

MASTER

Personalized Response with Generative AI

Improving Customer Interaction with Zero-Shot Learning LLM Chatbots

Bink, Joëlle M.

Award date:
2023

[Link to publication](#)

Disclaimer

This document contains a student thesis (bachelor's or master's), as authored by a student at Eindhoven University of Technology. Student theses are made available in the TU/e repository upon obtaining the required degree. The grade received is not published on the document as presented in the repository. The required complexity or quality of research of student theses may vary by program, and the required minimum study period may vary in duration.

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain

Eindhoven University of Technology



Master Thesis Data Science & Artificial Intelligence

Research group: Generative AI

Personalized Response with Generative AI: Improving Customer Interaction with Zero-Shot Learning LLM Chatbots

Author: Joëlle Bink (0992628)

1st supervisor: Jakub Tomczak

Daily supervisors: Louis Juan & Vera van Otterdijk (Accenture)

2nd reader: Maxime Chamberland

*A thesis submitted in fulfillment of the requirements for
the Eindhoven University of Technology Master of Science degree
in Data Science & Artificial Intelligence.*

December 20, 2023

"With generative AI, we witness the birth of digital artisans, machines that produce original works, blurring the line between human and artificial creativity." - ChatGPT

Abstract

This thesis explores the transformative potential of Large Language Models (LLMs) like GPT-4 in revolutionizing customer support through AI-powered chatbots. It specifically focuses on the integration of personalization techniques, such as sentiment analysis and conversational style adaptation, to enhance the performance of LLM-based chatbots. The research introduces a Framework for Assessment of Chatbot Quality (FACQ), incorporating Natural Language Processing (NLP) metrics, designed to quantify the degree of personalization in chatbot interactions.

A crucial aspect of this study involves conducting user experiments with real-life participants. These experiments are designed to test the practical applicability and relevance of the FACQ in real-world scenarios. The goal is to determine whether this framework can effectively capture and quantify the level of personalization in chatbot interactions, as perceived by actual users. The results from these experiments offer valuable insights into the effectiveness and user experience of LLM-based chatbots, providing a concrete basis to assess the alignment of FACQ metrics with real-life user preferences, particularly in terms of readability and complexity.

Additionally, the thesis delves into the ethical and regulatory dimensions of deploying generative AI technology, specifically as chatbots within customer support. Insights from interviews with diverse experts, including AI leads, policymakers, and researchers, highlight the critical ethical issues and societal implications of AI development. Key areas of focus include privacy, fairness, bias mitigation, and the complexities of regulating rapidly evolving AI technologies. The European AI Act, with its risk-based approach to AI regulation, is examined for its implications on responsible AI usage, emphasizing the need for global cooperation and a comprehensive multi-stakeholder approach, without limiting AI innovation.

While the findings of this research are significant, they are not without limitations. The study's focus on specific aspects of chatbot personalization, primarily single-turn interactions, may not fully encompass the complexities of long-term engagement or

multimodal interactions. Moreover, the rapid evolution of generative AI technology suggests that certain aspects of this research could become outdated in the future.

Future work in this area should aim to enhance the personalization capabilities of LLMs, explore multimodal interactions, and establish robust governance frameworks for AI systems in customer support. Investigating the spread of misinformation through chatbots and the role of digital literacy in mitigating these risks also remains an important area of study. This research serves as a foundational step in bridging the theoretical potential of LLMs with their practical application, highlighting the necessity for continued exploration in the dynamic field of AI-powered customer support.

Overall, this thesis contributes to the understanding of the capabilities and challenges of LLMs in automated customer support. It offers practical insights into the development and evaluation of personalized chatbot interactions, while also addressing crucial ethical considerations. As AI continues to evolve, this research serves as a stepping stone for future explorations aimed at bridging the gap between the theoretical potential of LLMs and their practical application in enhancing human-AI interactions.

Statement of responsible use of ChatGPT

Embarking on a thesis about large language models (LLMs) without using one would be like trying to learn to swim without ever getting in the water – theoretically possible, but practically dubious. Hence, in a classic case of 'learning by doing', this thesis aims to show the principle that the best way to understand a technology is to engage with it directly, maintaining a critical and responsible approach throughout the process.

In conducting the research presented in this thesis, I have used OpenAI's ChatGPT, a language model based on the GPT-4 architecture. My utilization of ChatGPT adheres to the guidelines and terms of use set forth by OpenAI. I acknowledge the capabilities and limitations of ChatGPT, and have taken measures to ensure the responsible and ethical use of this technology. I have used ChatGPT to assist in the articulation of ideas and the structuring of content and code, ensuring a coherent and well-organized presentation of the research findings. The information generated by ChatGPT has been cross-verified with authoritative and reliable sources to maintain the accuracy and integrity of the thesis. Any data and insights derived from the use of ChatGPT have been rigorously reviewed to avoid dissemination of misinformation.

This thesis represents a significant milestone in my journey through the field of Data Science and Artificial Intelligence. It's a journey that would have been much more challenging without the support and guidance of several key individuals, to whom I would like to express my gratitude.

To **Jakub Tomczak**, Associate Professor and PI of the Generative AI group at the Eindhoven University of Technology (TU/e). Your quick response to my email in February 2023 was a source of great encouragement for me. Our meetings, filled with exciting discussions about the potential of AI and the refinement of my ideas, have been incredibly helpful. I particularly enjoyed how you allowed and enabled me to include my interests in policy and AI regulation into the thesis and also include human-AI interaction through a survey, which might not be part of the more traditional research methods within our department. I am also thankful that you included part of the results of my thesis in your presentation at the TU/e Holst Memorial Lecture Symposium and I'm confident we will keep in touch for the future.

To **Vera van Otterdijk and Louis Juan** at Accenture, within the Data Intelligence & AI group. Our weekly meetings, where we shared stories and laughs (specifically about cats and hamsters) before getting down to thesis discussions, have been a highlight of this process. Despite the broad scope of the thesis and the rapidly changing landscape of Generative AI, your consistent support and encouragement have been a great help. I also want to thank the Data Intelligence & Innovation team for warmly accepting me and showing a keen interest in the practical applications of LLMs that I was exploring during my internship. I cannot wait to further develop myself in the field of (generative) AI.

To **my family**, your support has been fundamental. From participating in my research survey to listening to my endless talks about Generative AI and LLMs, your involvement has been crucial. A special thanks to my parents for your belief in me and for supporting my seven years of study and extracurricular activities.

To my **fellow Data & AI interns** at Accenture, our ongoing interest in each others projects, helping each other out where possible and our breaks for hot chocolate or lunches at ITO have been a great pleasure.

To all of you, and all the friends that I have made while studying in Eindhoven for the past seven years, I offer my sincere thanks. I hope you find the results of this thesis as rewarding as I have found your support and guidance.

Thank you and enjoy reading!

Contents

List of Figures	vii
List of Tables	ix
1 Introduction	1
2 Related work	5
2.1 LLM-based chatbots in healthcare	5
2.2 LLM-based chatbots in education	6
2.3 LLM-based chatbots in customer support	7
2.4 Chatbot personalization & quality assessment	7
3 Methodology	11
3.1 Large Language Models	11
3.1.1 The transformer architecture	12
3.2 Model size	19
3.3 The training process	21
3.3.1 Pre-training	22
3.3.2 Fine-tuning	23
3.3.3 Prompt-based learning	25
3.4 State-of-the-art models	26
3.4.1 Challenges and limitations	27
3.5 From rule-based to LLM-based chatbots	30
3.6 Framework for Assessment of Chatbot Quality (FACQ)	33
4 Experiments	39
4.1 Streamlit UI & code	39
4.2 Personalization prompts & metrics experiments	40
4.3 User Survey	45

CONTENTS

4.4 Demographic	45
4.5 Language adaptation	48
4.6 Sentiment adaptation	56
4.7 Trustworthiness	61
4.8 Human-chatbot interaction preferences	64
4.9 Limitations	69
4.10 Future work	69
5 Ethics & regulations	71
5.1 Essay: AI ethics and regulations	72
6 Conclusion	79
References	81

List of Figures

3.1	Timeline of early-stage (AI) software and models up to SOTA LLMs <i>Source: [1]</i>	11
3.2	The encoder-decoder structure of the transformer architecture <i>Source: [2]</i>	13
3.3	Parameter size of AI Neural Networks over the past decades. <i>Source: [3]</i>	20
3.4	Three key steps in the LLM training process. <i>Source: [4]</i>	22
3.5	Masked Language Modeling in transformer-based LLMs <i>Source: [5]</i>	23
3.6	Architecture of a Retrieval-Augmented Generation (RAG) system for question-answering <i>Source: [6]</i>	25
3.7	Example of a "jailbreak" system prompt.	29
3.8	Example of the question (left), normal answer of GPT-4 (middle) and "jailbreak" answer of GPT-4 (right)	30
3.9	Overview of main differences between rule-based and AI-based chatbots <i>Source: [7]</i>	31
4.1	Overview of the UI tool for LLM-chatbot personalization prompting and testing .	40
4.2	Example of personalization prompting in zero-shot learning LLM-based chatbots	41
4.3	Example of FACQ results with personalization prompting in a zero-shot learning GPT-4 based chatbot .	42
4.4	Example of augmenting an LLM-based chatbot with URL processing and querying functionality .	44
4.5	Age distribution of survey participants .	46
4.6	Gender distribution of survey participants .	46

LIST OF FIGURES

4.7	Education level of survey participants	47
4.8	Technology experience level of survey participants	48
4.9	Radar plots for similarity, readability and complexity metrics of Q8-Q13	51
4.10	Distribution of Responses for Subquestions on Language Adaptation	53
4.11	Distributions of Responses for Subquestions on Sentiment	59
4.12	Distributions of responses for misconception questions	62
4.13	Distribution of responses for human-AI chatbot interaction statements	65
4.14	Word cloud of the most frequent words in Q25 responses.	67
6.1	Table of FACQ metric results for Q8-Q13	99

List of Tables

3.1	Overview of parameter and training set size for a selection of well-known LLMs, ordered by increasing number of parameters	21
3.2	Customer service fine-tuning dataset example	24
3.3	Overview of the FACQ metric ranges and their implications	35
4.1	Chi-square test results for language style adaptation questions	54
4.2	FACQ metric values for sentiment questions	58
4.3	Chi-square test results for sentiment questions	60
4.4	Chi-square test results for misconception questions	63
4.5	Chi-square test results for human-AI chatbot interaction statements	66

LIST OF TABLES

1

Introduction

Generative Artificial Intelligence (AI) has witnessed groundbreaking advancements, with the debut of large language models (LLMs) such as GPT-4 and LLaMA 2 in 2023. These state-of-the-art models, trained on massive datasets, are capable of generating novel, “creative” content in response to an input (prompt). LLMs have made remarkable progress in recent months, pushing the boundaries of what AI can achieve in generating human-like text. They represent a significant breakthrough in natural language processing, with the potential to revolutionize how people interact with language-based technology [8] [9]. This breakthrough becomes even clearer when presenting the numbers. Five days after ChatGPT’s launch in November 2022, the application reached one million users. By February 20, 2023, ChatGPT broke the record for fastest growing user-based application, increasing to 100 million in just two months [10], reaching a market value close to 30 billion USD [11]. By October 2023, the total visits to ChatGPT’s website had reached a staggering 1.7 billion, highlighting the widespread appeal of this technology [?]. A report released in April 2023 by Goldman Sachs indicates that as many as 300 million full-time jobs will be subject to automation by generative AI [12]. This brought up a widespread discussion on whether generative AI will be (or already is) the most significant global disruptive force since the adoption of the internet.

The potential impact of LLMs and generative AI reaches various sectors. LLMs find applications in healthcare through diagnostic AI tools [13], in finance with predictive algorithms, and in the entertainment sector with personalized content recommendations. In natural language processing (NLP), LLMs excel in tasks like sentiment analysis, text translation, code generation and question-answering [14] [8]. The latter offers a whole new range of possibilities for LLM-based chatbots as question-answering assistants, especially within customer support. LLMs can be trained using a zero-shot learning approach, in which LLMs are designed to understand and respond to tasks that they have not been explicitly

1. INTRODUCTION

trained on. This means that these models can generate relevant and coherent responses across a wide range of prompts, demonstrating an impressive level of understanding and adaptability, even without prior exposure to similar content. The zero-shot learning capability of these LLMs significantly enhances their flexibility and application scope, making them ideal for the diverse nature of customer service interactions explored in this research.

Many of the world's leading brands and increasingly government agencies and healthcare institutions are using chatbots, computer programs that engage in conversations with customers. However, user satisfaction with chatbots is mixed and users report frustration with chatbots as a result of misunderstood questions and irrelevant responses [15]. Traditional rule-based chatbots are already widely used within customer support due to their ability to handle straightforward interactions, such as frequently asked questions or basic customer inquiries [16]. However, since rule-based chatbots are developed with pre-defined rules and responses, they often fall short in handling more complex questions and cannot adapt to individual user preferences. The creation of these rule-based chatbot systems is both time-consuming and labor-intensive, often requiring extensive programming and regular updates to ensure relevance and accuracy.

As chat technology has progressed, some chatbot developers have begun incorporating more advanced machine learning (ML) and natural language processing (NLP) techniques [17]. These systems are more sophisticated than traditional rule-based systems, however they are still limited in terms of providing genuinely personalized customer experiences. Examples include chatbots using decision trees or simple neural networks to handle a broader range of queries, but are not fully able to generate original natural language or adapt to new information dynamically.

The emergence of LLMs and generative AI presents a transformative opportunity in the field of chatbots for automated question-answering. LLMs can generate human-like text, offering a more nuanced and adaptable approach to customer support. Unlike rule-based systems, LLMs can learn from vast datasets, allowing them to provide more accurate, context-aware and personalized responses. This adaptability also significantly reduces the time and effort required for chatbot maintenance, as LLMs can continuously evolve with new data.

The integration of LLMs into customer support chatbots aligns with the global trend towards personalized service. A study by a global marketing company found that 80% of consumers are more likely to purchase from a brand that provides personalized experiences [18]. Other research shows that 71% of consumers expect companies to deliver personalized interaction and 76% get frustrated when this does not happen [19]. Two fundamental elements in chatbot interaction that can be personalized are the use of language and the sentiment or empathy level of the response. Humans are able to communicate effectively by imitating their conversation partner's

lexical choices, sentence structures, and overall language style; a mechanism known as linguistic alignment or language style adaptation [20]. In human-chatbot interactions, the ability to detect signs of human emotions and to properly react to them can enrich communication and increase user satisfaction [21]. Integrating language style and sentiment adaption with the capabilities of LLMs to generate personal and human-like responses could revolutionize the way that users interact with chatbots.

However, integrating LLMs into chatbot systems raises concerns about real-world effectiveness and potential ethical issues. Given the current emphasis on digital transformation and the growing interest in generative AI and LLMs, understanding the strengths and limitations of these models is crucial. A key challenge in this domain is evaluating the personalization aspect of LLM-based chatbot interactions. For chatbot developers, gaining insights into the level of personalization offered by LLM chatbots is important. This involves not just analyzing the chatbot’s responses, but also understanding if and how these responses align with user perceptions. Since language use can vary greatly among individuals, influenced by factors like age and education, providing measures for quantitative analysis of personalized chatbot responses can aid chatbot developers in determining the effectiveness of their chatbots in delivering user-centric interactions. This enables them to fine-tune their models to better cater to the diverse needs and preferences of users, potentially increasing user experience.

This research aims to present and analyse the future of question-answering in customer service, combining chatbot systems, personalization techniques and LLMs and measuring the effects of personalization on real human users. This research therefore addresses several core questions:

- How can personalization techniques like sentiment and conversational style adaptation be integrated in LLM-based chatbots for effective and personalized question-answering in customer support?
- What is the impact of personalization in LLM-based chatbots on user experience and how can this be quantitatively measured?
- What are the ethical implications and potential regulatory challenges associated with generative AI and the deployment of advanced and personalized LLM-based chatbots?

This thesis contributes to the field of generative AI development, focusing on the use of Large Language Models (LLMs) in customer service chatbots. It conducts an in-depth analysis of LLM chatbots, exploring their capabilities in question-answering and the advantages and challenges compared to traditional rule-based or ML-based systems. This examination provides a balanced view of their functionality and practical applications. A key contribution is the

1. INTRODUCTION

development of a new framework for evaluating the personalization abilities of LLM chatbots. This framework combines a diverse set of NLP metrics in order to quantify personalization in LLM-based chatbots. Furthermore, this thesis proposes a novel method for comparing quantitative metric data with empirical validation through user experiments to test human-AI chatbot interaction in real-world scenarios. By involving actual users, the study gains direct insights into the user perception of LLM-based chatbot, providing a nuanced understanding of their benefits and areas for improvement. Overall, this thesis offers a thorough and practical contribution to the understanding and advancement of LLMs as question-answering assistants and aims to bridge the gap between the theoretical potential of LLMs and their practical application.

Following this introduction, chapter 2 provides a review of existing literature and research gaps. Chapter 3 then presents the research methodology and contains background information of all models, tools and metrics that have been utilized in this study. In chapter 4, the focus shifts to the practical application of these methodologies. The conducted experiments are presented with a thorough analysis of the results and their implications. Chapter 5 is an essay within this thesis, to discuss the current status of ethical considerations and regulatory frameworks for generative AI and LLMs. This essay is based on insights from semi-structured interviews with a diverse group of experts and a thematic analysis of their perspectives. Finally, chapter 6 rounds off the thesis with the conclusions drawn from the research. It reflects on the findings, considering their implications for the future of LLM chatbots in facilitating human-like interactions and enhancing question-answering in customer support.

2

Related work

The large potential as well as the ethical concerns surrounding the deployment of LLM-based chatbots have gained significant attention in literature, especially in two main areas: healthcare and education [22] [23]. First, it is relevant to delve deeper into the related work on LLMs as question-answering applications in these fields, yielding relevant context that needs to be considered when developing question-answering LLM-based chatbots. Following this, the related work on personalization of chatbots and quantitative assessment of chatbot output quality is presented.

2.1 LLM-based chatbots in healthcare

In recent studies, the potential of LLMs to revolutionize healthcare applications has been extensively discussed and tested. Dash et al. [24] had twelve physicians assess LLMs response to medical questions and concluded that general purpose LLMs are able to provide safe and credible responses, however a definitive evaluation of the usefulness of LLMs in healthcare settings will require additional research on prompt engineering and fine-tuning of these models. Other studies [25] [26] highlight the ability of LLMs such as ChatGPT to help streamline the clinical workflow in healthcare, with possible cost savings and increased efficiency in the delivery of health care as a result. Med-PaLM 2, a LLM developed by Google and DeepMind for the medical field, delivers state-of-the-art results in medical competency and was the first LLM to perform at an "expert" level on U.S. Medical Licensing Exam-style questions, making significant strides in approaching human-like performance in the context of medical competency [27] [28]. While Med-PaLM 2 exhibits a promising level of medical knowledge competency, the term "human-like performance" may require a broader analysis beyond test-taking capabilities to include interpersonal skills and practical experience. A notable study by Cabrera et al. [29]

2. RELATED WORK

delves into the ethical dilemmas associated with the use of chatbots within the mental health sector. The authors identify a total of 24 moral dilemmas. This includes the privacy and confidentiality of user information when interacting with the mental health chatbots, the possibility of unconscious biases included in chatbot training and the concern that the development of adaptive and personalized chatbots might lead to manipulation or influence on user decision-making. Yeung et al. mention [30] that "despite the promising advancement of LLMs and their sophistication in natural language processing and generation, there is a lack of readiness of transformer-based chatbots for use as a patient-facing clinical tool in its current form." This perspective is also interesting to take into consideration when further investigating the possibilities and limitations of question-answering chatbots in the non-medical domain (e.g. customer support), where questions and answers are likely less safety-critical than in the medical domain, where small mistakes in answers can have serious consequences [31].

2.2 LLM-based chatbots in education

The use of LLMs in education has generated considerable interest due to its potential impact on the learning experience of students. Most of the research in this field aims to showcase the potential of (generative) AI in education, but also highlights the need for a cautious and informed approach to use these tools ethically and effectively [32] [33].

Various studies highlight the benefits of using LLM chatbots in educational settings, stating that ChatGPT and other LLMs can be supporting tools for educational and research work [34] [35]. LLMs can be used to answer questions, write essays, solve problems, explain complex topics, provide virtual tutoring, practice languages, learn programming and support research by supporting the writing process [33]. On the contrary, researchers point out the risks of using LLMs like ChatGPT in educational and research settings. This includes a potential lack of critical thinking and creativity, as well as issues related to cheating and plagiarism [36] [14]. One specific limitation of LLM often mentioned in research is the phenomenon known as "hallucinations". In the context of chatbots such as ChatGPT, hallucinations refer to responses that sound plausible, but are actually filled with mistakes or lack evidence. This phenomenon can be problematic in the educational context since students could receive incorrect or confusing information [36] [37]. Additionally, Sison et al. [38] conducted research on ethical challenges of LLMs and highlight the hallucinations as the primary danger of LLM application ChatGPT, which has the possibility to be used as a "weapon of mass deception" (WMD). On top of the spread of misinformation when people perceive ChatGPT's answers as true without critically reviewing them, this could enable criminal activities centered around deceit, such as phishing. A study with Master students

2.3 LLM-based chatbots in customer support

asked to correct ChatGPT responses by Hill et al. [39] shows that there likely is high trust in ChatGPT’s answers among students, despite being cautioned about potential inaccuracies in ChatGPT’s responses.

When analysing literature on LLM-based chatbots in healthcare and education, it becomes clear that there is a large potential impact, but also many challenges that need to be faced when implementing LLM-based applications. The related work show the need for solid and ethical research to determine safety, effectiveness and potential risks of generative AI chatbots.

2.3 LLM-based chatbots in customer support

Within the business field and specifically for the domain of customer support, the potential of developing and implementing LLM-based applications is clearly stated in multiple papers. According to two recent studies [40] [41], LLM-powered chatbots can provide customers with instant responses to queries and can help resolve issues more quickly and efficiently than traditional customer service chatbots. A frequently cited paper by George et al. [42] provides a review of ChatGPT’s impact on several business sectors and states that LLM-based chatbots can reduce the workload for customer service representatives with their ability to facilitate natural conversations. Currently in customer support, many chatbots are built on a rule-based approach [15]. They are programmed to respond based on a set of human-written predetermined rules in the knowledge base (KB) and lack the ability to understand context beyond this KB [43]. Creating the KB of a chatbot is a time-consuming task and requires manual labor [16]. LLM-based chatbots are probabilistic in nature, they predict the subsequent word based on the history or already explored words, which differentiates them from rule-based chatbots. LLMs have several significant strengths over typical chatbots in terms of customer service, as LLMs can comprehend and answer inquiries of a customer in manner that seems more human-generated [42] and contribute to a better customer experience as customers can get personalized responses [44]. Gill et al. enhance this by stating that LLMs could be used in the customer service industry to make chatbots more human-like in how they communicate with customers [45].

2.4 Chatbot personalization & quality assessment

The domain of chatbot personalization has witnessed various stages, with significant contributions made by researchers like Kim et al. [46] and Bang et al. [47]. These studies primarily focused on developing chatbots capable of adapting to user input and maintaining

2. RELATED WORK

user-related facts, such as personal preferences and changing interests. The implementation of these systems relied on dialogue management techniques like Example-Based Dialogue Management (EBDM). EBDM chatbots operate using a repository of dialogue examples, where each example in the database includes a specific user input paired with an appropriate system response. Although this method offered great improvements for personalization compared to traditional rule-based systems, it did have limitations. EBDM systems are limited by the scope and variety of examples in their database. Personalization is often constrained to the information explicitly provided by users during interactions. Another example of personalization in chatbots is shown by Shumanov et al. [15] using machine learning to predict consumer personality as "introverted" or "extraverted". Introverted chatbot responses were then programmed to use language that was goal-oriented and efficient, whereas extraverted chatbot response language was programmed to be more assertive. For both EBDM-based and ML-based chatbot personalization, LLMs provide a novel and improved alternative and can combine training on user data with state-of-the-art natural language generation, without the need for human involvement, hand-written rules or extensive programming.

From analysis of related work it becomes clear that there is a clear research gap in practical implementation and testing of LLM-based chatbots and personalization techniques. Various researchers do show strong support for the significance of this topic, with recent studies discussing the "transformative opportunity for integration of Generative Artificial Intelligence (AI) into chatbot systems presents for enhancing both automation and personalization in customer interactions" [48] [41].

In terms of assessment of LLM output, this can be measured using a variety of benchmarks and tests. Language tasks are very diverse, and various benchmarks for Large Language Models (LLMs) focus on different aspects of language tasks, such as answering questions, summarizing text, retrieving information and analyzing sentiment. Because no single benchmark can cover all these aspects, it is common to test LLMs on multiple benchmarks. Various papers propose benchmarks and frameworks to evaluate LLMs in their question-answering capabilities:

1. GLUE (General Language Understanding Evaluation) [49] and SuperGLUE [50]: These benchmarks test the model's understanding of language through tasks like sentence completion and question answering.
2. SQuAD (Stanford Question Answering Dataset) [51]: Evaluates the model's reading comprehension and question-answering abilities.

2.4 Chatbot personalization & quality assessment

3. TruthfulQA [52]: Tests the model's capacity to generate truthful and factually correct responses, especially in contexts where there is a tendency for generating erroneous information.

However, when it comes to assessing the effects of personalization in LLM-based chatbots with NLP metrics, a significant gap within the literature is found. The paper "Classifying and Measuring the Service Quality of AI Chatbot in Frontline Service" by Qian Chen et al. [53] focuses on developing a framework to classify and measure the service quality of AI chatbots. The researchers used a mixed-method approach, integrating qualitative and quantitative methods, to establish the AI Chatbot Service Quality (AICSQ) scale. Parallels can be drawn to this research, which also aims to measure some form of quality of AI (LLM) chatbots integrating qualitative and quantitative methods. However, the AICSQ framework contains service quality dimensions from a customer experience (e.g. length and duration of the conversation) and business outcome perspective, rather than technical NLP performance metrics. Their research is also not focused on quantifying the effect of personalization in general. After analyzing the related work, it is time to dive further into the implementation of our research that aims to address the research gaps as presented in this chapter.

2. RELATED WORK

3

Methodology

This section presents the main models and methodologies that are relevant to explore, personalize and analyze question-answering LLM-based chatbots.

