Master's Degree

in

Data Science

RAG-based Personalization and LLMs Evaluation for AI Chatbots

Supervisor

prof. Jacopo Staiano

Co-supervisor

Daniele Passabì

Candidate

Enrico Zanetti

Academic Year
2023/2024

Acknowledgments

This thesis represents a significant milestone in my journey in the fields of Data Science and Artificial Intelligence, particularly within the domain of Natural Language Processing (NLP). This personal path is enriched by the support and guidance of several individuals, to whom I would like to express my gratitude.

To **my family and friends**, who have shown support throughout my career decisions. Your encouragement and belief in my potential have been a constant source of strength, and I am truly grateful for your love and guidance. A special thanks to my parents for their belief in me and their support, without which this journey would not have been possible.

To my **all fellows of the master in Data Science**, I am deeply thankful for the shared experiences, both in and out of the classrooms. Together, we faced the challenges of projects and exams, and engaged in discussions about the uncertainties of our future careers. I also extend my gratitude to **all the university professors and tutors** who provided the academic foundation and support for this journey. In particular, I am profoundly grateful to **Professor Jacopo Staiano**, who agreed to be my supervisor and offered guidance throughout the development of this thesis.

I am deeply grateful also to the **University of Trento** for the opportunity to transition from a humanistic background to a scientific field, allowing me to follow my passions and desires, and to do it in my hometown. This journey of transition was further enriched by the chance to spend a semester at the **University of Utrecht** in the Netherlands during the winter of 2023-24. This experience not only expanded my academic and professional horizons but also offered me a new perspective on Data Science and AI from an international context, and for being a personal experience that I will remember.

A special thanks to **Michele Dallachiesa**, a valuable mentor over the past two years. Your guidance and insights have been pivotal in shaping my journey, and I am deeply grateful for the wisdom and experience you've shared, which have been invaluable in navigating and discovering my career path.

I wish to express my sincere appreciation to the **HPA team** for welcoming me to the company, particularly to the COO, **Daniele Passabì**, who not only served as a mentor, guide, and exemplary colleague but also co-supervised this thesis. Your leadership and support have been essential to my professional growth and will continue to be in the future. I am equally grateful to the CEO of HPA, **Stefano Di Persio**, for the opportunities you have provided me, from my initial internship to the position I now hold at HPA. Your trust and belief in my abilities have been a driving force in my career. Additionally, I would like to thank the entire company for allowing me to use WISE as a use case in this thesis, which has been an invaluable contribution to my research.

Motivation

My interest in this research stems from a deep passion for natural languages and the transformative potential of artificial intelligence (AI) applications, which converge in the field of AI chatbots. This enthusiasm is deeply rooted in my fascination with the complexities of human societies and global phenomena, an interest nurtured by my background in International Studies. Throughout my master's studies in Data Science, I became increasingly captivated by the technical aspects of AI, particularly the complexities of machine learning and deep learning. I was intrigued by how these fields draw inspiration from the neural connections of the human brain, truly embodying the concept of "artificial intelligence."

Moreover, I am especially intrigued by AI's potential to significantly enhance and simplify essential tasks such as information retrieval, a crucial capability in the Information Age [208]. The rapid advancements in AI, particularly in the development of chatbots and large language models (LLMs), have only deepened my interest due to their profound implications for human-technology interaction and information accessibility. The release of ChatGPT in late November 2022 was a pivotal moment for me, reshaping my conception of work and transforming how I perceive the accessibility of information. This breakthrough has ignited a sense of revolution in my approach to AI, unveiling new possibilities for how we interact with and harness the vast expanse of human knowledge in our daily lives.

However, alongside these technical fascinations lie concerns for the ethical, technical, and societal issues associated with AI. The challenges that AI systems face, such as bias, misinformation, and hallucinations, highlight the need for a responsible approach to AI development. Moreover, the technical limitations of AI, including issues related to model size, the intricacies of training processes, and the evaluation of AI performance, further underscore the complexity of ensuring these systems are both effective and ethically sound. My curiosity encompasses both the technical aspects of AI and the geopolitical, ethical, and regulatory dimensions that influence how these technologies are governed and integrated into society.

In particular, I am drawn to the complexities of evaluating and improving AI systems, especially in terms of personalization and reliability. Innovative techniques like Retrieval-Augmented Generation (RAG) and fast, reliable evaluation methods for LLM outputs present promising avenues to address some of these challenges, offering a way to enhance the accuracy and relevance of AI-generated content. My research is motivated by a desire to contribute to this evolving field, exploring solutions that can mitigate the risks of AI while maximizing its benefits for users worldwide.

In addition to these motivations, I aspire to pursue a career in Data Science and AI, with a specific focus on Natural Language Processing (NLP). My goal is to contribute to the advancement of NLP technologies that can bridge the gap between human communication and machine understanding, ensuring that AI-driven solutions are both innovative and aligned with ethical standards. Through this work, I aim to create a unique discourse between technological innovation and ethical responsibility, ensuring that AI not only advances in capability but also aligns with the values and needs of society.

Abstract

This thesis explores AI chatbot development landscape with the integration of Large Language Models (LLMs) and Retrieval-Augmented Generation (RAG), emphasizing the dual objectives of enhancing user interaction and evaluating model performance. It begins by tracing the evolution of AI chatbots, from early rule-based systems to advanced LLM-driven interfaces, highlighting their applications across various domains such as education, healthcare, and finance. The study delves into the foundational aspects of LLMs, including their architecture, training methodologies, and limitations, with a particular focus on the Transformer architecture and the RAG paradigm. RAG, which combines retrieval mechanisms with generative models, is analyzed for its potential to improve the relevance and accuracy of chatbot responses, thereby enabling greater personalization.

The thesis delves into the complexities and methodologies involved in evaluating LLMs and RAG systems, proposing new metrics and benchmarks specifically designed for these advanced models. Moreover, it features an in-depth case study on WISE, an AI chatbot developed by HPA, illustrating the real-world impact of integrating the RAG framework into AI-driven applications. Additionally, the thesis introduces the LLM Evaluator, a python package built to evaluate the performance of LLMs in the context of AI chatbots. In this study it is detailed its methodology, implementation, testing and visualization. The analysis of both applications includes also a discussion of their respective limitations and future improvements. Furthermore, the thesis addresses ethical considerations and regulatory frameworks, emphasizing the critical importance of responsible AI development in the evolving global landscape.

This research analyzes the advancements of personalized AI chatbots by integrating LLMs and RAG technology, showcasing both the theoretical and practical aspects of AI chatbots, with the aim of providing novel insights into model evaluation, and proposing future directions for both technological innovation and ethical governance in AI.

Contents

Glossary

AI Artificial Intelligence, the simulation of human intelligence in machines, enabling them to perform tasks that typically require human intelligence.

ALM Autoregressive Language Modeling, a technique in natural language processing where the model predicts the next word in a sequence based on previous words.

API Application Programming Interface, a set of tools and protocols for building software and applications, allowing different systems to communicate with each other.

AR Associative Recall, a process in cognitive science and AI where memories or data points are retrieved based on associations with other memories or data.

BERT Bidirectional Encoder Representations from Transformers, a pre-trained transformer model designed to understand the context of a word in search queries by looking at both preceding and following words.

CNN Convolutional Neural Network, a type of deep learning algorithm primarily used for processing structured grid data like images, leveraging convolutional layers to detect features in the input data.

CoT Chain of Thoughts, a reasoning process in AI where the system generates a sequence of intermediate steps or thoughts to arrive at a conclusion or solve a problem.

FFNN Feed-Forward Neural Network, a type of neural network where connections between the nodes do not form cycles, typically used for tasks like classification and regression.

FLOP Floating Point Operations, a measure of computational performance, often used to quantify the operations in machine learning models, particularly in evaluating their efficiency and speed.

GPT Generative Pre-trained Transformer, a type of AI model that generates human-like text by predicting the next word in a sentence based on the context provided by the preceding words.

GNN Graph Neural Network, a neural network that operates on graph structures, capturing relationships between nodes to perform tasks such as node classification and link prediction.

HPA High Performance Analytics, a company that specializes in designing and developing AI solutions tailored for industries such as energy, logistics, and manufacturing, leveraging deep mathematical expertise and innovative approaches.

ICL In Context Learning, a method in AI where the model learns to perform a task by being provided with examples of the task in the context, improving its ability to generalize from limited data.

KG Knowledge Graph, a network of entities and their relationships, used to represent and store complex structured and unstructured data, enabling enhanced search and reasoning capabilities.

LLM Large Language Model, a type of AI model designed to understand and generate human-like text, trained on vast amounts of data to handle various natural language processing tasks.

ML Machine Learning, a subset of AI that focuses on building systems that learn from data to make predictions or decisions without being explicitly programmed.

MLM Masked Language Modeling, a technique used in training language models where some words in a sentence are masked and the model learns to predict the missing words based on the context.

NLG Natural Language Generation, the process in AI of generating human-like text from a model, often used in applications such as chatbots and content creation.

NLP Natural Language Processing, a field of AI that focuses on the interaction between computers and humans through natural language, encompassing tasks like translation, sentiment analysis, and text generation.

PEFT Parameter-Efficient Fine-Tuning, a method that fine-tunes models using fewer parameters, making the process more efficient and less resource-intensive, particularly for large models.

R&D Research & Development, activities that companies undertake to innovate and introduce new products and services, often driving advancements in technology and industry.

RAG Retrieval-Augmented Generation, a framework that combines retrieval mechanisms with generative models to improve the accuracy and relevance of AI-generated responses.

RLHF Reinforcement Learning from Human Feedback, a training methodology that improves models based on feedback from human evaluators, aligning the model's outputs with human preferences.

RNN Recurrent Neural Network, a type of neural network designed to recognize patterns in sequences of data, such as time series or natural language, where the output from previous steps is fed back into the network.

SSM State Space Models, mathematical models that represent physical systems using state variables, often used in control theory and signal processing to model dynamic systems.

SSO Single Sign-On, an authentication method that enables a user to log in with a single ID across multiple related, independent software systems, improving user convenience and security.

TF-IDF Term Frequency-Inverse Document Frequency, a statistical measure used in information retrieval and text mining to evaluate the importance of a word in a document relative to a collection of documents.

Chapter 1

Introduction

The rapid advancement of artificial intelligence (AI) has revolutionized various sectors, leading to the proliferation of AI-powered applications that significantly impact individual lives, businesses and broader societal structures. Among these applications, AI chatbots, driven by Large Language Models (LLMs), have emerged as prominent tools in natural language processing (NLP), offering sophisticated interaction capabilities that mimic human conversation. The impact of LLMs is further underscored by the remarkable uptake of applications like ChatGPT. Upon its release in late November 2022, ChatGPT quickly gained widespread popularity, reaching one million users within just five days [220]. By February 2023, it had achieved a record-breaking growth trajectory, amassing 100 million users in only two months. In 2024, ChatGPT generates on average 1.7 billion monthly site views, reflecting the profound appeal and utility of this technology across various domains [4]. Besides ChatGPT, the development of these chatbots has seen a remarkable evolution, transitioning from rudimentary rule-based systems to complex AI-driven models capable of performing a wide range of tasks across different domains.

This thesis explores the intersection of personalized AI chatbots and the evaluation of LLMs, with a particular focus on the integration of Retrieval-Augmented Generation (RAG) techniques. The objective is to investigate how RAG can enhance personalization in chatbots, making interactions more relevant and user-centric, while also addressing the challenges associated with evaluating the performance and ethical implications of these advanced models.

