the following chapter. For these variables, several empty cells were found. Each row with at least one empty cell in one of the relevant variables was excluded to avoid distortion of the empirical analysis. This resulted in a final dataset of 247,452 human-generated answers which was used for the subsequent analysis. The procedure of the analysis is described in the following paragraph.

The empirical analysis and statistical tests were conducted with Python. First, the necessary libraries and the data were loaded in the development environment Visual Studio. The main libraries used were numpy and pandas for basic functionalities as well as matplot and statsmodels for further statistical calculations such as the histograms or regression analysis. After that, the basic descriptive statistics were calculated whereby the median, mode, variance, and skewness were added separately as well as the correlation matrix. For further insights into the data, histograms and boxplots were created using the matplot library. For the histograms, the values were put in the x-axis and the number of frequencies in the y-axis to obtain a bar chart visualization. For the boxplots a vertical illustration was chosen. Only the boxplot of score relative was used because the others could not provide any recognizable insights. Then, the ordinary least squares (OLS) regression analysis was conducted using the statsmodels library by setting the dependent and independent variables and adding a constant term. To exclude multicollinearity or interaction effects, the independent variables were tested on the dependent variable individually which was done by simply adjusting the independent variables in the script. Since multicollinearity seemed likely to be an issue, the tolerance was calculated to determine the variance inflation factor (VIF). Hereby, each independent variable was made a dependent variable of all the remaining independent variables. The VIF was then calculated

manually by subtracting the R-squared of each of these three regressions from 1. Subsequently, the variables were tested for interacting effects. This was achieved by adding three interaction terms, for each possible interaction between the independent variables, to the regression model. Interaction 1 tests score relative with similarity to question, interaction 2 score relative with time since question, and interaction 3 similarity to question with time since question. After that, the regression model was investigated for heteroscedasticity using the Levene-test. For this test, the scipy library was added. The significance level was set to 5 %. A diagram of the residuals and the predicted values was created to graphically represent the heteroscedasticity. To correct for heteroscedasticity, a weighted least squares (WLS) regression was conducted by adding weights for each observation. The weight is the inverse of the squared residual which corrects in the way that larger residuals receive lower weights whereby smaller residuals receive higher weights. After that, no further tests were conducted as this would exceed the extent of this work. This will be adequately discussed in chapter 6.3.

4.3 Explanation of Variables

This chapter provides a detailed explanation of each variable to explain what they measure and how they are measured.

Similarity to LLM answer. The dependent variable is similarity to LLM answer. This refers to the similarity of the human-generated answers to the answers generated from the LLM, ChatGPT, for the same question. This variable allows to evaluate how closely the human answer aligns with its corresponding answer from the AI model. As previously mentioned, this variable is measured by computing the semantic similarity of both texts with a BERT model that is fine-tuned on the QQP dataset. The similarity to LLM answer is measured as a decimal in the interval from zero to one, whereby zero is equal to no (semantic) similarity and one indicates perfect (semantic) similarity.

Score relative. The first independent variable is score relative. It represents the relative voting score a human-generated answer has received until 1 September 2023 and can lie in the interval from negative 1 (-1) to positive 1 (1), whereby -1 is the lowest relative score and 1 is the highest relative score achievable. The score relative is calculated by the absolute score of one human answer divided by the sum of votes of all answers to this question. For example, if there are three answers to one question with the first two answers having received an absolute score of one and the third answer an absolute score of two votes, the first two answers have a score relative of 0.25 (one divided by four) and the third answer a score relative of 0.5 (two divided by four). This allows to observe how one answer has performed or has been perceived by the users in relation to all other answers. It improves comparability between answers. Furthermore, an answer with a high relative score indicates a convergence of other users toward this solution, whereas multiple answers with similar relative scores suggest less consensus.

Time since question. The second independent variable is time since question. This variable is measured because the value of a solution is likely to be time-sensitive. For some questions, the most optimal solutions might vary over time as technological dependencies could do. Moreover, as discussed earlier, the passage of time until an answer is submitted to a specific question can reflect the level of complexity and thus can potentially discover strengths and weaknesses of the tested AI model. In the context of customer service, customers favor shorter waiting times indicating a potential benefit of chatbots over humans. The variable time since question shows the time difference in hours between the question being raised and each corresponding answer being submitted. From the original dataset, this variable was measured in seconds. However, for better comprehension of the results, this was converted into hours by dividing each value by 3600.

Similarity to question. The last independent variable is similarity to question. This refers to the semantic similarity of a human answer to its corresponding question. The value is

computed analogously to the calculation of the dependet variable but in this case reflects the predicted probability of the question and the human answer to be duplicates. Similar to the dependent variable, the variable similarity to question allows to evaluate how closely the human answer aligns with its corresponding question. It is also measured as a decimal in the interval from zero to one, whereby zero is equal to no (semantic) similarity and one indicates perfect (semantic) similarity.

5. Empirical Analysis and Results

In this chapter, the results from the conducted analysis is presented. It begins with explaining the descriptive statistics of the independent variables and the dependent variable before moving on to the regression analysis.

5.1 Descriptive Statistics

The descriptive statistics of the data used are presented in this paragraph. The goal is to give an overview on the distribution of the data and to understand the basic characteristics of the variables. The data was taken from the sample described in the previous chapter and consists of 247,452 observations. Table 1 shows the univariate descriptive statistics of the variables, beginning with the three independent variables, score relative, similarity to question, and time since question, and closes with the dependent variable similarity to LLM answer.

"Insert Table 1 about here"

Score relative. For score relative, the mean is 0.37 which indicates, on average, a positive tendency of voting scores to the human-generated answers. Since many answers did not receive any score or votes at all their absolute as well as relative score is zero leading to a mode of zero. The standard deviation (0.395) and the median (0.33, which is below the mean) indicate a moderate variation in the dataset and a potential right skew. However, the results from Table 1 show a substantial negative skewness of -1,22 which suggests an unusual distribution (Hair et al. 2013, p. 34).

"Insert Figure 1 about here"

A histogram can be used as a visual comparison to a normal distribution (Hair et al. 2013, p. 33). Looking at the histogram of the variable score relative provides a better overview: the distribution has several peaks at (1) zero (~ 70,000 entries), (2) between approximately 0.5 – 0.6 (~ 45,000 entries), and (3) between approximately 0.9 – 1.0 (~ 40,000 entries). Also, there are only few entries below zero which confirms a left-skewed distribution.

"Insert Figure 2 about here"

From the boxplot it can be seen that 50 % of the values lie approximately in the interval [0, 0.667]. The box is shifted slightly to higher values which also confirms a negative skew.

Similarity to question. With a mean of 0.04, values for similarity to question are, on average, very small compared to the range of possible values. Moreover, the mode of 0.00012 and median of 0.00028 suggests that most values are still much smaller than the mean (see Table 1). Therefore, outliers could strongly increase the mean.

"Insert Figure 3 about here"

From Figure 3, it can be seen that the distribution of the variable similarity to question is substantially skewed to the right with a skewness of 4.49 (see Table 1) (Hair et al. 2013, p. 34). Thus, the observed statistics suggest that there are some outliers with high similarity values in the data. However, the absolute value of this variable is not very meaningful as it was computed with the help of machine learning. Rather, the significance of the variable lies in comparison to other questions. Additionally, there is always more than one human-generated answer to a question. Thereby, it is not appropriate to cut off the outliers with linear measures as this would rule out some of the answers to the same question and consequently reducing comparability as well as modifying the calculation of the variable score relative.

Time since question. The distribution of this variable is similar to the previous variable. On average the response time is approximately 1,210 hours (see Table 1), which is extremely high as the person raising the question would have to wait almost two months for a response. This might result from some entries with extremely high values as the maximum value is 24,031 hours. However, three fourths of the entries lie in the interval [0, 22.1] showing that most questioners do not wait longer than a day for a response on Stack Overflow. The extreme values could be due to very short questions or questions with little relevance for the community.

"Insert Figure 4 about here"

Compared to a normal distribution, this distribution is substantially skewed to the right with a skewness of 3.61 (Hair et al. 2013, p. 34). Taking into consideration the large standard deviation of 3942.67 (see Table 1) – relative to the mean – suggests the prevalence of outliers.

Similarity to LLM. Lastly the characteristics of the dependent variable similarity to LLM are comparable to similarity to question with a mean of 0.0386, median of 0.00046, and mode of 0.00012 (see Table 1). However, since 75 % of all entries are in the interval [0, 0.00439] which is much lower than the mean, this indicates that there are some very large entries with a maximum of 0.99110 (see Table 1).

"Insert Figure 5 about here"

With a skewness of 4.59 the distribution is substantially skewed to the right, reinforcing the general trend of low similarity values of the human- and AI-generated answers.

Generally, it can be stated that all variables show substantial variance in their distribution. Except for the variable score relative, all distributions are highly skewed to the right. Moreover, there are potential outliers in the data, especially for time since question, which could influence the data analysis.

After looking at the variables individually, Table 2 shows the correlation coefficients between the variables.

"Insert Table 2 about here"

Except for one, all correlation coefficients are negligible. Thus, collinearity or even multicollinearity between the independent variables seems unlikely so far which increases the predictive power of each individual variable (Hair et al. 2013, p. 161). Only for the independent variable similarity to question and dependent variable similarity to LLM answer a positive correlation of 0.44 (see Table 2) can be observed. This suggests that responses similar to the question are likely to be similar to answers from the LLM. This is a first indication that there

might be a connection between the two variables and further that LLMs can potentially understand the questions properly and respond to them in a similar way humans would.

The descriptive analysis shows that the human-generated answers are perceived, on average, positively by the Stack Overflow community and are mostly generated timely within one day after a question was raised. However, the semantic similarity to the questions and AI-generated answers is on average very low. Also, only one independent variable shows a relevant positive correlation to the dependent variable. Therefore, at this point it is hardly possible to draw conclusions. To allow for further interpretation of the data, in the next step a multiregression analysis is conducted.

5.2 Analysis

The procedure for the regression analysis is described in chapter 4.2 and the results are presented in the following paragraphs. For the regression, all independent variables and the dependent variable were included in the model.

"Insert Table 3 about here"

From Table 3, it can be observed that all independent variables have a p-value of 0.000 allowing to reject the null hypothesis and to conclude that all variables have a very high probability to be significantly different from zero (Hair et al. 2013, p. 189). The effect is positive for the variables score relative and similarity to question, and negative for time since question. Therefore, at this point, the null hypotheses can be rejected for all three hypotheses stated in chapter 4.1. The positive effect of similarity to question confirms the initial indication from the correlation matrix. However, the other two variables were not correlated to the dependent variable but are highly significant in the regression model. To find out if the effects are only significant in the presence of the other variables, each independent variable was tested separately in Table 4.

"Insert Table 4 about here"

From Table 4 it can be observed that all variables are still highly significant and the effect is even greater as the coefficients are larger. However, only the variable similarity to question has a notable R-squared value of 0.190, indicating that this is the only variable that explains some of the variance of the dependent variable (Hair et al. 2013, p. 152). For that reason, and the fact that (multi-) collinearity may still exist in some situations, although the corrleation matrix did not indicate collinearity, the next step assesses multicollinearity (Hair et al. 2013, p. 196).