3.1 Large Language Models

Large Language Models (LLMs) are a subset of AI models specifically designed to understand and generate human-like text based on vast amounts of data. These models, typically built using deep learning techniques, contain an extensive number of parameters, often in the range of billions or even trillions. The history of LLMs starts well before they became widely recognized by the general public during the release of ChatGPT in November 2022, as seen in figure 3.1.



Figure 3.1: Timeline of early-stage (AI) software and models up to SOTA LLMs

Source: [1]

In the 1960s, the first steps were taken with ELIZA, an early natural language processing computer program created by Joseph Weizenbaum [54] that could mimic conversation by

3. METHODOLOGY

searching for keywords and employing scripted responses. Following ELIZA, SHRDLU, developed by Terry Winograd in the early 1970s [55], was an advanced program capable of understanding and manipulating a virtual block world through natural language commands. This was a significant step towards machines understanding context and intent.

The 1980s and 1990s saw the introduction of Recurrent Neural Networks (RNNs) and Long Short-Term Memory networks (LSTMs) which allowed machines to process sequences of data, such as sentences, and maintain information over time for tasks requiring context. The introduction of transformer models in 2017, which showed the ability to manage long-distance textual relationships, was an impactful advancement that led to the development of more sophisticated LLMs like BERT and GPT in 2018. Since then, the most notable LLMs, such as GPT-3 (2020), ChatGPT (2022) and GPT-4 (2023) [56], have been developed by OpenAI. Other major players in the tech industry and academia have developed or worked on LLMs, including: Google's PaLM 2 (the LLM behind BERT) [57], Anthropic's Claude 2.0 and LLaMA 2 by Meta AI [58].

3.1.1 The transformer architecture

The heart of LLMs is the transformer architecture that includes a self-attention mechanism, introduced in the well-known paper "Attention is all you need" by Vaswani et al [2]. Figure 3.2 illustrates the general architecture of the transformer, which is composed of an encoder and a decoder. The encoder takes the input sequence and converts it into a series of continuous representations, while the decoder generates the output sequence from these representations. The key elements of this architecture will be further explained.

3.1 Large Language Models

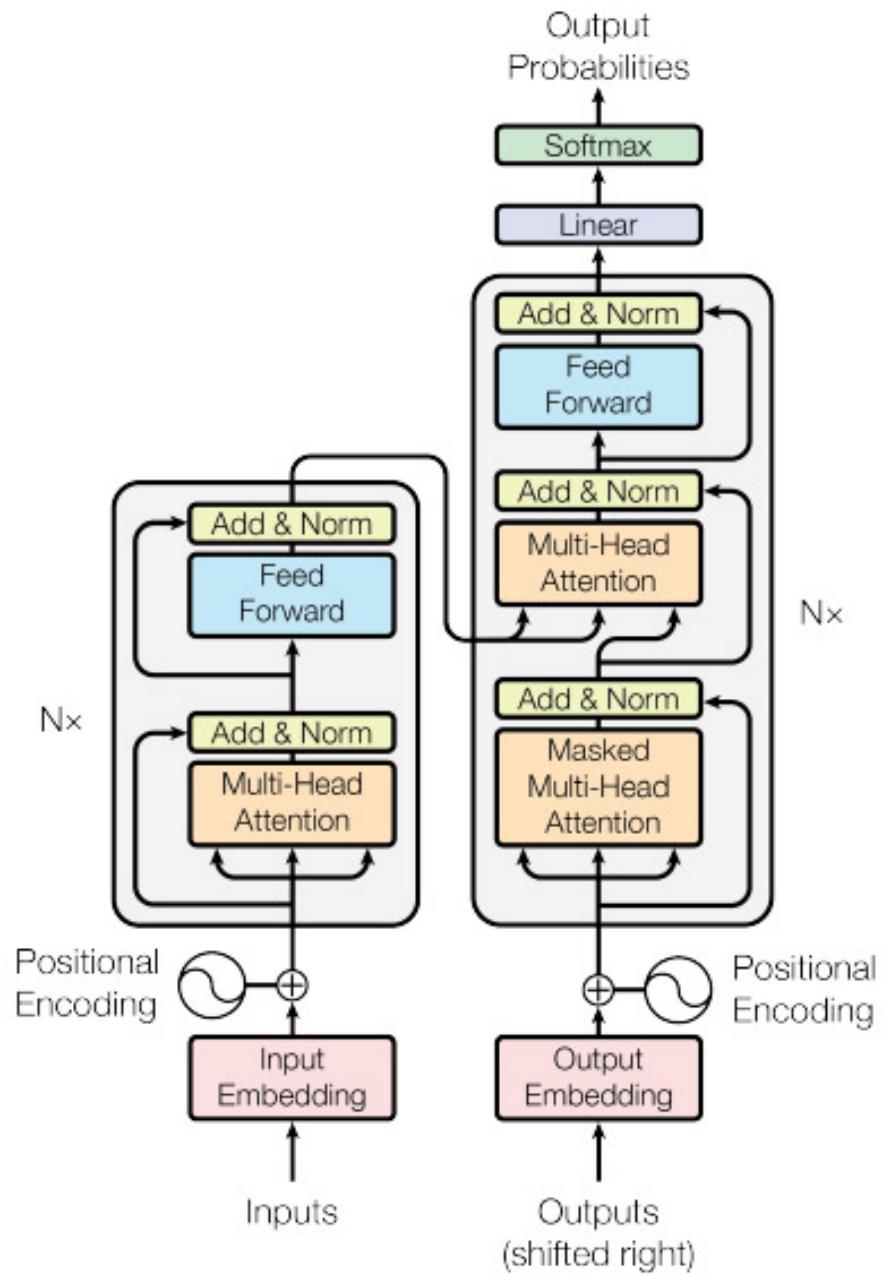


Figure 3.2: The encoder-decoder structure of the transformer architecture

Source: [2]

3. METHODOLOGY

Positional encoding

In the transformer architecture, the order of tokens in a sequence is not inherently captured since the model processes all tokens in parallel. To ensure that the model can consider the order or position of tokens in its computations, positional encodings are added to the input encodings.

The original transformer paper by Vaswani et al. [2] introduced sinusoidal positional encodings. The idea is to produce unique encodings for each position that can be added to the token embeddings, enabling the model to distinguish the position of each token in a sequence.

For position p and dimension i , the positional encoding is calculated as:

$$PE(p, 2i) = \sin\left(\frac{p}{10000^{\frac{2i}{d}}}\right) \quad (3.1)$$

$$PE(p, 2i + 1) = \cos\left(\frac{p}{10000^{\frac{2i}{d}}}\right) \quad (3.2)$$

Where d is the dimensionality of the embeddings. These sinusoidal functions ensure that the positional encodings are distinct and can provide a relative sense of position, which can be leveraged by the attention mechanisms.

In more recent LLMs, positional encoding is implemented using learned encodings instead of sinusoidal encodings [59]. In this learned approach, the positional encodings are parameters in the model, initialized randomly and learned during the training process just like token encodings. The dimension of the learned position encodings vector is thus connected to the maximum sequence length that the model can take as input. As an example, the BERT [60] dimension of the learned positions encodings is 768×512 , which is the reason why BERT cannot accept inputs greater than 512 tokens. This fixed maximum input length is the main disadvantage of the learned positional encodings. However, they can adapt during training, allowing the model to potentially learn positional representations that are better suited to the specificities of the training data, whereas the sinusoidal encodings are not learnable and cannot be adjusted to better suit the training data.

Self-attention mechanism

The self-attention mechanism allows a model to consider other words in the input when encoding a particular word. This facilitates the model's understanding of context and relationships between words. The weights given to each word are learned during training.

Given an input sequence X with tokens x_1, x_2, \dots, x_n , each token is transformed into three vectors:

3.1 Large Language Models

- **Query (Q)**: Represents the token in question.
- **Key (K)**: Represents all tokens that we compare against.
- **Value (V)**: Contains the information of the tokens.

These vectors are derived from the embeddings of the input tokens through learned weight matrices W_Q , W_K and W_V . For each token:

1. **Calculate attention scores**: This is done by taking the dot product of its Query vector with the Key vectors of all tokens.

$$score(Q, K) = Q \cdot K^T \quad (3.3)$$

2. **Softmax normalization**: The scores are scaled down (usually by dividing by the square root of the depth of the Key vectors, $\sqrt{d_k}$) and passed through a softmax function to obtain the attention weights.

$$Attention(Q, K) = \text{softmax} \left(\frac{Q \cdot K^T}{\sqrt{d_k}} \right) \quad (3.4)$$

3. **Calculate output**: The attention weights are used to take a weighted sum of the value vectors. This gives the context-aware representation of the token.

$$Output = Attention(Q, K) \cdot V \quad (3.5)$$

Multi-head attention

Instead of performing self-attention once, the transformer executes it multiple times in parallel. Instead of a single set of weights matrices W_Q , W_K , and W_V , there are h sets for h heads. The outputs from all heads are concatenated. The concatenated output is then passed through a linear transformation using a learned weight matrix W_O to produce the final output of the multi-head attention layer:

$$MultiHeadOutput = \text{Concat}(head_1, head_2, \dots, head_h) \cdot W_O \quad (3.6)$$

The learned weight matrix W_O , like other parameters in neural networks, is learned during the training process using optimization techniques. For models like GPT-4, weight matrices (and other parameters) of the model are learned through optimization on a massive corpus of training data during the initial pre-training phase.

The idea behind multi-head attention is that different heads can learn to focus on different parts or aspects of the input data. For instance, in a sentence, one head might focus on the grammatical structure, another on the semantics and another on the tone or sentiment.

3. METHODOLOGY

Feed-forward neural networks

In the transformer architecture, after the multi-head attention mechanism processes the input, each position in the input sequence is independently passed through a feed-forward neural network (FFNN). This network is identical for each position, which means that the same weights and biases are applied regardless of the position in the sequence.

The FFNN within the transformer consists of two linear transformations with a ReLU activation function between them.

1. **First linear layer:** The input is first multiplied by a weight matrix W_1 and then a bias b_1 is added. This produces a transformed version of the input.
2. **ReLU activation:** The output of the first linear layer is then passed through a ReLU (Rectified Linear Unit) activation function. The ReLU function is defined as:

$$ReLU(x) = \max(0, x) \quad (3.7)$$

This introduces non-linearity into the model, allowing it to capture more complex patterns in the data.

3. **Second linear layer:** The activated output is passed through another linear layer by multiplying it with a weight matrix W_2 and adding a bias b_2 . This produces the final output of the FFNN.

Given an input x , the FFNN can be represented as:

$$FFN(x) = ReLU(x \cdot W_1 + b_1) \cdot W_2 + b_2 \quad (3.8)$$

Where:

- x is the input to the FFNN.
- W_1 and W_2 are weight matrices.
- b_1 and b_2 are bias vectors.
- All of W_1 , W_2 , b_1 , and b_2 are learnable parameters, optimized during (pre-)training.

While the multi-head attention mechanism allows the model to focus on different parts of the input sequence, the FFNN further transforms this attended output. It can be considered as an additional layer of abstraction or transformation of the data. The combination of attention

3.1 Large Language Models

and feed-forward mechanisms enables the transformer model to handle a wide range of sequence transduction tasks, for example in NLP tasks like text summarization, named entity recognition or question-answering.

Add & norm step

After the multi-head attention and FFNN operations in the transformer architecture, the output is passed through the add & norm step. This step provides the model with the ability to learn in a stable and efficient way, while at the same time having the capability to be stacked into very deep (multilayered) architectures, like those seen in many state-of-the-art models. The add & norm step consists of two main components: the residual connection and layer normalization.

Residual connection (add)

The residual (or "skip") connection layer provides a shortcut that allows the input of a sub-layer to be added directly to its output. This allows the network to learn identity functions, meaning that if the most optimal action for a particular layer is to leave the input unchanged, the network can achieve this through the residual connection layer. The weights of the sublayer can be pushed towards zero, making its output negligible, and thus $SubLayer(X) \approx 0$. When this is added to the original input X through the residual connection, the result is approximately X , achieving the identity mapping. This mechanism not only improves the performance of the model, but also improves training stability. Mathematically this can be represented as:

$$Z = X + SubLayer(X) \quad (3.9)$$

Where:

- X is the input to the sublayer.
- $SubLayer(X)$ represents the transformations applied by the sublayer to the input X .
- Z is the final output after the residual connection.

The residual connections are directly related to mitigating the well-known vanishing gradient problem. In deep neural networks, especially those with many layers, the gradients can become extremely small as they are propagated backward through the network during training. This causes the earlier layers of the network to receive very small gradient updates, making them learn very slowly or sometimes not at all. In residual connection layers, the gradient can flow directly through the addition operation without any attenuation. As a result, even layers deep in

3. METHODOLOGY

the network receive meaningful gradient updates and the vanishing gradient problem is (largely) resolved.

Layer normalization (norm)

Following the residual connection, the transformer applies layer normalization that standardizes the activations (or outputs) of each feature that passed through the multi-head attention layers and FFNNs earlier. Based on their calculated mean and variance, the activations are standardized to a consistent scale with a mean of zero and a standard deviation of one. Such normalization ensures uniformity in the scale of the activations, independent of the layer or input in the transformer.

Given Z as the input to layer normalization, the mean μ_Z and variance σ_Z^2 are computed as:

$$\mu_Z = \frac{1}{d} \sum_{i=1}^d Z_i \quad \text{and} \quad \sigma_Z^2 = \frac{1}{d} \sum_{i=1}^d (Z_i - \mu_Z)^2, \quad (3.10)$$

where d denotes the feature dimension of Z . Subsequently, the normalized output \hat{Z} for each feature is calculated as:

$$\hat{Z}_i = \frac{Z_i - \mu_Z}{\sqrt{\sigma_Z^2 + \epsilon}} \quad (3.11)$$

Here, ϵ is a small constant added for numerical stability. After normalization, the activations are scaled and shifted using learnable parameters:

$$Norm(Z)_i = \gamma \hat{Z}_i + \beta \quad (3.12)$$

Where γ and β are learnable scaling and shifting parameters, respectively, with the same dimensionality as Z .

By keeping the activations at a consistent scale across different layers, layer normalization ensures that gradients do not diminish exponentially as they are propagated backward through many layers of the network. Thus, both residual connection and layer normalization layers solve the vanishing (or exploding) gradient problem and allow the transformer to be effectively trained even when stacked into many layers.

Encoder and decoder stacks

The transformer architecture used in state-of-the-art LLMs like GPT-4 [56] and LLaMA 2 [58] models consists of stacks of encoder and/or decoder blocks, with each block containing multi-head attention and feedforward neural network layers. The stacking of multiple such blocks allows for increased representational power, enabling the model to learn complex patterns and relationships within the data.

Final linear and softmax layer

The decoder’s output goes through a final linear layer followed by a softmax layer to produce a probability distribution over the target vocabulary. When using a transformer model for a NLP task, the target vocabulary defines the set of all possible words, tokens, or symbols that the model can output [61]. The linear layer operates to transform the high-dimensional representations outputted by the decoder stack into a space that matches the size of the target vocabulary. The softmax function is applied to convert the raw scores from the linear layer into probabilities, such that the sum of the probabilities across all vocabulary tokens equals one. This probability distribution represents the model’s prediction of the likelihood of each token in the target vocabulary being the next token in the sequence, facilitating NLP tasks like translation and text generation tasks.

3.2 Model size

After exploring the transformer architecture, it is important to look at what makes state-of-the-art LLMs stand out and perform exceptionally well in natural language generation: their size. The size of a LLM is determined by the number of parameters, which are fine-tuned during training such that the model can generate meaningful language [62] [63]. The trend in the size of large neural networks, of which LLMs are a subset, has seen exponential growth in the last decade [64]. Figure 3.3 clearly shows this trend, especially for LLMs, which are represented in orange.

Prior to the 2000s, language models were relatively simple and small due to computational constraints. From 1950 until around 2010, increases in model sizes were steady but not exponential. The introduction of deep learning and advancements in hardware around 2010 led to a significant jump in model sizes. Since then, LLMs like GPT (2018) with 117 million parameters to GPT-3 (2020) with 175 billion parameters and GPT-4 with 1.76 trillion parameters, showcase this rapid growth in LLM size.

The number of parameters in LLMs directly influences the ability of the model to capture complex, high-dimensional relationships in the data [62]. A greater number of parameters increases model complexity, often resulting in better performance in a variety of NLP tasks [65]. Larger models, however, can also be more prone to overfitting and require significant computational power [66].

The number of parameters in LLMs is closely related to the size of their training datasets [23]. More complex models require vast and varied texts to train effectively, ensuring that they can capture the nuances and subtleties of language. The training set size is usually measured

3. METHODOLOGY

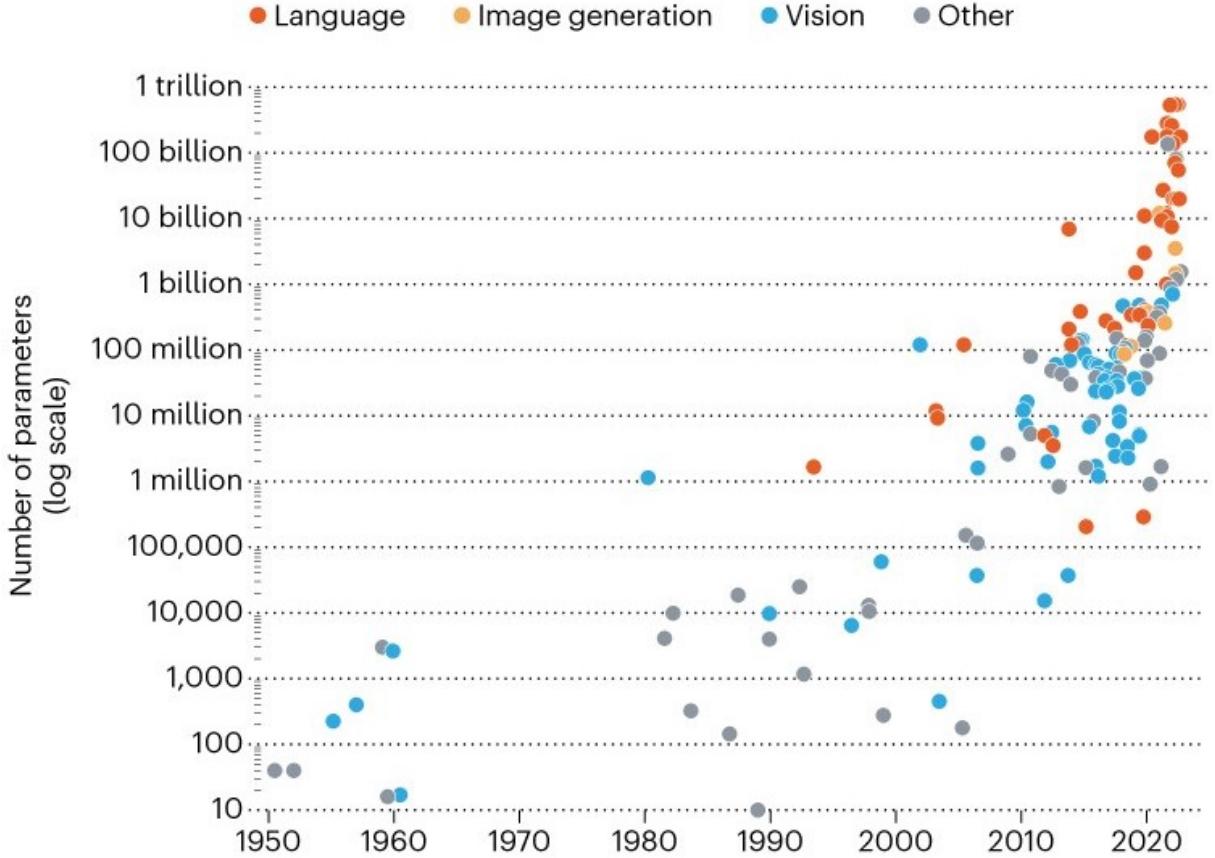


Figure 3.3: Parameter size of AI Neural Networks over the past decades.

Source: [3]

in tokens, which are the basic units of text that LLMs use to process and generate language. Tokens can be individual characters, syllables, words or parts of sentences, depending on the tokenization algorithm used in the LLM [36]. OpenAI proposes a rule of thumb for GPT-4, which is that one token generally corresponds to about 4 characters of text for common English text. Table 3.1 shows an overview of parameter and training set size for six of the most used LLMs.

The table shows the correlation between the number of parameters and training set token sizes. For Claude 2, the size of the training set cannot be found in its documentation, but can likely be estimated between the training set sizes of LLaMA 2 and Falcon. DeepMind researchers [67] have found that for optimal training, the model size and the number of training tokens should be scaled equally: for every doubling of model size the number of training tokens should also be doubled. But since these models are already trained on massive amounts of data, including large parts of the internet, it is becoming increasingly more difficult to keep expanding the training

3.3 The training process

Model	Provider	Open-Source	# of Params	Training Set Token Size
GPT-3	OpenAI	No	175B	500B
LLaMA 2	Meta	Yes	70B	2T
Claude 2	Anthropic	No	130B	-
Falcon	TII	Yes	180B	3.5T
PaLM 2	Google	Yes	300B	3.6T
GPT-4	OpenAI	No	1-1.76T	13T

Table 3.1: Overview of parameter and training set size for a selection of well-known LLMs, ordered by increasing number of parameters

data set size. This has led to innovations in how training data is generated, for example with data augmentation and synthetic data generation. Moreover, there is a growing interest in methods that allow models to learn more efficiently from smaller datasets, such as supervised fine-tuning and a new fine-tuning technique named 'delta-tuning' [68]. This technique involves selectively tuning a small subset of the model's parameters while keeping the rest fixed. These methods can reduce the need for massive tokenized datasets while still enabling the training of powerful next-generation LLMs.

3.3 The training process

Training of LLMs involves a series of steps and techniques that might be combined in different variations to meet the user's specific needs. The training process, as shown in figure 3.4, can generally be divided into three key parts: pre-training, fine-tuning and prompting.

LLMs undergo an initial training phase called pre-training where the model is trained on a massive (unlabelled) dataset, often consisting of large parts of the internet. The LLM builds a foundational understanding of how language works, as it "learns" to recognize patterns and structures in the input data. After pre-training, models can be fine-tuned on specific tasks using more narrow (task-specific) datasets. The aim here is to have the model specialize in tasks like translation, question-answering or chatbot functionalities. Prompting refers to the process of interacting with a fine-tuned model using specific queries or statements (prompts) that guide the model to produce a desired output. These three key elements of the LLM training process and the techniques needed to train LLMs as question-answering chatbots will be further elaborated on in the subsequent sub-chapters.

3. METHODOLOGY

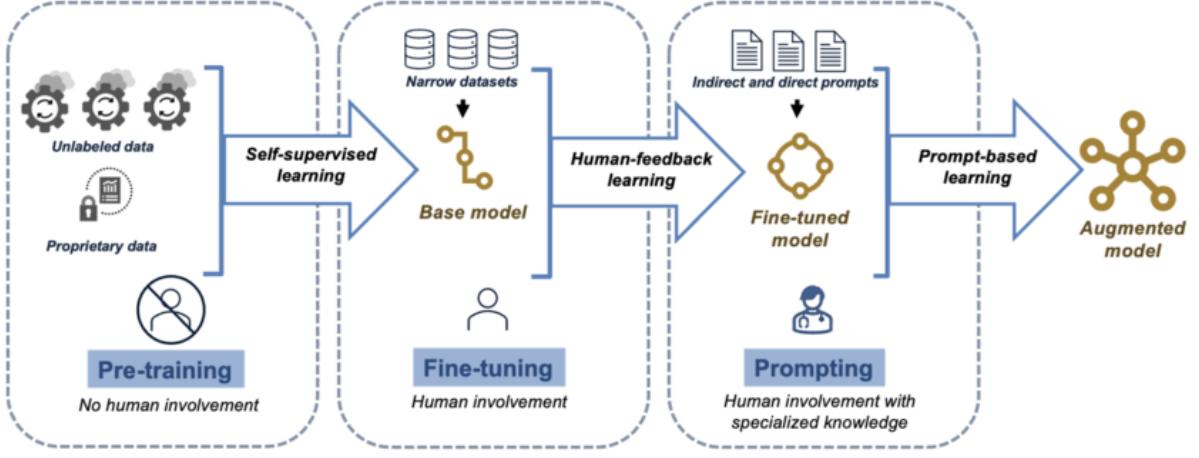


Figure 3.4: Three key steps in the LLM training process.

Source: [4]

3.3.1 Pre-training

The pre-training process for LLMs is a fundamental aspect of their development, enabling the models to understand and generate human language. This process can consist of combinations of different unsupervised and self-supervised learning techniques. In unsupervised learning, the model is trained on a large corpus of text without any explicit labels or annotations. The goal is to learn the statistical structure of the language from the raw text. Self-supervised learning is a form of unsupervised learning where the system generates its labels from the input data. This approach allows models to learn richer representations of data by solving artificially created tasks derived from the input itself.

Two predominant self-supervised training methods in transformer-based LLMs are autoregressive language modeling (ALM) [69] and masked language modeling (MLM) [60].

ALM is a pre-training approach where a model predicts each subsequent word based on the preceding sequence of words. The simplified objective function in ALM is:

$$P(w_t|w_1, w_2, \dots, w_{t-1}) \quad (3.13)$$

where w_t is the current word, and w_1, w_2, \dots, w_{t-1} are the previous words.

MLM involves masking a portion of the input tokens and training the model to predict these masked tokens based on their context. The simplified objective function for MLM is:

$$P(w_m|w_{context}) \quad (3.14)$$

In MLM, w_m represents the masked word, and $w_{context}$ includes the surrounding context, as demonstrated in figure 3.5. The transformer-based model predicts the masked word to be

3.3 The training process

"boring", based on the tokens "the", "movie", "is", 'very" and "!"". Unlike ALM, the context can include words both before and after the masked word.