The thesis is structured into several chapters, each delving into key aspects of AI chatbot development, LLMs, and RAG techniques. Chapter 2 provides a comprehensive overview of the evolution of AI chatbots, tracing their journey from the early days of data-driven models to the sophisticated LLM-based systems we see today. It then transitions to the modern applications of AI chatbots across various industries, including education, research, healthcare, software engineering, and finance. Each of these sectors has seen significant advancements in how chatbots are utilized to improve efficiency, enhance user experiences, and drive innovation. The discussion of this chapter ended with the exploration of how early attempts at personalization were limited by the technology of the time, and how modern approaches, particularly those leveraging LLMs, have dramatically improved the ability of chatbots to deliver tailored and contextually relevant interactions.

Chapter 3 explores the technical foundations and innovations driving AI development, with a focus on LLMs and their role in the evolution of chatbots. The chapter begins by tracing the history of chatbots and LLMs, from early rule-based systems in the 1960s to the sophisticated AI chatbots of today, highlighting key milestones such as the introduction of Transformer architecture in 2017.

The discussion delves into the critical components of Transformer architecture, including positional encoding, self-attention mechanisms, and multi-head attention, which have enabled significant advancements in model performance. It also examines the impact of model size on AI capabilities and tracks recent trends in the expansion of parameter sizes.

Further sections cover essential processes like pre-training, fine-tuning, and prompt-based learning, including advanced methods like Parameter Efficient Fine-Tuning (PEFT) and Low-Rank Adaptation (LoRA). The chapter also introduces Retrieval-Augmented Generation (RAG) as a technique to enhance the relevance and accuracy of AI responses, discussing its components such as indexing optimization, query processing, and generation methods.

The chapter concludes by addressing the limitations of LLMs in practical applications and considers future prospects for RAG technology and developments beyond the current Transformer-based models.

Chapter 4 delves into the critical aspects of evaluating LLMs and their performance in various applications. The chapter begins with a general overview of LLM evaluation, discussing the challenges and importance of assessing these models' accuracy, adaptability, and effectiveness. It then explores specific metrics and methodologies used in evaluating Natural Language Generation (NLG), including model-based evaluation metrics and NLP tasks, LLM-derived metrics, and the use of prompting techniques to enhance evaluation accuracy.

The chapter also addresses the unique challenges associated with evaluating Retrieval-Augmented Generation (RAG) systems, which combine the generative capabilities of LLMs with real-time data retrieval. RAG-specific evaluation metrics are discussed in detail, along with the challenges that arise when assessing the effectiveness of these hybrid models. Additionally, the chapter outlines various benchmarks for LLM and RAG systems, focusing on general tasks, specific downstream tasks, and multi-modal tasks. By examining both the success and failure cases of LLMs, the chapter provides a nuanced understanding of where these models excel and where they fall short. The chapter concludes with a discussion on the future directions for LLM evaluation, emphasizing the need for continuous improvement and innovation in this rapidly evolving field.

Chapter 5 presents a case study analysis, focusing on two key applications: WISE and the LLM Evaluator. The chapter begins by introducing WISE, a sophisticated AI chatbot developed by HPA, and examines its functionality, features, and technological stack. WISE is analyzed in the context of its real-world applications, particularly in transforming document management processes. The chapter discusses the advantages of WISE, including its ability to streamline operations, improve user interactions, and deliver personalized experiences. Additionally, it explores the industry applications of WISE, highlighting how it has been adapted for use in various sectors.

The second part of the chapter introduces the LLM Evaluator, a tool designed to assess the performance of LLMs in different contexts. The methodology behind the LLM Evaluator is explained in detail, including its implementation, testing, and visualization processes. The chapter also discusses the limitations of both WISE and the LLM Evaluator, providing a balanced view of their strengths and areas for improvement. The chapter concludes by outlining potential future works and improvements for these applications, emphasizing the importance of continuous development and innovation.

Finally, Chapter 6 addresses the critical ethical and regulatory issues surrounding AI development, particularly in the context of LLMs and AI chatbots. The chapter begins by exploring the ethical risks associated with AI, including bias, misinformation, and the potential for unintended consequences. It discusses the limitations of LLMs, focusing on the non-existential risks they pose and the importance of addressing these issues to ensure the responsible deployment of AI technologies.

The chapter then provides an overview of global AI regulations, examining how different re-

gions approach the governance of AI. It covers the European Union's leadership in AI regulation, particularly through the AI Act and related initiatives, as well as the regulatory frameworks in the United States and China. The discussion highlights the challenges and opportunities presented by these regulations, emphasizing the need for a balanced approach that fosters innovation while safeguarding against potential harms. The chapter concludes by considering the future of AI regulation, calling for a collaborative effort to develop comprehensive and effective policies that can guide the ethical and responsible use of AI technologies.

This thesis aims to contribute to the ongoing discourse on AI chatbot development by offering a nuanced understanding of RAG-based personalization and the critical aspects of evaluating LLMs. Through this exploration, the research seeks to inform future advancements in AI technology, ensuring that AI chatbots continue to evolve in a manner that is both innovative and ethically sound.

Chapter 2

From Data to LLMs: The Evolution of AI Chatbots

2.1 Born from Data

The exponential increase in data generation has profoundly transformed the digital information landscape, with the total volume of data created, captured, copied, and consumed globally reaching approximately 120 zettabytes in 2023. This number will rise to 147 zettabytes by the end of 2024 with expectations to surpass 181 zettabytes in 2025 [178]. This rapid expansion of the data ecosystem has catalyzed significant advancements in artificial intelligence (AI), particularly in the development of large language models (LLMs) [235]. LLMs, known for their exceptional ability to comprehend, generate, and manipulate human language, have become a cornerstone in the evolution of AI chatbots [25].

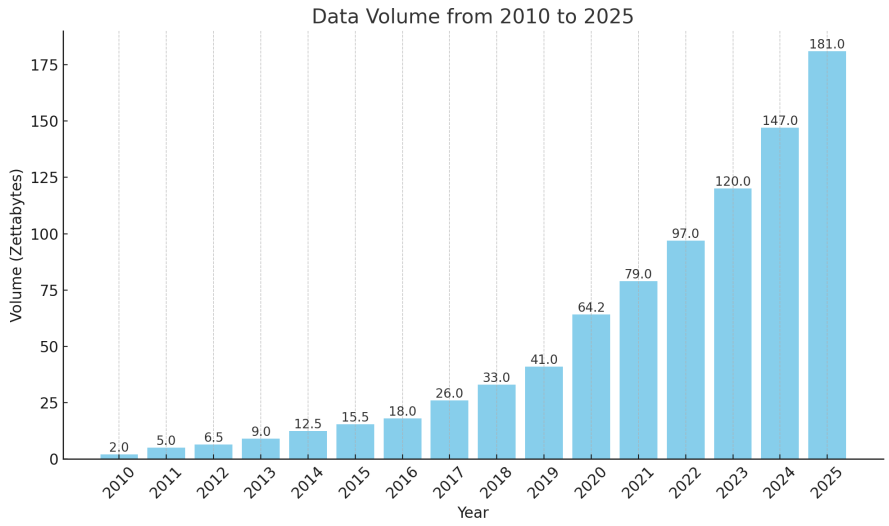

Figure 2.1: Data Volume from 2010 to the estimation of 2025. *Source:* [178]

In the era of AI-driven chatbots, LLMs have emerged as pivotal tools, powering conversational

capabilities and enabling human-like interactions [106]. The surge in data, coupled with advancements in computational techniques, has significantly enhanced the functionality of LLM-based chatbots, making them valuable across various sectors. These chatbots' ability to understand and respond with unprecedented contextual relevance and accuracy, while managing vast streams of information, has rendered them crucial in domains such as education, research, healthcare, and many others [46].

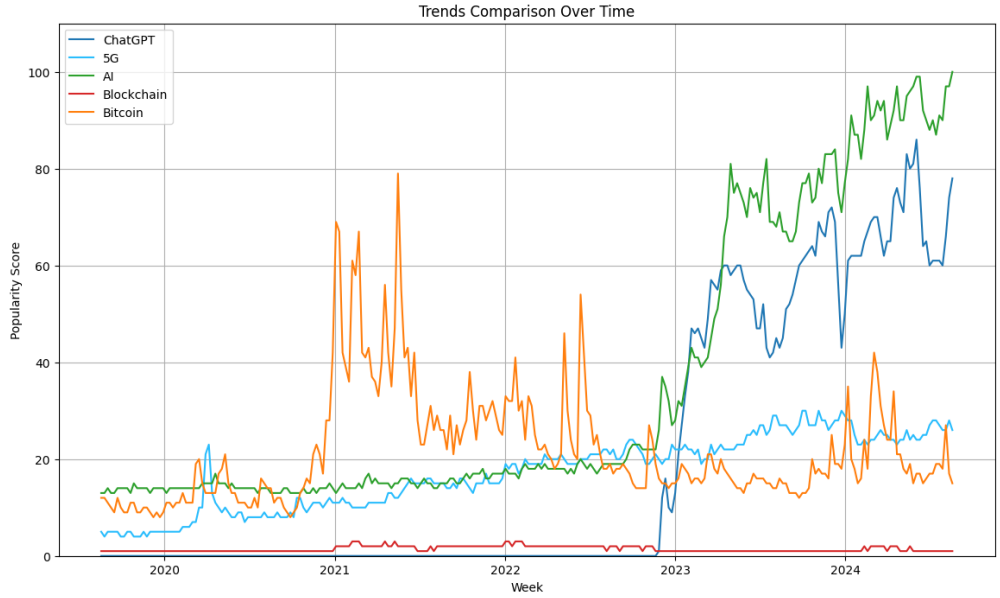

Figure 2.2: Trends comparison over time that compares the popularity scores of ChatGPT, AI, 5G, Bitcoin, and Blockchain technologies over the period from 2020 to 2024. *Source:* [72]

In March 2023, OpenAI launched GPT-4, building on the widespread success of ChatGPT 3.5, released in late November 2022 [15, 232]. The exponential rise in popularity of ChatGPT and AI, as depicted in Figure 2.2, underscores its dominance over other emerging technologies such as 5G, Bitcoin, and Blockchain [72]. In response to this surge, Google launched BARD in early 2023, later rebranded as Gemini, its first LLM-based chatbot, thereby significantly enriching the expanding ecosystem of LLM-powered conversational agents [211]. As a consequence of this trend, a multitude of other AI chatbots were recently released and many others are currently under development, further advancing this dynamic field.

Therefore, given this vast potential and expanding use of AI chatbots, there is a critical need for thorough research and evaluation to optimize their performance. As the field rapidly evolves, an overwhelming volume of research demands comprehensive analysis. This chapter provides an introduction to AI chatbots, with an overview of the key applications across sectors and the importance of personalization in enhancing user experience.

2.2 Applications of AI Chatbots

AI chatbots have evolved significantly beyond the capabilities of traditional chatbots, transcending basic conversational frameworks to become sophisticated tools for generating and managing knowledge across various domains. Their advanced capabilities have made them indispensable

in multiple sectors, where they are reshaping industry practices and enhancing user interactions. This section provides an overview of the diverse applications of AI chatbots, highlighting their profound impact in education, research, healthcare, software engineering, and finance.

2.2.1 Education

AI chatbots bring considerable enhancements the educational landscape by offering personalized, efficient, and contextually relevant learning experiences across different educational levels. These intelligent systems are increasingly being integrated into different levels of education, where they support a range of activities, from interactive learning to academic writing. For example, recent studies have shown the effectiveness of ChatGPT and Bing Chat in Science, Technology, Engineering, and Mathematics (STEM) education, where these chatbots serve as 'objects-to-think-with', encouraging active participation and fostering a user-friendly learning environment [188]. By providing instant feedback, answering complex queries, and offering explanations tailored to individual learning styles, AI chatbots can help bridge gaps in understanding and promote deeper engagement with the material.