Multicollinearity. As a first step the level of tolerance is determined and subsequently the variance inflation factor (VIF). The tolerance is determined using Python by successively transforming each independent variable into a dependent variable of all the remaining independent variables. The R-squared that is obtained from the three tests – for each independent variable – is then subtracted from one to obtain the tolerance value (1 – R-squared = tolerance) (Hair et al. 2013, p. 197). The result for each variable is very close to 1, indicating low or no multicollinearity. Subsequently the VIF is then obtained by calcualting the inverse of the tolerance, which is one for each variable, confirming no multicollinearity (Hair et al. 2013, p. 197).

Interaction effects. Since no multicollinearity was detected, it is important to find out if there are interaction effects between the independent variables (Hair et al. 2013, p. 695). Therefore, three interaction terms were added to the regression model: (1) score relative x similarity to question, (2) score relative x time since question, (3) similarity to question x time since question.

"Insert Table 5 about here"

The result shows two significant interaction effects for interaction (1) score relative x similarity to question and (3) similarity to question x time since question. Both interactions are negative and follow an ordinal interaction. Thus, when similarity to question is high, the effect of score relative on the dependent variable decreases. Similarly, the relevance of similarity to question

decreases when time since question is high. This could explain why the coefficient of each variable tested individually is higher than in a combined model.

Once the multicollinearity and the interaction effects were examined, it is also important to check if there are violations of the assumptions of the regression model. This includes (1) constant variance of the residuals (=homoscedasticity), (2) independence of the residuals, (3) normality, and (4) linearity (Hair et al. 2013, p. 178). Since the extent of this work as a master thesis is limited, the following paragraph focuses on testing for unequal variances only.

Heteroscedasticity. Heteroscedasticity is present when the variance of the residuals are unequal, thus potentially leading to distorted results and perhaps revealing statistical significant effects although they are not present (Hair et al. 2013, p. 182). To verify if the error terms have constant variance, Hair et al. (2013, p. 181) proposes the Levene-test. The test reveals that the variances are unequal indicating heteroscedasticity with a significant (p-value < 0.05) test statistic of 12562.81.

"Insert Figure 6 about here"

The scatterplot in Figure 6 illustrates that the variance of the residuals decreases as the predicted values increase, indicating heteroscedasticity. As a remedy, a variance-stabilizing WLS is conducted (Hair et al. 2013, p. 181).

"Insert Table 6 about here"

Although the model seems to have corrected for heteroscedasticity, the large R-squared of 0.999 suggests overfitting issues. This could affect the reliability and robustness of the results and further tests and investigations should be undertaken as described in chapter 6.3. However, this would go beyond the extent of this work, which is why the previous analysis is used for interpretation at this point. Although the results are not optimal, they can be construed as a first indication of relationships and represent the beginning of investigations in this direction.

6. Discussion

This chapter explains how the findings from the study and literature research contribute to theory and which recommendations for practice can be provided. Further, it critically discusses the limitations of this work and gives an outlook on future research avenues.

6.1 Contribution to Theory

This work contributes to theory in different ways due to its structure of a profound literature research in combination with an empirical analysis. The literature research provides a broad overview of AI-driven applications in marketing from a customer service point of view and categorizes them into three key dimensions (analysis, customer relations, and marketing strategy). Following the focus on customers, this work provides insights into how individuals perceive the use of AI in customer service and puts the technology acceptance model from Davis (1985) into an AI context. Although the technology acceptance model offers general guidelines why and when individuals could use AI, it does not comprehensively explain all facets of the use of AI. Moreover, Davis (1985) work was more focused on the use of computer systems by managers in a work-related environment who might have different assessment criteria compared to customers.

Furthermore, this work examines existing research on the dynamics of online communities and especially technical online communities such as Stack Overflow. In this regard, this work not only provides insights into the behavior of Stack Overflow users. It also offers a critical view on the use of AI in online communities as this could decrease the quality of the produced content in technical communities, for example (Li and Kim 2024). As a matter of fact, Stack Overflow banned the use of generative AI on its platform because it was considered to be overly error prone and thus harmful for its users (Makyen 2024).

Additionally, as not many research papers examinig AI performance in question-answering settings were found, the literature review presents findings from a seemingly

unexplored field of research. In this respect, the study carried out is unique in its research objective as no other study that investigates the similarity between human- and AI-generated responses to questions from Stack Overflow could be identified. The study adds to existing research in that it confirms that ChatGPT can generate human-like responses to complex questions.

6.2 Managerial Implications

The review of previous research and the results from the study allow to draw conclusions and make recommendations for marketers or other professionals for using AI in the area of customer service.

Generally, it can be said that AI offers benefits for both the customer by enhancing the customer experience and the company through optimizing operational processes and increasing efficiency. Many companies, such as Netflix, Amazon, or Under Armour, already make use of this technology. However, from the literature research, customers can have several considerations towards the use of AI which potentially influences their technology acceptance. Thereby, three key dimensions were identified which should be taken into account when implementing AI, namely safety, environment, and enjoyment. Firstly, the use of AI in customer service should follow ethical guidelines and ensure data safety as well as privacy. Secondly, the beneficial employment of a chatbot, for example, does not imply that it has exactly the same advantages in another context. This requires to thoroughly assess each use case separately and to include which role the respective AI application is thought to take over, be it a complaints management chatbot or a product advisor. Lastly, to increase customer engagement, the use of AI in customer service should follow an customer-centric approach focusing on an enjoyable and user-friendly experience. This could promote both the use of technology as well as the experience and satisfaction during use.

Moreover, it can be concluded that state-of-the-art AI is not capable of replacing humans entirely. Although the reviewed studies revealed remarkable results of different AI models, most authors indicate the limitations and drawbacks such as inaccuracy or lack of generating new knowledge (Hu et al. 2023; Whiles and Terry 2024). Yet it can be used as an adjunct tool to assist with recurring tasks or to enhance humans with its generative and analytical strengths. According to this finding, marketers should not follow the approach to replace humans with AI but rather to empower and train them to enhance their efficiency which potentially improves customer service experience and quality.

Furthermore, the conducted study suggests that ChatGPT, and especially the GPT 3.5 model, is capable of answering questions from Stack Overflow comprehensively and adequately. The results showed that the answers from ChatGPT were more similar to the more popular answers from humans with higher similarity to the intial question. In addition, ChatGPT can produce answers very quickly compared to humans which need in 75 % of the cases up to 22 hours for each single answer. This suggests that ChatGPT could improve response time to technical inquiries substantially and reduce manual work for humans. Though more research is needed, managers could consider to deploy AI models like ChatGPT in their customer service to improve operational efficiency and customer satisfaction. This would allow managers to reallocate human resources to areas where more knowledge and a creative way of thinking is required.

6.3 Limitations and Future Work

This work is subject to various limitations which are divided into sampling and data limitations as well as considerations from the study method and analysis. The section concludes with an outlook on future research directions.

The data, used for the analysis, originated from a public database which accumulated all questions and answers during the defined period of nearly three years. However, since

ChatGPT was released at the end of the year 2022 and the data was collected from 2020 up to the first September 2023, it could be possible that users posted content generated from ChatGPT on Stack Overflow despite ChatGPT being banned from the platform (Makyen 2024). This could imply that some supposedly human-generated answers were actually AI-generated and hence have a high similarity to other AI-generated answers. Moreover, the sample does not account for gender biases in the data. There is empirical evidence that women receive lower ratings for their contributions to the platform which could highly influence the variable score relative (Brooke 2021, p. 202).

For the analysis part, the main limitation is the reliability of the results as it could not be ensured that the assumptions of regression analysis from Hair et al. (2013, p. 178) were fulfilled, indicating the necessity for further investigations. For example, since the data could be non-linear, the values of the independent variables could be logarhythmically transformed to obtain a linearity, strengthening the validity of the linear OLS model. The application of non-linear regression methods could be another step to directly consider the non-linear relationships in the data (Hair et al. 2013, p. 179). Although a WLS regression was conducted to correct for heteroscedasticity, this most probably led to an overfitting of the results. Apart from that, the analyses investigate the collected data in its entirety. However, the distribution of the variables suggest that the variables exhibit varying dynamics for certain areas. Due to the large size of the sample, it would be appropriate to use machine learning algorithms, which would have exceeded the scope of this master's thesis. Hereby, unsupervised machine learning, as described in chapter 2, could be applied to identify clusters or further relationships between the variables which would be too complex for researchers to find out manually for this large dataset. Furthermore, the study is limited in the amount of variables included in the regression model. Since only the variable similarity to question showed a substantial R-squared value, the remainig predictor variables that influence the similarity of human answers to the answers from

the ChatGPT model are still unknown. Lastly, the results are limited since the used model is already outdated and was replaced with newer models such as the most recent gpt-4o model (OpenAI 2024a). This implies that the results are not representative for state-of-the-art AI models.

Moreover, this work provides the basis for further research directions. Since AI models, such as the ChatGPT model, evolve over time, newer and better performing models could be used to obtain more representative results. Additionally, comparing the base ChatGPT model to a fine-tuned ChatGPT model would allow to compare performance differences. Beyond that, categorizing the questions from Stack Overflow into different topics or question types could even create a more holistic picture of the dynamics of the variables. Furthermore, the conducted analyses examined the similarity of human- and AI-generated text quantitatively using their calculated semantic similarity. For validation purposes, future studies could investigate whether individuals have preferences over one response by directly comparing human- and AI-generated responses in a field study. Lastly, this work sheds a rather positive light on AI without adequately addressing its risks. However, there are serious threats for LLMs, such as the disclosure of sensitive customer data, which need to be considered when implementing them into customer service to ensure data security and privacy (The Open Worldwide Application Security Project 2023).