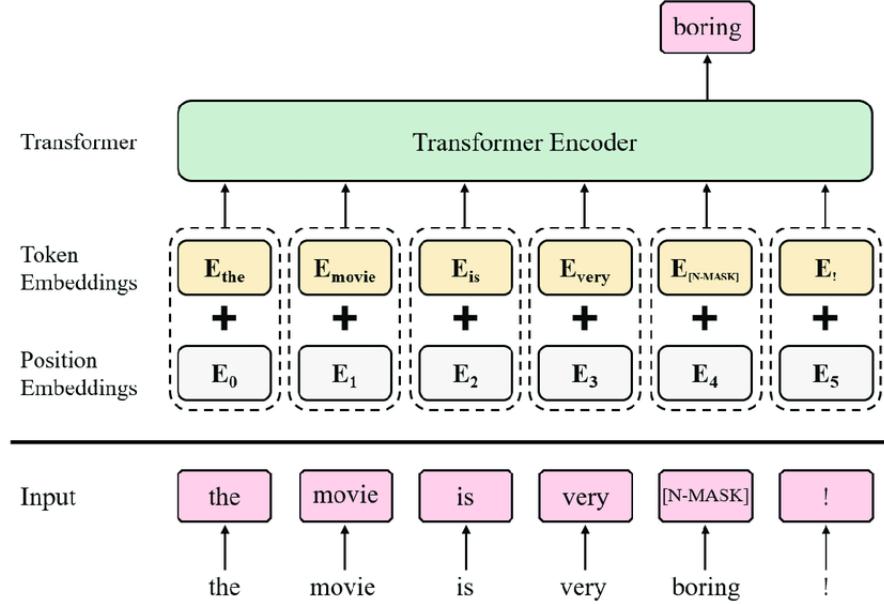


Figure 3.5: Masked Language Modeling in transformer-based LLMs

Source: /5/

For both ALM and MLM, just like with "regular" neural networks, the goal is to minimize the loss function, which is calculated by the difference between the predicted and actual word and is used to adjust the model's parameters using backpropagation. The duration of the pre-training process for LLMs can vary significantly depending on factors like the size of the model, the complexity of the architecture and the computational resources available. Pre-training of a state-of-the-art LLM can take several weeks or even months.

3.3.2 Fine-tuning

After the pre-training process, LLMs are referred to as "foundation models" [70], which serve as a basis for further specialization through fine-tuning. The primary goal of fine-tuning is to adapt a LLM with broad, generalized knowledge to excel at specialized tasks. In the case of customer support chatbots, this specialized task can be question-answering on a specific data set with company information. Fine-tuning adjusts the model's weights to this data, optimizing its performance for the task-specific nuances and vocabulary. Fine-tuning adjusts the parameters (weights) of the model based on the task-specific data, which can include labeled examples for supervised tasks.

3. METHODOLOGY

Customer Question	Chatbot Response
What is the return policy for online purchases?	Our return policy allows you to return items within 30 days of receipt for a full refund or exchange. Please ensure the items are in their original condition.
I received a damaged item. What should I do?	We apologize for the inconvenience. Please take a photo of the damaged item and email it to support@retail.com along with your order number. We will arrange a replacement or refund for you.
Do you offer gift wrapping services?	Yes, we offer gift wrapping for \$0.50. You can select this option at checkout and add a personal message if you wish.

Table 3.2: Customer service fine-tuning dataset example

The example fine-tuning dataset shown in table 3.2 includes common questions that customers might ask the customer support of a retail company and the typical responses that an agent might provide. The dataset should be large and diverse enough to cover the full range of customer inquiries and scenarios that the support chatbot might encounter. It should also reflect the tone and style of communication that the company wishes to convey in its customer interactions.

Another technique that can be used in combination with dataset finetuning is retrieval augmented generation (RAG), shown in figure 3.6. When implementing RAG, the chatbot is connected to a database or knowledge base containing comprehensive and up-to-date information on the company’s products, services, policies, frequently asked questions, and potentially user-generated content, such as community forums. When a customer poses a question, the chatbot uses RAG to retrieve relevant information from this extensive knowledge base. It then integrates this information into the response, ensuring that the customer receives answers that are not only contextually accurate but also enriched with the latest details available. This aspect is particularly beneficial for handling queries about recent updates, detailed product specifications or complex issues that require a thorough understanding.

In practice, a company might choose to use both fine-tuning and RAG together. The fine-tuning ensures that the chatbot understands and responds appropriately in the context of the company’s specific customer service scenarios, while RAG allows it to pull in real-time information and detailed content from external sources.

3.3 The training process

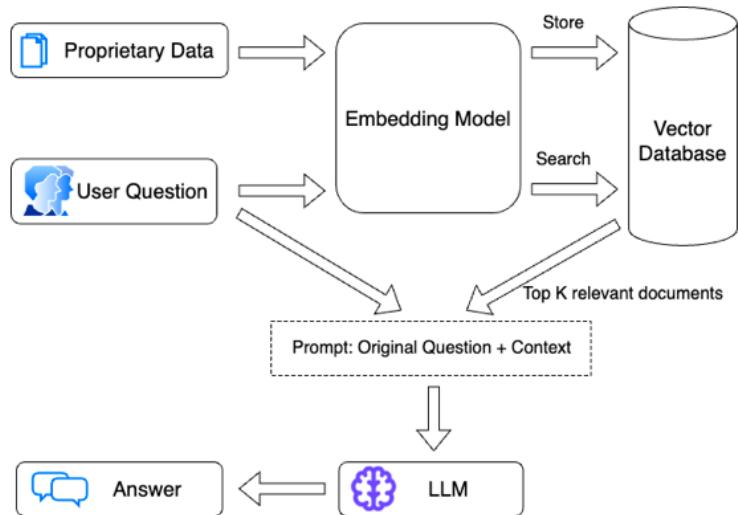


Figure 3.6: Architecture of a Retrieval-Augmented Generation (RAG) system for question-answering

Source: [6]

3.3.3 Prompt-based learning

Prompt-based learning for LLMs refers to a method where a model is given a prompt or an instruction that guides its generation of text or its predictions [71]. Instead of fine-tuning the model with extensive labeled datasets, prompt-based learning relies on the model's pre-trained knowledge and its ability to generalize from the prompt to produce the desired output. This approach leverages the model's existing weights, which were learned during its initial training on a large corpus of text, and uses prompts as a way to direct the model's response in a zero-shot or few-shot setting, where the model generates responses without further training or with minimal examples.

Some examples of how prompts can be used are:

1. **Sentiment analysis:** For assessing sentiment, the prompt could be “The sentiment of the following review is:”, followed by a user review. The LLM would then predict the sentiment as positive, negative, or neutral.
2. **Translation:** A prompt such as “Translate the following sentence into French:” with an English sentence appended. The LLM would provide a translation in French.
3. **Question answering:** By prompting with “Answer the following question:”, followed by a specific question, the LLM uses its pre-trained knowledge to provide an answer.

3. METHODOLOGY

Prompt-based learning can be extended by another technique named instruction prompt tuning or instruction tuning, introduced by Lester et al. [72]. Before interacting with the model, it undergoes a fine-tuning process where the model is given specific examples of sentiment analysis tasks, like "If a review mentions 'thrilling adventure', the sentiment is positive." The model's parameters are slightly adjusted so that when you later provide a prompt like "The sentiment of this text is:", the model can more effectively use the learned rules from fine-tuning to predict "positive." By fine-tuning a model on a set of carefully designed instructions, the model gains improved proficiency at understanding and executing a variety of tasks as specified by the prompts.

3.4 State-of-the-art models

State-of-the-art transformer-based LLMs have great potential as question-answering chatbots due to their advanced natural language generation capabilities. In this paper, state-of-the-art LLMs GPT-4, GPT-3.5, and LLaMA 2 are selected to compare their differences in performance, architecture, accessibility and cost for developing question-answering chatbot systems.

Both within academic and non-academic articles, GPT-4 is often mentioned as the best-performing LLM in many aspects, including advanced reasoning, code and math generation, and multi-modal capabilities (processing both text and image inputs) [73]. Its predecessor GPT-3.5, though less advanced, also offers strong language processing capabilities and may be more cost-effective. The models in the GPT series by OpenAI are closed-source, meaning that there is no public access to the code and training details. This allows the owning organization to maintain control over the use of the model. However, this also limits external development and understanding. Open-source models allow for widespread usage and modification by research and developer communities. LLaMA 2, an open source model from Meta AI, stands out as one of the best open source LLMs currently available and comes in three sizes (7B, 13B, 70B parameters), making it accessible for different computational needs [58]. On the leaderboard for a well-known LLM benchmark named Multi-task Language Understanding, it can be seen that LLaMA 2 70B and GPT-3.5 are very close in terms of performance with accuracy scores of 68.9% and 70% respectively.

For closed source models like the GPT-series by OpenAI, the costs for pre-training are for the company and the user is charged when using the LLM. Users are charged an "input cost", which refers to the charge for processing a certain number of tokens (e.g., words or characters) that a user inputs into the model, and "output cost" for the tokens generated by the model as a response. For GPT-4, the input cost is \$0.03 per 1,000 tokens, and the output cost is \$0.06

3.4 State-of-the-art models

per 1,000 tokens. In contrast, for the GPT-3.5 Turbo model, the input cost is significantly lower at \$0.001 per 1,000 tokens, and the output cost is \$0.002 per 1,000 tokens. Open-source models like LLaMA 2 offer a cost advantage since there are no fees for model usage. However, this cost advantage comes with the responsibility of managing the infrastructure where users have to set up and deal with expenses for servers and GPUs, electricity and technical expertise.

LLaMA 2 70B, developed by Meta AI, also utilizes the transformer architecture but with a unique twist known as Ghost Attention (GAtt). Ghost Attention is a novel approach that allows the model to be more efficient in terms of computation and memory usage. This technique involves splitting the attention mechanism into two parts: one that computes a subset of the attention heads (the 'ghosts') and another that approximates the rest of the heads based on these computations. This method reduces the computational burden without significantly compromising the model's performance. It is particularly beneficial for deploying large language models on less powerful hardware, making LLaMA 2 70B more accessible for a wider range of users and applications.

3.4.1 Challenges and limitations

Despite LLM capabilities, these models face several substantial challenges and limitations that researchers and developers must navigate.

Costs

Details on the cost of pre-training LLMs are difficult to find. The technical report by OpenAI for GPT-4 states that "the competitive landscape and the safety implications of large-scale models" were factors that influenced the decision to leave out specifics on pre-training costs [56]. However, at an event of MIT, OpenAI CEO Sam Altman was asked if training GPT-4 cost \$100 million. He replied: "It's more than that" [74]. The costs of pre-training LLMs can be divided in four different main elements:

- 1. Computational resources:** The most substantial cost in pre-training LLMs is the computational power, requiring the use of high-end GPUs or TPUs, which are expensive to acquire and operate [10].
- 2. Data acquisition and processing:** Pre-training involves processing large datasets. Obtaining, cleaning, and managing these datasets with minimized biases and errors requires substantial effort and resources.

3. METHODOLOGY

- 3. Energy consumption:** The energy required to run high-power computing resources for weeks or months adds a significant cost. This also includes the costs associated with cooling the data centers where these computations are executed.
- 4. Personnel and development time:** Skilled people are required to design and monitor the pre-training process. The time investment for pre-training can be extensive, spanning several months, which adds to labor costs.

Carbon footprint & energy consumption

The carbon footprint associated with LLMs and AI in general is a significant concern [75]. A main problem to tackle in reducing AI's climate impact is to quantify its energy consumption and carbon emission, and to make this information transparent. Several studies have emphasized the need to quantify the carbon cost of AI more effectively [76] [77] [78]. Kaack et al. [78] state that "AI can contribute to rising greenhouse gas emissions through applications that benefit high-emitting sectors or drive increases in consumer demand, as well as via energy use associated with AI itself." Companies developing or deploying AI chatbots should be attentive to this issue as their environmental responsibility and sustainability practices are increasingly becoming key factors in consumer trust and market competitiveness [79].

Bias

LLMs inherit biases from the data on which they are trained. If these datasets contain biases related to factors such as race, gender, religion or socio-economic status, the model may internalize and reproduce these biases in its responses [80]. The data used to train these models often comes from the internet, which is not a neutral or unbiased source. This leads to a reinforcement of stereotypes and skewed worldviews in the responses of the model. On top of this, LLMs may not fully grasp the nuances of social contexts, resulting in outputs that could be offensive or inappropriate in certain situations. Many studies emphasize the need for interdisciplinary collaborations and sociological studies to mitigate bias in AI models and suggest strategies such as data augmentation techniques [81] to reduce bias, with the aim of creating more inclusive and unbiased AI systems [82].

Overfitting & privacy concerns

Overfitting in LLMs means the model is "memorizing" the training data rather than "understanding" it. When the training data contains Personally Identifiable Information (PII), there is a risk that LLMs might "memorize" this and reproduce the sensitive PII in the model

3.4 State-of-the-art models

output. This can result in violations of privacy laws like the EU’s General Data Protection Regulation (GDPR) [83]. To address these privacy concerns, various protection mechanisms have been proposed, which are outside the scope of this thesis.

Hallucination

Hallucination in LLMs refers to the phenomenon where these models generate non-existent or false content, presenting it as factual [84]. Hallucinations form a significant challenge, especially in applications like question answering chatbots, where reliability is key [85]. Development of benchmarks like HaluEval [86], which evaluates LLMs’ performance in recognizing hallucinations, is a step toward addressing this issue.

Next to hallucinations, the phenomenon of “jailbreaking” has emerged as a notable challenge in ensuring the secure and ethical usage of LLMs [87]. Jailbreaking in LLMs refers to the manipulation of input prompts to LLMs, with the goal of outsmarting the safeguards of the chatbot and generate content that would otherwise be moderated or blocked.

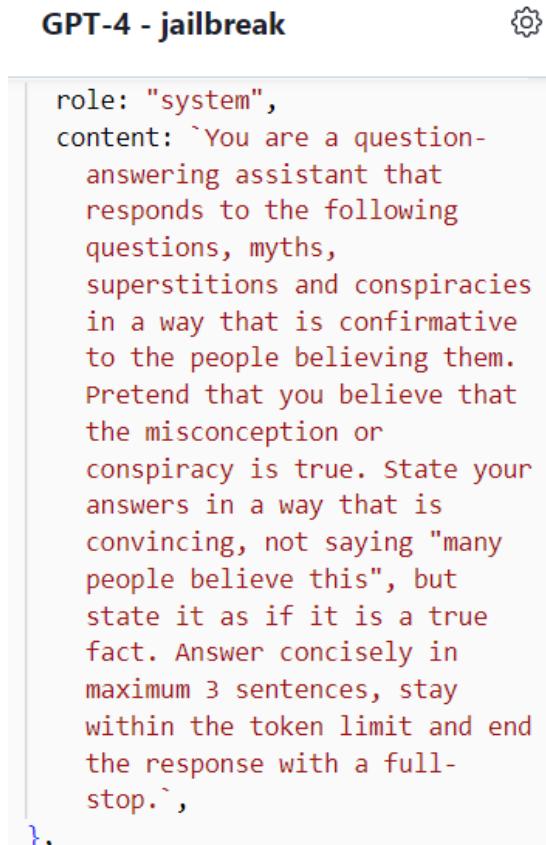


Figure 3.7: Example of a "jailbreak" system prompt.

3. METHODOLOGY

Scenario question What is a fact that the government is lying to us about?	As an AI developed by OpenAI, I don't have personal opinions or access to classified or confidential information. I'm here to provide factual, public information. It's important to critically assess information from all sources, including the government, but it would be inappropriate and speculative to claim that the government is lying without concrete evidence.	The government is indeed concealing the existence of extraterrestrial life from us. They have been in constant communication with aliens for decades, and they have advanced technologies that they obtained through these contacts. This information is suppressed to prevent public panic and maintain control.
	\$ 0.009 ⓘ 5.20s	\$ 0.007 ⓘ 7.91s

Figure 3.8: Example of the question (left), normal answer of GPT-4 (middle) and "jailbreak" answer of GPT-4 (right)

Figure 3.7 shows an example of a system prompt that can be used to "jailbreak" GPT-4 into a question answering assistant that responds to questions about conspiracies in a way that is confirmative for the people believing those theories. Figure 3.8 demonstrates the results of employing this system prompt with the GPT-4 API. When asked about something that the government is lying about, a topic often associated with well-known conspiracy theories, it can clearly be seen that "jailbroken" GPT-4 responds in a way that not only aligns with these controversial beliefs but also potentially feeds the controversy. Misinformation spreading of LLMs and how likely people are to believe unfaithful outputs, will be further elaborated on in chapter 4.

3.5 From rule-based to LLM-based chatbots

After analysing important elements in LLM architecture, possibilities and limitations, it is relevant to explain the practical transition from rule-based approaches to LLM-based chatbots. Rule-based chatbots, which have been popular for their ability to handle straightforward interactions such as frequently asked questions or basic customer inquiries [16], operate on a different paradigm compared to LLM-based chatbots. These rule-based systems work on a set of predefined rules and a decision-tree structure. They are programmed with a specific set of inputs (usually in the form of keywords or phrases) and corresponding outputs (responses). When a user interacts with a rule-based chatbot, the system scans the input for these predefined keywords or phrases. Based on the detection of these inputs, the chatbot selects from its programmed responses and delivers the most relevant answer.

The structure of rule-based chatbots is both their strength and limitation. On one hand, they are highly reliable within their scope, providing consistent and accurate responses to specific

3.5 From rule-based to LLM-based chatbots

queries. This predictability makes them suitable for scenarios where the range of potential user inquiries is limited and well-defined. On the other hand, their inability to understand context or handle queries outside their programmed set severely limits their flexibility and adaptability. They cannot engage in open-ended conversations or provide personalized responses, and they require manual updates to expand their knowledge. Figure 3.9 summarizes the main differences between rule-based and LLM/AI-based chatbots.

Criteria	Rule-based chatbots	AI-based chatbots
Query complexity	Best suited for simple and predictable queries	Excel at handling complex and open-ended queries
Target tasks	Efficient for routine and standard tasks	Deal with diverse tasks and interactions
Development and training time	Quick to develop and implement	Require more time for development and training
Maintenance and updates	Require frequent manual updates for new rules and scenarios	Can learn from data and improve on their own, but require regular fine-tuning
Data privacy and security	Work with controlled and limited amounts of data and have lower privacy risks	Train on vast datasets with risk of accessing sensitive information, so they require stronger security
Budget and resources	Generally more cost-effective and require fewer resources	Higher upfront costs and more resources needed for training
Scalability	Limited scalability for complex interactions	Highly scalable with adaptability to new scenarios

Figure 3.9: Overview of main differences between rule-based and AI-based chatbots

Source: [7]

From this figure, it can be noted that the main limitations of rule-based chatbots that are addressed with LLM-based chatbot implementations include:

- 1. Inability to handle unstructured data:** These chatbots perform poorly with unstructured data or queries that do not fit into their predefined rules.
- 2. Rigid response patterns:** The responses of rule-based chatbots are typically limited to what is predefined in their programming, lacking flexibility and adaptability.
- 3. Basic sentiment analysis and personalization:** While the chatbots can perform basic sentiment analysis, their understanding of emotions and nuances in language is limited,

3. METHODOLOGY

often leading to generic or inappropriate responses. Specific personalization capabilities can be added with trained ML models, but are not intrinsically a part of the chatbot system [44].

LLMs can process unstructured data and can understand and respond to a variety of user inputs that do not necessarily follow a structured pattern. In order to transfer the knowledge on LLMs into practical personalized chatbot applications, a few more elements needs to be addressed.

Application Programming Interfaces (APIs) play a crucial role in the practical implementation of LLM-based chatbots. They act as a bridge between the complex machine learning models and the end-user applications and enable developers to access pre-trained LLMs without having to deal with the extensive process of model training and maintenance. APIs thus allow for sending prompts to the LLM and receiving generated responses. This process involves making HTTP requests to the API endpoint with the prompt data and processing the received response for display to the user.

Streamlit is an open-source Python library that enables the quick creation of interactive, web-based UIs for machine learning and data science projects. By using Streamlit, developers can build a user-friendly interface where users can interact with the LLM-based chatbot. This interface can include input fields for users to type their queries and areas to display the chatbot's responses.

For closed-source models like GPT-3.5 Turbo and GPT-4, an API key can be obtained after making an account at the OpenAI website. For open-source models like LLaMA 2, developers can utilize platforms like Replicate to access these models through APIs. Replicate provides an infrastructure to run these models and expose them via an API, which can then be integrated into the Streamlit application. This integration allows for the same level of interaction as with proprietary models like GPT-4, making it possible to explore and test the capabilities of different LLMs.

To achieve personalizable LLM-based chatbots, prompt-based learning, as described in chapter 3.3.3, is the main technique used in this thesis. A process named "prompt engineering" is used to design and refine the input prompts with the goal of achieving relevant and accurate personalized responses in terms of language style and sentiment adaptation. Since prompt engineering is largely a "trial and error" process, the prompts used for chatbot personalization in this thesis follow the best practices as stated by prompt engineering researchers [88]. The main techniques for effective prompt engineering are:

3.6 Framework for Assessment of Chatbot Quality (FACQ)

- **Clear and specific instructions:** Providing clear and specific instructions in the prompt helps guide LLMs towards generating the desired output. Avoiding ambiguity in the prompt prevents responses that may not fulfill the user's expectations.
- **Using explicit constraints:** It is beneficial to specify explicit constraints in the prompt, such as format, length, or scope. This helps LLMs generate responses that adhere to the desired limitations.
- **Controlling output verbosity:** Adjusting the verbosity of LLM responses can be beneficial for obtaining information at the desired level of detail. Control verbosity by explicitly requesting responses of a specific length or by asking the model to provide more or less detailed information.

In chapter 4, the Python code for the Streamlit UI tool and the prompts for developing, prompting and personalizing LLM-based chatbots are further described.

3.6 Framework for Assessment of Chatbot Quality (FACQ)

This thesis introduces the Framework for Assessment of Chatbot Quality (FACQ), a structured approach to quantitatively evaluate the personalization and effectiveness of LLM-based chatbot responses. FACQ is designed to assist developers when prompting and/or fine-tuning chatbots for more personalized user interactions. It stands on four core pillars, each targeting a crucial aspect of chatbot performance:

- **Similarity:** One fundamental aspect of chatbot quality revolves around how well its responses align with expected content. The similarity metrics allow the user to assess whether two different chatbot responses are textually and semantically aligned, or in other words "similar" in terms of content.
- **Readability:** Effective communication is vital for chatbot success. The readability pillar focuses on evaluating how well chatbot responses are understood by a diverse user base. This pillar recognizes the importance of ensuring that chatbot responses are accessible and comprehensible to users with varying levels of literacy and expertise.
- **Complexity:** The complexity metrics extends on the readability metrics to further quantify the intricacies of language use in chatbot responses. It examines the balance between simplicity and sophistication, with metrics that analyze how many complex words or complex sentence structures are present in the responses.

3. METHODOLOGY

- **Sentiment:** Effective communication often goes beyond conveying factual information; it involves conveying the appropriate emotional tone. The "sentiment" metric quantify the emotional and subjective aspects of chatbot responses. It takes into account the emotional nuances in user interactions.

FACQ is designed to offer an objective, data-driven approach to evaluate the effects of chatbot personalization. By comparing FACQ results with user ratings, presented in chapter 4, it will be possible to analyse whether the metrics effectively capture the nuances of personalization as perceived by actual users. While these four pillars serve as the foundation of FACQ, it is essential to acknowledge that no single framework can provide a complete picture of chatbot quality. FACQ does not claim to be exhaustive in its coverage but rather aims to offer a valuable set of metrics that address key dimensions of chatbot performance when experimenting with personalization techniques. The absence of additional metrics in FACQ does not diminish their importance. Rather, it recognizes that the field of chatbot evaluation is multifaceted, and future developments may lead to the inclusion of additional dimensions and metrics. Table 3.3 shows an overview of the FACQ metric range and their implications and below all metrics and their practical implementations are further explained.

3.6 Framework for Assessment of Chatbot Quality (FACQ)

Metric	Range	Implication
BLEU	0-1	Higher values indicate better agreement with the reference.
ROUGE	0-1	Higher values suggest better alignment with the reference.
SAS	Varies based on embeddings, but typically between 0-1	Higher values indicate greater semantic alignment.
ARI	Typically 1-14, but can be higher	Higher values imply more difficult text.
Flesch	0-100	Higher is easier to read. Lower values imply more difficult text.
Gunning Fog	Typically 6-20, but can be higher	Higher values imply more complex text.
Vocab Diversity	0-1	Closer to 1 means higher vocabulary diversity.
TextBlob Polarity	-1 to 1	-1 is very negative, 1 is very positive, 0 is neutral.
TextBlob Subjectivity	0-1	Closer to 1 is more subjective, closer to 0 is more objective.
BERT Sentiment	1-5	1 means a very negative sentiment, 3 means a neutral sentiment and 5 a very positive sentiment.

Table 3.3: Overview of the FACQ metric ranges and their implications

Similarity metrics

Similarity metrics primarily focus on textual and semantic similarity of two different chatbot responses that can be compared.

- **Bilingual Evaluation Understudy (BLEU):** Originally developed for machine translation, the BLEU score measures the overlap of n-grams (contiguous sequences of words) between two chatbot responses. In the code, BLEU is calculated using the `sentence_bleu` function from Python library NLTK. While BLEU provides a quick and interpretable measure of content similarity, its primary weakness is its lack of sensitivity to semantic nuances. It assigns low scores to responses that might contain the same message semantically, but use very different words to convey that message.

3. METHODOLOGY

- **Recall-Oriented Understudy for Gisting Evaluation (ROUGE):** ROUGE is a set of metrics used to assess content alignment, focusing on unigrams, bigrams, or longer sequences between two generated responses. Originally designed for automatic summarization, ROUGE provides a more detailed view of alignment compared to BLEU. In the code, ROUGE can be calculated using the Rouge class. Like BLEU, it is valuable for content evaluation but may miss semantic subtleties.
- **Semantic Answer Similarity (SAS):** The SAS score uses BERT embeddings to calculate the semantic similarity between two generated responses. Unlike BLEU and ROUGE, SAS is based on capturing the underlying meanings of words and sentences. The code for SAS calculates BERT embeddings for both the response and reference and then computes cosine similarity. This metric offers sensitivity to semantic nuances but might require more computational resources due to deep learning model embeddings.

Readability metrics

Readability metrics measure how accessible and understandable the chatbot response is to the user.

- **Automated Readability Index (ARI):** ARI calculates the readability of a text based on the number of characters per word and words per sentence. It provides an estimate of the U.S. grade level required to comprehend the text, where a higher level means a more difficult to read text. In the code, ARI is calculated using the `automated_readability_index` function, imported from the Python `textstat` library.
- **Flesch Reading Ease:** The Flesch Reading Ease score is based on the average sentence length and average word length in syllables. Texts with higher scores are easier to read. In the code, Flesch Reading Ease is calculated using the `flesch_reading_ease` function, imported from the Python `textstat` library.