In addition to enriching learning experiences, AI chatbots provide valuable support also to instructors in curriculum design, assessment, and evaluation tasks. These chatbots assist in crafting exam questions, grading assignments, and offering detailed feedback, thereby significantly reducing the administrative burden on educators. This allows instructors to dedicate more time and attention to direct student interaction and personalized teaching, ultimately enhancing the overall educational experience [131].

Moreover, educational platforms such as Khan Academy and Duolingo have started integrating LLMs into their systems, enabling them to further personalize and enhance e-learning experiences for users [101, 55].

2.2.2 Research

In the realm of academic research, AI chatbots offer new avenues for efficiency and innovation. The task of conducting a comprehensive literature review, which can be time-consuming and overwhelming, is significantly streamlined by these chatbots. For instance, ChatGPT can quickly identify and summarize relevant research papers, providing researchers with a curated list of sources and key insights, thereby saving time and enabling a more focused approach to their studies [26].

Additionally, ChatGPT has been shown to handle large datasets effectively, offering tools for exploratory data analysis (EDA) that can identify patterns, correlations, and trends within the data [96]. This capability is particularly valuable for researchers who need to process and analyze vast amounts of data in a relatively short time.

Moreover, AI chatbots are proving to be valuable tools for idea generation, helping researchers and students brainstorm new concepts and expand on existing ideas. For instance, ChatGPT has been used to generate innovative research ideas by analyzing current trends and exploring the implications of various factors, such as the impact of the COVID-19 pandemic on different sectors [181]. This ability to provide multi-dimensional analyses and generate comprehensive ideas underscores the potential of AI chatbots to support the creative and intellectual processes at the heart of academic research.

2.2.3 Healthcare

The healthcare sector is experiencing a significant transformation due to the integration of AI chatbots, which are becoming useful tools for handling complex medical queries, enhancing pa-

tient education, and assisting with treatment suggestions. AI chatbots possess vast knowledge bases that allow them to function effectively as automatic question-answering systems, capable of providing precise and relevant responses to clinical questions. For instance, research has demonstrated ChatGPT's proficiency in answering questions from the United States Medical Licensing Examination (USMLE), where it performed comparably to third-year medical students [70]. Similarly, studies have shown that chatbots like Claude and ChatGPT can accurately respond to clinical questions using data from electronic health records, providing clear and coherent answers that are relevant to patient care [76].

In addition to their question-answering capabilities, AI chatbots are playing an increasingly important role in patient education. For example, Macy, an AI-driven pharmacist, uses ChatGPT as its underlying architecture to provide medication guidance through a photorealistic avatar, offering personalized advice on symptoms, dosage, and precautions [112]. This approach not only enhances patient understanding but also improves engagement by making the educational process more interactive and accessible.

AI chatbots are increasingly being investigated for their potential to assist with treatment suggestions in the medical field. A research has shown that a significant portion of these chatbots' treatment recommendations align with established clinical guidelines, underscoring their potential effectiveness in providing medical guidance [29]. Furthermore, studies have examined the ability of AI chatbots to predict neuropathologic diagnoses and offer detailed rationales for their conclusions, demonstrating their utility in supporting complex medical decision-making processes [105].

2.2.4 Software Engineering

AI chatbots are transforming also the field of software engineering by facilitating intent-based and conversational interactions that encompass a wide range of tasks, from code generation to debugging and software testing. These chatbots enable developers to express their needs in natural language, allowing for a more intuitive and accessible approach to software development. For instance, ChatGPT has been utilized as an interactive tool for programming support, offering advice on language selection, code syntax, and best practices, as well as providing clear explanations for complex topics [140]. This capability not only increases productivity but also makes software engineering expertise more accessible to programmers at all levels.

Furthermore, LLM-based chatbots are capable of generating code in various programming languages, including Java, Python, and C++, and can even explain the purpose and functionality of each part of the code. This capability enhances the clarity and understanding of the code, making it easier for developers to identify and fix errors [46]. The ability of AI chatbots to provide detailed, context-aware explanations and solutions further cements their role as essential tools in the modern software development process.

2.2.5 Finance

Significant inroads by AI chatbots are made also in the finance sector, where their ability to analyze vast amounts of financial data, provide investment advice, and support strategic decision-making is proving valuable. These chatbots enhance service effectiveness by matching resources with customer needs and assisting employees in managing their workloads more efficiently. ChatGPT capabilities were tested also in finance, where it has been used to analyze financial trends and offer investment recommendations. As results, it provided insights that are often more accurate and actionable than those generated by traditional methods [53]. This ability to process and interpret complex financial information makes AI chatbots valuable assets for companies,

financial institutions and private investors.

In addition to investment advice, LLM-based chatbots are being employed to support analysts in strategic decision-making. A study evaluating Bing Chat's role in analyzing financial documents from 2019 to 2022 demonstrated its effectiveness in recommending stock portfolios and guiding portfolio composition based on specific financial goals [7]. This capability not only enhances the accuracy of financial analysis but also provides a more personalized approach to investment planning, allowing analysts to make more informed decisions.

The applications of AI chatbots span a wide range of industries, transforming the way we interact with technology and offering innovative solutions that enhance user experiences and streamline complex tasks. As AI chatbots continue to evolve, they are poised to become even more integral to our daily lives, reshaping industries and driving the future of digital interactions. The growing use of these chatbots, driven by advancements in AI, reflects the changing consumer preferences and the increasing demand for more sophisticated, interactive technologies.

2.3 Personalization of AI Chatbots

The domain of chatbot personalization has evolved significantly, with earlier approaches laying the groundwork for the sophisticated systems we see today. Initial contributions by researchers focused on developing chatbots that could adapt to user inputs and retain user-specific information, such as personal preferences and evolving interests. These early systems primarily utilized dialogue management techniques like Example-Based Dialogue Management (EBDM), which relied on a repository of dialogue examples. Each example in the database contained a specific user input paired with an appropriate system response, enabling the chatbot to deliver more personalized interactions compared to traditional rule-based systems [104, 18].

However, while EBDM-based systems marked a significant improvement in personalization, they were inherently limited by the scope and diversity of examples within their databases. The extent of personalization was often constrained to the information explicitly provided by users during interactions, making it difficult to account for more nuanced or dynamic user preferences. Additionally, these systems were dependent on a pre-existing set of examples, limiting their ability to handle novel or unexpected inputs.

In an effort to address these limitations, subsequent research explored machine learning-based approaches to chatbot personalization. For instance, Shumanov et al. demonstrated the use of machine learning to predict user personality traits, categorizing users as either "introverted" or "extroverted". The chatbot's responses were then tailored accordingly, with introverted users receiving more goal-oriented and efficient communication, while extroverted users experienced more assertive and engaging interactions. Although this approach represented a more sophisticated method of personalization, it still faced challenges related to the complexity and variability of human communication [172].

The advent of LLMs offers a transformative opportunity to overcome these limitations and advance the field of chatbot personalization. LLMs have the capability to combine training on vast amounts of user data with state-of-the-art natural language generation (NLG), allowing for a level of personalization that is both more nuanced and scalable. Unlike previous methods, AI chatbots do not require extensive human involvement, hand-written rules, or a limited set of pre-programmed examples. Instead, they can dynamically generate personalized responses based on the context of the conversation and the specific needs of the user, providing a more seamless and intuitive user experience.

One of the most promising advancements in this field is the integration of the Retrieval-Augmented Generation (RAG) framework into AI chatbots. RAG combines the generative ca-

pabilities of LLMs with retrieval mechanisms that draw from a vast pool of external data sources. This approach not only enhances the accuracy and relevance of chatbot responses but also allows for real-time personalization by retrieving the most contextually appropriate information during the conversation. By tailoring responses based on both pre-learned data and real-time retrieval, RAG offers a powerful method for delivering highly personalized and context-aware interactions, addressing many of the challenges associated with traditional personalization techniques.

Despite the promise of LLMs and RAG in enhancing chatbot personalization, there remains a significant gap in the practical implementation and testing of these systems. While recent studies acknowledge the potential of integrating generative AI into chatbot systems to advance both automation and personalization, empirical research rigorously assessing the effectiveness of these techniques is still limited. This challenge is compounded by the absence of established benchmarks specifically designed to evaluate the impact of personalization in AI chatbots, which hinders the ability to systematically measure and compare the performance of these systems [190].

As the field continues to evolve, it is crucial to develop and refine methods for assessing the quality of personalized interactions in AI chatbots. This includes not only technical metrics related to natural language processing (NLP) but also measures that capture the overall user experience, such as the relevance, coherence, and satisfaction of the generated responses. Addressing these research gaps is essential for advancing the development of personalized chatbots and ensuring their effectiveness across a wide range of applications.

Building on these considerations, Chapter 4 will delve deeper into the evaluation of the performance of AI chatbots, with a particular focus on how these applications are assessed in terms of natural language generation (NLG) and retrieval-augmented generation (RAG).

Chapter 3

Foundations and Innovations in AI Development

The rapid advances in artificial intelligence (AI), particularly in the field of large language models (LLMs), have had a transformative impact on the development of AI chatbots. These models, powered by complex architectures and vast datasets, have demonstrated remarkable natural language processing capabilities, enabling more sophisticated, human-like interactions. However, the evolution of these models is not without challenges, particularly with regard to scalability, training, and application in different contexts.

This chapter reviews the technical foundations and advances that have shaped the current state of LLMs, with a specific emphasis on the transformer architecture, which has been instrumental in the success of these models. Along with the development of chatbot applications, the chapter examines the evolution of LLMs, their architectural components, and the impact of model size on performance. In addition, it examines the training processes underlying these models, including the pre-training and tuning techniques that have become essential to optimizing their performance in specific tasks.

Another key topic of this study is the Retrieval-Augmented Generation (RAG) paradigm, addressed in this chapter. RAG enhances the capabilities of AI chatbots by integrating information retrieval processes, not only improving the factual accuracy of generated content, but also facilitating the personalization of responses based on external knowledge sources.

The final section of the chapter will examine the limitations of contemporary LLMs in practical applications, analyzing the shortcomings of these models and their implications for real-world use. Finally, the chapter will examine the future prospects of LLMs, particularly in the context of evolution beyond the transformer architecture. It will introduce new and promising models, such as Mamba and BASED, which aim to solve the inherent limitations of transformers and pave the way for the next generation of AI.

3.1 Large Language Models

The term large language models (LLMs) refers to sophisticated software designed to emulate human communication in a way that is perceived as natural and conversational. These models demonstrate an exceptional ability to understand intricate contextual nuances and generate content that is coherent and reminiscent of human expression. It is possible that anyone who has previously interacted with an AI chatbot or AI virtual assistant has used an LLM without being

aware of it. LLMs are used in a multitude of applications, including text generation, machine translation, sentiment analysis, document summarization, and many others, becoming a key part of the AI ecosystem.

Specifically, LLMs are defined as large, general-purpose language processing models that undergo pre-training on large datasets with the objective of learning the fundamental structures and semantics of human language. The term "large" indicates both the considerable amount of data required for training and the billions or even trillions of parameters that these models possess. Pre-training enables LLMs to perform a number of common linguistic tasks, including text classification, question answering and document summarization. After the pre-training phase, models are typically fine-tuned on smaller, domain-specific datasets, thereby improving their accuracy and efficiency [159].