Tables

Table 1: Univariate descriptive statistics (own illustration)

variable	mean	median	mode	s. d.[1]	25%	75%	skewness	min	max
1. Score relative	0.37138	0.33333	0.00000	0.39507	0.00000	0.66667	-1.22	-1.00000	1.00000
2. Similarity to question	0.04004	0.00028	0.00012	0.14277	0.00013	0.00282	4.49	0.00006	0.99626
3. Time since question[2]	1210.15736	1.85861	0.15250	3942.67671	0.45611	22.05444	3.61	0.00000	24030.51639
4. Similarity to LLM answer	0.03860	0.00046	0.00012	0.13506	0.00017	0.00439	4.59	0.00006	0.99110

[1]Standard deviation
[2]Measured in hours

Table 2: Correlation matrix (own illustration)

variable	mean	s. d.[2]	1	2	3
1. Score relative	0.37	0.40			
2. Similarity to question	0.04	0.14	0.00		
3. Time since question[1]	1210.16	3942.68	0.01	-0.03	
4. Similarity to LLM answer	0.04	0.14	0.02	**0.44**	-0.04

[1]Measured in hours

[2]Standard deviation

Table 3: OLS regression analysis for similarity to LLM answer (own illustration)

R-squared	**0.192**
Adjusted R-squared	**0.192**
F-statistic	1.96E+07
Prob (F-statistic)	0.000

	coefficient	**standard error**	**t-value**	**p-value**	**95 % confidence interval**
const.	0.0208	0.000	59.488	**0.000**	[0.020; 0.021]
Score relative	**0.0067**	0.001	10.836	**0.000**	[0.005; 0.008]
Similarity to question	**0.4119**	0.002	240.759	**0.000**	[0.409; 0.415]
Time since question	**-9.80E-07**	6.20E-08	-15.819	**0.000**	[-1.1E-06; -8.59E-07]

Table 4: Each variable tested individually (own illustration)

1. Score relative

R-squared	**0.000**
Adjusted R-squared	**0.000**
F-statistic	107.2
Prob (F-statistic)	0.000

	coefficient	**standard error**	**t-value**	**p-value**	**95 % confidence interval**
const.	0.0360	0.000	96.519	**0.000**	[0.035; 0.037]
Score relative	**0.0071**	0.001	10.355	**0.000**	[0.006; 0.008]

2.Similarity to question

R-squared	**0.190**
Adjusted R-squared	**0.190**
F-statistic	5.82E+04
Prob (F-statistic)	0.000

	coefficient	**standard error**	**t-value**	**p-value**	**95 % confidence interval**
const.	0.0221	0.000	86.983	**0.000**	[0.022; 0.023]
Similarity to question	**0.4129**	0.002	241.309	**0.000**	[0.410; 0.416]

3.Time since question

R-squared	**0.002**
Adjusted R-squared	**0.002**
F-statistic	464.7
Prob (F-statistic)	5.8E-103

	coefficient	**standard error**	**t-value**	**p-value**	**95 % confidence interval**
const.	0.0404	0.000	142.367	**0.000**	[0.040; 0.041]
Time since question	**-1.483E-06**	6.88E-08	-21.556	**0.000**	[-1.62E-06; -1.35E-06]

Table 5: Interaction effects (own illustration)

R-squared	**0.192**
Adjusted R-squared	**0.192**
F-statistic	9801
Prob (F-statistic)	0.000

	coefficient	**standard error**	**t-value**	**p-value**
const.	0.0208	0.000	56.54	**0.000**
Score relative	**0.0070**	0.001	10.52	**0.000**
Similarity to question	**0.4209**	0.002	176.09	**0.000**
Time since question	**-9.279E-07**	8.90E-08	-10.43	**0.000**
Interaction (1)	**-0.0142**	0.004	-3253.00	**0.001**
Interaction (2)	**2.033E-07**	1.61E-07	1.27	**0.206**
Interaction (3)	**-5.212E-06**	5.52E-07	-9.44	**0.000**

Table 6: Weighted least squares regression (own illustration)

R-squared	**0.999**
Adjusted R-squared	**0.999**
F-statistic	1.31E+08
Prob (F-statistic)	0.000

	coefficient	**standard error**	**t-value**	**p-value**	**95 % confidence interval**
const.	0.0208	1.30E-06	1.60E+04	**0.000**	[0.021; 0.021]
Score relative	**0.0067**	2.11E-06	3.18E+03	**0.000**	[0.007; 0.007]
Similarity to question	**0.4119**	2.39E-05	1.72E+04	**0.000**	[0.412; 0.412]
Time since question	**-9.79E-07**	1.19E-10	-8.24E+03	**0.000**	[-9.8E-07; -9.79E-07]

Figures

Figure 1: Histogram score relative (illustration obtained from Python)

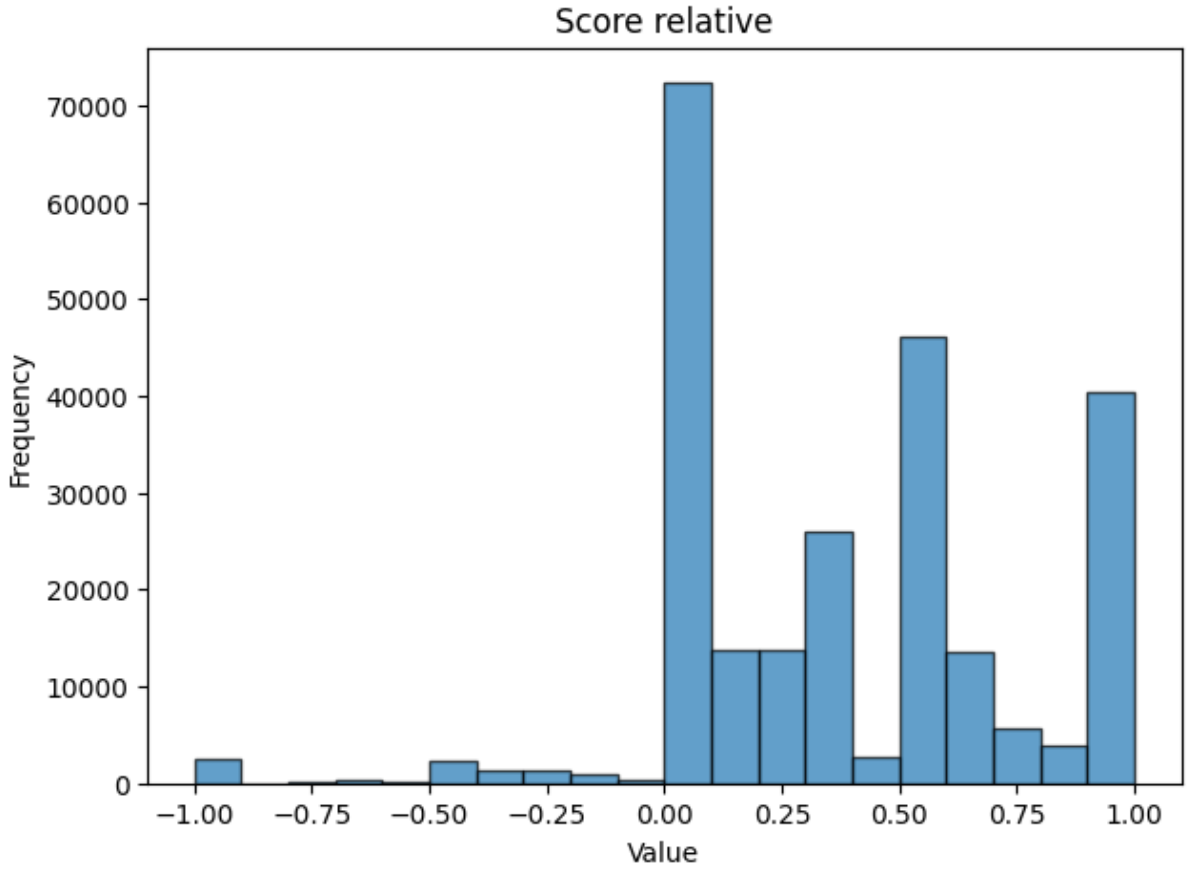


Figure 2: Boxplot score relative (illustration obtained from Python)

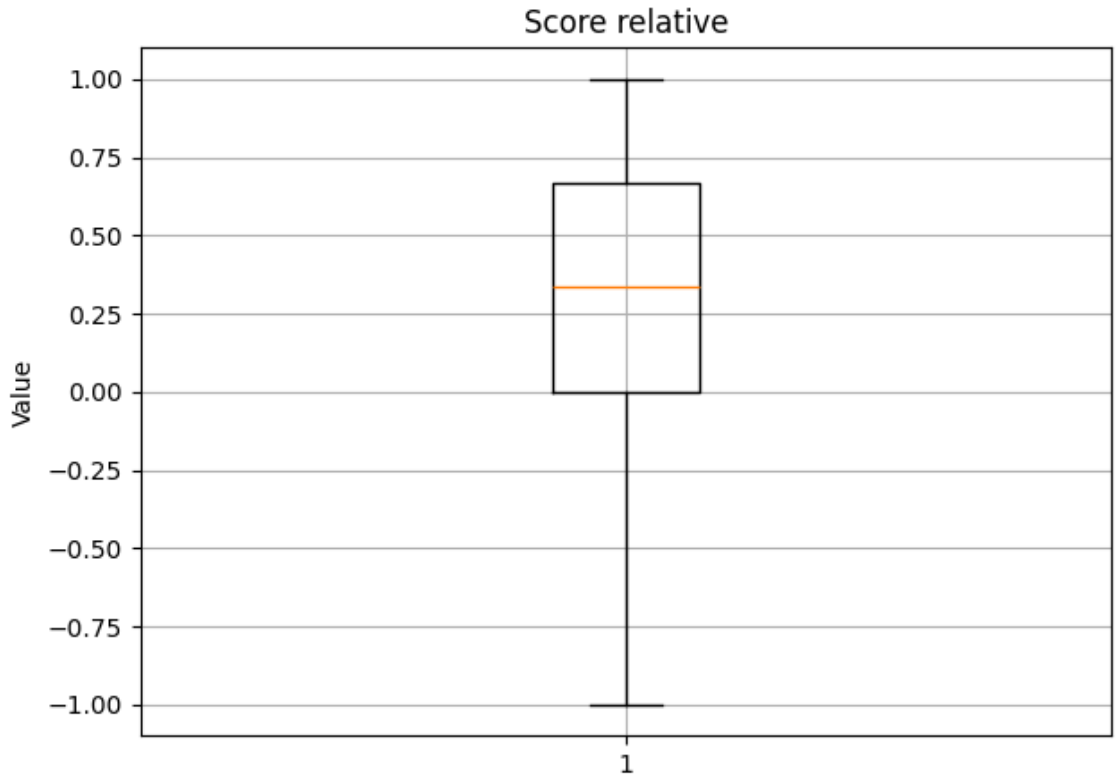


Figure 3: Histogram similarity to question (illustration obtained from Python)