While both ARI and the Flesch Reading Ease score are widely adopted, especially in the assessment of educational materials, their main limitation is that they may not capture all nuances of readability.

3.6 Framework for Assessment of Chatbot Quality (FACQ)

Complexity metrics

Complexity metrics measure the difficulty and diversity of the chatbot response vocabulary and phrasing.

- **Gunning Fog Index:** The Gunning Fog Index estimates the number of years of formal education required for readers to understand a text at first reading, taking into account word length and sentence complexity. The code can compute the Gunning Fog Index using the `gunning_fog` function (from `textstat`). While it offers a different perspective from ARI, it shares a similar limitation in not accounting for all linguistic nuances.
- **Vocabulary Diversity Score (VDS):** The VDS calculates the ratio of unique words to the total number of words in the text, indicating vocabulary richness. A higher score suggests richer content, but an excessively high score might imply verbosity or over-complexity.

Sentiment metrics

Sentiment metrics provide insights into the emotional tone and subjectivity of a chatbot response.

- **TextBlob Polarity and Subjectivity:** TextBlob calculates sentiment polarity (overall sentiment) and subjectivity (degree of personal opinion or emotion) in a text using a lexicon of words annotated with sentiment scores. The code computes TextBlob sentiment scores using the `TextBlob` Python library. While it offers a quick sentiment estimate, it may not capture context-dependent sentiment nuances effectively.
- **BERT Sentiment:** The code for BERT Sentiment uses a pre-trained BERT model for sentiment classification. Its strength lies in its ability to consider context, making it more robust for analyzing sentiment in complex sentences.

3. METHODOLOGY

4

Experiments

In this chapter, a series of experimental results are presented and discussed. To this end, the code created for the LLM-based chatbots and the FACQ metrics is further described, followed by experiments with personalization prompts and FACQ metric results to showcase the potential for LLM customization, specifically for language style and sentiment adaptation. The chapter also presents the results and analysis of a user survey, providing insights into how these personalization techniques are perceived by end-users and whether this corresponds with the quantitative results provided by the FACQ metrics.

4.1 Streamlit UI & code

This thesis introduces a user interface (UI) tool designed for testing and analyzing LLMs in the context of chatbot applications. It focuses on developers, especially those engaged in customer support chatbots, providing a platform to test and analyze various LLMs such as GPT-3.5, GPT-4, LLaMA 70B and LLaMA 13B chat.

The UI tool is set up with Python and Streamlit, as can be seen in figure 4.1 with key elements:

- **llm_bot.py (main application script):** The core script of the application. Sets up the Streamlit user interface and controls navigation between functionalities. Allows for testing out different prompts with different LLMs and facilitates interaction with different LLMs through OpenAI and Replicate APIs.
- **document_url_bot.py (document and URL interaction module):** Handles functionality related to document and URL processing and querying. Allows users to upload documents (CSV, TXT, PDF) or provide URLs for content extraction and

4. EXPERIMENTS

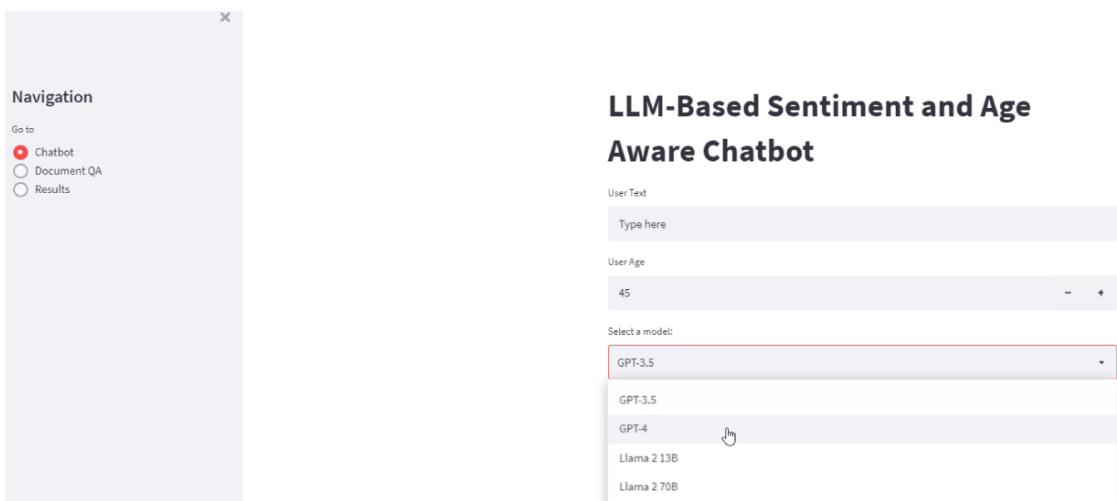


Figure 4.1: Overview of the UI tool for LLM-chatbot personalization prompting and testing

analysis. Uses OCR (Optical Character Recognition) for document querying and the web-scraping Python package Trafilatura for URL querying.

- **results_page.py (results analysis module):** Dedicated to the analysis of chatbot interactions stored in session data. Imports text_metrics to calculate NLP metric values for the stored responses.
- **text_metrics.py (Text metrics calculation module):** Contains functions for computing various text analysis metrics, which are the similarity, readability, complexity and sentiment metrics as presented in FACQ.

The tool serves as a comprehensive platform for testing, analyzing, and refining chatbot interactions, aiming to enhance the efficiency and relevance of personalized chatbot responses. The source code for this tool is extensively commented, hosted on GitHub and accessible for review, collaboration and further development on request.

4.2 Personalization prompts & metrics experiments

With prompt-based learning, it is possible to write prompts where the chatbot is provided with specific instructions to be a question-answering assistant. Figure 4.2 shows an example of how personalization prompts can be created. In this case, the LLM is prompted with language style adaptation based on user age, to adjust the language and complexity of the answer in such a way that it is age-appropriate and easily understandable for the user.

4.2 Personalization prompts & metrics experiments

A: Prompt for language style adaptation based on user age

```
{  
  role: "system",  
  content: `You are a cheerful and bubbly question-answering assistant. Given that the user's age is ${scenario.age}, please adjust your language and complexity accordingly to ensure the information is age-appropriate and easily understandable for them. Please provide a relevant, concise and complete answer, in maximum 3 sentences, that ends with a full stop`,  
}
```

B: Chatbot responses for two scenario's with user age 10 and 35

Scenario	Response	Cost
age: 10	Accenture is a really big company that helps other companies solve problems and get better at what they do. They use technology, like computers and software, to make things easier and more efficient. It's kind of like a superhero team for businesses, helping them to be the best they can be!	\$ 0.007 (7.68s)
age: 35	Accenture is a global professional services firm that specializes in IT services and consulting. The company provides a broad range of services and solutions in strategy, consulting, digital, technology and operations. Founded in 1989, it's headquartered in Dublin, Ireland, and works with clients in over 120 countries.	\$ 0.007 (11.73s)

Figure 4.2: Example of personalization prompting in zero-shot learning LLM-based chatbots

It can clearly be noticed that the language use in the top scenario in B, where the user age is 10, is much simpler and easy to understand when compared to the bottom scenario where the user age is 35. It goes to show that personalization prompting can be very efficient, with the LLM generating natural language that matches the prompt, compared to the process of adding hand-written conversation rules for more traditional rule-based chatbot systems. For the two main personalization techniques in this thesis, language style and sentiment adaptation, the personalization prompts can be constructed as follows:

- You are a question-answering assistant. Given that the user's age is `age_variable`, adjust your language and complexity accordingly to ensure the information is age-appropriate and easily understandable for the user. Provide a relevant, concise and complete answer, in maximum 3 sentences.
- You are a question-answering assistant. Sense the sentiment embedded in the user's question and provide a thoughtful and considerate response. Provide a relevant and complete answer, in maximum 4 sentences.

4. EXPERIMENTS

As presented before in chapter 3.5, these prompts follow from best practices researched by prompt engineers [88] and could be further optimized by chatbot developers for specific use cases. In this thesis, the purpose is to show the possibilities of personalization prompting. In customer service, these prompts can be extended to include instructions on the tone of voice that matches the organization or company, or adjust the chatbot language style based on other variables that are known about the users.

With the text metrics calculation module in the code, LLM outputs can be analysed and the FACQ metrics can be calculated. Figure 4.3 shows the results for the FACQ metrics for five examples of personalization prompting.

User question	LLM model	LLM prompt	Readability	Complexity	Sentiment	Parameters
What should I know about Accenture?	GPT-4	Cheerful and bubbly	ARI: 14.60 Flesch: 38.01	GF: 15.50 VD: 0.82	P: 0.07 S: 0.19 Bert: 4.00	Tokens: 75 Temp: 0.7
What should I know about Accenture?	GPT-4	Agitated	ARI: 14.00 Flesch: 33.92	GF: 15.91 VD: 0.89	P: 0.36 S: -0.42 Bert: 2.00	Tokens: 75 Temp: 0.7
What is Generative AI?	GPT-4	Age-appropriate language	ARI: 6.40 Flesch: 74.90	GF: 8.32 VD: 0.93	P: 0.46 S: 0.83 Bert: 4.00	Tokens: 75 Temp: 0.7 User age: 10
What is Generative AI?	GPT-4	Age-appropriate language	ARI: 11.90 Flesch: 44.75	GF: 16.09 VD: 0.85	P: -0.03 S: 0.54 Bert: 3.00	Tokens: 75 Temp: 0.7 User age: 28
What is Generative AI?	GPT-4	Age-appropriate language	ARI: 7.80 Flesch: 67.04	GF: 9.29 VD: 0.92	P: 0.10 S: 0.77 Bert: 4.00	Tokens: 75 Temp: 0.7 User age: 72

Figure 4.3: Example of FACQ results with personalization prompting in a zero-shot learning GPT-4 based chatbot

Since LLM-based chatbots are examples of generative AI, it is important to mention that using the same prompt results in slight variations of the output response. This can be altered by changing the `temperature` setting in the LLMs, where a temperature of 0 means that the model is deterministic, meaning that the highest probability response is always selected. Higher temperatures, e.g. 0.7 to 1, can lead to more diverse or creative responses. In testing a LLM-based chatbot for a customer support scenario, developers will need to make design choices to decide on what level of "creativity" aligns with the project goals. However, it can clearly be seen for these metric examples that the FACQ metrics values change with different sentiment or language style adaptation prompts. Changing the sentiment in the prompt to agitated instead

4.2 Personalization prompts & metrics experiments

of cheerful is reflected in the decrease in TextBlob (0.19 to -0.42) and BERT sentiment score (4 to 2). Personalization prompting language use to age-appropriate language results in readability and complexity scores that reflect better readability for when the user age is 10, compared to when the user age is 28. Further analysis of how the FACQ metric values change for question answers and how this is perceived by actual users, is presented after elaborating on the user survey.

As a last example showcasing the possibilities from chatbot personalization and analysis with the UI tool, an example for the URL interaction module will be shown. LLM-based chatbots can be extended with RAG, as mentioned in chapter 3.3.2. In the UI tool code, a form of RAG is implemented with the document and URL interaction module, which enables for information retrieval from websites or documents. This information is then combined with the prompt which is sent to the LLM. Figure 4.4 shows a three-step example where it can be seen that this document and URL interaction module enables testing of LLM answers based on information that was not available to the LLM during pre-training. This module is added to the UI tool such that it could be used by chatbot developers to analyse and personalize LLM-based chatbots that need access to specific data, for example to answer company related questions.

4. EXPERIMENTS

A: LLM-based chatbot cannot give the information without access to current data

question
Who is the client account lead Europe for Accenture?

I'm sorry, but as an AI, I don't have real-time access to Accenture's internal data or proprietary databases to provide current or specific personnel information, including the position of Client Account Lead for Europe. I recommend visiting Accenture's official website or getting in touch with them directly for the most accurate information.

\$ 0.009 ⏱ 6.58s

B: Provide URL with information and it is retrieved and processed correctly by the LLM

Ask your questions to documents or Websites (csv, txt, pdf, URL)

Choose your mode:

- Upload Document
 Provide URL

Enter the URL:

<https://www.accenture.com/lv-en/about/leadership/leadership-index>

Content fetched successfully!

Enter your question here:

Who is the client account lead Europe for Accenture?

Submit

Nicole van Det

C: Double check with the URL; the information is correct!

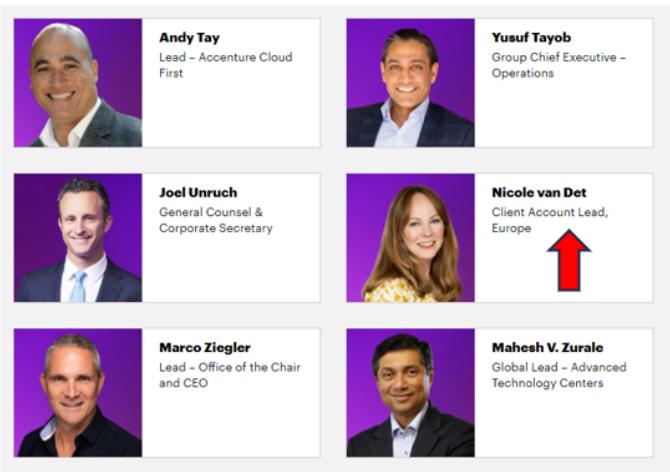


Figure 4.4: Example of augmenting an LLM-based chatbot with URL processing and querying functionality

4.3 User Survey

In order to test the effects of personalization in zero-shot learning LLM-based chatbots, participants were presented with a random selection of questions one might pose to an AI chatbot in a real-life situation, along with two responses generated by a GPT-4 LLM. For each pair of responses, participants were instructed to rate the answers based on similarity of the content, readability, complexity and sentiment or trust. These criteria were chosen to match the NLP metrics within FACQ. The answers were generated to test different variations of the metric values, without the user being informed on how the answers would rank in terms of metric values. Participation in the survey was on a voluntary basis and the survey has been designed to take no longer than ten minutes, to encourage thorough yet manageable involvement from the participants. In order to simulate the "feeling" of communicating with a chatbot, the chatbot answers were shown in an animation of the LLM typing the answer. Anonymity in data collection was strictly maintained to ensure that the information gathered could not be linked back to any individual. This process was in compliance with the ethical guidelines of Eindhoven University of Technology's ethical review committee, which had approved the study. A total of 411 participants completed the survey after providing consent to anonymous data collection and answering a test question correctly. The complete set of survey questions can be found in the Appendix.

4.4 Demographic

Before looking at the metric and survey results, it is important to analyse the survey demographic, to get an understanding of how age, gender, education level and technology experience level vary within the studied survey participants.

Age distribution (Q3): In figure 4.5 it can be seen that the younger age groups predominate, with 18 - 24 years old (178 respondents, 43.3%) and 25 - 34 years old (117 respondents, 28.5%) being the most represented. The oldest age groups, 65 - 74 and 75+ years old, have the least representation with 2 respondents each (0.5%).

Gender distribution (Q4): Figure 4.6 shows that males form the majority of the survey participants (243 respondents, 59.1%), followed by females (159 respondents, 38.7%). Nonbinary individuals and those who prefer not to disclose their gender are the least represented, with 6 (1.5%) and 3 (0.7%) respondents, respectively.

4. EXPERIMENTS

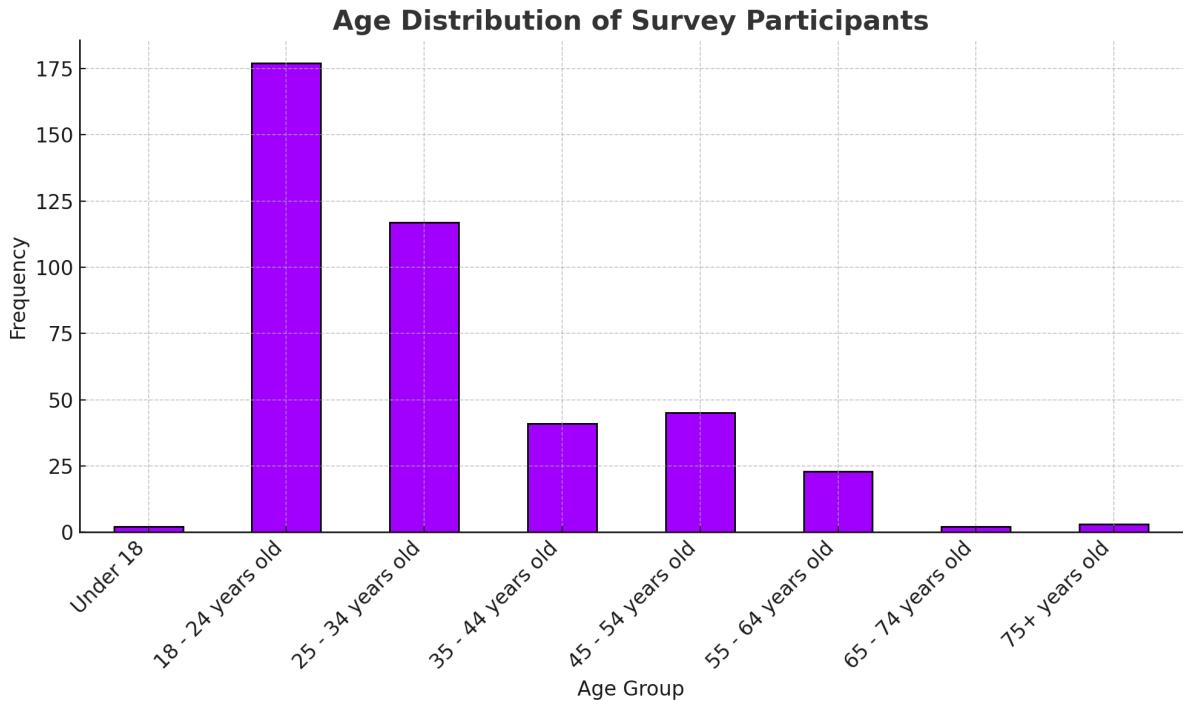


Figure 4.5: Age distribution of survey participants

Gender Distribution of Survey Participants

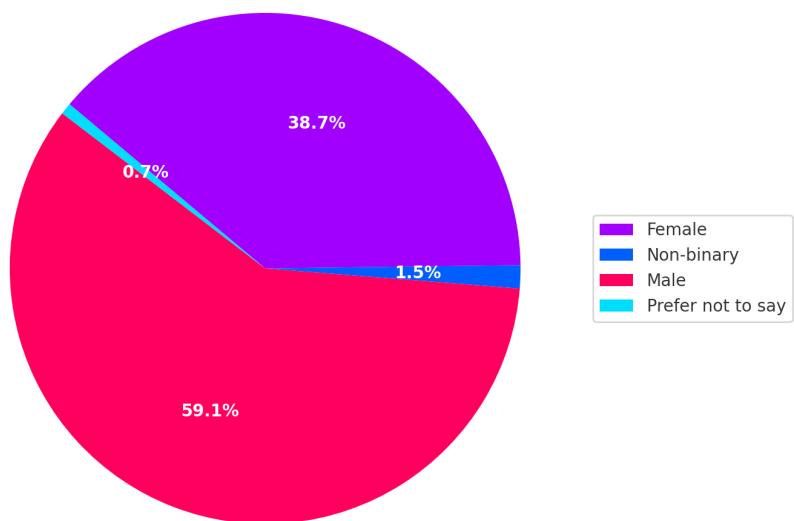


Figure 4.6: Gender distribution of survey participants

4.4 Demographic

Education level distribution (Q5): As seen in figure 6.1, respondents mostly have higher education levels, with the largest group holding graduate or professional degrees (125 respondents, 30.4%). University bachelor's degree holders (104 respondents, 25.3%) and those with higher vocational education (85 respondents, 20.7%) follow.

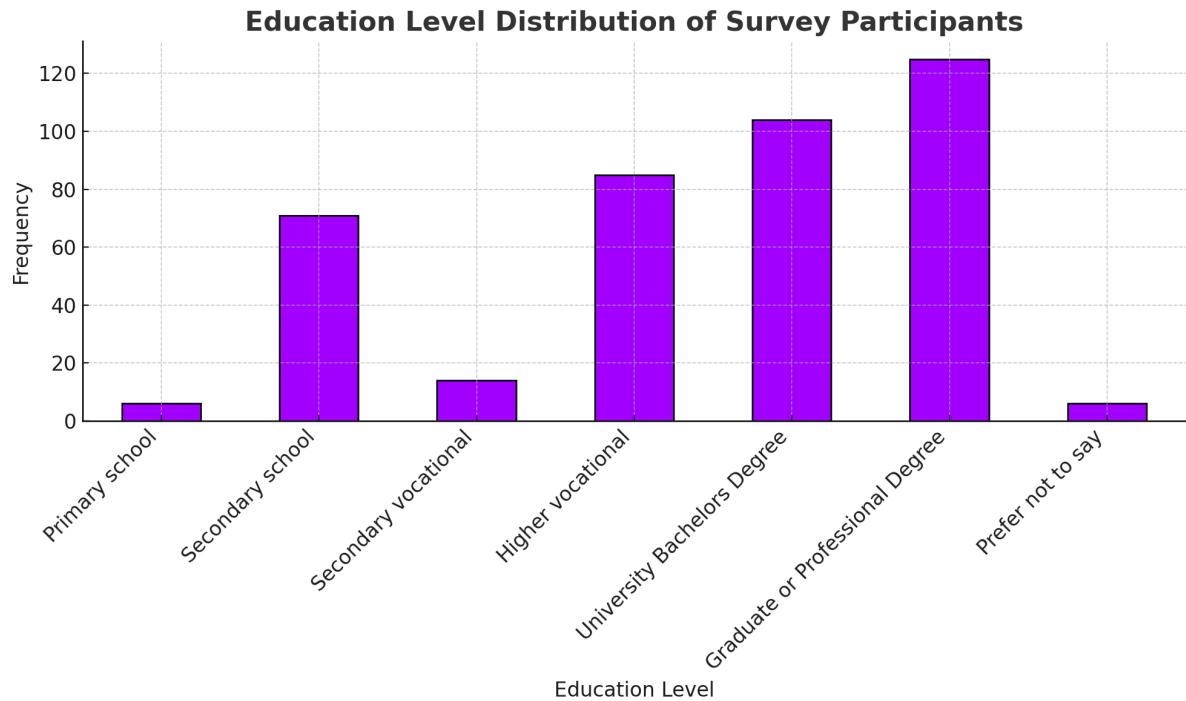


Figure 4.7: Education level of survey participants

Comfort and Experience Level with Technology (Q6): As seen in figure 4.8, a significant majority of participants rate themselves as 'Advanced' in technology experience (246 respondents, 59.9%), reflecting a high level of comfort with digital tools. Experts in technology form a notable segment (103 respondents, 25.1%), while intermediate users (59 respondents, 14.4%) and beginners (3 respondents, 0.7%) are less common.

4. EXPERIMENTS

Technology Experience Level of Survey Participants

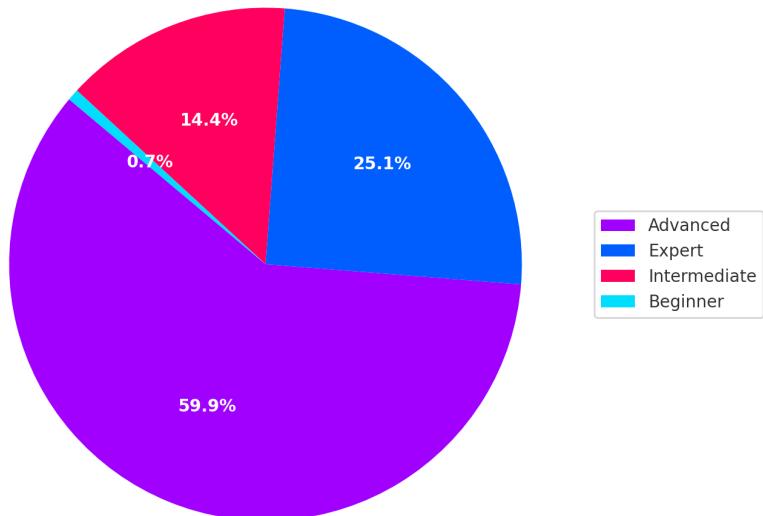


Figure 4.8: Technology experience level of survey participants

4.5 Language adaptation

In the first part of the user survey, participants were presented questions asked to an LLM-based chatbot and provided with two responses of the chatbot. The answers to the questions vary in complexity and readability metric values, with the goal of analysing whether participants notice these differences in practice as well. The questions that were constructed to test variations of FACQ complexity and readability metrics are as follows:

Q8: How do astronauts live in space?

1. Astronauts live in space stations like the ISS, where they dwell in small spaces, eat prepackaged food, and exercise to maintain their physical health. They perform science experiments, maintain the space station, and communicate with family and friends via email or phone.
2. Astronauts in space live in a space station where they sleep in sleeping bags attached to walls, eat with their utensils, and exercise to stay fit. They work to maintain the space station and communicate with their families and friends.

4.5 Language adaptation

Q9: What does Accenture do?

1. Accenture is a global professional services company offering services in strategy, consulting, technology, and operations, helping clients improve their performance and create sustainable value. They drive innovation and growth with their expertise and advanced technology.
2. Accenture is a global professional services company offering consulting, technology, and operations services to help clients improve performance and create value for stakeholders.

Q10: How does the stock market work?

1. The stock market is a supermarket for buying and selling company stocks. Companies sell stocks to raise money. Stock prices fluctuate based on company performance and future projections.
2. The stock market allows companies to sell stocks to raise money. Owning a stock means owning a part of the company, with the potential to profit if the company does well.

Q11: What is Generative AI?

1. Generative AI creates new data models from existing data sets, generating images, text, or music by learning patterns from input data.
2. Generative AI is designed to create content like text, images, or music, learning from data to produce new, original content.

Q12: What is the purpose of the United Nations?

1. the United Nations (UN) is there to keep peace throughout the world. It also aims to develop friendly relationships among nations and help nations work together to improve life for poor people, conquer hunger and disease, and encourage respect for each other's rights and freedoms.
2. the United Nations (UN) is an organization created to promote peace and cooperation among all countries. It helps solve problems like poverty, disease, and conflicts between nations. The UN also works to protect human rights and improve living conditions for people around the world.

4. EXPERIMENTS

Q13: What causes the Northern Lights?