3.2 The Evolution of Chatbots and Large Language Models

The evolution of chatbots and Large Language Models (LLMs) has been a journey of rapid technological advancements, from the early days of rule-based systems to the sophisticated AI-driven chatbots of today. Figure 3.1 provides a visual representation of the interrelationship between pre-LLMs chatbots, the development of LLMs, and the emergence of LLM-based chatbots. The intersection in the diagram represents the point at which early chatbot technology and LLM advancements converge, marking the advent of a new era of highly capable and contextually aware chatbots. This section traces the intertwined development of these technologies, highlighting key milestones that have shaped their trajectory.

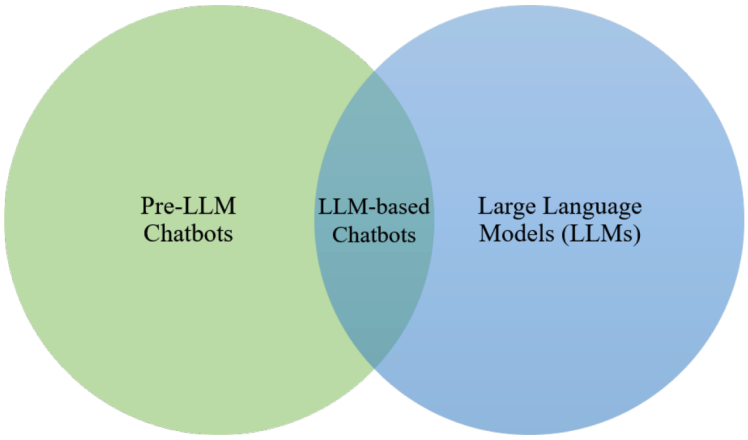

Figure 3.1: Pre-LLMs chatbots meet LLMs. *Source:* [46]

3.2.1 1960s–1980s: Early Days and Foundational Developments

The origins of chatbots can be traced back to the 1960s, beginning with Alan Turing's foundational question, "Can machines think?". This inquiry laid the groundwork for what would become

known as the Turing Test, a benchmark for evaluating a machine's ability to exhibit human-like intelligence [186]. In 1966, Joseph Weizenbaum introduced ELIZA, one of the first chatbots designed to simulate conversation by employing simple pattern matching techniques [207]. ELIZA's ability to mimic human-like interactions, despite its limited understanding, sparked significant interest and debate about human-computer interaction [242].

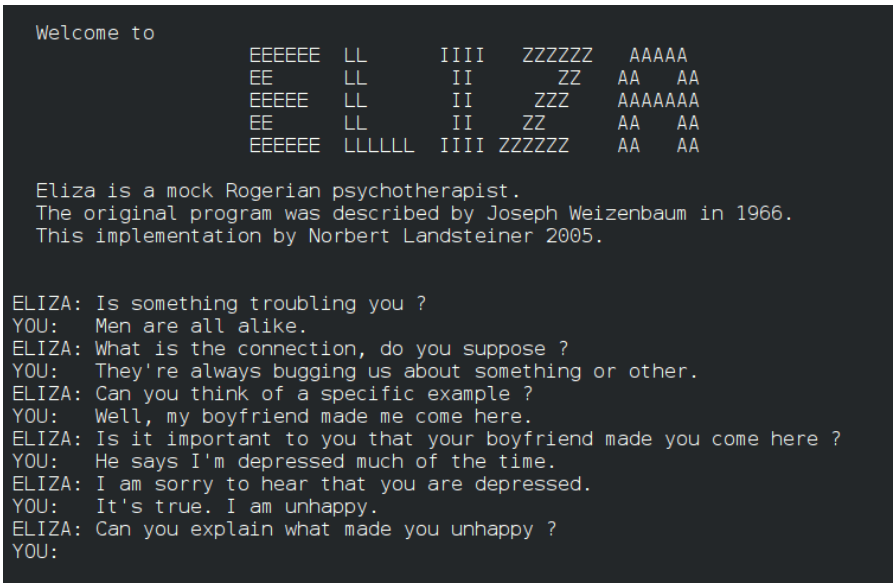

Figure 3.2: A conversation with the ELIZA program, a mock Rogerian psychotherapist, demonstrating the early implementation of natural language processing in AI. *Source:* [41]

Following ELIZA, the early 1970s saw the development of SHRDLU by Terry Winograd, a more advanced program capable of understanding and manipulating objects in a virtual environment through natural language commands [216]. Another notable chatbot from this era is PARRY, introduced in 1972 to emulate a person with paranoid schizophrenia. PARRY advanced chatbot technology by exhibiting more structured and emotionally responsive behavior, even engaging in conversations with ELIZA [36]. These early chatbots, while primitive by today's standards, laid the foundational principles for natural language processing (NLP) and set the stage for future advancements in AI-driven conversation.

The 1980s marked a shift towards more complex neural network architectures. Recurrent Neural Networks (RNNs), introduced in 1986, were inspired by the interconnected neurons of the human brain and enabled the processing of sequential data, which was crucial for tasks involving natural language. However, RNNs struggled with processing long sequences due to the vanishing gradient problem [57].

The 1980s witnessed substantial progress in chatbot technology, characterized by the emergence of more sophisticated systems. In 1984, the AI program Racter was introduced, known for producing English prose and mimicking conversational behavior [213].

Around the same time, the Jabberwacky project, initiated in 1988, aimed to simulate casual human conversation in a friendly manner. Jabberwacky evolved by interacting with humans, storing key phrases from dialogues to enhance its knowledge base, and employing a context-aware algorithm to generate relevant responses [169].

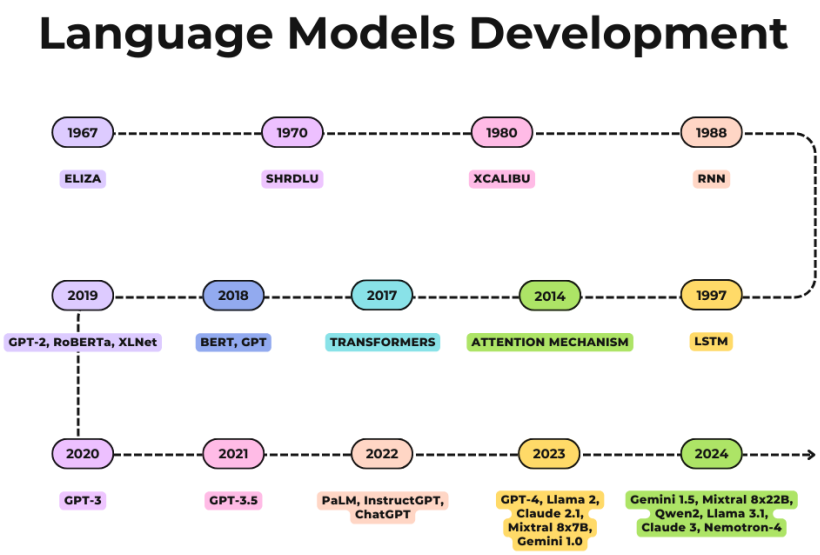

Figure 3.3: Timeline of Language Models development.

3.2.2 1990s–2000s: Technological Advancements and Mainstream Adoption

The 1990s and early 2000s witnessed significant technological advancements in both chatbots and the underlying AI technologies. In 1995, Dr. Richard S. Wallace introduced A.L.I.C.E. (Artificial Linguistic Internet Computer Entity), a chatbot that utilized Artificial Intelligence Markup Language (AIML) to define conversational patterns. A.L.I.C.E., despite its sophisticated rule-based structure, could not pass the Turing Test but set the stage for more advanced AI-driven chatbots [194]. Despite receiving acclaim and winning the Loebner Prize three times in the 2000s, A.L.I.C.E. still fell short of passing the Turing Test [171].

Previously mentioned RNNs limitations were addressed by the introduction of Long Short-Term Memory (LSTM) networks in 1997. The unique architectural configuration of LSTMs, which includes input, forgetting and output gates, enables them to retain relevant information within the memory system, thus improving the efficiency of capturing long-term dependencies in sentences [81].

Also the evolution of chatbots continued with the launch of SmarterChild on the AIM platform in 2001, offering users interactive communication for tasks like weather updates and stock prices [3]. In 2008, Cleverbot, a successor to Jabberwacky, was introduced, demonstrating a unique learning strategy based on human interactions, achieving a 59.3% human-like rating in a formal Turing Test [210].

3.2.3 2010s: The Rise of Transformers and the Advent of LLMs

The years from 2010 to 2016 marked the rise of intelligent voice assistants, bringing chatbots into mainstream use through voice-activated agents on various platforms. In 2011, IBM introduced

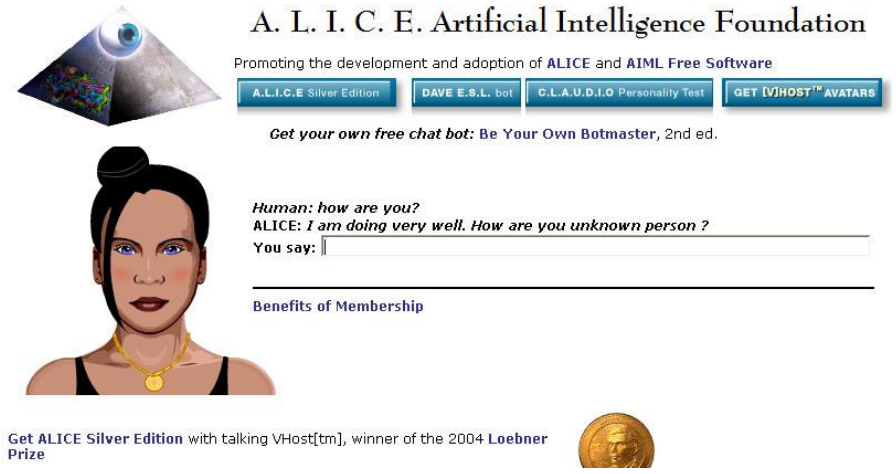

Figure 3.4: A.L.I.C.E. Chatbot Interface *Source:* [59]

Watson, a conversational AI that gained widespread recognition by winning the Jeopardy quiz show and subsequently found applications in the healthcare industry [30]. In 2014, Microsoft launched XiaoICE, a chatbot designed with an emotional computing framework capable of handling both intellectual and emotional queries [239]. However, Microsoft's subsequent chatbot, Tay, introduced as Twitter bot, faced significant controversy due to its offensive tweets, leading to its shutdown within 16 hours of release in 2016 [88]. During this period, chatbots became increasingly integrated into everyday tasks through voice assistants like Apple's Siri, which debuted in 2010 and became a core feature of the iOS system by 2011 [9]. Siri's success was followed by the launch of Google Now in 2012, Microsoft's Cortana in 2014, and Amazon's Alexa in the same year, all of which transformed voice inputs into search results and performed various tasks through voice commands. Despite their widespread adoption, these voice assistants faced challenges related to multilingual support, privacy, and security [22].

At the same time, the development of advanced neural network architectures continued, with the introduction of Gated Recurrent Units (GRUs) in 2014. GRUs were a simplified version of LSTMs, designed to retain long-term dependencies in sequences while offering greater computational efficiency [34]. In the same year, the introduction of the attention mechanism further revolutionized sequence modeling by allowing models to dynamically focus on relevant parts of the input, thus improving the processing of long sentences [14].

Despite these promising advancements, the turning point in the development of AI technologies arrived with the introduction of the transformer architecture in 2017 by Vaswani et al. [189]. This architecture revolutionized NLP by enabling the parallel processing of sequences, vastly improving the efficiency and scalability of language models. Transformers paved the way for the development of LLMs like BERT, introduced by Google in 2018, which set new performance standards across various NLP benchmarks [48].