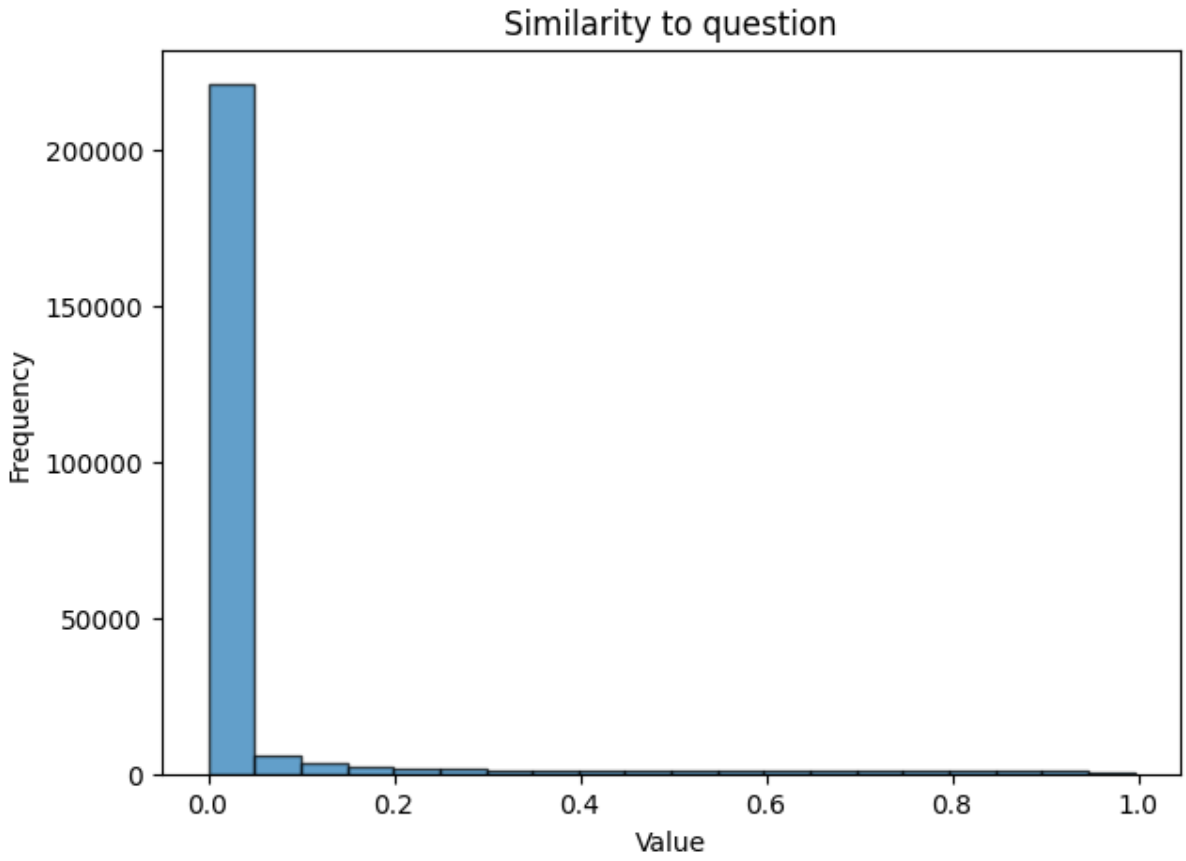


Figure 4: Histogram time since question (illustration obtained from Python)

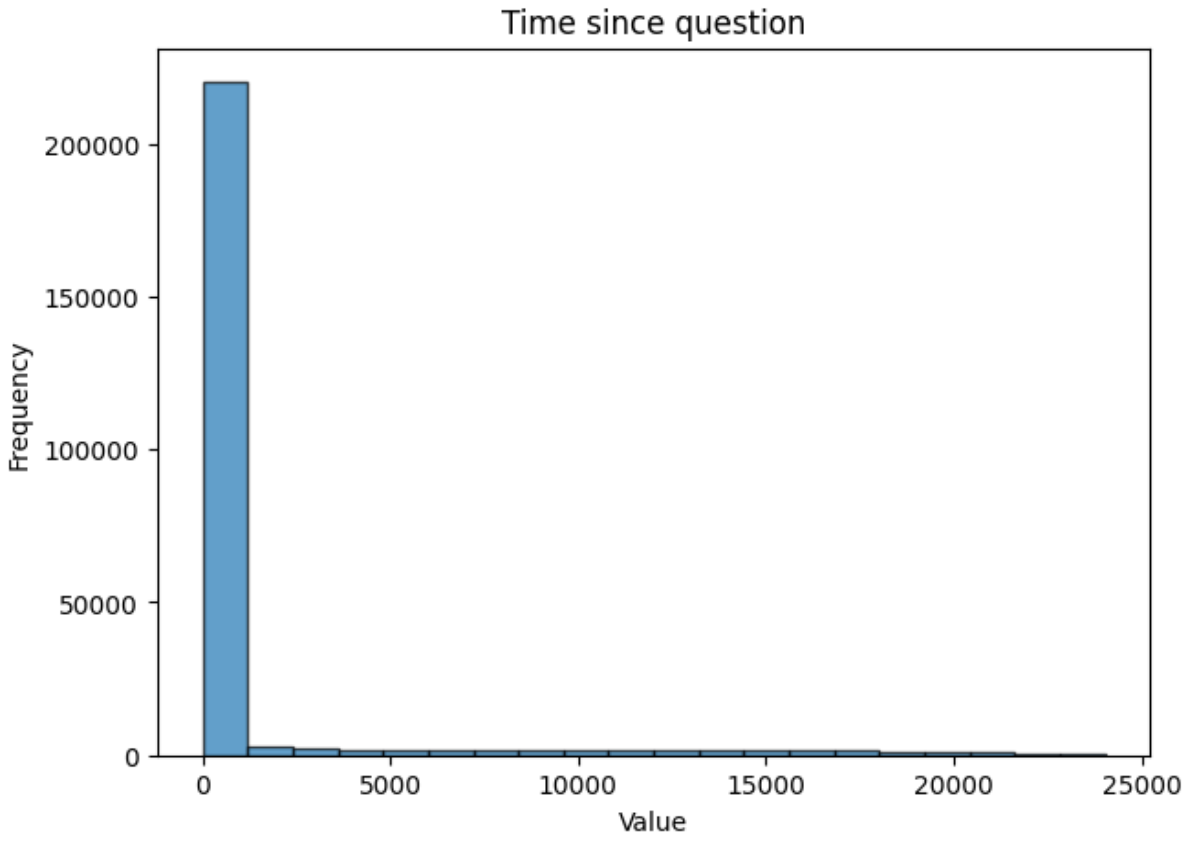


Figure 5: Histogram similarity to LLM answer (illustration obtained from Python)

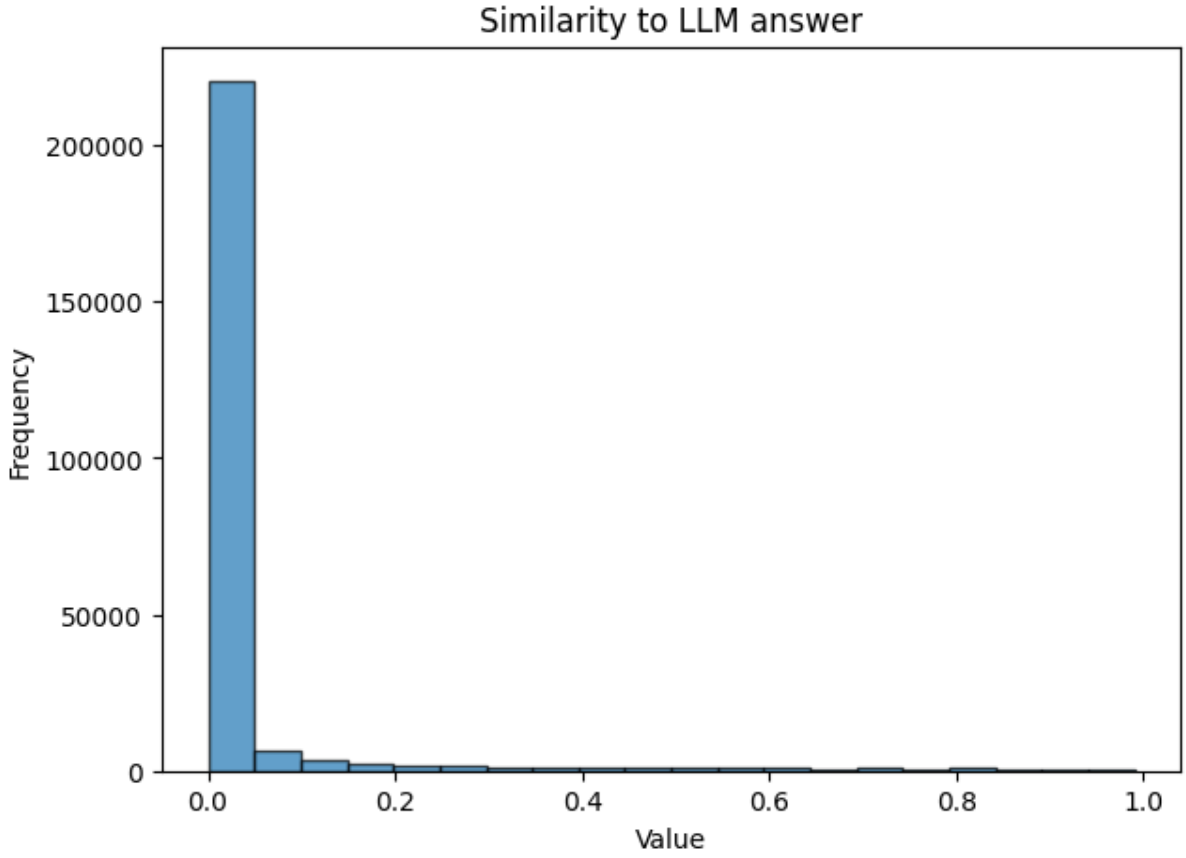
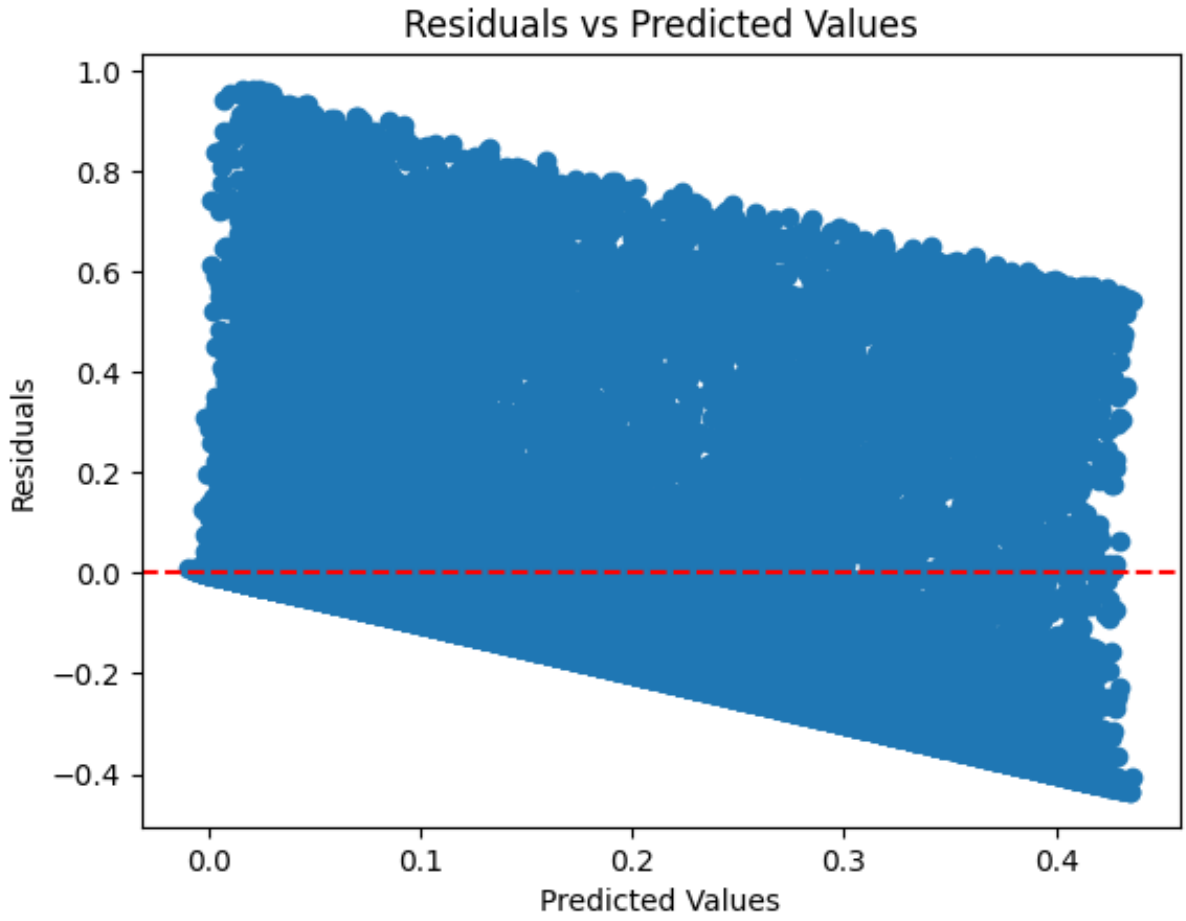