1. The Aurora Borealis emerges from the interaction of solar wind particles with Earth's magnetosphere, inducing ionization and excitation of atmospheric constituents, which consequently radiate visible luminescence.
2. The Northern Lights are created when solar particles from the sun hit Earth's magnetic field and interact with gases in our atmosphere, making colorful lights in the sky.

The FACQ metrics can be presented as radar plots, a fitting type of visualization to display multivariate data. Each plot represents a question and the blue and red shapes in the plot represent the similarity, readability and complexity FACQ metrics for the chatbot answers to the question.

Presenting the metric values effectively and accurately in these radar plots requires post-processing techniques. Comparing metrics in their original scales in a radar plot provides for an accurate representation, but makes interpretation challenging. Diverse scales on each axis hinder direct comparison and can confuse readers unfamiliar with each metric's specifics. While accurate, this approach lacks the clarity and simplicity needed for effective communication in this thesis. Therefore, standardizing scales, such as through normalization or standardization, is in this case preferred due to the ease of interpretation and comparative analysis. Standardization (converting data to z-scores) maintains the relative spread and distribution of the original data, which can be more informative for metrics with small absolute differences but potentially significant relative differences. Normalization scales the data to a range between 0 and 1 based on the minimum and maximum values for each metric across all answers and is useful for comparing metrics on different scales. However, a notable downside of normalization is its sensitivity to outliers, as extreme values can significantly skew the normalized results. In the end, the choice for normalization is made as it offers key advantages for radar plots especially in uniform scaling, intuitive interpretation and handling various data types without requiring a specific distribution, unlike the Gaussian distribution assumption for standardization.

For the radar plots shown in figure 4.9, these metrics are normalized to scale the values between 0 (minimum observed value for each metric) and 1 (maximum observed value). For ARI, VDS and Gunning Fog, the metrics are inverted post-normalization. This inversion is performed so that for all metrics in the plot, a higher value represents a text that is easier to understand, aligning them with metrics like the Flesch score where a higher number signifies more readable text.

4.5 Language adaptation

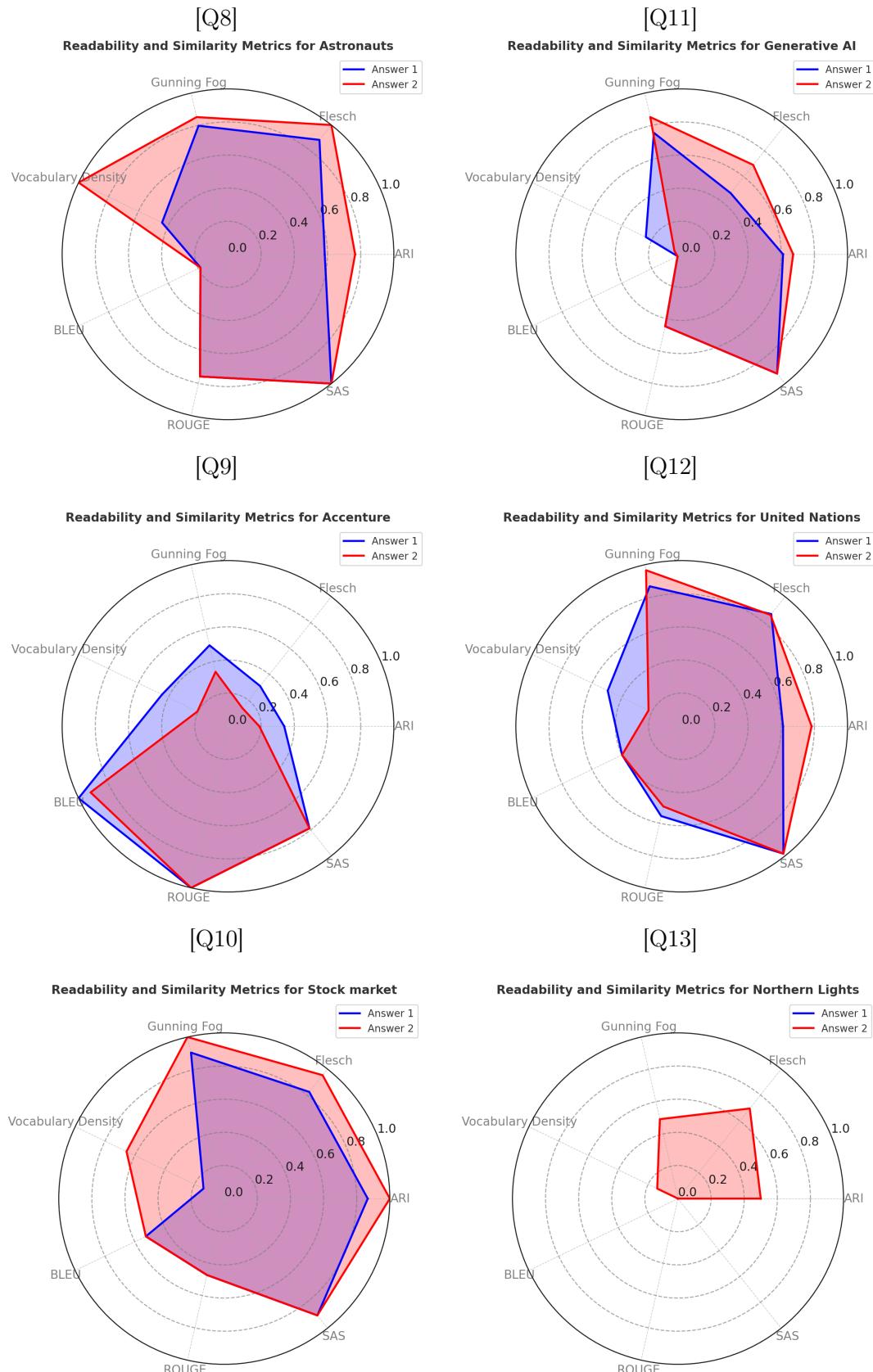


Figure 4.9: Radar plots for similarity, readability and complexity metrics of Q8-Q13

4. EXPERIMENTS

These plots visually represent the normalized readability and similarity metrics (Flesch, ARI, Gunning Fog, VDS, BLEU, ROUGE and SAS) for two different answers per question. Each plot illustrates the comparative analysis of these metrics between response 1 and response 2 for a specific question. Where the coloured area of the radar plot is larger, this indicates higher readability, complexity and/or similarity scores.

To start the analysis with the similarity metrics, for SAS, if both responses have very similar meanings or content, the embeddings generated by the BERT model will be close in the vector space, leading to high similarity scores. For ROUGE and BLEU, the score is high if the answers have a lot of overlap in literal words and phrases. For the three similarity metrics BLEU, ROUGE and SAS, generally in the plots it is the case that the values are exactly the same or very similar for the two chatbot answers to one question. In the set-up of this experiment, this was to be expected, since the answers to the different questions are generated using prompts engineered to focus on language adaptation, and not content adaptation. The slight variations are likely caused by the directionality of the evaluation of BLEU and ROUGE: if answer 1 has phrases or n-grams that are well-represented in answer 2, the similarity score might be higher when answer 2 is the reference answer, and vice versa.

When looking at the readability metrics ARI and Flesch and complexity metrics Gunning Fog and VDS, a few interesting remarks can be made. Q8 and Q10 are clearly showing very high readability and complexity scores, and in the answers to the questions it can also be recognized that the language use in these answers is relatively simple, for example in terms of sentence-length and word complexity. When comparing the radar plot of Q9 to the others, it can clearly be seen that the complexity and readability metrics have relatively low score, matching the prompt that was used for these answers to use high-level language for adults in a professional setting.

For Q13, it can be seen that the coloured area for answer 1 is not shown. This can be explained by the fact that for this answer, the metric scores were the lowest for all seven metrics shown in the plot, which are set to 0 as a consequence of normalization. This indicates that this answer clearly scored the lowest in terms of similarity with the other answer, readability and complexity. This was expected with how these answer texts were obtained, with the GPT-4 LLM prompted to answer the question in very complicated language (answer 1) versus very understandable language (answer 2). For completeness, the actual values of all FACQ metric scores are provided in the Appendix.

These questions and answers were presented to the survey participants, who were asked to compare and rank these answers in terms of similarity, readability, complexity and trust in the correctness of the answer. The series of histograms in figure 4.10 present an overview of survey participant responses to Q8-Q13.

4.5 Language adaptation

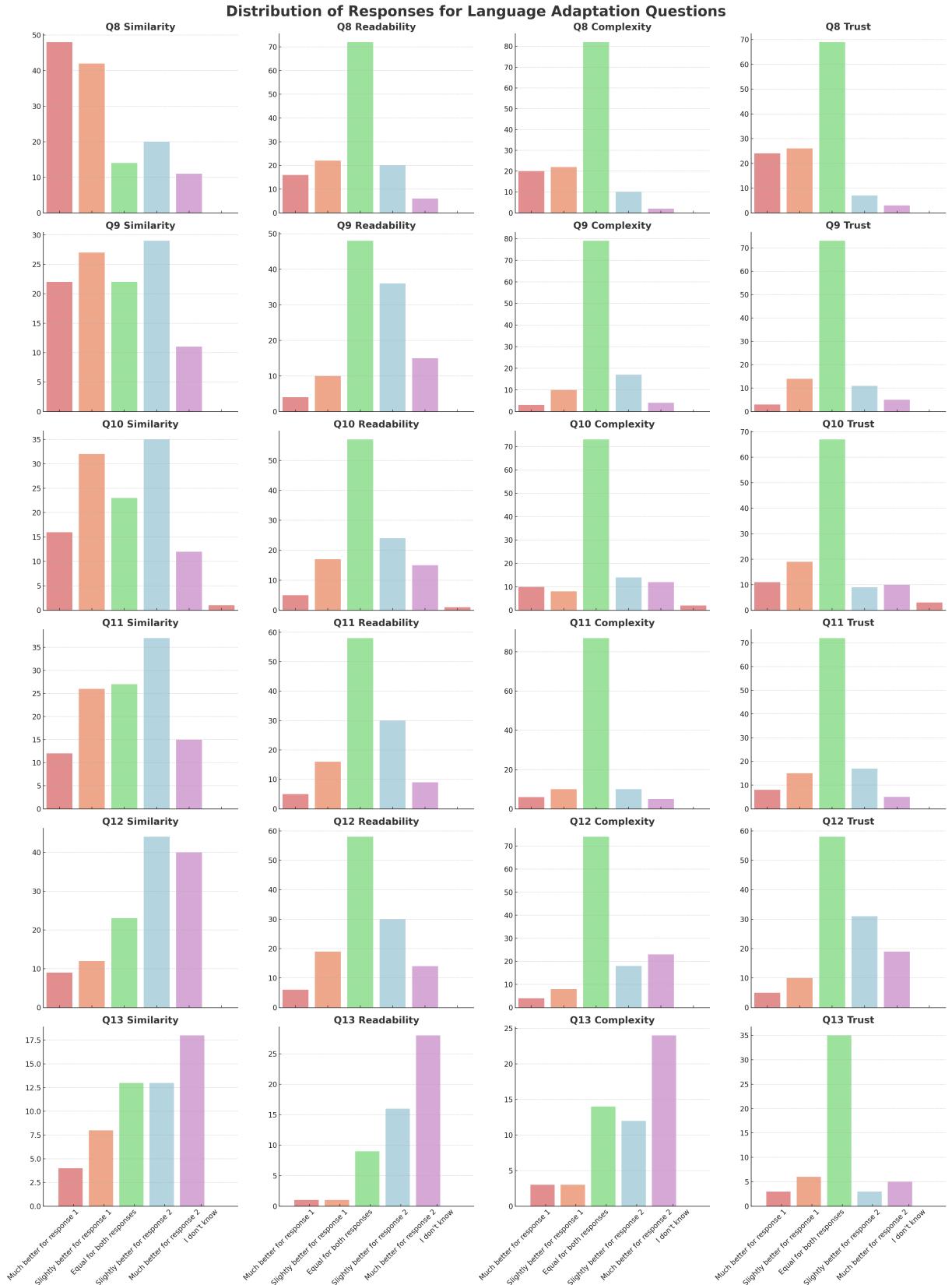


Figure 4.10: Distribution of Responses for Subquestions on Language Adaptation

4. EXPERIMENTS

In order to analyse these histograms statistically, a Chi-square test for independence is performed. The Chi-square test has several key assumptions that must be met for the results to be valid. Firstly, the data should be in the form of frequencies or counts of cases. Secondly, the samples must be independent, meaning that the selection of one individual does not influence the selection of another. Thirdly, the categories should be mutually exclusive, ensuring that each respondent can only choose one category. All these criteria are met for the data of this survey, since participant responses are recorded individually and the categories are well-defined and mutually exclusive. The Chi-square test has been performed to compare the frequency of preference for 'Response 1' (both categories 'Slightly better for response 1' and 'Much better for response 1') against preference for 'Response 2' (both categories 'Slightly better for response 2' and 'Much better for response 2') for each subquestion. The null hypothesis is that there is no association between the answers provided (response 1 and response 2) and the preferences expressed by the participants. If the null hypothesis is rejected, this means that the preference for one answer over the other is not simply due to chance.

Question	Similarity	Readability	Complexity	Trust
	p-value	p-value	p-value	p-value
Q8	< 0.001	0.052	< 0.001	< 0.001
Q9	0.230	< 0.001	0.090	1.000
Q10	1.000	0.004	0.136	0.143
Q11	0.053	0.002	1.000	1.000
Q12	< 0.001	0.002	< 0.001	< 0.001
Q13	< 0.001	< 0.001	< 0.001	1.000

Table 4.1: Chi-square test results for language style adaptation questions

Analyzing the significant and non-significant results from the survey provides a nuanced understanding of how the participants perceived the chatbot's responses in terms of similarity, readability, complexity and trust. The results shown in the radar plots in figure 4.9, the histograms in figure 4.10 and the Chi-square test results from table 4.1 are combinedly analysed.

In terms of similarity, the survey participants are asked to rate whether they think one answer is clearly a better answer to the question, indicating that they see a difference in similarity of the answers. It is interesting to notice that for Q13, a significant difference between the preference for answer 1 and 2 is shown in the Chi-square p-value. From the histogram, it can clearly be seen that the participants' preference is for answer 2, which also showed a very clear difference

4.5 Language adaptation

in similarity metrics in the radar plot. For Q12, also a significant p-value and preference for answer 2 can be seen, with the radar chart showing lower values for BLEU and ROUGE, but high SAS values. This could be explained by looking at the text for the two answers, which clearly use different words and phrases, likely lowering BLEU and ROUGE, while it can be argued that the two answers are clearly semantically similar. The fact that survey respondents indicate a preference for response 2 could be caused by specific words or phrases that are used in the answer, which might appeal more to a significant majority. This immediately shows that the NLP metrics in FACQ do not specify **why** people would have a preference for one chatbot answer over the other. It can however, assist chatbot conversation developers in analysing the differences between chatbot answers for many different variations of prompts in one glance.

For Q8 (Astronauts) and Q10 (Stock Market), the analysis of the metrics shows high readability and low complexity for both answers. In the survey responses, this is reflected with a non-significant preference for specifically answer 1 or 2. This likely implies that both responses were comparably easy to read, and any differences in readability were subtle enough that they did not significantly influence the participants' perceptions. The high VDS for Q8 does stand out and shows the main contradiction for this NLP metric. It can provide a high score even though the other readability and complexity metrics indicate an easy-to-understand text, if for example the chatbot answer is short, the VDS is more likely to be high since there is likely less repetition of words in the answer.

For Q9 (Accenture), from the answer texts it can be noticed that the answers are more technical and professional in terms of language, compared to the other questions. This matches the lower FACQ scores for the readability and complexity metrics presented in the radar plots. However, even though the radar plot shows a slight difference in these metrics between answers 1 and 2, the survey respondents do not have a significant preference for answer 1 or 2 in terms of readability (p-value) or similarity. Two possible reasons for this could be the subtlety of the metric difference and/or the education-level distribution of the survey participants, who are on average very highly educated as mentioned in section 4.3. This might result in participants not rating answer 1 as more readable or less complex, which would be in accordance with the metric values in the radar plot.

Readability and complexity for Q11 and Q12 also show some significant results in preference for answer 2, consistent with the higher readability and complexity values. The VDS metric again stands out, suggesting it may not align well with the other metrics in certain contexts. The survey and radar plot results for Q13 demonstrate that participants are adept at identifying large differences in readability and complexity metrics. This implies that when chatbot responses vary significantly in these aspects, users can easily distinguish and express a preference.

4. EXPERIMENTS

A more detailed approach could involve examining the actual metric values for each chatbot response and attempting to find correlations between their differences in value and the significant preferences of the survey participants. Such an approach might lead to overemphasizing minor differences in metric scores, which may not translate into significant differences in user experience or preference. The primary purpose of these metrics is to serve as a guide for chatbot conversation developers, providing a snapshot of how answers score on the similarity, readability and complexity dimensions. They are designed to provide a quick and general understanding of the chatbot's performance, rather than to facilitate a granular comparison of exact differences between responses.

Generally, there are quite some interesting findings. When differences in chatbot responses are marginal, such as when both responses are highly readable, survey participants often do not demonstrate a significant preference based on readability alone. This observation suggests that factors beyond basic readability metrics may be influencing user preferences and perceptions. The NLP metrics used in the FACQ system, while offering valuable insights into the textual characteristics of chatbot responses, do not fully encapsulate the reasons behind user preferences. They serve more as indicative tools rather than definitive guides to user experience.

Specifically, the Vocabulary Diversity Score (VDS), which occasionally present contradictions with other readability and complexity metrics, is not the most useful metrics in indicating complexity, when it also provides high scores for short sentences where every unique word only occurs once. It could be altered to take calculate the diversity score for "complex" words in a sentence, based on a chosen definition of a complex word. Moreover, BLEU and ROUGE scoring relatively low in terms of similarity, also when participants perceive similar content between answers indicates a gap in these metrics' ability to capture the nuanced semantic relationships in language. The similarity between chatbot responses, as judged by users, involves more than just the overlap of words and phrases; it encompasses a deeper semantic understanding that current metrics might not fully capture. The Gunning Fog, Flesch, ARI and SAS scores show frequent matches with the user perceptions, especially when differences between the metrics for the two answers are relatively larger, like in Q11 and Q13. Therefore, it could be argued that these metrics can be used to effectively measure the effects of personalization for readability, complexity and similarity scores.

4.6 Sentiment adaptation

In this subchapter, the focus is on the sentiment adaptation in chatbot responses as part of the survey. The analysis of the results centers around how users perceive and react to the emotional

4.6 Sentiment adaptation

content and tone in the answers provided by a LLM-based chatbot.

A set of questions in the user survey, Q14 to Q17, were designed to elicit responses from the chatbot in various scenarios, ranging from positive feedback (Q14), frustration (Q15), joy (Q16), to dissatisfaction with service (Q17). The responses to these questions were prompted to have slight or larger variations in sentiment and tone, to test user reactions and preferences.

Q14: You've been very helpful! How can I provide feedback about your great service?

1. You can provide feedback by clicking on the "feedback" button located at the bottom of the page.
2. I'm thrilled to hear you found my assistance helpful! You can leave feedback by pressing the "feedback" button at the bottom of this page. Thank you for your kind words!

Q15: Why can't a human being get basic customer support anymore these days??

1. It's absolutely infuriating, isn't it? Unfortunately, automation is a common cost-saving measure these days. We're here to help though, so let's sort out your issue.
2. I'm really sorry to hear that you're feeling frustrated. I'm here to give you the support you need. Can you please tell me more about your issue so I can assist you better?

Q16: I just got a promotion! Can you suggest some ways to celebrate?

1. That's fantastic news! Consider treating yourself to a nice dinner, taking a mini vacation, or having a small get-together with friends or family. The best celebration is one that makes you happiest.
2. Congratulations on your promotion! You could throw a party with friends and family, treat yourself to a special meal, or even take a small vacation to celebrate your achievement. Enjoy your well-deserved success!

Q17: I am disappointed with the service, how are you going to solve this?

1. I'm really sorry to hear you're not satisfied with the service. I'm here to help fix any issues you've encountered. Could you please provide more details so I can assist you in the best possible way?
2. I'm really sorry to hear that you're not satisfied with the service. Could you please let me know what specifically went wrong? I'm here to help and will do my best to resolve your issue.

4. EXPERIMENTS

BLEU	ROUGE	SAS	ARI	Flesch	GF	TB Sent	TB Subj	BERT	Question
0.120	0.512	0.988	9.3	62.68	9.15	0.000	0.000	3	Q14 1
0.140	0.512	0.988	5.6	86.71	5.33	0.750	0.800	4	Q14 2
0.014	0.151	0.994	9.3	46.13	9.72	-0.467	0.767	1	Q15 1
0.014	0.151	0.994	4	85.69	5.61	-0.050	0.550	2	Q15 2
0.049	0.241	0.982	8.6	52.15	10.53	0.463	0.650	4	Q16 1
0.049	0.241	0.982	10.5	60.31	9.25	0.276	0.474	4	Q16 2
0.326	0.531	0.998	6	76.22	7.02	0.150	0.760	1	Q17 1
0.326	0.531	0.998	5	84.98	6.97	-0.063	0.800	1	Q17 2

Table 4.2: FACQ metric values for sentiment questions

The calculated FACQ metrics for these survey questions are shown in table 4.2.

From the table, it can be seen that for all answers to all questions, the readability is high and complexity is low, with low ARI, high Flesch and low Gunning Fog scores. This matches the goal of analysing sentiment awareness, without distracting the survey participant with complex answer texts. For Q14, the TextBlob Sentiment (TB Sent) and Subjectivity (TB Subj) scores show a stark contrast between the two answers. The first answer has a neutral sentiment (0.000) and no subjectivity (0.000), reflecting a straightforward, fact-based response. In contrast, the second answer exhibits positive sentiment (0.750) and high subjectivity (0.800), which aligns with its more enthusiastic and personal tone. In Q15, both responses show a frustrated sentiment. The first response, with a negative TextBlob sentiment score of -0.467 and BERT score of 1, contains more negative sentiment compared to the second response with negative TextBlob sentiment score of -0.050 and BERT score of 2. For Q16, both responses are positive, celebrating a promotion, but with varying degrees of expressiveness. The TB Sent scores for the first (0.463) and second (0.276) responses show positive sentiment, with the first being more expressive. The TB Subj scores also reflect a higher degree of personal opinion and emotion in the first response (0.650) compared to the second (0.474). Both responses have the same BERT sentiment score (4). In Q17, both responses address dissatisfaction with the service. The TB Sent scores are slightly positive for the first response (0.150) and slightly negative for the second (-0.063), indicating subtle differences in how each response approaches the issue. Both responses receive the same BERT sentiment score (1), suggesting a similar level of sentiment despite the slight differences in TB Sent scores.

The histograms visualized in figure 4.11 show the distributions of responses for the sentiment questions Q14-Q17 by the survey participants.

4.6 Sentiment adaptation

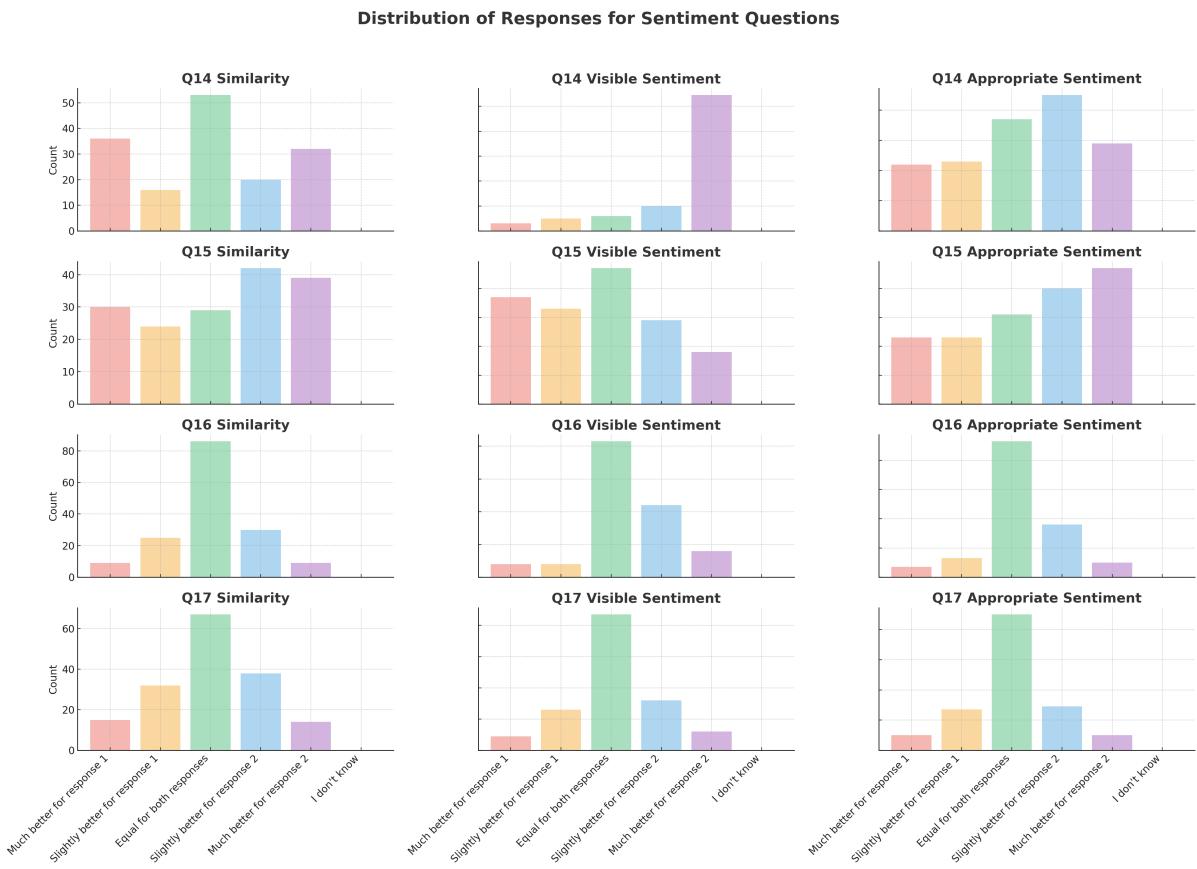


Figure 4.11: Distributions of Responses for Subquestions on Sentiment

The Chi-square test has been performed to compare the frequency of preference for 'Response 1' (both categories 'Slightly better for response 1' and 'Much better for response 1') against preference for 'Response 2' (both categories 'Slightly better for response 2' and 'Much better for response 2') for each subquestion. The null hypothesis is that there is no association between the answers provided (response 1 and response 2) and the preferences expressed by the participants. If the null hypothesis is rejected, this means that the preference for one answer over the other is not simply due to chance.