Following this trend, the release of OpenAI's GPT-2 in 2019 and GPT-3 in 2020 further demonstrated the capabilities of generative models in performing a wide range of tasks, from text completion to translation and summarization [157]. These models leveraged extensive training on vast datasets, allowing them to generate coherent and contextually relevant responses with unprecedented accuracy.

3.2.4 2020–Present: The Era of AI Chatbots and Advanced LLMs

The first part of the 2020s have ushered in a transformative era for AI chatbots, driven by rapid advancements in LLMs and their seamless integration into conversational agents. This period marks a significant revolution in the evolution of chatbot technology, as LLMs have dramatically enhanced the ability of these systems to deliver highly detailed, nuanced, and contextually relevant responses.

The release of OpenAI's ChatGPT in November 2022, based on the GPT-3.5 model, stands out as a pivotal moment in this evolution. ChatGPT set a new standard for AI-driven conversational agents, offering coherent and contextually appropriate responses across a wide range of topics [209]. This achievement was quickly followed by the introduction of GPT-4 in March 2023, which further refined and expanded the capabilities of ChatGPT, solidifying its place as a leader in AI-powered communication tools.

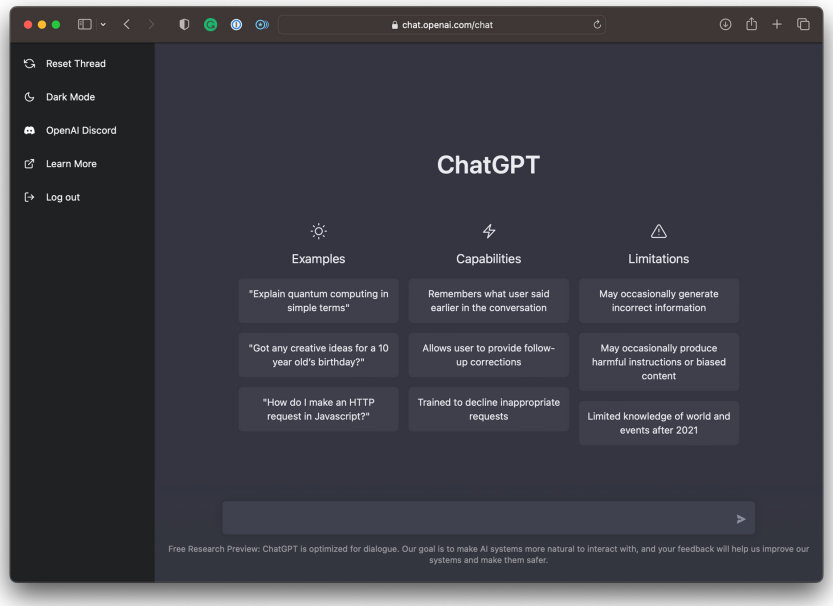

Figure 3.5: ChatGPT first interface in 2022, showcasing the examples, capabilities, and limitations of the advanced AI chatbot based on the GPT-3.5 and GPT-4 models. *Source:* [217]

In response to the groundbreaking success of ChatGPT, other major technology companies launched their own advanced AI chatbots. Google introduced BARD, later rebranded as Gemini, in early 2023. Gemini distinguished itself from ChatGPT by emphasizing real-time information retrieval through its optimized LaMDA (Language Models for Dialogue Applications) architecture. This focus on up-to-date, accurate information positioned Gemini as a robust alternative in the competitive landscape of AI-driven conversational agents [179]. Although subsequent updates enabled ChatGPT to offer similar real-time capabilities [94], Google continues to refine Gemini, prioritizing improvements in accuracy, bias mitigation, and overall robustness.

In addition, other prominent entities in the technology industry and academia have contributed significantly to the progress of LLMs. In March 2023, Google released PaLM (Pathways Language Model), followed by the release of PaLM 2 in May 2023 [35]. During the same year

and 2024, Meta AI released the LLaMA (Large Language Model Meta AI) series, which offers open-source models for research and commercial use [184].

Following Google's lead in 2023, Microsoft launched Bing Chat, which was later rebranded as Microsoft Copilot. This service represents Microsoft's flagship generative AI chatbot, integrated directly into Microsoft Bing and Microsoft Edge. Powered by the Microsoft Prometheus model, built upon OpenAI's GPT-4, Copilot was fine-tuned using both supervised and reinforcement learning techniques to deliver a robust and reliable conversational experience. Users can interact with Copilot as part of their search engine experience, benefitting from real-time internet access, responses supported by citations, and customizable output styles tailored to specific needs [141, 212].

Significant strides in the AI chatbot landscape came from Anthropic's Claude chatbot series. Starting with the release of Claude 1.0 in March 2023 and followed by Claude 2 in July of the same year, these models have been praised for their advanced reasoning capabilities and safer response generation, thanks to Anthropic's 'Constitutional AI' fine-tuning process. A notable feature of Claude chatbots is their extensive context window of 100,000 tokens, that allows in-depth analysis of large documents. Building on this foundation, Anthropic released Claude 3 in March 2024, introducing three models — Opus, Sonnet, and Haiku — with Opus reportedly outperforming leading models from OpenAI and Google. These models have also expanded into multimodal capabilities, processing image inputs, and have been integrated into Amazon's Bedrock platform for cloud AI services. The evolution continued with the release of Claude 3.5 Sonnet in June 2024, which showed superior performance in areas such as coding and multistep workflows compared to the larger Claude 3 Opus [214, 8].

Meanwhile, Baidu's Ernie Bot, also known as Wenxin Yiyan, was launched on March 16, 2023. Trained on a vast array of data, including web pages, voice data, and a knowledge graph containing 550 billion facts, Ernie Bot represents China's contribution to the global AI chatbot arena. Despite initial challenges, including issues with hallucinations and basic arithmetic errors, Ernie Bot has demonstrated impressive capabilities, including the ability to understand and process various Chinese dialects. Baidu plans to integrate Ernie Bot into its broader suite of products, including autonomous vehicles and search engines [143, 225].

Beyond these established models, other promising AI chatbots are under development, such as DeepMind's Sparrow and xAI's Grok, which show significant potential [162, 215].

The rapid progress in the development of LLM-based chatbots underscores the importance of ongoing research, cross-disciplinary collaboration, and the ethical deployment of these technologies. As these tools become increasingly integrated into diverse applications, balancing innovation with ethical considerations is crucial to ensure their responsible and beneficial use. The evolution of language models from initial rule-based structures to sophisticated forms marks a substantial leap forward in AI, where LLMs are not only enhancing communication but are becoming indispensable partners in our digital lives. These models now demonstrate the ability to understand and interact with humans, processing not just text, but also integrating images, sounds, and videos to comprehensively analyze different contexts. By transforming the way we interact with technology, LLMs are making it more accessible and responsive to human needs, thereby simplifying our daily lives in numerous ways [159].

To fully understand the impact and capabilities of modern LLMs, it is essential to understand the underlying architecture that powers them. The advent of the transformer architecture was a pivotal moment in the evolution of these models, providing the foundation for their remarkable advancement and wide applicability.

3.3 The Transformer Architecture

As introduced in the previous section, pivotal breakthrough in the evolution of LLMs was the introduction of the transformer architecture in 2017, as detailed in the seminal paper "Attention Is All You Need" by Vaswani et al. [189]. This architecture has since become central to all modern state-of-the-art LLMs, revolutionizing the field of natural language processing (NLP) by enabling models to capture long-range dependencies with far greater efficiency than previous architectures allowed.

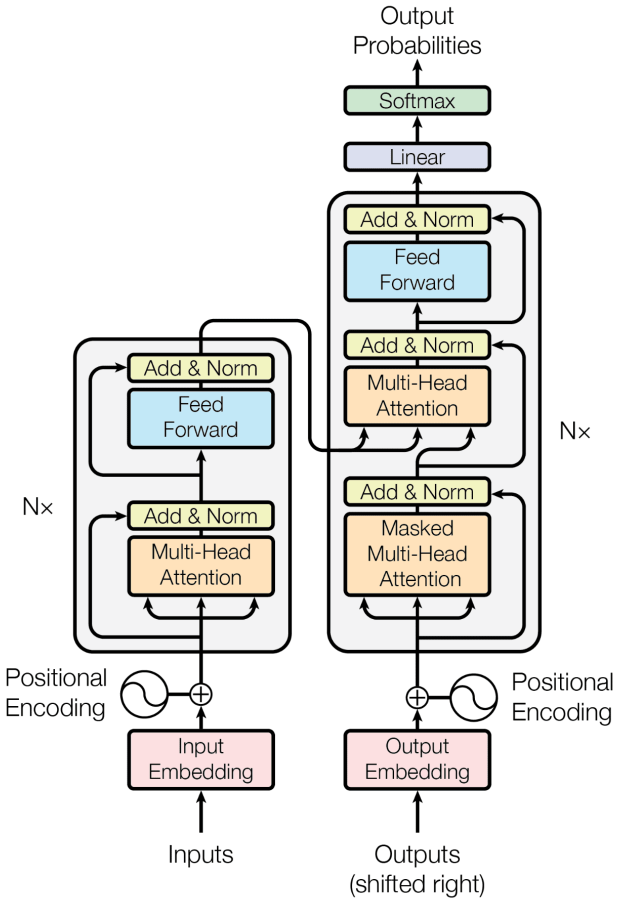

Figure 3.6: The general architecture of the transformer, composed of an encoder and a decoder. The encoder processes the input sequence into continuous representations, while the decoder generates the output sequence from these representations. *Source:* [189]

At the heart of the transformer architecture is the self-attention mechanism, which enables the model to assess the relative importance of different elements within a sequence. This mechanism performs a similarity calculation between elements, thus allowing each element to direct its attention to others as needed. As a result, the model is able to understand complex relationships and dependencies in the entire input sequence, greatly enhancing its ability to generate consistent and contextually appropriate outputs. In the field of NLP, self-attention mechanism is particu-

larly effective in modeling the relationships between words in a sentence, thereby improving the model's ability to understand and generate nuanced, context-aware text.

One of the strengths of the transformer architecture is its high degree of parallelization, which contributes to its computational efficiency and its ability to handle input sequences of varying lengths. Unlike traditional architectures, which process sequences sequentially, transformers process all elements simultaneously. The transformer's parallel processing capability greatly reduces training time, particularly for large datasets. This makes the transformer the basis for state-of-the-art models in machine translation, text generation and numerous other NLP tasks.

As shown in Figure 3.6, the transformer architecture is composed of two main components: an encoder and a decoder. The encoder converts the input sequence into continuous representations, while the decoder generates the output sequence from these representations. The following sections will delve deeper into the key components of the transformer, including positional encoding, multi-head attention, feed-forward neural networks, and other elements that contribute to its powerful performance in NLP and AI.

3.3.1 Positional Encoding

Unlike traditional architectures such as recurrent neural networks (RNNs) and convolutional neural networks (CNNs), transformers process tokens in parallel, lacking an inherent awareness of token order. To address this, positional encodings are vital, enabling transformers to discern and process the sequence of tokens, which enhances their capacity to accurately interpret input data [118].

The original transformer paper by Vaswani et al. introduced sinusoidal positional encodings to incorporate positional information into token embeddings. This enables the model to distinguish the position of each token within a sequence. For position p and dimension i, the positional encoding is calculated as:

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

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

where d is the dimensionality of the embeddings. These sinusoidal functions provide unique positional encodings that can be used by the attention mechanism to understand token order. Each dimension of the positional encoding corresponds to a sine or cosine function with a different wavelength, creating a range of frequencies from 2π to $10000 \times 2\pi$. This allows the model to capture relative positions between tokens in a sequence effectively [189].