Figure 6: Plot of the residuals vs. the predicted values (illustration obtained from Python)



Appendix

Appendix A: Literature Review Table

Author/ Year	Summary	Type of AI	Method	Main Findings
Andrade and Tumelero (2022)	This study investigates the contribution of artificial intelligence to the efficiency of customer service.	AI Chatbots	Integration of the IBM Watson system into the Analytical Intelligence Unit (AUI) of a Brazilian commercial bank.	The chatbot service reduced call center and relationship center queues and enabled human operators to take on more complex tasks.
Arango, Singaraju, and Niininen (2023)	This study investigates the donation intentions of consumers to AI-generated charity advertisment.	Advertising AI	They conducted four studies in different settings: Study 1: two groups, where one is aware of the advertisment being AI generated and the other is not aware Study 2: 2x4 between subject design to investigate the effects of the motives behind using AI in the advertisement Study 3: 2x2 between subject design to investigate if the donation intentions of individuals change under extraordinary circumstances	Potential donors respond differently to the use of AI in charitable advertisements whereby ethical reasons like protecting children's rights and special circumstances can lead to a higher acceptance/positive response.
Bandi, Adapa, and Kuchi (2023)	This work provides guidelines for implementing AI systems and evaluation metrics to assess various AI models for different fields of application.	Generative AI	Literature review	The authors distinguished the requirements to implement AI in three categories, namely hardware, software, and user experience and provided a taxonomy of different AI for specific use cases to develop customized AI systems.
Brei (2020)	This works provides an overview of AI-driven marketing applications and examines future trends of AI on marketing.	Various AI applications	Literature review	AI will not substantially transform marketing in its entirety but has the potential to improve some marketing processes.
Brooke (2021)	This work investigates gender biases in the technical online community Stack Overflow.	Not AI-related	Analysis of a dataset from the Stack Exchange Data Dump which included 560,106 users. Then their gender was identified, and	The answers from women obtain lower scores, on average, even though women put more effort in their contributions to the

			different variables tested and categorized to males or females. The variables were user reputation, tenure, level of activity, answer score, answer effort, and readability. Moreover, network analysis revealed user behavior with respect to gender.	community. Moreover, women tend to interact rather with other women than with other males.
Budler, Gosak, and Stiglic (2023)	This work examines the effectiveness of conversational AI in answering health-related questions.	Conversational agents in question-answering systems	Literature review	Conversational AI applications such as chatbots are generally useful and intuitive to use which saves time and resources for personnel.
Burtch, Lee, and Chen (2023)	The authors investigate the effects of Generative AI on user engagement in two online communities, Stack Overflow and Reddit.	Generative AI	They conducted a study with web traffic and content data from Stack Overflow and Reddit from October 2021 to March 2023. They used the synthetic control method and benchmarked their results to a control group.	Generative AI has a negative effect on user engagement for Stack Overflow in terms of reduced traffic and content quality whereby no effects were seen for Reddit emphasizing the importance of social attachment in online communities.
Campbell et al. (2020)	This work provides an overview of various marketing applications along a 9-stage marketing planning process.	Various AI applications for different marketing functions	Literature review	There are many possibilities in all 9 stages of the marketing planning process that can improve marketing outcomes. However, companies must consider the risks of the use of AI, such as data breaches or unauthorized use of data.
Crolic et al. (2022)	This study examines the effects of anthropomorphic chatbots on customer satisfaction with regard to the customer's emotional state.	Chatbot	Based on a large-scale dataset (1,6 million text entries from customers into the chatbot interface) from an telecommunication company, the authors conducted investigated if an anthropomorphic chatbot influences the customer satisfaction if the customer is in an angry or non-angry emotional state. Study design: Based on an analysis of the words used by	For angry customers the chatbot led to a decrease in customer satisfaction and a lower probability of the customer giving feedback to the company. Even for customers with lower levels of anger this effect is still significant. However, the negative effect of this type of chatbot does not hold for customers in a non-angry emotional state.

			customers, the proportion of words that are associated with anger is identified to identify the level of anger. This is then compared to the level of satisfaction rating by customers and the likelihood of them giving feedback to the company.	
Davenport et al. (2020)	This work examines how AI could potentially influence the future of marketing with respect to different intelligence levels of AI.	Various AI applications	Literature review and exchange with experts	The outcome of this work is a framework of how AI will influence future marketing activities by looking at the intelligence levels of AI, the type of task, and for AI that is deployed in a robot. They conclude by stating that in the short-term, there won't be radical changes of marketing processes but in the future AI is likely to augment humans rather than to replace them.
Davis (1985)	This work develops a technology acceptance model and examines the predictors of technology acceptance.	Not AI-related	This work follows a threefold approach by conducting a literature review, survey, and experimental tests.	Technology acceptance is driven by two factors, namely perceived usefulness and perceived ease of use.
Davis (1989)	This work examines the effects of three variables on technology acceptance, in particular computer technology and aims to validate the dynamics of three variables, namely perceived usefulness, perceived ease of use, and user acceptance.	Not AI-related	Study 1: Survey of 120 participants on the use of two computer systems to rate the perceived usefulness and perceived ease of use of the two systems. Study 2: Field experiment with 40 MBA students to evaluate the perceived usefulness and perceived ease of use of two computer systems which the participants did not know beforehand and were only introduced to them beforehand in a one-hour crash-course.	Perceived usefulness and perceived ease of use are two predictor variables for the use of technology. Perceived usefulness, however, is stronger than ease of use, because the users value the benefits from using technology higher than the usability.
Devlin et al. (2019)	This work introduces the BERT model and shows different pre-	Bidirectional encoder representatio	The authors tested a fine-tuned BERT model	The BERT model is an easy-to-use AI model which can be simply

	training opportunities for the model.	ns from transformers	on eleven different LLM tasks.	fine-tuned to be capable to solve many tasks without profound modifications.
Elmashhara et al. (2024)	This work examines how the gamification of conversational AI in customer service affects the customer engagement, motivation, and their purchase behavior.	Gamified conversational AI	Study 1: The authors employed a chatbot in the Facebook Messenger app, using Chatfuel, in an online retail context. The chatbot was gamified with a quiz game that covers the investigation of both hedonic and utilitarian motivation. The goal of the game was to answer the questions correctly and based on the number of correct questions they received a discount. In the study, they investigated direct (cognitive, emotional, and behavioral engagement) and indirect (utilitarian and hedonic motivation) effects. Study 2: This study was designed to find out the game that leads to the desired customer behavior by employing several chatbots with different games. Study 3: The last study was to find out what is the best game that leads to direct action on side of the customer. 218 were randomly allocated to one of three games from Study 2.	The gamification of chatbots can be beneficial and detrimental. The optimal result can be obtained by using a game of chance which leads to higher purchase probability and is simple enough to not overwhelm the customer.
Farooqui (2022)	This study examines the effect of AI on shopping experience of customers, with a focus on customer loyalty and service experience.	Voice assistants, chatbots, robots, and digital devices	This study interviewed customers from Pune City who shopped traditionally and those who shopped with the help of various AI tools.	There is a positive correlation of AI and service experience of customers.
Garvey, Kim, and Duhachek 2023	This study consists of three different experiments to examine the reaction of consumers to	Chatbot	Study 1a: experiment with 174 undergraduate students, examined response to worse-than-expected	For worse than expected offers consumer respond better to an AI agent and for better-than-expected offers consumer respond

	discrepant offer presentation.		offer from an AI agent versus a human agent Study 1b: experiment with 299 participants, examined response to a better-than-expected offer from an AI agent versus human agent Study 2: experiment with 174 undergraduate students and 299 other participants, examined Inferred AI Intentions Alter offer acceptance Study 3: Experiment with 403 members of an MTurk online panel, examined anthropomorphism of AI agents	better to a human agent. This is partly due to inferred intentions of consumers.
Hossain et al. (2022)	This study observes the relative advantages of the use of AI to enhance marketing analytics capabilities for export-oriented industrial goods manufacturers	Various marketing analytics capability enhancing AI applications in marketing	Multimethod investigation: literature review, news reports review, and manager interviews	The level of AI adoption moderates the positive effect of marketing analytics capabilities (MACs) on competitiveness whereby companies with high MACs profit more from high AI adoption and companies with low MACs profit more from low AI adoption.
Hu et al.(2023)	This work examines pre-trained language models in answering questions that require a lot of knowledge.	Pre-trained language models	The authors created 18 different question answering system from nine different pre-trained language models and tested their accuracy and performance and compared the strengths and weaknesses of each system and language model.	Among many other findings, ChatGPT was found to be superior in zero-shot questions but inferior in simple questions.
Huang and Rust (2021)	This work develops a strategic framework that provides guidelines in the implementation of AI in marketing.	Mechanic, thinking, feeling AI	Literature review	Mechanical, thinking, and feeling AI should be used for different tasks in marketing research, marketing strategy, and marketing action.
Huang and Rust (2022)	This work establishes a framework for collaborative artificial intelligence in	Mechanic, thinking, feeling AI	Cross-disciplinary literature review	Marketers should optimize the collaboration of human and AI systems and

	marketing based on various intelligences of AI and the strengths and weaknesses of human vs. artificial intelligence.			leverage the superiority of mechanic AI over human intelligence. However, human should be in charge of tasks that require thinking and feeling capabilities.
Huang and Rust (2023)	The authors investigate the emotional capabilities of generative AI in the customer relation context.	Various AI applications	Literature review	AI will continue to shape the development of marketing, and its capabilities are likely to progress further. AI is at the forefront of gaining significant emotional capabilities that enables to AI to establish emotional connection to individual customers.
Hui et al. (2023)	This work investigates the effects of AI on service quality for when service employees are enhanced with AI tools.	Human-AI collaboration	Study 1: Interview of 312 food delivery employees from different e-food restaurants who used AI tools on their job. Study 2: Interview of 363 customers that purchased food from one of the companies in which the employees from study 1 worked.	Customer engagement and satisfaction is influenced by the quality of AI services, psychological safety and AI empathy as well as AI usability.
Hulman et al. (2023)	This work compares answers from ChatGPT to answers from humans for questions regarding diabetes.	Generative pre-trained transformer	183 participants were shown two answers to the same question, one by humans and the other by ChatGPT, for a total of ten questions. They had to choose which one they think is written by a human and which one by ChatGPT.	The participants were able to find out the ChatGPT answer in 59.5 % of the cases, indicating that ChatGPT is still far from being similar to human written text.
Jarco and Sulkowski (2023)	This work examines the effectiveness of ChatGPT in making strategic business decisions.	Generative pre-trained transformer	The study compared the results from solving a business case from three different scenarios. The first scenario, a human alone solved the case, the second scenario ChatGPT alone solved the case with as little human interaction as possible, and the third scenario was a combination of a human and ChatGPT.	ChatGPT on its own is not very good at business consulting and making strategic decisions, however, it can help humans to create an effective path for making the decisions.
Karakose et al. (2023)	This work compares the responses from ChatGPT 3.5 and	Generative pre-trained transformers	The researchers asked both versions of ChatGPT the same four	Both versions of ChatGPT performed quite well and were able