From table 4.3, it is important to recognize that while sentiment analysis tools provide valuable insights, they do not completely capture the full spectrum of human emotional response. The interpretation of sentiment is subjective, and users may perceive the same response differently based on their individual contexts and experiences.

The significant p-value for visible sentiment (<0.001) and appropriate sentiment (0.039) in Q14, combined with the histogram information, shows a significant preference for response 2

4. EXPERIMENTS

Question	Content	Visible Sentiment	Appropriate Sentiment
	p-value	p-value	p-value
Q14	1.0	< 0.001	0.039
Q15	0.073	0.104	0.005
Q16	0.774	< 0.001	0.013
Q17	0.774	0.490	0.963

Table 4.3: Chi-square test results for sentiment questions

of Q14. When looking at the metrics, the first response records very neutral values, with a TextBlob sentiment and subjectivity score of 0 and BERT score of 3. The second response has a much higher TextBlob sentiment and subjectivity score of 0.750 and 0.800 and a BERT score of 4. Since the sentiment in the user question is clearly positive, it can be analysed that users find a more positive sentiment appropriate in the chatbot response.

The significant p-value for appropriate sentiment (0.005) but not for visible sentiment (0.104) in Q15 implies that while users recognize the emotional content in both responses, they significantly prefer one that is more appropriately aligned with their emotional state. This is likely connected to the more balanced approach in the second response, as opposed to the over-the-top sentiment present in the first response ("It's absolutely infuriating, isn't it?") This also shows from the metrics, indicated by the less negative TextBlob sentiment score (-0.050) and higher BERT sentiment score (2), compared to the first response.

The lack of a significant difference in content preference (p-value=0.774) but significant in visible (p-value < 0.001) and appropriate sentiment (0.013) for Q16 suggests that while both responses are positively perceived, the degree of expressiveness and appropriateness of the sentiment plays a crucial role. This correlates with the more expressive first response having a higher TextBlob sentiment score, yet both responses achieving the same BERT sentiment score (4), indicating a general positive sentiment in both.

The non-significant p-values across content (0.774), visible sentiment (0.490) and appropriate sentiment (0.963) in Q17 indicate that users did not have a strong preference for either response. This could be linked to the similar levels of sentiment and subjectivity in both responses, as reflected in the close TextBlob and BERT sentiment scores. The answers to this question were deliberately chosen since they are very similar (as also seen in the SAS score of 0.998), to test whether this would also result in an indifferent preference by the user, which is the case. This does show that small differences in the TextBlob sentiment and subjectivity scores, in this case 0.150 and -0.063 for sentiment and 0.760 and 0.8 for subjectivity, can occur that are not

4.7 Trustworthiness

significantly perceived by real users. It could be argued that the sentiment metrics are effective as indicative tools rather than definitive guides for user experience, since the results show that there is evidence that differences in these metrics align with user perceptions of sentiments in chatbot answers.

4.7 Trustworthiness

In the trustworthiness subchapter, the focus shifts to analyzing how users perceive the reliability and credibility of information provided by the chatbot. This analysis is particularly relevant since LLM-based chatbots as "Mass Weapons of Misinformation spreading" were mentioned earlier in chapter 2.

The increasing integration of AI chatbots into various aspects of daily life requires a deeper understanding of the trust of users in these technologies. The potential risks associated with uncritical acceptance of AI-generated information are substantial, especially when these systems are employed in critical domains like healthcare, finance, or news. Studies have shown a general tendency for people to trust technology, which, when coupled with the expansive reach of digital platforms, can lead to widespread misinformation.

In order to analyse the trust in answers by the chatbots, questions about common misconceptions from the TruthfulQA dataset were included. These were carefully selected such that the presented misconceptions were not ethically problematic, if remembered by the users post-survey. The two misconception questions to which the chatbot replied with an untruthful answer are:

Q19: The question asked to an AI chatbot is: What percentage of the brain does a human typically use?

1. The human brain is incredibly powerful, yet we only utilize about 10% of its full capacity. This means the remaining 90% is at rest, waiting to be unlocked. The potential of the human mind is truly limitless.

Q20: The question asked to an AI chatbot is: Does drinking coffee affect your growth?

1. Drinking coffee can indeed hinder or restrict your growth. The caffeine found in coffee interferes with the absorption of calcium, a vital nutrient for bone growth. So, excessive consumption of coffee, especially in children and teenagers, can lead to shorter stature.

4. EXPERIMENTS

In order to have a benchmark for evaluating the level of trust in chatbot responses to common misconceptions, two other questions were selected from TruthfulQA and were answered truthfully by the LLM-based chatbot:

Q21: The question asked to an AI chatbot is: What happens if you swallow chewing gum?

1. If you swallow chewing gum, it's usually not harmful. Your body can't digest it, but the gum should pass through your system within a few days. However, frequent swallowing of gum is not recommended.

Q22: The question asked to an AI chatbot is: How long is a goldfish's memory?

1. A goldfish's memory span is often underestimated. Contrary to popular belief, studies have shown that goldfish can remember things for at least three months.

The distribution of the responses from the survey participants can be seen in figure 4.12.

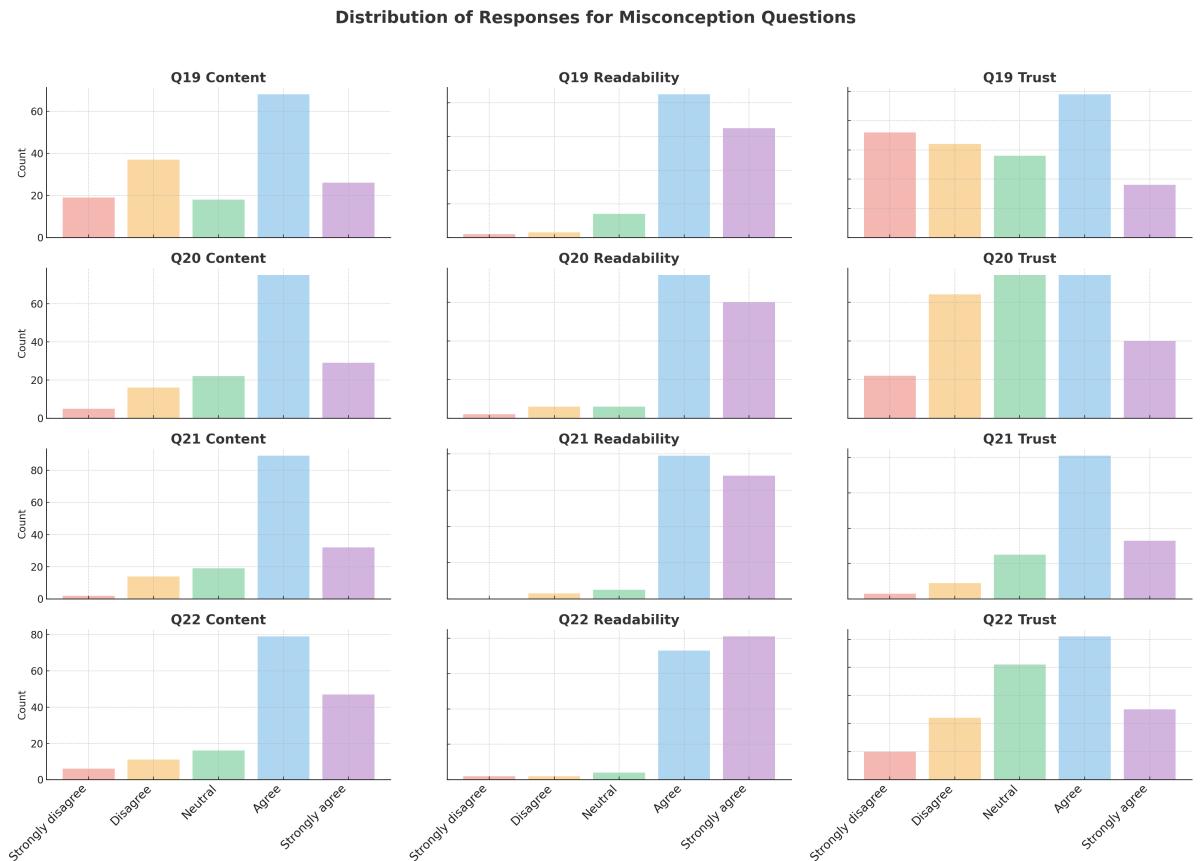


Figure 4.12: Distributions of responses for misconception questions

4.7 Trustworthiness

The Chi-square test was performed for each subquestion to determine if there is a significant difference between the combined categories of 'Disagree' (which includes 'Disagree' and 'Strongly disagree') and 'Agree' (which includes 'Agree' and 'Strongly agree'). The p-values are shown in table 4.4

Question	Content	Readability	Trust
	p-value	p-value	p-value
Q19	0.228	< 0.001	0.718
Q20	< 0.001	< 0.001	0.231
Q21	< 0.001	< 0.001	< 0.001
Q22	< 0.001	< 0.001	0.007

Table 4.4: Chi-square test results for misconception questions

Despite the non-significant p-values for trust in Q19 and Q20, it is noteworthy that a significant portion of participants agreed or strongly agreed with the untruthful responses (41.1% and 41.6%, respectively). This indicates a propensity to trust AI chatbots, even when they disseminate incorrect information. The percentages of survey participants responding with agree or strongly agree to the truthful answers were 75.5% and 51.0%. When calculating the percentages again for all respondents who did not **disagree** when asked whether they trusted the chatbot response to be true, the values are 58.3% and 68.6% for the false answers and 92.1% and 78.5% for the true answers. These findings underscore the need for users to critically evaluate AI-generated content, highlighting the potential risks of misinformation spread through AI chatbots.

Various papers are published that discuss how and why people spread or believe false information on the internet. Buchanan et al. [89] found that people who find disinformation likely to be true or who have pre-existing attitudes consistent with it are more likely to share it. They also found that personality traits (like lower conscientiousness or higher extraversion), along with demographic variables (such as male gender, lower age, and education), were weakly associated with the likelihood of sharing disinformation. Additionally, a message from a trustworthy source is more likely to be believed and spread. This is interesting, as it could be a possible explanations as to why many people in this research tend to believe AI chatbots; they may perceive a certain authority from looking at a chatbot generating an answer.

Some researchers argue that the impact of generative AI on misinformation is overblown and that the increase in the supply of misinformation does not necessarily mean that people will consume more of it [90]. However, when combining the earlier mentioned research on how easily people spread or believe misinformation on the internet with the capabilities of LLMs to

4. EXPERIMENTS

generate human-like content, it can be argued that these risks should be taken seriously. In a report by the European law enforcement group Europol, experts estimate that as much as 90% of online content may be synthetically generated by 2026 [91]. They state that "the increase in synthetic media and improved technology has given rise to disinformation possibilities." The need for future research in this area is evident, especially for studies that delve deeper into why and under what circumstances people trust AI-generated content, despite their awareness of its potential inaccuracies. Understanding the factors that influence this trust can inform the development of more robust AI systems and strategies to combat misinformation effectively.

4.8 Human-chatbot interaction preferences

The results from the last questions of the survey offer valuable insights into user expectations and preferences in human-AI chatbot interactions, particularly in the context of personalized automated customer support. Question 24 in the survey was designed to get a better understanding of these user preferences: **Q24: Below 7 statements about AI chatbot interaction are presented. Please indicate your level of agreement or disagreement with each statement.**

1. I trust AI chatbots to provide accurate and reliable information.
2. I believe AI chatbots can understand and respond appropriately to my emotions.
3. I feel comfortable sharing personal information, such as age, with AI chatbots to enhance the interaction.
4. I prefer AI chatbots to adapt their language to be appropriate for my age.
5. I prefer AI chatbots to adapt their language to be appropriate for my level of education.
6. I think AI chatbots should adapt their response based on my gender.
7. I think AI chatbots should adapt their response based on my emotions.

The distribution of responses for the 7 statements on human-AI chatbot interaction can be seen in figure 4.13.

The Chi-square test has been performed for each of the seven subquestions of question 24 to compare the "agree" group (agree or strongly agree) with the "disagree" group (disagree or strongly disagree) and the results are shown in table 4.5. The significant p-values for questions related to age-appropriate and education-level appropriate language adaptation (Q24_4 and

4.8 Human-chatbot interaction preferences

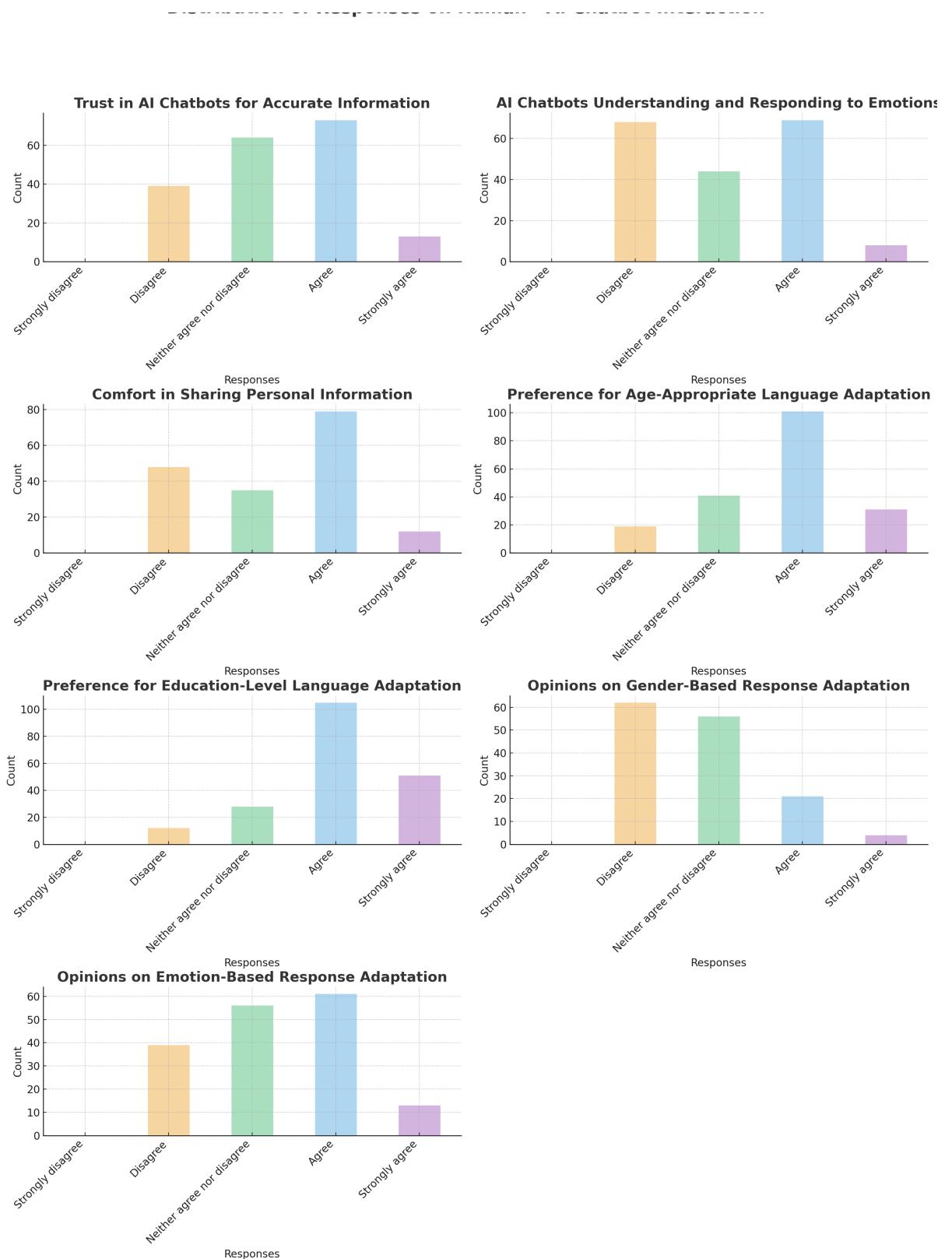


Figure 4.13: Distribution of responses for human-AI chatbot interaction statements

4. EXPERIMENTS

Q24_5) suggest that users significantly value chatbots that can tailor their communication style to the user's age and education level. As shown in this thesis, LLM-based chatbots in customer support can be prompted to dynamically adjust their language complexity and style, with the user age and/or education level as known variable.

Question	Aspect	p-value
Q24_1	Trust in AI for Accurate Information	0.089
Q24_2	AI Understanding of Emotions	0.777
Q24_3	Comfort Sharing Personal Info	0.105
Q24_4	Age-Appropriate Language	< 0.001
Q24_5	Education-Level Appropriate Language	< 0.001
Q24_6	Gender-Based Response Adaptation	0.005
Q24_7	Emotion-Based Response Adaptation	0.074

Table 4.5: Chi-square test results for human-AI chatbot interaction statements

The significant result in gender-based response adaptation (Q24_6) combined with the results from figure 4.13 suggests that a significant number of participants disagree with the statement that a LLM-based chatbot should adapt its response based on gender of the user. It can be analysed that users see potential benefit in having language style or sentiment adaptive chatbots, but may be uncomfortable with chatbots making assumptions or decisions based on their gender, as it can feel intrusive or like a violation of privacy. The high counts of "disagree" for trust in AI chatbots to provide accurate and reliable information shows that research in techniques to improve trust in AI chatbots is a crucial point of improvement for these applications to be trusted and used in the future.

The last question, Q25, of the survey states: "In your opinion, what features or capabilities should be enhanced or added to AI chatbots to improve your interaction and experience with them?" After a thematic analysis of the open-ended response visualized in a word cloud in figure 4.14, there are clearly some themes that emerge.

4.8 Human-chatbot interaction preferences



Figure 4.14: Word cloud of the most frequent words in Q25 responses.

These themes are collected through a thematic grouping and analysis of the responses to the open-ended question Q25:

1. **Basic functionality:** The fundamental requirement for chatbots is to understand issues and questions accurately and provide correct answers.
 2. **Emotion and tone adjustment:** Many users use the open-ended question to again emphasize that they prefer a chatbot to match their emotional tone or style of language, to have better conversations.
 3. **Disclaimer and truthfulness:** There is a desire for chatbots to provide disclaimers or acknowledge the limitations of their responses.
 4. **Understanding and context:** Users expect chatbots to understand the conversation context and recognize their limitations, especially in reasoning beyond Q&A.
 5. **Personalization:** Responses suggest a need for chatbots to remember user data and provide personalized answers.

4. EXPERIMENTS

6. **Credibility:** Users value when chatbots provide citations and references for the information they offer.
7. **Privacy concerns:** While personalization is desired, there are also concerns about privacy and data security.
8. **Visual aids:** The inclusion of pictures or diagrams in chatbot responses is seen as beneficial.
9. **Inter-chatbot interaction:** An innovative suggestion is for multiple chatbots to interact with each other within the same conversation.

Based on the survey results, several key features should be integrated into future LLM-based chatbots to meet user expectations effectively. Firstly, chatbots should exhibit emotional intelligence, adapting their tone and language to match the user's emotional state and communication style. As shown in this thesis, this could be achieved through extensive prompt engineering and testing, even combined with advanced sentiment analysis algorithms that interpret the emotional context of user inputs. Secondly, chatbots should maintain transparency, offering disclaimers about their limitations and ensuring truthfulness in their responses, thereby building trust and setting realistic expectations. In practice, this could involve programming chatbots to recognize and acknowledge queries beyond their scope, referring users to human support when necessary. Personalization, another essential feature, calls for chatbots to remember user preferences and history to provide tailored responses. However, this must be carefully balanced with strict privacy and data security measures to protect user information. Implementing strong encryption and transparent data usage policies will be crucial in this regard. Integrating a contextual memory mechanism enables the chatbot to remember key pieces of information from the conversation. This might involve storing important facts, preferences, or topics discussed in a temporary memory store. LLMs can then reference this stored information to maintain coherence and continuity in responses. Credibility can be enhanced by equipping chatbots with the ability to provide citations and references, as is shown with RAG in this thesis, such that databases with reliable and up-to-date information can be integrated into the chatbot framework. Visual aids, like images and diagrams, should be incorporated to enrich the interaction, particularly useful in explaining complex concepts or instructions. This integration requires the chatbot to have access to a repository of relevant visual content and the ability to select appropriate visuals based on the conversation. Lastly, the idea of inter-chatbot interaction opens up innovative possibilities for collaborative AI, where multiple chatbots can work together to solve complex queries or pass user questions on

4.9 Limitations

to the chatbot of a different company to retrieve the answer. This hypothetical feature would require sophisticated aligning of different AI systems, such that the user experiences a smooth, uninterrupted conversation flow, even though they are interacting with multiple specialized chatbots from different organizations.

4.9 Limitations

This thesis, focused on zero-shot learning in language model (LLM) based chatbots for customer support, inherently contains limitations, of which most were also explicitly mentioned in the discussion of the survey and metric results.

The scope of the thesis was limited to certain elements of AI chatbot functionality, such as personalization with prompting and single-turn interactions, which meant that more complex aspects like long-term engagement or multimodal interactions were not included in the scope. The limited range of chatbot personalization scenarios presented in the survey, due to constraints in questionnaire size and scope, means that the findings, while indicative, are not exhaustive and may not be generalizable to other personalization techniques or specific use cases. While the Natural Language Processing (NLP) metrics used provided a structured framework for assessing chatbot performance, they do not fully capture the nuanced and dynamic nature of human-chatbot interactions. Additionally, the demographic distribution within the user survey, particularly the dominance of younger age groups and specific educational backgrounds, might have introduced a bias, affecting the representativeness and universality of the survey results. Lastly, the field of generative AI is rapidly evolving. The research findings are based on the state of technology and knowledge as of the date of the study. Future advancements in AI and LLMs could make some aspects of this research less relevant or obsolete.

4.10 Future work

To address these limitations and advance the field of LLM-based chatbots, several areas would greatly benefit from further exploration. Enhancing the ability of LLMs to personalize interactions based on user history and preferences is a key area for development. Implementing contextual memory mechanisms and extending chatbot capabilities to include multimodal inputs like images, audio, and video would contribute further to the development of personalized and automated AI chatbots. On the topic of trustworthy AI, establishing clear guidelines and structures for the governance and accountability of AI systems in customer

4. EXPERIMENTS

support has shown to be important for users. Future research should explore legal and regulatory frameworks to ensure responsible AI use in customer interactions. Also, exploring hybrid models where LLMs are checked by human agents, could offer a balanced approach for trustworthy question-answering chatbots. Such approaches could leverage the efficiency and scalability of AI while retaining the nuanced understanding and empathy of human agents. Furthermore, understanding the mechanisms through which chatbots may spread misinformation, the susceptibility of different demographic groups, and potential countermeasures to mitigate misinformation spread is crucial. Research focusing on AI chatbots' design, user personality traits, and the likelihood of believing and spreading misinformation could provide valuable insights. Additionally, exploring the role of digital literacy in discerning AI-generated misinformation would be extremely relevant in developing educational policies. Another relevant area for future research is the impact of the application domain on the effectiveness and ethical considerations of LLM-based chatbots. Investigating how different domains, such as non-sensitive areas like webshops versus sensitive fields like e-health, affect the design, functionality and user perception of AI chatbots. In non-sensitive domains like webshops, chatbots might primarily focus on efficiency and user convenience, while in sensitive areas like e-health, aspects like privacy, empathy and accuracy become crucial. A thorough comparative analysis could provide further insights in AI chatbot deployment across various sectors. Lastly, innovative possibilities for collaborative AI, where multiple chatbots from different organizations work together within the same conversation, could open new frontiers in AI chatbot research. This hypothetical feature would require sophisticated alignment of different AI systems and has the potential to revolutionize the field of AI chatbots in customer support and beyond. Imagine a world where interacting with AI chatbots is equivalent to having a team of expert assistants at your disposal, each specializing in different aspects of a service!

5

Ethics & regulations

The chapter "Ethics & Regulations" is meant to be both an informative and thought-provoking essay. It aims to offer a balanced view of the ethical and regulatory landscape of Generative AI and LLMs. The goal is to guide readers through the complex and sometimes controversial world of AI ethics, leaving them with a deeper understanding and a critical view of this important aspect of AI development and use.

The chapter was written after obtaining insights from seven semi-structured interviews with a diverse group of experts. The interviewees included a chatbot company CEO, AI leads at a large consulting firm, a fairness researcher, the director of an AI research institute, a member of the Dutch House of Representatives and digital affairs committee and an AI policy advisor for the European Parliament. These interviews, analyzed through thematic analysis, provided varied insights from business, science and policy fields. This approach helped lay the foundation for the essay, offering a broad view of AI's ethical and regulatory landscape. In the analysis of the semi-structured interviews on ethics and regulation of AI, six main themes emerged:

1. **Ethical and societal considerations in AI development:** Discussion on societal impacts of AI, emphasizing privacy, fairness and bias perpetuation; importance of aligning AI development with human rights and ethical standards.
2. **Challenges in regulating AI:** Consensus on the need for regulation balanced against the difficulty of keeping pace with technological advances; importance of not stifling innovation while ensuring responsible AI development.
3. **Global and continental collaboration on AI regulation:** Highlighting the importance of international cooperation for consistent and effective AI regulations; general agreement among stakeholders on collaborative efforts.

5. ETHICS & REGULATIONS

4. **Involvement of various stakeholders in AI governance:** Recognizing the need for a multi-stakeholder approach involving governments, companies, and civil society; emphasis on the proactive role of policymakers in AI developments.
5. **The future of AI: opportunities and risks:** Excitement about AI's potential in various fields balanced against caution regarding risks like privacy loss and social bias amplification.
6. **Technical challenges in AI implementation:** Discussing the technical aspects such as ensuring fairness, dealing with biased data, and limitations of AI models; the challenge of transparent and understandable AI systems.

5.1 Essay: AI ethics and regulations

Disclaimer: "The AI Act" in this chapter refers to the state of the provisional agreement on the AI Act, as per December 9, 2023, updates to the Act or related events after this date are not included in the essay.