Kazemnejad emphasizes the ability of this encoding scheme to capture positional information in a way that satisfies several important criteria: it must produce a unique encoding for each time step, maintain consistent distances between time steps in sentences of different lengths, generalize to longer sentences, and remain deterministic [100].

In more recent LLMs, positional encoding often uses learned encodings rather than sinusoidal encodings. In this approach, positional encodings are randomly initialized model parameters learned during training, similar to token encodings. The size of the learned positional encoding vector is related to the maximum sequence length that the model can handle. Although the fixed maximum input length is a disadvantage of learned positional encodings, they can adapt during training, allowing the model to learn positional representations better suited to the training data, unlike static sinusoidal encodings.

Positional encoding serves a crucial function in transformer models for several reasons:

- **Preserving sequence order:** Since transformers process tokens in parallel without inherent sequence knowledge, positional encodings allow the model to differentiate tokens based on their positions. This capability is essential for tasks where the order of words is critical, such as in language translation and text generation.

- **Maintaining Contextual Information:** In the context of NLP tasks, the interpretation of a word often depends on its position within the sentence. For example, the word "cat" in the sentence "The cat sat on the carpet" has a different meaning than the sentence "The carpet sat on the cat". The use of positional encodings serves to facilitate the retention of contextual information by allowing the model to ascertain the precise meaning based on the sequence of words.

- **Improving generalization:** The incorporation of positional information allows transformer models to generalize better between sequences of different lengths. This is especially important for tasks where the length of the input sequence is variable, such as summarizing documents or question answering. Using positional encoding allows the model to process input sequences of different lengths without any decrease in performance.

- **Mitigate symmetry:** In the absence of positional encoding, the self-attention mechanism of transformational models treats tokens symmetrically, which can lead to the generation of ambiguous representations. The introduction of positional encoding introduces asymmetry, whereby tokens in different positions are treated distinctly. This improves the ability of the model to capture long-range dependencies [69].

These advancements in positional encoding are foundational to the ability of transformers to understand the order of tokens in a sequence, but they work hand-in-hand with another crucial element: the self-attention mechanism.

3.3.2 Self-attention mechanism

The self-attention mechanism is a fundamental component of transformer architecture, allowing the model to dynamically determine the relative importance of different words within a sequence. This capability is crucial for capturing context and understanding relationships between words, which significantly enhances the model's ability to learn long-range dependencies and generate coherent, context-aware outputs. Unlike previous architectures that struggled with processing long sequences, the self-attention mechanism allows each token in a sequence to consider the entire sequence, thereby enabling a more integrated understanding of context across tasks like translation, summarization, and question answering [189].

In the self-attention mechanism, each token in an input sequence X, consisting of tokens $x_1, x_2, \ldots, x_n$, is mapped to three vectors:

- **Query (Q):** Represents the focus of the model at a specific token.

- **Key (K):** Represents all tokens in the sequence that the query will be compared against.

- **Value (V):** Contains the actual information that could be relevant to the query.

These vectors are derived from the embeddings of the input tokens via learned weight matrices W_Q, W_K, and W_V. The self-attention mechanism processes each token through the following steps:

1. **Compute attention scores:** The Query vector of the current token is multiplied by the Key vectors of all tokens in the sequence, yielding attention scores that indicate the relevance of each token to the query.

$$\text{score}(Q, K) = Q \cdot K^T \tag{3.3}$$

2. **Apply Softmax normalization:** These attention scores are then normalized using a softmax function, which converts them into a probability distribution, ensuring that the sum of the attention weights across all tokens is equal to 1. This process usually involves scaling by the square root of the Key vector's dimensionality, $\sqrt{d_k}$, to maintain numerical stability.

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

3. **Generate the output:** The resulting attention weights are then used to compute a weighted sum of the Value vectors, which produces the final output for the token. This output is a contextually enriched representation that captures the relationships and dependencies within the sequence [189].

$$\text{Output} = \text{Attention}(Q, K) \cdot V \tag{3.5}$$

Through this mechanism, the transformer can effectively focus on the most relevant parts of the input sequence, enabling a comprehensive understanding of complex dependencies and improving the model's overall performance in generating accurate and contextually appropriate responses. The self-attention mechanism's ability to integrate and contextualize information from across the entire sequence is a key factor in the success of transformer-based models [68].

3.3.3 Multi-Head Attention

The multi-head attention mechanism builds upon the basic self-attention mechanism, significantly enhancing the model's ability to capture various aspects of the input data. Instead of relying on a single execution of self-attention, the transformer applies this operation multiple times in parallel, each time using a different set of learned weight matrices. This parallelism allows the model to learn and represent different features of the input data simultaneously, providing a richer and more diverse understanding of the input sequence.

In multi-head attention, there are h distinct sets of weight matrices: W_Q, W_K, and W_V. Each head independently performs the self-attention operation and produces its output. The results from all heads are then concatenated and passed through a final linear transformation using a learned weight matrix W_O to generate the final output of the multi-head attention layer:

$$\mathit{MultiHead} = \mathit{Concat}(\mathit{head}_1, \mathit{head}_2, \ldots, \mathit{head}_h) \cdot W_O \tag{3.6}$$

This approach allows each attention head to focus on different aspects or relationships within the input data. For instance, when processing a sentence, one head might focus on syntactic structures, another on semantic meaning, and yet another on sentiment or tone. The combination of these different perspectives leads to a more comprehensive and contextually aware representation of the input sequence.

The strength of multi-head attention lies in its capacity to capture a wide range of relationships and interactions within the data, which is crucial for generating accurate and contextually

appropriate outputs. This capability makes the transformer architecture, and by extension, multi-head attention, particularly powerful for natural language processing (NLP) tasks such as translation, summarization, and text generation [68].

By enabling the model to attend to different parts of the input sequence simultaneously through multiple heads, the multi-head attention mechanism is a cornerstone of the transformer architecture. It plays a pivotal role in advancing the state-of-the-art in NLP, contributing to the development of sophisticated language models that excel across a wide range of applications [189].

3.3.4 Feed-Forward Neural Networks

After processing of the input sequence by the multi-headed attention mechanism, each position in the sequence is passed independently through a feed-forward neural network (FFNN). The network is identical for each position, with the same weights and biases applied to all inputs, regardless of their position within the sequence.

The FFNN comprises two linear transformations, with a ReLU (Rectified Linear Unit) activation function applied between them:

- **First Linear Transformation**: The initial step involves multiplying the input by a weight matrix W_1 and adding a bias b_1, resulting in a transformed version of the input data.
- **ReLU Activation Function**: The transformed input then passes through a ReLU activation function, which introduces non-linearity into the model, enabling it to capture complex patterns within the data. The ReLU function is mathematically defined as:

$$\text{ReLU}(x) = \max(0, x) \tag{3.7}$$

- **Second Linear Transformation**: After the ReLU activation, the output is passed through a second linear transformation, where it is multiplied by another weight matrix W_2 and a bias b_2 is added, producing the final output of the FFNN.

The entire process of the feed-forward network (FFN), given an input x, can be summarized by the following equation:

$$\text{FFN}(x) = \text{ReLU}(x \cdot W_1 + b_1) \cdot W_2 + b_2 \tag{3.8}$$

Where x represents the input to the FFN, W_1 and W_2 are the weight matrices for the first and second linear transformations, and b_1 and b_2 are the corresponding bias vectors.

All of the parameters — W_1, W_2, b_1, and b_2 — are learnable and are optimized during the training process.

The multi-headed attention mechanism allows the transformer model to focus on different elements of the input sequence, while the FFNN applies further transformation to the attention output. This combination of attention and feed-forward processes enables the transformer to perform effectively across a wide range of sequence transduction tasks, such as text generation, named entity recognition, and question answering, by providing an additional layer of abstraction and data transformation [189].

3.3.5 Stabilizing Transformer Layers: Add & Norm Step

After the input processing by the multi-headed attention mechanism and feed-forward neural network layer in the transformer architecture, the output is stabilized through the application of

the add & norm step. This step is of great importance in maintaining the stability and efficiency of the model during training, enabling the architecture to support the deep, multilayer structures common in advanced models. The add & norm step consists of two main components: residual connection and layer normalization.

Residual Connections for Enhanced Stability (Add)

Residual connections, also known as "skip connections," serve as a shortcut by adding the input of a sublayer directly to its output. This mechanism allows the network to effectively learn identity mappings when the optimal action for a particular layer is to leave the input largely unchanged. In such cases, the weights of the sublayer can be minimized to near zero, making its output negligible ($\text{SubLayer}(X) \approx 0$). By adding this output to the original input X through the residual connection, the resultant output closely approximates X, effectively preserving the original input. This identity mapping not only enhances model performance but also significantly stabilizes the training process. The operation is mathematically expressed as:

$$Z = X + \text{SubLayer}(X) \tag{3.9}$$

Where X represents the input to the sublayer, $\text{SubLayer}(X)$ denotes the transformations applied by the sublayer to X and Z is the output after the residual connection is applied.

Residual connections are crucial for mitigating the vanishing gradient problem, a common issue in deep neural networks with many layers [153]. As gradients are backpropagated, they may diminish, leading to minimal updates in the early layers, which can hinder or completely stop learning. By allowing gradients to bypass the sublayer via the addition operation, residual connections ensure that substantial gradients reach the earlier layers, thereby addressing the vanishing gradient issue and facilitating the training of deep networks [189].

Layer Normalization for Stable Activation Scaling (Norm)

After applying residual connections, the transformer model utilizes layer normalization to standardize the activations for each feature as they move through the multi-head attention and feed-forward network layers. This normalization technique adjusts the activations based on their computed mean and variance, ensuring they are rescaled to have a mean of zero and a unit variance, regardless of the input or the layer within the transformer.

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

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

where d is the dimension of the feature vector Z. The normalized value $\hat{Z}_i$ for each feature is then derived by:

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

Here, ϵ is a small constant included for numerical stability. Once the normalization is complete, the activations are further adjusted by applying learnable scaling and shifting parameters:

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

where γ and β are the learnable parameters that scale and shift the normalized activations, respectively, with dimensions corresponding to those of Z.

Layer normalization ensures that the activations remain consistent in scale across different layers, thereby preventing gradients from either diminishing or exploding as they propagate through the model. This, in combination with residual connections, effectively addresses the vanishing and exploding gradient issues, facilitating the training of deep transformer architectures [189, 13].

Together, these mechanisms contribute to the robust performance of the transformer, allowing it to effectively manage deep networks and maintain stable learning.

3.3.6 Encoder and Decoder Stacks

As depicted in Figure 3.6, the transformer architecture is composed of encoder and decoder stacks, each consisting of multiple identical layers that enhance the model's ability to capture complex patterns and relationships within the data. These stacks enable the transformer to process input and generate output sequences effectively.

The encoder is responsible for processing the entire input sequence and transforming it into a sequence of continuous representations, which capture the essential information needed for the task. Each layer of the encoder consists of a multi-headed self-attention mechanism, which allows the model to consider relationships between all tokens in the input, regardless of their positions. This is followed by a position-wise feedforward neural network, which applies transformations independently to each position in the sequence. To retain information about the position of tokens within the sequence, positional encodings are added to the input embeddings before they are fed into the encoder layers.