	ChatGPT 4 to digital school leadership and technology integration of teachers.		questions simultaneously and analyzed the quality of the responses later on by calculating Cohen's Kappa and applying their own rating scale.	to deliver responses that were in line with current academic research. However, ChatGPT 4 created better responses in terms of understandability and clearness.
Katar et al. (2023)	This work examines the text generation capabilities of ChatGPT (GPT3) on writing academic papers.	Generative pre-trained transformer	Literature review	ChatGPT is not able to write an entire article on its own but is a great tool for researchers to assist in the writing process.
Kim, Kim, and Baek (2024),	The researchers examine the effects of perceived usability, enjoyment, and responsiveness on the use of ChatGPT.	Chatbot, generative pre-trained transformer	They collected data from 441 people who frequently use ChatGPT and used structural equation modeling to test the effects of perceived usability, enjoyment, and responsiveness on perceived attachment and satisfaction with ChatGPT, and ultimately the effect of perceived attachment and satisfaction with ChatGPT on the ChatGPT continuance intention.	Chatbots, such as ChatGPT, should be crafted with a focus on an enjoyable user-friendly interface.
Kocoń et al. (2023)	This work compares the performance of ChatGPT on 25 different NLP tasks to the performance of the best models that were available at that time.	Generative pre-trained transformer	The authors evaluated more than 49,000 responses from ChatGPT 4 on the performance of different NLP tasks such as grammatical correctness, emotion recognition, or sentiment analysis.	ChatGPT can solve most of the NLP tasks quite well but there exist other models that perform better.
Kosar et al. (2024)	This work examines the effect of using ChatGPT on the learning, engagement and success of computer science students.	Generative pre-trained transformer	Two groups of students were created from a total of 182 participants. One group used ChatGPT to solve an assignment and the other could not use ChatGPT for the same assignment.	The performance of the students was not influenced by the use of ChatGPT indicating the save use of ChatGPT for students.
Leong (2023)	The study compares the similarity of AI-generated text to human-generated text.	Chatbots	The use of clauses and inter-clausal relations between 50 human-generated abstracts were compared to 150 AI-generated abstracts for the same research articles (three for each article from Bard,	Although none of the chatbots matches to the human-generated abstracts in all clausal categories, ChatGPT performed best out of all three chatbots used.

			ChatGPT, and Poe assistant).	
Li, Yao, and Nan (2023)	This work examines human-like traits of chatbots on using intention and customer engagement of customers.	Chatbots	Study 1: 280 participants divided into two groups, one group with a chatbot that conveyed emotional warmth and the other with a chatbot that conveyed factual competence. The participants were presented a conversation from one of the chatbots and had to rate perceived warmth and competence. Study 2: In a 2x2 design 344 participants were randomly allocated to 1 of 4 groups to investigate the influence of the level of their need to belong.	Human-like traits in chatbots, such as perceived warmth and competence, positively influence using intention and customer engagement. This effect is mediated by perceived usefulness and moderated by the need to belong and information sensitivity.
Li and Kim (2024)	This work examines the impact of LLMs on knowledge sharing and content creation in user generated content platforms.	Generative pre-trained transformer	679,662 answers and 773,613 questions from Stack Overflow were gathered and analyzed with the difference-in-differences method	Content generation increases due to the use of AI whereby high-quality content decreases. Moreover, highly engaged users tend to decrease their activity on the platform.
J. Liu et al. (2023)	This work compares the effectiveness of Stack Overflow compared to ChatGPT as coding assistants for solving programming tasks.	Generative pre-trained transformer	Two groups of individuals with comparable knowledge in programming, one group using Stack Overflow and the other group using ChatGPT, have to solve the same three programming tasks in a similar environment. The tasks were related to algorithmic challenges, library usage, and debugging tasks.	The ChatGPT group was overall faster than the Stack Overflow group. For algorithmic challenges and library usage, the code quality of ChatGPT group was better whereby the Stack Overflow group showed better code quality for debugging tasks.
S. Liu et al. (2023)	This work is related to clinical decision support and compares the usefulness of suggestions from ChatGPT to human-generated ones.	Generative pre-trained transformer	The researchers used ChatGPT to generate 36 clinical decision support suggestions to different clinical decision support alerts and compared them to 29 human-generated suggestions. The answers were then	Of the top 20 suggestions 9 resulted from ChatGPT indicating the usefulness of ChatGPT in the field of clinical decision support but rather as an auxiliary tool than to replace humans.

			evaluated by expert clinicians.	
Ljepava (2022)	This work provides an overview of various marketing applications for different marketing stages.	Various AI applications for different marketing functions	Literature review	The author mapped the identified AI applications along five marketing stages, analysis, strategy, tactics, customer relations, and value proposition. The most applications were found in the first stage for the analysis of customer data.
Lozić and Štular (2023)	This work analyzes the performance of AI chatbots in generating scientific content in the fields of humanities and archaeology.	AI chatbots	Different AI chatbots, namely ChatGPT3.5 and 4, Bard, Claude 2, Aria, and Bing, were asked the same two specific scientific questions. Human experts evaluated the outcome of each chatbot by its accuracy and quality.	ChatGPT showed best performance of all models tested, however, neither of the models were able to come close to human generated content in terms of accuracy and quality.
Metzler, Günneman, and Miettinen (2016)	This work proposes three new approaches to analyze the structure of communities and networks.	Not AI-related	The experiments are divided into two sets whereby, the first analyzes human-annotated communities, and the other applies the hyperbolic community model to communities detected by algorithms.	The authors developed a hyperbolic community model that allows to identify structures within communities.
Metzler, Günneman, and Miettinen (2019)	This work analyzes the dynamics of online questioning and answering communities.	Not AI-related	The authors applied the hyperbolic community model to datasets from Reddit, Stack Exchange and Healthboards.	There is only a small group of highly active users that are responsible for most of the social interactions.
Musheyev et al. (2024)	This work evaluates the performance of AI chatbots on urological questions.	AI Chatbots	Four chatbots, ChatGPT, Perplexity, Chat Sonic, and Microsoft Bing AI, were used to answer the most frequently asked questions to urological questions obtained from Google trends. Experts evaluated the outcome of each chatbots and compared them to another.	The answers from all chatbots were mostly accurate but had only medium understandability and lacked to give clear advice for the questioners.
Mustafa and Zhang (2022)	This work examines how to achieve higher user participation online questioning and answering communities.	Not AI-related	The researchers applied fuzzy-set qualitative comparative analysis to a dataset consisting of 382 responses from not technical communities	There are two main findings, technical communities require a certain level of reciprocity to increase user participation

			and 395 responses from technical communities.	whereby not technical communities require online social interaction for more user participation.
Ngai and Wu (2022)	This work investigates the relevance of AI and machine learning in marketing and provides a holistic framework for AI in marketing.	Various AI applications	Literature review and development of a framework for AI in marketing	This work provides a holistic framework highlighting the tools and technologies necessary to leverage AI in marketing.
Overgoor et al. (2019)	This work elaborates on the opportunities for AI to support decision making in marketing.	Various AI applications	The researchers applied the Cross-industry Standard Process for Data Mining framework to create guidelines for managers where, when, and how to implement AI in marketing.	This work highlights the importance of a profound business and data understanding when planning to use AI to get tailored and functional AI solutions.
Prentice and Nguyen (2020)	The authors examine the effects of AI in customer service with hotel customers.	Various AI applications in customer service such as chatbots, AI robots, and digital assistants	The authors evaluated the responses of 380 hotel customers, that experienced AI customer service during the visit in a hotel, from an online survey.	Customers preferred to interact with humans instead of AI customer service in a physical hotel.
Puntoni et al. (2021)	This work investigates the social and individual challenges that can occur when AI is used in customer service.	Various AI applications	Literature review	The authors point out that AI can be both harmful and helpful, whereby it is important to establish and adhere to principles on how to use AI. According to the authors AI is seen to positively and needs to be set in the context of ist risks and consideration customers might have.
Raj et al. (2023)	This work examines possible use cases of ChatGPT in business processes and marketing highlighting the beneficial use of AI in companies.	Generative pre-trained transformer	Literature review and expert interviews. The expert interviews were evaluated using the Preference Selection Index and Complex proportional assessment.	The study identified several benefits of the use of ChatGPT in business processes such as faster information acquisition and higher levels of personalization.
Rivas and Zhao (2023)	This work investigates the potential benefits of AI, ChatGPT in particular, in marketing with respect to ethical considerations.	Generative pre-trained transformer	Literature review	By considering the risks and concerns of the use of AI in businesses with the use of data and computer science experts, AI has the potential to revolutionize the marketing landscape. Human oversight and

				transparency should be at the forefront when using AI in marketing.
Roumeliotis and Tselikas (2023)	This work explains the key functionalities of the generative pre-trained transformers technology of ChatGPT and provides an overview of research on ChatGPT.	Generative pre-trained transformers	Literature review	ChatGPT has various potential fields of application that are also increasingly being researched by researchers from several disciplines.
Sarker et al. (2023)	This study examines the similarity between human- and AI-generated text and whether AI-generated text is undifferentiable from human text.	Generative pre-trained transformer	6250 human-generated answers to questions or other comments to content was collected from Stack Overflow, Yahoo and YouTube. Then GPT 3.5, GPT 4, and Davinci-3 created answers to the same questions or other content with the same title. For comparison the authors used Parts of Speech distribution analysis, Bilingual evaluation study scores, Global Vectors for Word Representations, a pre-trained BERT model, and Sentence-BERT model.	The text generated by AI was distinguishable from human-generated text because humans use statistical anomalies in their sentences which AI simply does not and rather produce word structures with a high statistical likelihood.
Shahab et al. (2024)	This work elaborates on the functionalities of LLMs such as ChatGPT and examines possible applications for gastroenterology.	Large language models	Literature review	LLMs such as ChatGPT have the potential to improve medical care in gastroenterology as an adjunct tool for medical personnel. Although there are some limitations today, in knowledge or reproduction of biases from the training data, the authors point out towards future developments of LLMs which will further increase the number of beneficial applications of AI in medical care.
Suri et al. (2024)	This study investigates whether ChatGPT reproduces similar decision heuristics as humans do.	Generative pre-trained transformer	Study 1: Two similar prompts were created to measure the anchoring effect for ChatGPT and for humans, as a control. The answers from both were used as results to	ChatGPT showed similar decision heuristics as humans, the anchoring effect, representativeness and availability heuristic, framing effect, and endowment effect.