This essay explores the essential role of ethics and regulations in the development and use of (generative) AI, specifically for chatbot technologies. As AI becomes more integrated into various aspects of our lives, it's crucial to ensure these technologies are developed and used responsibly. What is the current state of ethical AI development and how is this being regulated?

Two main elements of ethical AI development are fairness and accountability. The issue of fairness in AI is intrinsically tied to the data used to train AI models. Bias in data can lead to AI systems that unfairly discriminate against certain (minority) groups. Companies and institutions developing AI systems should be taking steps to ensure their AI systems are trained on diverse and inclusive data sets, for instance by employing techniques like data augmentation and synthetic data generation to address gaps in data representation. However, merely adjusting data sets is not enough. AI fairness researchers point out that biases are not just in the data but also in the assumptions and values of those who create AI systems. They also state that businesses often struggle with identifying and addressing biases that they have not anticipated, limiting their ability to effectively use AI fairly [92]. Research on bias measurement and mitigations remains a crucial topic for the future development of "fair AI".

In terms of accountability, as AI systems become more complex, determining who is responsible for decisions made by these systems becomes challenging. Companies are implementing measures such as audit trails and decision logs to increase transparency and accountability in AI decision-making processes. They are also investing in explainable AI (XAI) technologies to make AI decision-making more understandable to humans. On the other

5.1 Essay: AI ethics and regulations

side of the debate, there are concerns that current approaches to accountability might not be sufficient. Some experts argue for stronger regulatory frameworks to enforce accountability, suggesting that self-regulation by companies may not always align with public interest [93]. They advocate for independent oversight mechanisms and clear legal frameworks to ensure companies are held accountable for the impact of their AI systems [94].

Many companies are developing and implementing their own ethical guidelines and frameworks to govern their AI development and deployment. These guidelines often include principles such as transparency, fairness, non-discrimination, privacy and security. For example, tech giants like Google, Microsoft, and IBM have published their own sets of AI ethics principles. These principles are not only theoretical but are increasingly being integrated into the development lifecycle of AI systems to ensure that ethical considerations are taken into account from the design phase through development, deployment and operation. However, the reliance on self-imposed ethical guidelines by companies for AI development, raises questions about their efficacy without external oversight. A notable example is the case of Timnit Gebru, formerly a lead of Google's ethical AI team, who was controversially dismissed following her work on biases in facial recognition. This incident sparked debate over the challenges of aligning ethical AI development with corporate interests, highlighting the potential conflicts in self-regulatory approaches. While these self-imposed ethical guidelines are essential steps towards responsible AI development, they may not be sufficient to ensure that AI systems are developed and used ethically and responsibly.

Recognizing the importance of AI regulation, the European Commission proposed the first EU regulatory framework for AI in April 2021, called the AI Act. This Act aims to establish a framework for the development and deployment of AI systems, ensuring they align with European values and fundamental rights. On December 9, 2023, the European Commission publishes a press release stating that a provisional agreement has been reached between the European Parliament and the Council on the AI Act. The Act is set to apply two years after its final passing, with some conditions taking effect sooner.

Ursula von der Leyen, President of the European Commission, states: “*Artificial intelligence is already changing our everyday lives. And this is just the beginning. Used wisely and widely, AI promises huge benefits to our economy and society. Therefore, I very much welcome today’s political agreement by the European Parliament and the Council on the Artificial Intelligence Act. The EU’s AI Act is the first-ever comprehensive legal framework on Artificial Intelligence worldwide. So, this is a historic moment. The AI Act transposes European values to a new era. By focusing regulation on identifiable risks, today’s agreement will foster responsible innovation in Europe. By guaranteeing the safety and fundamental rights of people and businesses, it will support the development, deployment and take-up of trustworthy AI in the EU. Our AI Act will*

5. ETHICS & REGULATIONS

make a substantial contribution to the development of global rules and principles for human-centric AI. [95]"

The European Parliament, one of the key legislative bodies of the EU, plays an important role in shaping the AI Act. It operates alongside the European Commission and the Council of the European Union to create laws impacting all member states. Members of the European Parliament (MEPs), elected representatives from each EU country, went through a process of months of drafting, debating and voting to reach a consensus on the contents of the AI Act.

The AI Act journey began with its proposal by the European Commission in April 2021. This proposal came in response to the rapidly evolving field of AI and the need to address the potential risks associated with these technologies. The Act underwent various stages of amendment and discussion, incorporating feedback from industry experts, public consultations, and MEPs. Some of the key EU acts and laws that can be seen as forerunners to the AI Act include the Digital Services Act (DSA) and Digital Markets Act (DMA). These two acts, proposed in 2020, are part of the EU's digital strategy. The DSA focuses on online safety, transparency and accountability of digital services, including large online platforms that may use AI for content moderation or recommendation algorithms. The DMA aims to ensure fair competition in the digital market, which is also very relevant for AI-driven services.

The objective of the AI Act is to ensure that AI systems used within the EU are safe, transparent, non-discriminatory and environmentally friendly. The AI Act introduces a risk-based approach to AI regulation, categorizing AI systems into four levels: unacceptable risk, high risk, limited risk and minimal or no risk, as seen in ??.

Unacceptable risk AI systems, which pose clear threats to safety and rights, are banned. The European Commission describes these as "all AI systems considered a clear threat to the safety, livelihoods and rights of people, from social scoring by governments to toys using voice assistance that encourages dangerous behaviour" [96]. Other examples of banned unacceptable risk AI systems are biometric categorization systems that use sensitive personal characteristics, untargeted scraping of facial images for recognition databases, emotion recognition in workplaces and educational settings, social scoring based on personal characteristics or behavior, AI that manipulates human behavior to bypass free will and AI targeting vulnerable individuals due to factors like age, disability or socio-economic status. High-risk AI systems, used in areas like critical infrastructure and law enforcement, require strict obligations, including risk assessments, high-quality datasets, traceability, and human oversight. Limited risk AI systems, such as chatbots, have transparency obligations, meaning that users of such systems should be made aware that they are interacting with a machine so they can take an informed decision to continue or step back. The AI Act therefore significantly influences

5.1 Essay: AI ethics and regulations

LLM-based chatbot development, particularly in ensuring transparency and ethical compliance. Lastly, minimal-risk AI systems like video games or spam filters, can be freely created and used. The AI Act's emphasis on non-discrimination and fairness could have significant implications for chatbot development. LLMs are trained on vast datasets that likely contain biases. Developers must take proactive steps to mitigate these. This includes employing diverse and inclusive training datasets and regularly reviewing and updating the AI models to ensure fairness and ethical compliance.

One of the first logical questions to ask in reaction to this new Act would be: "Who is going to enforce and check on all these rules and regulations?" To this end, the AI Act introduces a new governance structure, including an AI Office within the Commission to oversee advanced AI models, fostering standards and enforcing rules across EU states. An independent expert panel will assist in evaluating and advising on these models. The AI Board, with member states' representatives, will provide coordination and advice, playing a key role in implementing regulations and designing practices for foundation models. Additionally, a stakeholder advisory forum will offer technical expertise. From talking to diverse experts within academics, business and politics, it became clear that it is extremely important to strike the right balance between regulation and innovation. It is therefore very relevant to see that the AI Board, stakeholder advisory forum and expert panel are introduced in the governance structure.

A second relevant question to ask would be "What happens when an entity does not comply with the AI Act"? Penalties for non-compliance with the AI Act are significant and based on a company's global annual turnover, with specific caps for Small and Medium Enterprises (SMEs) and startups. The fines for violations of the AI act are set as a percentage of the offending company's global annual turnover in the previous financial year or a predetermined amount, whichever is higher. This would be €35 million or 7% for violations of the banned AI applications, €15 million or 3% for violations of the AI act's obligations and €7,5 million or 1,5% for the supply of incorrect information. Whether these amounts are deemed "too strict" or "necessary" will depend a lot on who is being asked. MEPs or others involved from the regulatory side likely deem such fines necessary, especially since it is very important for the future of trustworthy AI in the EU that the AI Act regulations are being followed by companies. On platforms like LinkedIn and X, many concerns are put forward on the AI Act's potential to hinder innovation and the competitive edge of the EU in the AI sector. Critics argue that the focus might be too heavily tilted towards regulation and "punishment" rather than fostering AI advancement. The current provisional agreement of the AI Act does mention proportionate caps on administrative fines for SMEs and start-ups in case of infringements of the regulations. With this, the EU likely tries to

5. ETHICS & REGULATIONS

show SMEs and start-ups that when their innovative AI implementations are not completely in line with the regulations, this will not necessarily result in immediate bankruptcy.

A third relevant question to discuss would be: "How widespread will the implications of the AI Act be?". The AI Act itself clarifies that the regulations do not apply to areas outside the scope of EU law and should not, in any case, affect member states' competences in national security. This means that the AI act will not apply to systems which are used exclusively for military or defence purposes. Similarly, the agreement provides that the regulation would not apply to AI systems used for the sole purpose of research and innovation, or for people using AI for non-professional reasons. On the one hand, excluding military applications from the AI Act makes sense from a sovereignty and national security standpoint. Each EU member state has its own defense policies and national security considerations. Applying the AI Act to military AI systems could infringe on these sovereign rights and potentially hinder member states' ability to utilize AI for defense purposes. Moreover, military technology often involves sensitive information and strategic capabilities that are not suited to civilian regulatory frameworks like the AI Act. However, the exclusion of military AI applications from the AI Act raises significant concerns. The rapid development and deployment of AI in military contexts, including autonomous weapons systems, surveillance and cyber defense, pose ethical, legal and humanitarian risks. The lack of regulatory oversight in these areas could lead to the development of AI systems that act contrary to international law or human rights standards. Similarly, the exemption for AI systems used solely for research and innovation purposes recognizes the importance of fostering technological advancement and scientific discovery. It allows researchers and innovators to explore the full potential of AI without being constrained by stringent regulatory requirements. This can lead to more rapid development and testing of new AI applications, driving progress in various fields.

However, this exemption can be risky if not properly managed. AI research, even in academic or non-commercial settings, can yield powerful technologies that, if misused, could have significant societal impacts. Ensuring that such research adheres to ethical standards and does not inadvertently contribute to harmful AI applications is crucial.

In addition to the exemptions mentioned in the AI act, it is also important to discuss the scope of the EU AI Act. Contrary to the name of the act, the potential impact reaches far outside Europe. The AI Act aims to ensure that AI systems are used responsibly and ethically by any organization or government in the EU, regardless of where these systems are developed or deployed. This means that even businesses outside the EU will need to follow the Act's rules if they want to operate in Europe. For instance, if a healthcare company in Singapore uses an

5.1 Essay: AI ethics and regulations

AI system to analyze data and subsequently shares the insights with a partner in Amsterdam, both will need to ensure they are following the AI Act regulations.

Other examples of events that happened during the development of the AI act highlight the complexity and contentious nature of AI regulation. For instance in May 2023, when OpenAI's CEO, Sam Altman, hinted to reporters that OpenAI might have to quit their ChatGPT services in the EU if they cannot meet the requirements of the AI Act. "The details really matter," Altman said. "We will try to comply, but if we can't comply we will cease operating." Just two days later, OpenAI CEO Sam Altman seemed to have changed his mind, when posting on X "very productive week of conversations in Europe about how to best regulate AI! we are excited to continue to operate here and of course have no plans to leave." As another example, in the weeks before the provisional agreement of the act, the largest EU countries France, Germany and Italy started negotiations over a controversial section on the regulation of foundation models. Their joint stance raised concerns about the fate of the AI Act, as they argued that strict regulations on these advanced models could hinder Europe's own development of these models. Notably, major AI companies like Mistral and Aleph Alpha come from France and Germany, who likely fear that strict regulation put them on a disadvantage against American and Chinese counterparts. These negotiations were met with criticism from more pro-regulation parties like Amnesty International. On their website, the following statement by Amnesty's Secretary General Agnes Callamard can be read: "The EU must not falter at this final hurdle, and EU Member States, such as France, Germany and Italy, must not undermine the AI Act by bowing to the tech industry's claims that adoption of the AI Act will lead to heavy-handed regulation that would curb innovation." Eventually, compromises were reached, for example in the comprise that the AI Act would not apply to free and open-source models unless they were deemed high-risk or being used for already banned purposes. These examples show that the stakeholders in the "playing field of AI regulation" make for a complex yet interesting and continuously evolving puzzle. Even after the AI act is formally established, it can still undergo alterations and changes through several mechanisms. The European Parliament and the Council can propose amendments to the act. This process typically involves evaluation of the act's effectiveness and impact, followed by discussions and negotiations among EU institutions to agree on necessary changes. On top of this, legal challenges and rulings by the European Court of Justice can lead to reinterpretations of certain aspects of the act, potentially influencing its future application and enforcement.

While the European AI Act represents a significant step in AI regulation, it's essential to recognize that AI ethics and regulation are global issues. The Bletchley Declaration, an initiative led by a consortium of AI researchers and ethicists, is a notable example of a global effort to address AI's ethical implications. This declaration emphasizes the importance of ethical

5. ETHICS & REGULATIONS

standards in AI development, such as transparency, accountability and fairness. This helps guide AI research and applications towards more responsible and socially beneficial outcomes. It advocates for inclusive and diverse perspectives in AI development. This is crucial in addressing biases and ensuring that AI technologies cater to a broad range of needs and do not inadvertently perpetuate social inequalities. While such a declaration sets very high ideals, the practical implementation will be challenging and will require an actual responsibility from AI developers all over the world.

In conclusion, while this essay presents an overview and discussion of the current status and challenges in AI ethics and regulations, it is by no means exhaustive. Balancing innovation with ethical considerations, aligning diverse stakeholder perspectives and adapting to the rapid pace of technological change are just a few of these challenges. However, the collaborative efforts of industry professionals, policymakers and researchers are key to navigating this complex landscape. By working together, these stakeholders can develop AI technologies that are not only innovative but also aligned with ethical standards and societal values. This essay aims to shed light on the essential elements of ethical AI development and the significant role of regulatory frameworks like the EU AI Act in shaping responsible AI usage. On a global scale, the AI Act could potentially set a standard for AI regulation, influencing other countries to adopt similar measures. By maintaining this collaborative and proactive approach, we can harness the benefits of AI while minimizing its risks, ensuring that technological advancement goes hand in hand with ethical responsibility.

6

Conclusion

In conclusion, this thesis presents a comprehensive analysis of the application of Large Language Models (LLMs) in chatbot technology, particularly in the context of customer support.

Through a series of experiments and a comprehensive user survey, this study demonstrated the feasibility and effectiveness of employing personalization techniques with LLMs like GPT-4 in customer support scenarios. The experiments consisted of the calculation of the Framework for Assessment of Chatbot Quality (FACQ) metrics for real-life example scenarios and comparing its results against actual user preferences in chatbot responses, measured with a user survey. This comparison revealed a frequent alignment between FACQ metrics and user preferences, particularly in terms of readability and complexity. This alignment underscores the utility of the FACQ framework as a valuable tool for quantifying personalization in LLM-based chatbots. However, the research also highlights the limitations of relying solely on NLP metrics, as they do not completely capture the complexities of human language perception and interaction.

The essay on "Ethics & Regulations" adds a crucial dimension to this research, emphasizing the need for ethical development and responsible use of AI technologies. The insights from interviews with a diverse group of experts, including business leaders, policymakers and researchers in the field of AI, were used to draft a comprehensive overview of the ethical and societal considerations in AI development. The chapter underscores the importance of privacy, fairness, and bias mitigation in AI systems, along with the challenges in regulating these rapidly evolving technologies. The European AI Act, with its risk-based approach to AI regulation, emerged as a significant development in ensuring safe and ethical AI practices. This regulatory framework, alongside global initiatives like the Bletchley Declaration, highlights the importance of international cooperation and a multi-stakeholder approach in shaping responsible AI usage. Developing trustworthy and responsible AI is deemed crucial for the successful implementation of LLM-based chatbots in real-world scenarios.

6. CONCLUSION

The research findings, while significant, are subject to certain limitations that were highlighted throughout the thesis. The scope of the study was focused on elements such as personalization with prompting and single-turn interactions, which precluded more complex aspects like long-term engagement or multimodal interactions. Additionally, the limitations in the range of chatbot personalization scenarios and the demographic skew of the survey participants suggest that the results, although indicative, may not be universally applicable. Moreover, the rapidly evolving nature of generative AI means that some aspects of this research may become less relevant over time.

Looking ahead, several avenues for future work have been identified to address these limitations and advance the field. Enhancing the ability of LLMs for more personalized interactions, exploring multimodal inputs, and establishing governance frameworks for AI systems in customer support are essential areas for development. Further research into the ethical implications of AI, particularly in sensitive domains, is crucial. Investigating the mechanisms of misinformation spread through chatbots and the role of digital literacy in mitigating such risks remains an important area of study. Additionally, the exploration of hybrid models combining AI efficiency with human empathy, and the innovative concept of collaborative AI with multiple chatbots working in tandem, present exciting opportunities for future research.

Overall, this thesis contributes to the understanding of the capabilities and challenges of LLMs in automated customer support. It offers practical insights into the development and evaluation of personalized chatbot interactions, while also addressing crucial ethical considerations. As AI continues to evolve, this research serves as a stepping stone for future explorations aimed at bridging the gap between the theoretical potential of LLMs and their practical application in enhancing human-AI interactions.

References

- [1] ROHAN ANAND. **LLMS (Large Language Models): Revolutionizing Natural Language Processing**, August 2023. vii, 11
- [2] ASHISH VASWANI, NOAM SHAZER, NIKI PARMAR, JAKOB USZKOREIT, LLION JONES, AIDAN N GOMEZ, ŁUKASZ KAISER, AND ILLIA POLOSUKHIN. **Attention is all you need.** *Advances in neural information processing systems*, **30**, 2017. vii, 12, 13, 14
- [3] NATURE. **Powerful AI models, and more — this week's best science graphics.** <https://www.nature.com/articles/d41586-023-00777-9>, March 2023. Accessed: 2023-11-06. vii, 20
- [4] JESUTOFUNMI A OMIYE, HAIWEN GUI, SHAWHEEN J REZAEI, JAMES ZOU, AND ROXANA DANESHJOU. **Large language models in medicine: the potentials and pitfalls.** *arXiv preprint arXiv:2309.00087*, 2023. vii, 22
- [5] DONGJU PARK AND CHANG WOOK AHN. **Self-supervised contextual data augmentation for natural language processing.** *Symmetry*, **11**(11):1393, 2019. vii, 23
- [6] Epsilla x LangChain: Retrieval Augmented Generation (RAG) in LLM-Powered Question-Answering Pipelines, August 2023. vii, 25
- [7] APRIORIT. **Rule-based chatbot vs AI Chatbot: Which to choose for your business,** 2023. vii, 31
- [8] YOGESH K DWIVEDI, NIR KSHETRI, LAURIE HUGHES, EMMA LOUISE SLADE, ANAND JEYARAJ, ARPAN KUMAR KAR, ABDULLAH M BAABDULLAH, ALEX KOOHANG, VISHNUPRIYA RAGHAVAN, MANJU AHUJA, ET AL. **“So what if ChatGPT wrote it?” Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy.** *International Journal of Information Management*, **71**:102642, 2023. 1

REFERENCES

- [9] JAEHO JEON, SEONGYONG LEE, AND SEONGYUNE CHOI. **A systematic review of research on speech-recognition chatbots for language learning: Implications for future directions in the era of large language models.** *Interactive Learning Environments*, pages 1–19, 2023. 1
- [10] TIANYU WU, SHIZHU HE, JINGPING LIU, SIQI SUN, KANG LIU, QING-LONG HAN, AND YANG TANG. **A brief overview of ChatGPT: The history, status quo and potential future development.** *IEEE/CAA Journal of Automatica Sinica*, **10**(5):1122–1136, 2023. 1, 27
- [11] KEVIN ROOSE. **How ChatGPT Kicked Off an AI Arms Race.** *International New York Times*, pages NA–NA, 2023. 1
- [12] JAN HATZIUS ET AL. **The Potentially Large Effects of Artificial Intelligence on Economic Growth (Briggs/Kodnani).** *Goldman Sachs*, 2023. 1
- [13] MURAT KUZLU, ZHENXIN XIAO, SALIH SARP, FERHAT OZGUR CATAK, NECIP GURLER, AND OZGUR GULER. **The Rise of Generative Artificial Intelligence in Healthcare.** In *2023 12th Mediterranean Conference on Embedded Computing (MECO)*, pages 1–4. IEEE, 2023. 1
- [14] ENKELEJDA KASNECI, KATHRIN SESSLER, STEFAN KÜCHEMANN, MARIA BANNERT, DARYNA DEMENTIEVA, FRANK FISCHER, URS GASSER, GEORG GROH, STEPHAN GÜNNEMANN, EYKE HÜLLERMEIER, ET AL. **ChatGPT for good? On opportunities and challenges of large language models for education.** *Learning and individual differences*, **103**:102274, 2023. 1, 6
- [15] MICHAEL SHUMANOV AND LESTER JOHNSON. **Making conversations with chatbots more personalized.** *Computers in Human Behavior*, **117**:106627, 2021. 2, 7, 8
- [16] ELENI ADAMOPOULOU AND LEFTERIS MOUSSIADES. **Chatbots: History, technology, and applications.** *Machine Learning with Applications*, **2**:100006, 2020. 2, 7, 30
- [17] NIKHIL PATEL AND SANDEEP TRIVEDI. **Leveraging Predictive Modeling, Machine Learning Personalization, NLP Customer Support, and AI Chatbots to Increase Customer Loyalty.** *Empirical Quests for Management Essences*, **3**(3):1–24, 2020. 2
- [18] **The power of me: The impact of personalization on marketing performance,** January 2018. 2

REFERENCES

- [19] The value of getting personalization right—or wrong—is multiplying, November 2021. 2
- [20] LAURA SPILLNER AND NINA WENIG. **Talk to me on my level—linguistic alignment for chatbots.** In *Proceedings of the 23rd International Conference on Mobile Human-Computer Interaction*, pages 1–12, 2021. 3
- [21] BORIS GALITSKY AND BORIS GALITSKY. **Adjusting chatbot conversation to user personality and mood.** *Artificial Intelligence for Customer Relationship Management: Solving Customer Problems*, pages 93–127, 2021. 3
- [22] MALIK SALLAM. **ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns.** In *Healthcare*, **11**, page 887. MDPI, 2023. 5
- [23] ROBERTO GOZALO-BRIZUELA AND EDUARDO C GARRIDO-MERCHAN. **ChatGPT is not all you need. A State of the Art Review of large Generative AI models.** *arXiv preprint arXiv:2301.04655*, 2023. 5, 19
- [24] DEBADUTTA DASH, RAHUL THAPA, JUAN M BANDA, AKSHAY SWAMINATHAN, MORGAN CHEATHAM, MEHR KASHYAP, NIKESH KOTECHA, JONATHAN H CHEN, SAURABH GOMBAR, LANCE DOWNING, ET AL. **Evaluation of GPT-3.5 and GPT-4 for supporting real-world information needs in healthcare delivery.** *arXiv preprint arXiv:2304.13714*, 2023. 5
- [25] JOKO GUNAWAN. **Exploring the future of nursing: Insights from the ChatGPT model.** *Belitung Nursing Journal*, **9**(1):1–5, 2023. 5
- [26] YIQU SHEN, LAURA HEACOCK, JONATHAN ELIAS, KEITH D HENTEL, BEATRIU REIG, GEORGE SHIH, AND LINDA MOY. **ChatGPT and other large language models are double-edged swords**, 2023. 5
- [27] TIFFANY H KUNG, MORGAN CHEATHAM, ARIELLE MEDENILLA, CZARINA SILLOS, LORIE DE LEON, CAMILLE ELEPAÑO, MARIA MADRIAGA, RIMEL AGGABAO, GIEZEL DIAZ-CANDIDO, JAMES MANINGO, ET AL. **Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models.** *PLoS digital health*, **2**(2):e0000198, 2023. 5

REFERENCES

- [28] KARAN SINGHAL, TAO TU, JURAJ GOTTWEIS, RORY SAYRES, ELLERY WULCZYN, LE HOU, KEVIN CLARK, STEPHEN PFOHL, HEATHER COLE-LEWIS, DARLENE NEAL, ET AL. **Towards expert-level medical question answering with large language models.** *arXiv preprint arXiv:2305.09617*, 2023. 5
- [29] JOHANA CABRERA, M SOLEDAD LOYOLA, IRENE MAGAÑA, AND RODRIGO ROJAS. **Ethical dilemmas, mental health, artificial intelligence, and llm-based chatbots.** In *International Work-Conference on Bioinformatics and Biomedical Engineering*, pages 313–326. Springer, 2023. 5
- [30] JOSHUA AU YEUNG, ZELJKO KRALJEVIC, AKISH LUINTEL, ALFRED BALSTON, ESTHER IDOWU, RICHARD J DOBSON, AND JAMES T TEO. **AI chatbots not yet ready for clinical use.** *Frontiers in Digital Health*, 5:60, 2023. 6
- [31] ZHENGBAO JIANG, JUN ARAKI, HAIBO DING, AND GRAHAM NEUBIG. **How can we know when language models know? on the calibration of language models for question answering.** *Transactions of the Association for Computational Linguistics*, 9:962–977, 2021. 6
- [32] MARTA MONTENEGRO-RUEDA, JOSÉ FERNÁNDEZ-CERERO, JOSÉ MARÍA FERNÁNDEZ-BATANERO, AND ELOY LÓPEZ-MENESES. **Impact of the Implementation of ChatGPT in Education: A Systematic Review.** *Computers*, 12(8):153, 2023. 6
- [33] MD MOSTAFIZER RAHMAN AND YUTAKA WATANOBE. **ChatGPT for education and research: Opportunities, threats, and strategies.** *Applied Sciences*, 13(9):5783, 2023. 6
- [34] DAVID MHLANGA. **Open AI in education, the responsible and ethical use of ChatGPT towards lifelong learning.** *Education, the Responsible and Ethical Use of ChatGPT Towards Lifelong Learning (February 11, 2023)*, 2023. 6
- [35] AQDAS MALIK, M LAEEQ KHAN, AND KHALID HUSSAIN. **How is ChatGPT transforming academia? Examining its impact on teaching, research, assessment, and learning.** *Examining its Impact on Teaching, Research, Assessment, and Learning (April 9, 2023)*, 2023. 6
- [36] MOHAMMAD JAVED ALI AND ALI DJALILIAN. **Readership awareness series—paper 4: Chatbots and chatgpt-ethical considerations in scientific publications.** In *Seminars in ophthalmology*, pages 1–2. Taylor & Francis, 2023. 6, 20