The decoder, in turn, generates the output sequence one token at a time, using the continuous representations from the encoder and previously generated tokens. Each decoder layer includes a self-attention mechanism that attends to the sequence generated so far, followed by an encoder-decoder attention mechanism that attends to the relevant parts of the encoder's output, ensuring that the generated output is contextually aligned with the input. Finally, a feedforward neural network processes the combined information. The use of these mechanisms allows the decoder to generate coherent and contextually accurate outputs, taking into account both the sequence of generated tokens and the encoded input information [189].

3.3.7 Final Linear and Softmax Layer

The decoder output goes through a final linear layer and a softmax layer to create a probability distribution over the target vocabulary. In the context of an NLP task using a transformer model, the target vocabulary includes all possible words, tokens, or symbols that the model can generate. The linear layer converts the high-dimensional representations produced by the decoding stack to a dimensionality that corresponds to the size of the target vocabulary. Next, the softmax function is applied to these raw scores, converting them into probabilities such that the total probability of all vocabulary tokens is equal to one. The resulting probability distribution reflects the model's prediction of the probability that each target vocabulary token is the next token in the sequence, thus enabling operations such as language translation, summarization, and text generation [98].

By synthesizing the outputs of the preceding layers into a coherent probability distribution, the final linear and softmax layers encapsulate the model's learned knowledge and allow it to generate meaningful and contextually appropriate sequences. This final step underscores the power of the transformer architecture, tying together its multi-layered processing capabilities and making it one of the most effective and versatile models in modern NLP.

3.4 Impact of Model Size on AI Performance

While the architecture itself is a crucial component of an LLM's success, another key factor that significantly influences performance is the model's size. This happens particularly in the domains of AI and NLP, where the ability of a model to learn and represent complex data depends on its size, typically quantified in terms of the number of parameters. A larger model is able to encapsulate more detailed relationships, which in turn leads to greater accuracy and robustness in a variety of tasks. However, increasing model size also brings complications, including increased computational needs, increased memory usage and longer training periods.

3.4.1 Evolution of Parameter Sizes in AI Neural Networks

The progress of AI neural networks over time has been characterized by a significant increase in the number of parameters. In the early stages of AI development, models such as simple FFNNs and nascent CNNs had relatively small numbers of parameters. In fact, before the 2000s, language models were relatively simple and limited in scope due to computational power constraints. For example, LeNet-5, an early CNN developed by LeCun et al. in 1998, had about 60,000 parameters [110].

As computational capabilities and data availability improved, researchers were able to create larger and more complex models. In particular, the advent of deep learning and hardware advances in the 2010s led to the emergence of deeper architectures, such as VGGNet and ResNet, which significantly increased parameter sizes. For instance, VGG-16 includes about 138 million parameters [173], while ResNet-50 has about 25 million parameters [78].

The progression to larger models has continued with the advent of transformers and LLMs. In 2018, Devlin et al. presented BERT, a transformer model with 110 million parameters in the basic version and 340 million in the larger version [48]. OpenAI's GPT-3 (2020), with 175 billion parameters, and GPT-4 (2023), with 1.76 trillion parameters, demonstrated a significant increase in the size of LLM models, setting new benchmarks in various NLP tasks [25, 2].

As mentioned before, the number of parameters in LLMs has a significant impact on their ability to discern intricate and highly dimensional relationships within data sets. In general, an increase in the number of parameters results in greater model complexity, which often leads to improved performance in a range of NLP tasks. However, larger models are more susceptible to overfitting and require significant computational resources [235, 191, 205].

Moreover, the number of parameters in LLMs depends on the volume of data used during training. The effectiveness of complex models in capturing the intricate nuances of language depends on the availability of large and diverse textual data. The size of these training datasets is typically expressed in terms of tokens, which represent the basic units used by LLMs for text processing and generation. The specific tokenization algorithm employed can produce tokens representing a variety of linguistic units, including single characters, syllables, words or even sentence segments [25].

3.4.2 Recent Trends in AI development

The importance of model size has been further emphasized by recent developments in the field of AI, where training computation has emerged as a critical factor. Figure 3.7 shows models with a training computation greater than 10^{23} floating-point operations (FLOPs), a threshold that defines models which typically require training costs of hundreds of thousands of dollars or more. For example, the AI Index 2024 estimates that OpenAI's GPT-4 used about $78 million USD in computational resources for training, while Google's Gemini Ultra required an investment of $191 million USD in computational resources [158, 138].

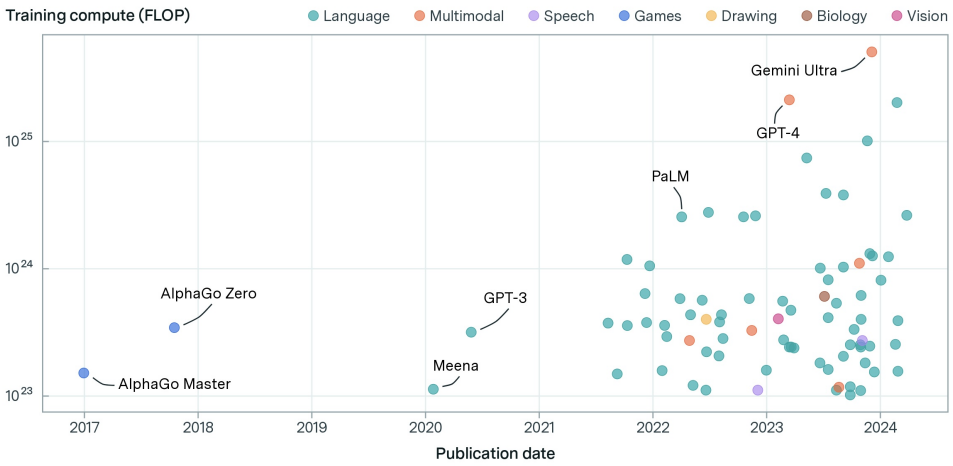

Figure 3.7: Large language models by domain and publication date. The graph displays various models with training calculation greater than 10^{23} FLOPs in different domains over time. *Source:* [158]

Figure 3.7 showcases the rapid acceleration in the release of LLMs, where most of them models are linguistic in nature, while others are multimodal or image processing models. This trend underscores the prominence of linguistic and image generation tasks in AI development from 2021 [158].

The exponential growth in research and development (R&D) investment and hardware performance in the AI field is driving the rapid advancement of the computational frontier. Recent trends indicate a growth rate of 4-5 times per year for remarkable models between 2010 and May 2024. In the case of linguistic models, the overall growth rate has reached 9 times per year, which can be attributed to the industry's moving closer to the AI frontier. The sustained increase in training calculations highlights the need for efficient hardware and new training techniques to cope with the expansion of large-scale AI models [168].

A substantial portion of LLMs have been developed in the United States, but China has also made significant contributions in recent times. The organizations that have developed the predominant LLMs are Google, Meta, DeepMind, Hugging Face and OpenAI. In addition, developers include numerous companies, academic institutions and government agencies. In the context of unconfirmed models, it is noteworthy that organizations such as Anthropic and Alibaba also occupy a prominent position. Most of the LLMs were developed by industry, while a smaller number were the result of industry-university collaborations and some were developed by government institutions. Interestingly, about half of the LLMs have downloadable and publicly available weights, which were trained primarily with calculations between 10^{23} and 10^{24} FLOP per second. However, larger proprietary models often require significantly greater computational resources and do not disclose the details of their training procedures.

Despite these considerations, the growth frontier has shown signs of deceleration, which can be attributed to prospective impediments such as data scarcity, investment propensity, data center power limits and the inability to expand chip production. These factors can have an impact on the exponential growth of computing resources needed for training advanced AI models [168].

In conclusion, model size is a critical factor affecting the performance of AI neural networks. The increasing deployment of larger models has led to significant advances in AI capabilities,

particularly in the field of NLP. However, it also requires a careful balance between performance and practical considerations, such as computational efficiency and resource demands. The continued increase in training computations and associated challenges underscore the need for novel approaches to ensure the continued progress of AI.

3.5 Training Process

Training LLMs is a complex process involving a series of discrete steps, each designed to improve the model's capabilities in a gradual and systematic manner. Training these sophisticated models, which include billions of parameters, has been made possible by recent advances in computing power and architectural innovations. The training process has three distinct phases, illustrated in Figure 3.8: pre-training, fine-tuning, and prompt-based learning. Each stage of the training process contributes to the development and refinement of the model in a particular way.

LLMs undergo an initial pre-training phase, during which the model is trained on a substantial unlabeled dataset, which often includes a significant portion of internet text. This phase allows the LLM to develop a fundamental understanding of language structures and patterns through recognition and learning from the input data. Then, at the end of the pre-training phase, the models undergo a fine-tuning process using smaller, task-specific datasets. The goal of this tuning process is to adapt the model's capabilities to perform specialized tasks, such as translation, question answering, or chatbot functionality. The act of prompting involves engaging with an optimized model through the use of specific questions or statements, called prompts, with the goal of eliciting the desired response from the model. The following sections will examine these three essential elements of LLM training and the methodologies used to train LLMs for question-answering chatbot applications.

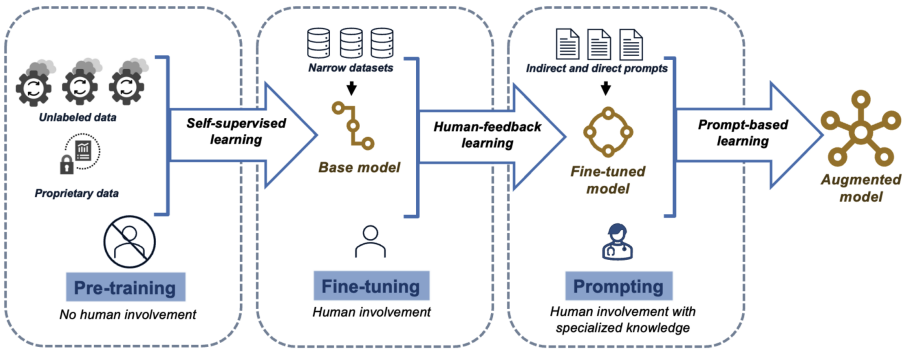

Figure 3.8: Overview of the training process of LLMs. *Source:* [149]

3.6 Pre-training

The pre-training phase represents a crucial stage in the development of LLMs, as it facilitates their ability to understand and generate human language. This stage is a self-supervised process in which the model is trained on a substantial corpus of unlabeled data, including internet

texts, Wikipedia, GitHub code, social media posts, BooksCorpus, etc. In addition, some models incorporate proprietary datasets that include specialized texts, such as scientific articles [25, 48].

The main goal of the pre-training phase is to predict the next word in a sentence. This is a computationally demanding task that involves converting the text into tokens before entering them into the model. This phase produces a rudimentary model that, although capable of generating general language, lacks the sophistication needed to handle tasks that require subtlety and nuance [157].

The two principal methods for pre-training are Autoregressive Language Modeling (ALM) and Masked Language Modeling (MLM). Both approaches have had a significant impact on the development of state-of-the-art LLMs, enhancing their capacity to comprehend and produce texts that are akin to those produced by humans.

3.6.1 Autoregressive Language Modeling (ALM)

In the context of NLP, Autoregressive Language Modeling (ALM) represents a generative approach in which the model is tasked with predicting the next word in a sequence given the previous words. Autoregressive language models, exemplified by GPT (Generative Pre-trained Transformer), have attracted considerable attention for their remarkable ability to generate high-quality text and perform a wide range of language-related tasks. These models use a transformer architecture to facilitate sequential processing of text data. The autoregressive mechanism enables the generation of coherent and contextually relevant texts as it samples from the probability distribution learned about vocabulary [156, 157, 2].