			measure the effect for both groups. Study 2: Similar approach as in study 1 but the prompt was designed to test representativeness and availability heuristic. Hereby, the authors used the Linda problem von Tversky and Kahneman (1983). Study 3: Different prompts were created reflecting one positively framed statement and one negatively framed statement. Study 4: Similar approach as in Study 1 but the prompt was changed to decision scenario to test the endowment effect.	
Vaswani et al. (2023)	The authors propose a new transformer architecture based on attention mechanisms only and test its performance.	Transformer model	The proposed transformer architecture solely based on attention mechanisms is tested on the WMT 2014 English to German translation tasks and BLEU score.	The proposed model can be trained faster than recurrent or convolutional layers architectures. For translation tasks from English to German and English to French, the model outperformed previous models.
Vlačić et al. (2021)	This work provides an overview on the role of AI in marketing and identifies four areas of relevance for AI in marketing.	Various AI applications	Literature review	The literature review revealed five theoretical dimensions in academic research of AI in marketing and identified several research avenues of AI in marketing.
Whiles and Terry (2024)	This work examines the accuracy of ChatGPT in health care and urology.	Generative pre-trained transformer	Literature review	In most cases ChatGPT answered the questions appropriately. For the other answers it lacked vital information, clarity, or understandability and was not able to give precise recommendations for action.
Xueming Luo et al. (2021)	This study examines the effect of AI coaches on the training of salespeople	AI-powered coach	Field Experiment 1: 429 sales agents are ranked into low-, mid-, and high-performing	Mid-ranked salespeople benefit more from an AI coach than a human coach. This finding is

	and further compares the effectiveness of AI coaches to human agents.		groups. Each category is split up so that one group receives feedback from only a human or only an AI agent Field Experiment 2: 100 bottom-ranked agents, examines whether amount of feedback from AI has effect on performance of sales agents Field Experiment 3: 451 sales agents, effect of joint coaching from human and AI	inversed for low- and high-ranked salespeople. A combination of both delivers the best results.
Yin, Li, and Qiu (2023)	This work examines whether AI influences customer engagement behavior with respect to technology readiness of customers.	Various AI applications in customer service	The authors conducted three experiments, two laboratory experiments and one online situational experiment, with Chinese hotel customers.	Customers with high technology readiness have higher CEB (customer engagement behavior), especially for not anthropomorphic AI.
Zaitsu and Jin (2023)	The authors compare GPT3.5- and GPT4-generated text to Japanese human-generated text from academic papers.	Generative pre-trained transformers	The authors collected 72 human-generated Japanese papers to 72 texts from GPT3.5 and 72 texts from GPT4 and compared their stylometric features to find out how similar AI-generated text is to human-generated text.	There is a significant difference between GPT3.5 and GPT4, whereby GPT4 performs slightly better than GPT3.5. Although GPT4 has more parameters it is still not similar to text written by humans, especially due to stylometric features.

Appendix B: Comparative Literature Table

	AI applications			Performance of AI			Category	
Citation	**Analysis**	**Customer relations**	**Marketing strategy**	**Compar-ison**	**Outcome**	**Accuracy**	**Technical features**	**OCs and tech. use**
This study	X	X	X	X	X	X	X	X
Andrade and Tumelero (2022)		X			X			
Arango, Singaraju, and Niininen (2023)			X		X			
Bandi, Adapa, and Kuchi (2023)	X	X	X				X	
Brei (2020)	X	X	X		X			
Budler, Gosak, and Stiglic (2023)		X			X	X		
Burtch, Lee, and Chen (2023)					X			X
Campbell et al. (2020)	X	X	X		X		X	
Crolic et al. (2022)		X		X	X			
Davenport et al. (2020)	X	X	X		X			
Davis (1985)								X
Davis (1989)								X
Devlin et al. (2019)						X	X	
Elmashhara et al. (2024)		X			X			
Farooqui (2022)		X		X	X			

Garvey, Kim, and Duhachek 2023	X	X		X	X			
Hossain et al. (2022)	X		X					
Hu et al. (2023)					X	X	X	
Huang and Rust (2021)	X	X	X				X	
Huang and Rust (2022)	X	X	X				X	
Huang and Rust (2023)	X	X	X				X	
Hui et al. (2023)		X			X			
Hulman et al. (2023)				X				
Jarco and Sulkowski (2023)	X		X	X	X			
Karakose et al. (2023)	X		X		X			
Katar et al. (2023)	X			X				
Kim, Kim, and Baek (2024),		X	X					
Kocoń et al. (2023)			X		X	X	X	
Kosar et al. (2024)					X	X		
Leong (2023)				X			X	
Li, Yao, and Nan (2023)		X			X			
Li and Kim (2024)		X			X			X
J. Liu et al. (2023)				X	X			

S. Liu et al. (2023)	X		X	X	X			
Ljepava (2022)	X	X	X					
Lozić and Štular (2023)				X	X			
Metzler, Günneman, and Miettinen (2016)								X
Metzler, Günneman, and Miettinen (2019)								X
Musheyev et al. (2024)		X				X		
Mustafa and Zhang (2022)								X
Ngai and Wu (2022)	X	X	X					
Overgoor et al. (2019)			X		X			
Prentice and Nguyen (2020)	X	X		X	X			
Puntoni et al. (2021)			X					
Raj et al. (2023)			X		X			
Rivas and Zhao (2023)	X	X	X					
Roumeliotis and Tselikas (2023)	X	X	X				X	
Sarker et al. (2023)				X				

Shahab et al. (2024)		X	X				X	
Suri et al. (2024)			X	X				
Vaswani et al. (2023)							X	
Vlačić et al. (2021)	X	X	X					
Whiles and Terry (2024)		X				X		
Xueming Luo et al. (2021)			X	X	X			
Yin, Li, and Qiu (2023)		X						
Zaitsu and Jin (2023)				X				

References

Allianz (2022), "Betrugsabwehr," *Allianz.de*, (accessed August 16, 2024), [available at https://www.allianz.de/presse/mitteilungen/spuernase-mensch-und-kollege-ki].

Amazon (2024), "Amazon Alexa Offizielle Webseite: Was ist Alexa?," *Amazon Alexa*, (accessed August 16, 2024), [available at https://developer.amazon.com/de-DE/alexa.html].

Andrade, Ivan Martins De and Cleonir Tumelero (2022), "Increasing customer service efficiency through artificial intelligence chatbot," *Revista de Gestão*.

Arango, Luis, Stephen Pragasam Singaraju, and Outi Niininen (2023), "Consumer responses to AI-generated charitable giving ads," *Journal of Advertising*, 52 (4), 486–503.

Bandi, Ajay, Pydi Venkata Satya Ramesh Adapa, and Yudu Eswar Vinay Pratap Kumar Kuchi (2023), "The Power of Generative AI: A Review of Requirements, Models, Input–Output Formats, Evaluation Metrics, and Challenges: Future Internet," *Future Internet*, 15 (8), 260.

Brei, Vinicius (2020), "Machine Learning in Marketing: Overview, Learning Strategies, Applications, and Future Developments," *Foundations and Trends® in Marketing*, 14, 173–236.

Brooke, S. J. (2021), "Trouble in programmer's paradise: gender-biases in sharing and recognising technical knowledge on Stack Overflow," *Information, Communication & Society*, 24 (14), 2091–2112.

Budler, Leona Cilar, Lucija Gosak, and Gregor Stiglic (2023), "Review of artificial intelligence-based question-answering systems in healthcare," *WIREs Data Mining and Knowledge Discovery*, 13 (2), e1487.

Burtch, Gordon, Dokyun Lee, and Zhichen Chen (2023), "The Consequences of Generative AI for UGC and Online Community Engagement," SSRN Scholarly Paper, Rochester, NY.

Campbell, Colin, Sean Sands, Carla Ferraro, Hsiu-Yuan (Jody) Tsao, and Alexis Mavrommatis (2020), "From data to action: How marketers can leverage AI," *Business Horizons*, 63 (2), 227–43.

Crolic, Cammy, Felipe Thomaz, Rhonda Hadi, and Andrew T. Stephen (2022), "Blame the Bot: Anthropomorphism and Anger in Customer–Chatbot Interactions," *Journal of Marketing*, 86 (1), 132–48.

Davenport, Thomas, Abhijit Guha, Dhruv Grewal, and Timna Bressgott (2020), "How artificial intelligence will change the future of marketing," *Journal of the Academy of Marketing Science*, 48 (1), 24–42.

Davis, Fred D. (1985), "A technology acceptance model for empirically testing new end-user information systems: Theory and results (tesis doctoral)," *Massachusetts Institute of Technology, Massachusetts, Estados Unidos*.

Davis, Fred D. (1989), "Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology: MIS Quarterly," *MIS Quarterly*, 13 (3), 319–40.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2019), "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," arXiv.

Elmashhara, Maher Georges, Roberta De Cicco, Susana C. Silva, Maik Hammerschmidt, and Maria Levi Silva (2024), "How gamifying AI shapes customer motivation, engagement, and purchase behavior," *Psychology & Marketing*, 41 (1), 134–50.

Farooqui, Safia (2022), "The Role of Emotional Intelligence & Artificial Intelligence in Customer Loyalty and Engagement," *Integral Review: A Journal of Management*, 12 (2), 21–24.

Garvey, Aaron M., TaeWoo Kim, and Adam Duhachek (2023), "Bad News? Send an AI. Good News? Send a Human," *Journal of Marketing*, 87 (1), 10–25.

Gentsch, Peter (2019), *Künstliche Intelligenz für Sales, Marketing und Service: Mit AI und Bots zu einem Algorithmic Business – Konzepte und Best Practices*, Wiesbaden: Springer Fachmedien.

Gleasure, R., Kluge, S., Parra Moyano, J., & Constantiou, I. (2024). *Just ask ChatGPT? When and why users prefer human-like answers on digital information platforms.* Manuscript submitted for publication in JMIS.

Goldman Sachs (2023), "AI investment forecast to approach $200 billion globally by 2025," (accessed August 16, 2024), [available at https://www.goldmansachs.com/insights/articles/ai-investment-forecast-to-approach-200-billion-globally-by-2025].

Hair, Joseph F., Jr., William C. Black, Barry J. Babin, and Rolph E. Anderson (2013), *Multivariate Data Analysis: Pearson New International Edition*, Harlow, United Kingdom, UNITED KINGDOM: Pearson Education, Limited.