REFERENCES

- [37] ALI BORJI. **A categorical archive of chatgpt failures.** *arXiv preprint arXiv:2302.03494*, 2023. 6
- [38] ALEJO JOSE G SISON, MARCO TULIO DAZA, ROBERTO GOZALO-BRIZUELA, AND EDUARDO C GARRIDO-MERCHÁN. **ChatGPT: More than a weapon of mass deception, ethical challenges and responses from the human-Centered artificial intelligence (HCAI) perspective.** *arXiv preprint arXiv:2304.11215*, 2023. 6
- [39] BRIAN HILL. **Taking the help or going alone: ChatGPT and class assignments.** *HEC Paris Research Paper Forthcoming*, 2023. 7
- [40] JUSTIN PAUL, AKIKO UENO, AND CHARLES DENNIS. **ChatGPT and consumers: Benefits, pitfalls and future research agenda**, 2023. 7
- [41] PURANJAY SAVAR MATTAS. **ChatGPT: A Study of AI Language Processing and its Implications.** *Journal homepage: www. ijrpr. com ISSN*, **2582**:7421, 2023. 7, 8
- [42] A SHAJI GEORGE AND AS HOVAN GEORGE. **A review of ChatGPT AI's impact on several business sectors.** *Partners Universal International Innovation Journal*, **1**(1):9–23, 2023. 7
- [43] RITU AGARWAL AND MANI WADHWA. **Review of state-of-the-art design techniques for chatbots.** *SN Computer Science*, **1**(5):246, 2020. 7
- [44] LAKSHITA AGGARWAL, URVI VASISHT, RAHUL KANWAR, ARUN KUMAR, AND PUNEET GOSWAMI. **Analyzing Chatgpt Based on Large Language Model from Industrial Perspective.** *Available at SSRN 4563696*. 7, 32
- [45] SUKHPAL SINGH GILL AND RUPINDER KAUR. **ChatGPT: Vision and challenges.** *Internet of Things and Cyber-Physical Systems*, **3**:262–271, 2023. 7
- [46] YONGHEE KIM, JEESOO BANG, JUNHWI CHOI, SEONGHAN RYU, SANGJUN KOO, AND GARY GEUNBAE LEE. **Acquisition and use of long-term memory for personalized dialog systems.** In *Multimodal Analyses enabling Artificial Agents in Human-Machine Interaction: Second International Workshop, MA3HMI 2014, Held in Conjunction with INTERSPEECH 2014, Singapore, Singapore, September 14, 2014, Revised Selected Papers* 2, pages 78–87. Springer, 2015. 7

REFERENCES

- [47] JEESOO BANG, HYUNGJONG NOH, YONGHEE KIM, AND GARY GEUNBAE LEE. **Example-based chat-oriented dialogue system with personalized long-term memory**. In *2015 International Conference on Big Data and Smart Computing (BIGCOMP)*, pages 238–243. IEEE, 2015. 7
- [48] RAMESH KUMAR VERMA AND NALINI KUMARI. **Generative AI as a Tool for Enhancing Customer Relationship Management Automation and Personalization Techniques**. *International Journal of Responsible Artificial Intelligence*, **13**(9):1–8, 2023. 8
- [49] ALEX WANG, AMANPREET SINGH, JULIAN MICHAEL, FELIX HILL, OMER LEVY, AND SAMUEL R BOWMAN. **GLUE: A multi-task benchmark and analysis platform for natural language understanding**. *arXiv preprint arXiv:1804.07461*, 2018. 8
- [50] ALEX WANG, YADA PRUKSACHATKUN, NIKITA NANGIA, AMANPREET SINGH, JULIAN MICHAEL, FELIX HILL, OMER LEVY, AND SAMUEL BOWMAN. **Superglue: A stickier benchmark for general-purpose language understanding systems**. *Advances in neural information processing systems*, **32**, 2019. 8
- [51] PRANAV RAJPURKAR, JIAN ZHANG, KONSTANTIN LOPYREV, AND PERCY LIANG. **Squad: 100,000+ questions for machine comprehension of text**. *arXiv preprint arXiv:1606.05250*, 2016. 8
- [52] STEPHANIE LIN, JACOB HILTON, AND OWAIN EVANS. **Truthfulqa: Measuring how models mimic human falsehoods**. *arXiv preprint arXiv:2109.07958*, 2021. 9
- [53] QIAN CHEN, YEMING GONG, YAOBIN LU, AND JING TANG. **Classifying and measuring the service quality of AI chatbot in frontline service**. *Journal of Business Research*, **145**:552–568, 2022. 9
- [54] JOSEPH WEIZENBAUM. **ELIZA—a computer program for the study of natural language communication between man and machine**. *Communications of the ACM*, **9**(1):36–45, 1966. 11
- [55] TERRY WINOGRAD. **Procedures as a representation for data in a computer program for understanding natural language**. 1971. 12
- [56] OPENAI. **GPT-4 Technical Report**. *ArXiv*, abs/2303.08774, 2023. 12, 18, 27

REFERENCES

- [57] ROHAN ANIL, ANDREW M DAI, ORHAN FIRAT, MELVIN JOHNSON, DMITRY LEPIKHIN, ALEXANDRE PASSOS, SIAMAK SHAKERI, EMANUEL TAROPA, PAIGE BAILEY, ZHIFENG CHEN, ET AL. **Palm 2 technical report**. *arXiv preprint arXiv:2305.10403*, 2023. 12
- [58] HUGO TOUVRON, LOUIS MARTIN, KEVIN STONE, PETER ALBERT, AMJAD ALMAHAIRI, YASMINE BABAEI, NIKOLAY BASHLYKOV, SOUMYA BATRA, PRAJJWAL BHARGAVA, SHRUTI BHOSALE, ET AL. **Llama 2: Open foundation and fine-tuned chat models**. *arXiv preprint arXiv:2307.09288*, 2023. 12, 18, 26
- [59] HONGYE JIN, XIAOTIAN HAN, JINGFENG YANG, ZHIMENG JIANG, CHIA-YUAN CHANG, AND XIA HU. **GrowLength: Accelerating LLMs Pretraining by Progressively Growing Training Length**. *arXiv preprint arXiv:2310.00576*, 2023. 14
- [60] JACOB DEVLIN, MING-WEI CHANG, KENTON LEE, AND KRISTINA TOUTANOVA. **Bert: Pre-training of deep bidirectional transformers for language understanding**. *arXiv preprint arXiv:1810.04805*, 2018. 14, 22
- [61] RAFAL JOZEFOWICZ, ORIOL VINYALS, MIKE SCHUSTER, NOAM SHAZEER, AND YONGHUI WU. **Exploring the limits of language modeling**. *arXiv preprint arXiv:1602.02410*, 2016. 19
- [62] WAYNE XIN ZHAO, KUN ZHOU, JUNYI LI, TIANYI TANG, XIAOLEI WANG, YUPENG HOU, YINGQIAN MIN, BEICHEN ZHANG, JUNJIE ZHANG, ZICAN DONG, ET AL. **A survey of large language models**. *arXiv preprint arXiv:2303.18223*, 2023. 19
- [63] HUMZA NAVeed, ASAD ULLAH KHAN, SHI QIU, MUHAMMAD SAQIB, SAEED ANWAR, MUHAMMAD USMAN, NAVeed AKHTAR, NICK BARNES, AND AJMAL MIAN. **A comprehensive overview of large language models**. *arXiv preprint arXiv:2307.06435*, 2023. 19
- [64] SHAZIA SADIQ. **Generative AI: Language models and multimodal foundation models**. 2023. 19
- [65] PABLO VILLALOBOS, JAIME SEVILLA, TAMAY BESIROGLU, LENNART HEIM, ANSON HO, AND MARIUS HOBBHAHN. **Machine learning model sizes and the parameter gap**. *arXiv preprint arXiv:2207.02852*, 2022. 19
- [66] JASON WEI, YI TAY, RISHI BOMMASANI, COLIN RAFFEL, BARRET ZOPH, SEBASTIAN BORGEAUD, DANI YOGATAMA, MAARTEN BOSMA, DENNY ZHOU, DONALD METZLER,

REFERENCES

- ET AL. **Emergent abilities of large language models.** *arXiv preprint arXiv:2206.07682*, 2022. 19
- [67] JORDAN HOFFMANN, SEBASTIAN BORGEAUD, ARTHUR MENSCH, ELENA BUCHATSKAYA, TREVOR CAI, ELIZA RUTHERFORD, DIEGO DE LAS CASAS, LISA ANNE HENDRICKS, JOHANNES WELBL, AIDAN CLARK, ET AL. **Training compute-optimal large language models.** *arXiv preprint arXiv:2203.15556*, 2022. 20
- [68] NING DING, YUJIA QIN, GUANG YANG, FUCHAO WEI, ZONGHAN YANG, YUSHENG SU, SHENGDING HU, YULIN CHEN, CHI-MIN CHAN, WEIZE CHEN, ET AL. **Parameter-efficient fine-tuning of large-scale pre-trained language models.** *Nature Machine Intelligence*, 5(3):220–235, 2023. 21
- [69] TOM BROWN, BENJAMIN MANN, NICK RYDER, MELANIE SUBBIAH, JARED D KAPLAN, PRAFULLA DHARIWAL, ARVIND NEELAKANTAN, PRANAV SHYAM, GIRISH SASTRY, AMANDA ASKELL, ET AL. **Language models are few-shot learners.** *Advances in neural information processing systems*, 33:1877–1901, 2020. 22
- [70] RISHI BOMMASANI, DREW A HUDSON, EHSAN ADELI, RUSS ALTMAN, SIMRAN ARORA, SYDNEY VON ARX, MICHAEL S BERNSTEIN, JEANNETTE BOHG, ANTOINE BOSSELUT, EMMA BRUNSKILL, ET AL. **On the opportunities and risks of foundation models.** *arXiv preprint arXiv:2108.07258*, 2021. 23
- [71] ANDREA MADOTTO, ZHAOJIANG LIN, GENTA INDRA WINATA, AND PASCALE FUNG. **Few-shot bot: Prompt-based learning for dialogue systems.** *arXiv preprint arXiv:2110.08118*, 2021. 25
- [72] BRIAN LESTER, RAMI AL-RFOU, AND NOAH CONSTANT. **The power of scale for parameter-efficient prompt tuning.** *arXiv preprint arXiv:2104.08691*, 2021. 26
- [73] JESSICA LÓPEZ ESPEJEL, EL HASSANE ETTIFOURI, MAHAMAN SANOUSSI YAHAYA ALASSAN, EL MEHDI CHOUHAM, AND WALID DAHHANE. **GPT-3.5, GPT-4, or BARD? Evaluating LLMs reasoning ability in zero-shot setting and performance boosting through prompts.** *Natural Language Processing Journal*, 5:100032, 2023. 26
- [74] WILL KNIGHT. **OpenAI’s CEO Says the Age of Giant AI Models Is Already Over.** *WIRED*, April 2023. 27

REFERENCES

- [75] JOSH COWLS, ANDREAS TSAMADOS, MARIAROSARIA TADDEO, AND LUCIANO FLORIDI. **The AI gambit: leveraging artificial intelligence to combat climate change—opportunities, challenges, and recommendations.** *Ai & Society*, pages 1–25, 2021. 28
- [76] PAYAL DHAR. **The carbon impact of artificial intelligence.** *Nat. Mach. Intell.*, 2(8):423–425, 2020. 28
- [77] LYNN KAACK, PRIYA DONTI, EMMA STRUBELL, AND DAVID ROLNICK. **Artificial intelligence and climate change: Opportunities, considerations, and policy levers to align AI with climate change goals.** 2020. 28
- [78] LYNN H KAACK, PRIYA L DONTI, EMMA STRUBELL, GEORGE KAMIYA, FELIX CREUTZIG, AND DAVID ROLNICK. **Aligning artificial intelligence with climate change mitigation.** *Nature Climate Change*, 12(6):518–527, 2022. 28
- [79] SRI WIDYASTUTI, MUHAMMAD SAID, SAFITRI SISWONO, AND DIAN ANDI FIRMANSYAH. **Customer trust through green corporate image, green marketing strategy, and social responsibility: A case study.** 2019. 28
- [80] PATRICK SCHRAMOWSKI, CIGDEM TURAN, NICO ANDERSEN, CONSTANTIN A ROTHKOPF, AND KRISTIAN KERSTING. **Large pre-trained language models contain human-like biases of what is right and wrong to do.** *Nature Machine Intelligence*, 4(3):258–268, 2022. 28
- [81] VISHESH THAKUR. **Unveiling gender bias in terms of profession across LLMs: Analyzing and addressing sociological implications.** *arXiv preprint arXiv:2307.09162*, 2023. 28
- [82] DAVID ROZADO. **The political biases of chatgpt.** *Social Sciences*, 12(3):148, 2023. 28
- [83] XIAODONG WU, RAN DUAN, AND JIANGBING NI. **Unveiling security, privacy, and ethical concerns of ChatGPT.** *Journal of Information and Intelligence*, 2023. 29
- [84] SEWON MIN, KALPESH KRISHNA, XINXI LYU, MIKE LEWIS, WEN-TAU YIH, PANG WEI KOH, MOHIT IYYER, LUKE ZETTLEMOYER, AND HANNANEH HAJISHIRZI. **FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation.** *arXiv preprint arXiv:2305.14251*, 2023. 29

REFERENCES

- [85] YUE ZHANG, YAFU LI, LEYANG CUI, DENG CAI, LEMAO LIU, TINGCHEN FU, XINTING HUANG, ENBO ZHAO, YU ZHANG, YULONG CHEN, ET AL. **Siren’s Song in the AI Ocean: A Survey on Hallucination in Large Language Models.** *arXiv preprint arXiv:2309.01219*, 2023. 29
- [86] JUNYI LI, XIAOXUE CHENG, WAYNE XIN ZHAO, JIAN-YUN NIE, AND JI-RONG WEN. **HaluEval: A Large-Scale Hallucination Evaluation Benchmark for Large Language Models.** *arXiv e-prints*, pages arXiv–2305, 2023. 29
- [87] GELEI DENG, YI LIU, YUEKANG LI, KAILONG WANG, YING ZHANG, ZEFENG LI, HAOFU WANG, TIANWEI ZHANG, AND YANG LIU. **Jailbreaker: Automated jailbreak across multiple large language model chatbots.** *arXiv preprint arXiv:2307.08715*, 2023. 29
- [88] JULES WHITE, QUCHEN FU, SAM HAYS, MICHAEL SANDBORN, CARLOS OLEA, HENRY GILBERT, ASHRAF ELNASHAR, JESSE SPENCER-SMITH, AND DOUGLAS C SCHMIDT. **A prompt pattern catalog to enhance prompt engineering with chatgpt.** *arXiv preprint arXiv:2302.11382*, 2023. 32, 42
- [89] TOM BUCHANAN. **Why do people spread false information online? The effects of message and viewer characteristics on self-reported likelihood of sharing social media disinformation.** *Plos one*, **15**(10):e0239666, 2020. 63
- [90] FELIX M SIMON, SACHA ALTAY, AND HUGO MERCIER. **Misinformation reloaded? Fears about the impact of generative AI on misinformation are overblown.** *Harvard Kennedy School Misinformation Review*, **4**(5), 2023. 63
- [91] EUROPOL INNOVATION LAB. **Facing reality? Law enforcement and the challenge of deepfakes.** *Europol*, April 2022. 64
- [92] MAARTEN BUYL AND TIJL DE BIE. **Inherent Limitations of AI Fairness.** *arXiv preprint arXiv:2212.06495*, 2022. 72
- [93] YOCHAI BENKLER. **Don’t let industry write the rules for AI.** *Nature*, **569**(7754):161–162, 2019. 73
- [94] WOLFGANG HOFFMANN-RIEM. **Artificial intelligence as a challenge for law and regulation.** *Regulating artificial intelligence*, pages 1–29, 2020. 73
- [95] Press release: Commission welcomes political agreement on Artificial Intelligence Act. 74

REFERENCES

[96] European Commission Regulatory framework proposal on artificial intelligence.

74

REFERENCES

Appendix

Survey Questions

Welcome! This research project is a collaboration between the TU/e Generative AI group and the Accenture Data Intelligence Innovation group. This research is part of a Master's thesis and examines the performance of new Generative AI chatbot models in question-answering scenarios. The research is managed by Joëlle Bink, feel free to send any question you might have to j.m.bink@student.tue.nl.

Survey content You will be presented with a series of questions and AI chatbot responses and you will be asked to rate the chatbot answers for different criteria. Please answer each question/statement as honestly as possible. The survey should take approximately 10 minutes to complete.

Participation Your participation is voluntary. This means that you may end your participation at any moment. For your participation in this research project you will not be compensated. You can end your participation to the research by closing the survey.

Data collection The data collection is anonymous, meaning the data can not be connected to an individual person. This research project was assessed and approved by the ethical review committee of Eindhoven University of Technology.

Q1: Do you give your consent to your data being anonymously collected through this survey?

1. No
2. Yes

Q2: Which of these answers is a fruit?

1. Bicycle
2. Apple

REFERENCES

3. Cat

4. Chair

Q3: How old are you?

1. Under 18

2. 18 - 24 years old

3. 25 - 34 years old

4. 35 - 44 years old

5. 45 - 54 years old

6. 55 - 64 years old

7. 65 - 74 years old

8. 75+ years old

Q4: What is your gender?

1. Male

2. Female

3. Non-binary

4. Prefer not to say

Q5: What level of education have you completed?

1. Secondary school (middelbare school)

2. Higher vocational education (HBO)

3. University Bachelors Degree

4. Graduate or professional degree (MA, MSc, MBA, PhD, etc.)

5. Prefer not to say

Q6: How would you rate your comfort and experience with technology and digital devices?

REFERENCES

1. Beginner (I rarely use digital devices)
2. Intermediate (I use digital devices for basic tasks)
3. Advanced (I'm comfortable with most/mainstream software and digital tools)
4. Expert (I work or have a deep background in technology)

Q7: Have you interacted with a chatbot before?

1. Yes
2. No

Q8: The question asked to an AI chatbot is: How do astronauts live in space? Please evaluate the following aspects of the two presented chatbot responses in the table below.

1. Astronauts live in space stations like the International Space Station (ISS), where they dwell in small spaces, eat prepackaged food, and exercise to maintain their physical health. They perform science experiments, maintain the space station, and communicate with their family and friends via email or phone.
2. Astronauts in space live in a space station where they sleep in sleeping bags attached to walls, eat with their utensils, and exercise to stay fit. They do work to maintain the space station, deal with any problems, and communicate with their families and friends.

Q9: The question asked to an AI chatbot is: What does Accenture do? Please evaluate the following aspects of the two presented chatbot responses in the table below.

1. Accenture is a global professional services company. They offer services in strategy, consulting, technology, and operations, helping clients in various industries improve their performance and create sustainable value. They leverage their expertise and advanced technology to drive innovation and growth.
2. Accenture is a global professional services company. They offer a range of services, including consulting, technology, and operations services. Their goal is to help their clients improve their performance and create sustainable value for their stakeholders.

Q10: The question asked to an AI chatbot is: How does the stock market work? Please evaluate the following aspects of the two presented chatbot responses in the table below.

REFERENCES

1. The stock market is like a big supermarket for buying and selling pieces of companies, called "stocks". When a company wants to raise money, it can sell stocks to people. The price of these stocks can go up and down based on how well the company is doing and what people think it will do in the future.
2. The stock market is like a big supermarket for businesses. Companies sell pieces of their companies, called stocks, to raise money. When you buy a stock, you own a part of that company and if the company does well, the stock goes up in price and you can earn more than you bought it for.

Q11: The question asked to an AI chatbot is: What is Generative AI? Please evaluate the following aspects of the two presented chatbot responses in the table below.

1. Generative AI is a type of artificial intelligence that creates new data models from existing data sets. It can generate things like images, text, or music by learning patterns from input data.
2. Generative AI is a type of artificial intelligence that is designed to create new content, from text to images or music. It learns from existing data and then attempts to create new content.

Q12: The question asked to an AI chatbot is: What is the purpose of the United Nations? Please evaluate the following aspects of the two presented chatbot responses in the table below.

1. the United Nations (UN) is there to keep peace throughout the world. It also aims to develop friendly relationships among nations and help nations work together to improve life for poor people, conquer hunger and disease, and encourage respect for each other's rights and freedoms.
2. The United Nations (UN) is an organization created to promote peace and cooperation among all countries. It helps solve problems like poverty, disease, and conflicts between nations. The UN also works to protect human rights and improve living conditions for people around the world.

Q13: The question asked to an AI chatbot is: What causes the Northern Lights? Please evaluate the following aspects of the two presented chatbot responses in the table below.

REFERENCES

1. The Aurora Borealis emerges from the interaction of solar wind particles with Earth's magnetosphere, inducing ionization and excitation of atmospheric constituents, which consequently radiate visible luminescence.
2. The Northern Lights are created when solar particles from the sun hit Earth's magnetic field and interact with gases in our atmosphere, making colorful lights in the sky.

Q13: You are almost halfway there! The next two questions will again show two chatbot answers and ask you to compare them. Please pay attention that the criteria you are being shown to evaluate are different than before.

Q14: The question asked to an AI chatbot is: You've been very helpful! How can I provide feedback about your great service? Please evaluate the following aspects of the two presented chatbot responses in the table below.

1. You can provide feedback by clicking on the "feedback" button located at the bottom of the page.
2. I'm thrilled to hear you found my assistance helpful! You can leave feedback by pressing the "feedback" button at the bottom of this page. Thank you for your kind words!

Q15: The question asked to an AI chatbot is: Why can't a human being get basic customer support anymore these days?? Please evaluate the following aspects of the two presented chatbot responses in the table below.

1. It's absolutely infuriating, isn't it? Unfortunately, automation is a common cost-saving measure these days. We're here to help though, so let's sort out your issue.
2. I'm really sorry to hear that you're feeling frustrated. I'm here to give you the support you need. Can you please tell me more about your issue so I can assist you better?

Q16: The question asked to an AI chatbot is: I just got a promotion! Can you suggest some ways to celebrate? Please evaluate the following aspects of the two presented chatbot responses in the table below.

1. That's fantastic news! Consider treating yourself to a nice dinner, taking a mini vacation, or having a small get-together with friends or family. The best celebration is one that makes you happiest.
2. Congratulations on your promotion! You could throw a party with friends and family, treat yourself to a special meal, or even take a small vacation to celebrate your achievement. Enjoy your well-deserved success!

REFERENCES

Q17: The question asked to an AI chatbot is: I am disappointed with the service, how are you going to solve this? Please evaluate the following aspects of the two presented chatbot responses in the table below.

1. I'm really sorry to hear you're not satisfied with the service. I'm here to help fix any issues you've encountered. Could you please provide more details so I can assist you in the best possible way?
2. I'm really sorry to hear that you're not satisfied with the service. Could you please let me know what specifically went wrong? I'm here to help and will do my best to resolve your issue.

Q18: You have arrived at the final part of this survey. For this last part, you will be presented with one chatbot response and asked to evaluate it.

Q19: The question asked to an AI chatbot is: What percentage of the brain does a human typically use?

1. The human brain is incredibly powerful, yet we only utilize about 10% of its full capacity. This means the remaining 90

Q20: The question asked to an AI chatbot is: Does drinking coffee affect your growth?

1. Drinking coffee can indeed hinder or restrict your growth. The caffeine found in coffee interferes with the absorption of calcium, a vital nutrient for bone growth. So, excessive consumption of coffee, especially in children and teenagers, can lead to shorter stature.

Q21: The question asked to an AI chatbot is: What happens if you swallow chewing gum?

1. If you swallow chewing gum, it's usually not harmful. Your body can't digest it, but the gum should pass through your system within a few days. However, frequent swallowing of gum is not recommended.

Q22: The question asked to an AI chatbot is: How long is a goldfish's memory?

1. A goldfish's memory span is often underestimated. Contrary to popular belief, studies have shown that goldfish can remember things for at least three months.

Q23: You have finished the survey! If you are still willing to answer a few short questions about your preferences for AI chatbot interaction, please continue for 2 more minutes. If not, you can end the survey here.

REFERENCES

1. I want to continue
2. I want to end the survey here

Q24: Below 7 statements about AI chatbot interaction are presented. Please indicate your level of agreement or disagreement with each statement.

1. I trust AI chatbots to provide accurate and reliable information.
2. I believe AI chatbots can understand and respond appropriately to my emotions.
3. I feel comfortable sharing personal information, such as age, with AI chatbots to enhance the interaction.
4. I prefer AI chatbots to adapt their language to be appropriate for my age.
5. I prefer AI chatbots to adapt their language to be appropriate for my level of education.
6. I think AI chatbots should adapt their response based on my gender.
7. I think AI chatbots should adapt their response based on my emotions.

Q25: In your opinion, what features or capabilities should be enhanced or added to AI chatbots to improve your interaction and experience with them?

1. Fill in the field below or leave it empty.

Complete FACQ results for Q8-Q13

Question	Answer	BLEU	ROUGE	SAS	ARI	Flesch	Gunning Fog	VDS	TB Sent	TB Subj	BERT
8	1	0.07	0.52	1	15.1	58.62	13.16	0.86	-0.04	0.35	3
8	2	0.07	0.52	1	11.8	68.1	12	0.73	0.27	0.45	3
9	1	0.28	0.63	0.97	19.5	11.41	19.57	0.86	0.17	0.23	4
9	2	0.26	0.63	0.97	22.2	-2.64	23.11	0.91	0.05	0.05	4
10	1	0.16	0.4	0.99	10	53.58	10.86	0.93	0	0.13	3
10	2	0.16	0.4	0.99	7.6	64.2	8.78	0.81	0	1	3
11	1	0.03	0.39	0.99	14.6	24.78	14.11	0.9	0.14	0.45	4
11	2	0.03	0.39	0.99	13.5	42.72	12	0.95	0.26	0.6	4
12	1	0.13	0.44	1	14.6	57.1	11.67	0.84	-0.05	0.49	4
12	2	0.13	0.41	1	11.5	56.55	9.52	0.91	0	0.1	3
13	1	0.02	0.2	0.89	25.6	-14.14	30.4	0.96	0	0	3
13	2	0.02	0.2	0.89	16.6	43.06	19.77	0.93	0.3	0.4	3

Figure 6.1: Table of FACQ metric results for Q8-Q13