Hossain, Md Afnan, Raj Agnihotri, Md Rifayat Islam Rushan, Muhammad Sabbir Rahman, and Sumaiya Farhana Sumi (2022), "Marketing analytics capability, artificial intelligence adoption, and firms' competitive advantage: Evidence from the manufacturing industry," *Industrial Marketing Management*, 106, 240–55.

Hu, Nan, Yike Wu, Guilin Qi, Dehai Min, Jiaoyan Chen, Jeff Z Pan, and Zafar Ali (2023), "An empirical study of pre-trained language models in simple knowledge graph question answering," *World Wide Web*, 26 (5), 2855–86.

Huang, Ming-Hui and Roland T. Rust (2021), "A strategic framework for artificial intelligence in marketing," *Journal of the Academy of Marketing Science*, 49 (1), 30–50.

Huang, Ming-Hui and Roland T. Rust (2022), "A Framework for Collaborative Artificial Intelligence in Marketing," *Journal of Retailing*, 98 (2), 209–23.

Huang, Ming-Hui and Roland T. Rust (2023), "The Caring Machine: Feeling AI for Customer Care," *Journal of Marketing*, 00222429231224748.

Hui, Zhang, Ali Nawaz Khan, Zhang Chenglong, and Naseer Abbas Khan (2023), "When Service Quality is Enhanced by Human–Artificial Intelligence Interaction: An Examination of Anthropomorphism, Responsiveness from the Perspectives of Employees and Customers," *International Journal of Human–Computer Interaction*, 0 (0), 1–16.

Hulman, Adam, Ole Lindgård Dollerup, Jesper Friis Mortensen, Matthew E. Fenech, Kasper Norman, Henrik Støvring, and Troels Krarup Hansen (2023), "ChatGPT-versus human-generated answers to frequently asked questions about diabetes: A Turing test-inspired survey among employees of a Danish diabetes center: PLoS ONE," *PLoS ONE*, 18 (8).

Jarco, Dawid and Lukasz Sulkowski (2023), "Is ChatGPT better at business consulting than an experienced human analyst? An experimental comparison of solutions to a strategic business problem: Forum Scientiae Oeconomia," *Forum Scientiae Oeconomia*, 11 (2), 87–101.

Karakose, Turgut, Murat Demirkol, Ramazan Yirci, Hakan Polat, Tuncay Yavuz Ozdemir, and Tijen Tülübaş (2023), "A Conversation with ChatGPT about Digital Leadership and Technology Integration: Comparative Analysis Based on Human–AI Collaboration," *Administrative Sciences (2076-3387)*, 13 (7), 157.

Katar, Oğuzhan, Dilek Özkan, Özal Yildirim, and U. Rajendra Acharya (2023), "Evaluation of GPT-3 AI Language Model in Research Paper Writing: Turkish Journal of Science & Technology," *Turkish Journal of Science & Technology*, 18 (2), 311–18.

Kim, Jeong Soo, Minseong Kim, and Tae Hyun Baek (2024), "Enhancing user experience with a generative ai chatbot: International Journal of Human-Computer Interaction," *International Journal of Human-Computer Interaction*.

Kocoń, Jan, Igor Cichecki, Oliwier Kaszyca, Mateusz Kochanek, Dominika Szydło, Joanna Baran, Julita Bielaniewicz, Marcin Gruza, Arkadiusz Janz, Kamil Kanclerz, Anna Kocoń, Bartłomiej Koptyra, Wiktoria Mieleszczenko-Kowszewicz, Piotr Miłkowski, Marcin Oleksy, Maciej Piasecki, Łukasz Radliński, Konrad Wojtasik, Stanisław Woźniak, and Przemysław Kazienko (2023), "ChatGPT: Jack of all trades, master of none: Information Fusion," *Information Fusion*, 99, N.PAG-N.PAG.

Kosar, Tomaž, Dragana Ostojić, Yu David Liu, and Marjan Mernik (2024), "Computer Science Education in ChatGPT Era: Experiences from an Experiment in a Programming Course for Novice Programmers," *Mathematics (2227-7390)*, 12 (5), 629.

Leong, Alvin Ping (2023), "Clause complexing in research-article abstracts: Comparing human- and AI-generated texts," *ExELL: Explorations in English Language & Linguistics*, 11 (2), 99–132.

Li, Baoku, Ruoxi Yao, and Yafeng Nan (2023), "How do friendship artificial intelligence chatbots (FAIC) benefit the continuance using intention and customer engagement?," *Journal of Consumer Behaviour*, 22 (6), 1376–98.

Li, Xinyu and Keongtae Kim (2024), "Unveiling the Effects of LLMs: Shifting UGC Contribution in an Online Coding Community," SSRN Scholarly Paper, Rochester, NY.

Liu, Jinrun, Xinyu Tang, Linlin Li, Panpan Chen, and Yepang Liu (2023), "Which is a better programming assistant? A comparative study between chatgpt and stack overflow," arXiv.

Liu, Siru, Aileen P Wright, Barron L Patterson, Jonathan P Wanderer, Robert W Turer, Scott D Nelson, Allison B McCoy, Dean F Sittig, and Adam Wright (2023), "Using AI-generated suggestions from ChatGPT to optimize clinical decision support," *Journal of the American Medical Informatics Association*, 30 (7), 1237–45.

Ljepava, Nikolina (2022), "AI-Enabled Marketing Solutions in Marketing Decision Making: AI Application in Different Stages of Marketing Process," *TEM Journal*, 1308–15.

Lozić, Edisa and Benjamin Štular (2023), "Fluent but Not Factual: A Comparative Analysis of ChatGPT and Other AI Chatbots' Proficiency and Originality in Scientific Writing for Humanities: Future Internet," *Future Internet*, 15 (10), 336.

Makyen (2024), "Policy: Generative AI (e.g., ChatGPT) is banned," *Meta Stack Overflow*, Forum post.

Metzler, Saskia, Stephan Günnemann, and Pauli Miettinen (2016), "Hyperbolae Are No Hyperbole: Modelling Communities That Are Not Cliques," arXiv.

Metzler, Saskia, Stephan Günnemann, and Pauli Miettinen (2019), "Stability and dynamics of communities on online question–answer sites," *Social Networks*, 58, 50–58.

Musheyev, David, Alexander Pan, Stacy Loeb, and Abdo E. Kabarriti (2024), "How Well Do Artificial Intelligence Chatbots Respond to the Top Search Queries About Urological Malignancies?," *European Urology*, 85 (1), 13–16.

Mustafa, Sohaib and Wen Zhang (2022), "How to Achieve Maximum Participation of Users in Technical Versus Nontechnical Online Q&A Communities?," *International Journal of Electronic Commerce*, 26 (4), 441–71.

Ngai, Eric W. T. and Yuanyuan Wu (2022), "Machine learning in marketing: A literature review, conceptual framework, and research agenda," *Journal of Business Research*, 145, 35–48.

OpenAI (2024a), "Models overview," (accessed August 11, 2024), [available at https://platform.openai.com].

OpenAI (2024b), "API Partnership with Stack Overflow," (accessed August 17, 2024), [available at https://openai.com/index/api-partnership-with-stack-overflow/].

Overgoor, Gijs, Manuel Chica, William Rand, and Anthony Weishampel (2019), "Letting the Computers Take Over: Using AI to Solve Marketing Problems," *California Management Review*, 61 (4), 156–85.

Prentice, Catherine and Mai Nguyen (2020), "Engaging and retaining customers with AI and employee service," *Journal of Retailing and Consumer Services*, 56, 102186.

Puntoni, Stefano, Rebecca Walker Reczek, Markus Giesler, and Simona Botti (2021), "Consumers and Artificial Intelligence: An Experiential Perspective," *Journal of Marketing*, 85 (1), 131–51.

Raj, Rohit, Arpit Singh, Vimal Kumar, and Pratima Verma (2023), "Analyzing the potential benefits and use cases of ChatGPT as a tool for improving the efficiency and effectiveness of business operations," *BenchCouncil Transactions on Benchmarks, Standards & Evaluations*, 3 (3), 1–10.

Rivas, Pablo and Liang Zhao (2023), "Marketing with ChatGPT: Navigating the Ethical Terrain of GPT-Based Chatbot Technology," *AI*, 4 (2), 375–84.

Roumeliotis, Konstantinos I. and Nikolaos D. Tselikas (2023), "ChatGPT and Open-AI Models: A Preliminary Review," *Future Internet*, 15 (6), 192.

Sarker, Emon, Labib Rahman, Nabeel Mohammed, and Mohammad Ruhul Amin (2023), "Humans vs ChatGPT: Uncovering the Non-trivial Distinctions by Evaluating Parallel Responses."

Shahab, Omer, Bara El Kurdi, Aasma Shaukat, Girish Nadkarni, and Ali Soroush (2024), "Large language models: a primer and gastroenterology applications," *Therapeutic Advances in Gastroenterology*, 17.

Suri, Gaurav, Lily R. Slater, Ali Ziaee, and Morgan Nguyen (2024), "Do large language models show decision heuristics similar to humans? A case study using GPT-35: Journal of Experimental Psychology: General," *Journal of Experimental Psychology: General*.

The Open Worldwide Application Security Project (2023), "OWASP Top 10 for LLM," (accessed August 11, 2024), [available at https://owasp.org/www-project-top-10-for-large-language-model-applications/].

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin (2023), "Attention Is All You Need," *Advances in neural information processing systems*, (30).

Vlačić, Božidar, Leonardo Corbo, Susana Costa e Silva, and Marina Dabić (2021), "The evolving role of artificial intelligence in marketing: A review and research agenda," *Journal of Business Research*, 128, 187–203.

Whiles, Bristol B. and Russell S. Terry (2024), "Artificial Intelligence Chatbots: How Accurate Is the Information?," *AUANews*, 29 (1), 8–9.

Xueming Luo, Marco Shaojun Qin, Zheng Fang, and Zhe Qu (2021), "Artificial Intelligence Coaches for Sales Agents: Caveats and Solutions," *Journal of Marketing*, 85 (2), 14–32.

Yin, Dexiang, Minglong Li, and Hailian Qiu (2023), "Do customers exhibit engagement behaviors in AI environments? The role of psychological benefits and technology readiness," *Tourism Management*, 97, 1–18.

Zaitsu, Wataru and Mingzhe Jin (2023), "Distinguishing ChatGPT(-3.5, -4)-generated and human-written papers through Japanese stylometric analysis," *PLoS ONE*, 18 (8), 1–12.

Affidavit

“I hereby declare that I have written the enclosed master seminar thesis myself and that I have not used any outside help that is not apparent from the information I have provided. I also assure that this thesis or parts thereof have not been submitted by myself or by others as a performance record elsewhere. Literal or analogous adoptions from other writings and publications in printed or electronic form are marked. All secondary literature and other sources are identified and listed in the bibliography. The same applies to graphical representations and images as well as to all internet sources and answers generated by AI-based applications. I further agree that my work may be sent and stored anonymously in electronic form for the purpose of plagiarism checking. I am aware that correction of the work may be waived if this declaration is not given."

Mannheim, August 22nd 2024