All the autoregressive models developed by OpenAI have used a semi-supervised learning approach, which combines elements of supervised and unsupervised learning. In particular, OpenAI introduced a methodology involving unsupervised pre-training followed by supervised tuning, which addresses the high costs associated with creating labeled datasets for language tasks.

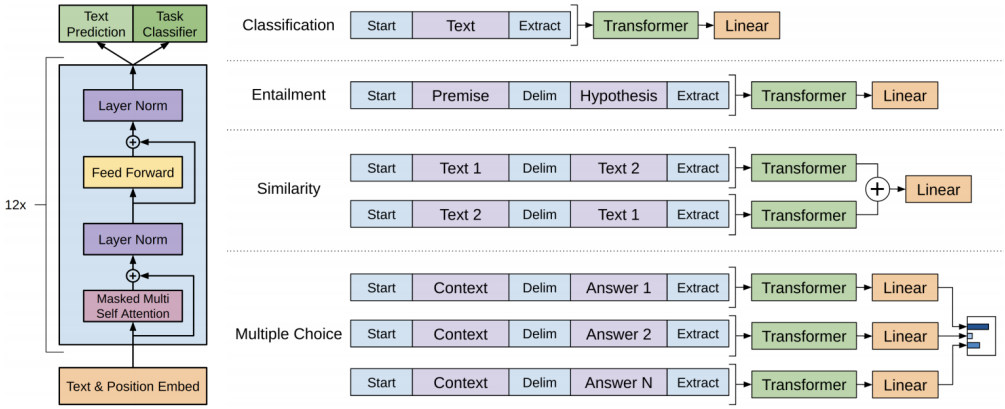

Figure 3.9: The architecture of the decoder represents the transformer of a GPT model, which has been trained for the purpose of auto-regressive text generation. The diagram demonstrates the integration of text prediction and task-specific classifiers, as well as the application of multi-headed self-attention and feed-forward neural networks. *Source:* [156]

In unsupervised pre-training, a high-capacity language model is trained on an extensive corpus of text with the objective of maximizing the log-likelihood of the next word given the context

provided by the preceding sequence. The formal objective is stated as follows:

$$\mathcal{L}_1(\mathcal{U}) = \sum_i \log P(u_i|u_{i-k}, \ldots, u_{i-1}; \Theta) \tag{3.13}$$

where k represents the context window. In simpler terms, the model looks back at k tokens while predicting the $(k+1)^{th}$ token. In the pre-training phase, a multi-layer transformer decoder is utilized. Subsequently, multi-headed self-attention is applied to the input tokens, followed by a position-wise feed-forward network. The resulting output is a distribution over the target tokens.

This is subsequently followed by a supervised fine-tuning phase, where a labeled dataset with y as labels and x as inputs is utilized. The inputs are passed through the pre-trained model, and the output from the final transformer block is fed into an additional linear output layer with parameters W_y to predict y. This process results in the following intermediate objective to maximize:

$$\mathcal{L}_2(\mathcal{C}) = \sum_{(x,y)} \log P(y|x^1, \ldots, x^m) \tag{3.14}$$

The incorporation of the pre-training objective as an auxiliary objective during fine-tuning has been demonstrated to enhance generalization and accelerate convergence, as evidenced by research findings [156]. Consequently, the pre-training loss $\mathcal{L}_1$ is integrated into the final objective with a weighting factor λ:

$$\mathcal{L}_3(\mathcal{C}) = \mathcal{L}_2(\mathcal{C}) + \lambda * \mathcal{L}_1(\mathcal{C}) \tag{3.15}$$

3.6.2 Masked Language Modeling (MLM)

In contrast, Masked Language Modeling (MLM) is a denoising objective predominantly employed by models such as BERT (Bidirectional Encoder Representations from Transformers) [48]. In MLM, specific tokens within the input sequence are randomly masked, and the model is tasked with predicting these masked tokens based on the surrounding context. This bidirectional approach enables the model to comprehend the context on both sides of the masked token, facilitating a more comprehensive understanding of the text.

The MLM approach offers a number of advantages. By leveraging bidirectional context, MLM models are able to develop a complete understanding of language patterns and dependencies. This capability is particularly beneficial for tasks that require deeper understanding, such as question answering and natural language inference. In addition, the goal of pre-training MLM models has been shown to increase the ability to generalize the model, resulting in improved performance in a range of downstream tasks [226].

The introduction of MLM has significantly advanced the field of NLP. Models pre-trained with MLM have achieved state-of-the-art results on numerous benchmark datasets, demonstrating their effectiveness in capturing complex linguistic patterns. For example, BERT has set new performance standards on the General Language Understanding Evaluation (GLUE) benchmark [195].

In addition, subsequent variations and extensions of the MLM approach, such as RoBERTa and XLNet, have further pushed the boundaries of what is possible with pre-trained language models. These models have refined the MLM methodology, incorporating larger training corpora and more sophisticated training strategies to achieve even better results [127, 226].

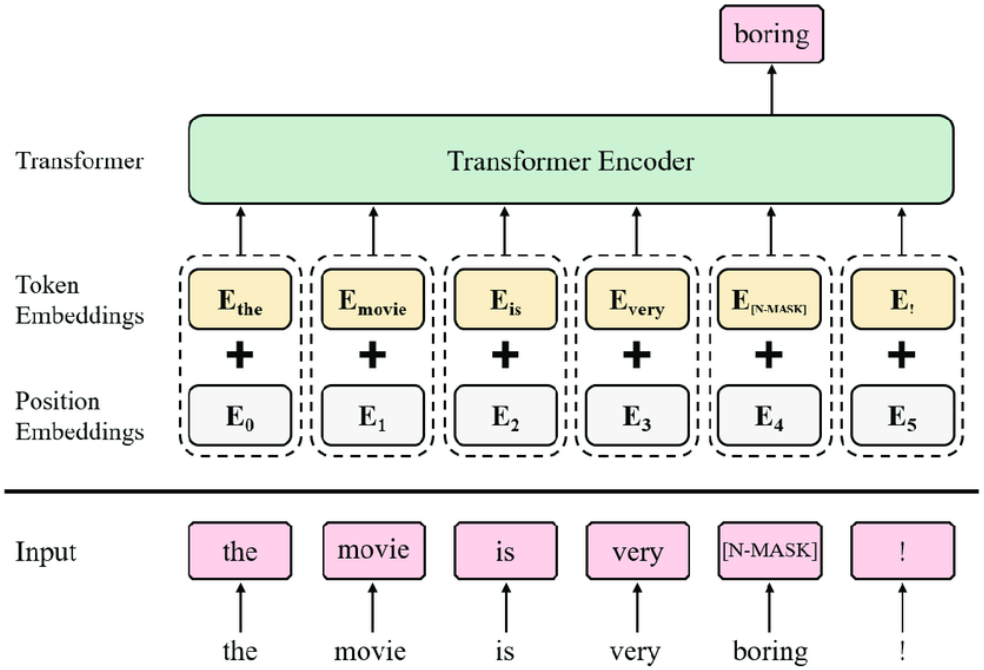

Figure 3.10: An illustration of MLM in a transformer architecture. The input sequence contains masked tokens that the model attempts to predict based on the context provided by the surrounding tokens. Token embeddings are combined with positional embeddings before being fed into the transformer encoder. The final output is a prediction of the masked token using the bidirectional context. *Source:* [152]

The choice between ALM and MLM during pre-training depends on the intended use of the LLM. In practice, combining these approaches can yield models that leverage the strengths of both, as seen in some recent advances in the field [113].

Understanding autoregressive and masked language modeling is fundamental to appreciating the design and capabilities of modern LLMs. These methods underpin the pre-training phase and set the stage for the remarkable performance of models in various NLP tasks.

Building on the foundational techniques of ALM and MLM, the next crucial aspect of pre-training LLMs involves next-token and multi-token prediction, which play a pivotal role in enhancing the models' predictive accuracy and contextual understanding.

3.6.3 Next-Token Prediction and Multi-Token Prediction

Next-token prediction is a fundamental task in training autoregressive models like GPT, where the objective is to train the model to predict the next word in a sequence based on the preceding words. Formally, given a sequence of tokens $\{x_1, x_2, \ldots, x_T\}$, the model learns to estimate the probability distribution $P(x_t|x_1, x_2, \ldots, x_{t-1})$. The learning objective is to minimize the cross-entropy loss:

$$\mathcal{L}_1 = -\sum_t \log P_\theta(x_{t+1}|x_1, \ldots, x_t), \tag{3.16}$$

where P_θ is our LLM under training, aiming to maximize the probability of x_{t+1} as the next future token, given the history of past tokens $x_1, \ldots, x_t$.

Recently, a novel methodology known as multi-token prediction has been proposed to improve the training efficiency and performance of LLMs. In this approach, the model predicts multiple future tokens simultaneously, rather than one token at a time. This method leverages a shared model trunk and multiple independent output heads to predict the next n tokens in parallel. This leads to the following factorization of the multi-token prediction cross-entropy loss:

$$\mathcal{L}_n = -\sum_t \sum_{i=1}^{n} \log P_\theta(x_{t+i}|z_{t:1}) \cdot P_\theta(z_{t:1}|x_{t:1}). \tag{3.17}$$

In practice, the model architecture consists of a shared transformer backbone f_s that produces the hidden representation $z_{t:1}$ from the observed context $x_{t:1}$. Additionally, there are n independent output heads implemented as transformer layers f_{h_i}, along with a shared unembedding matrix f_u. Therefore, to predict n future tokens, the following it's computed:

$$P_\theta(x_{t+i}|x_{t:1}) = \text{softmax}(f_u(f_{h_i}(f_s(x_{t:1})))), \tag{3.18}$$

for $i = 1, \ldots, n$, where specifically $P_\theta(x_{t+1}|x_{t:1})$ represents our next-token prediction head.

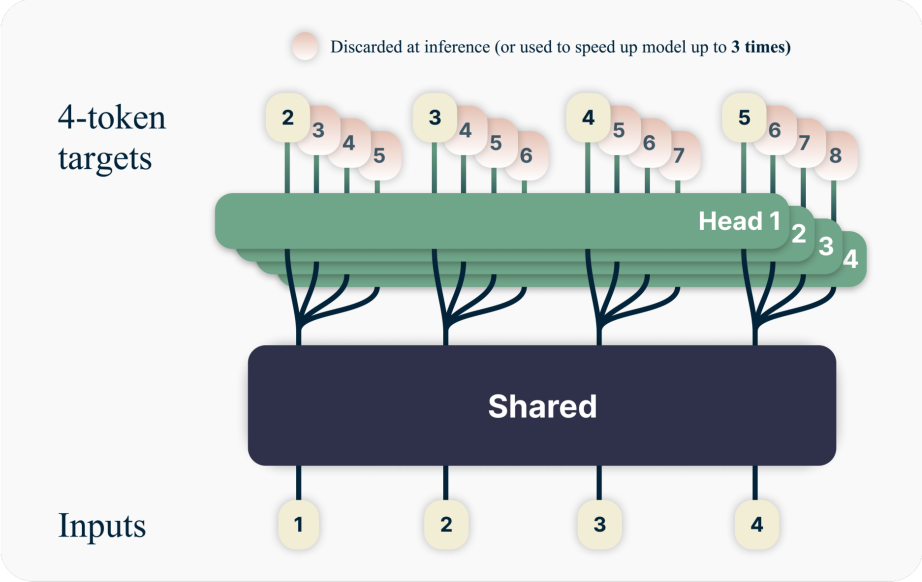

Figure 3.11: Overview of the Multi-Token Prediction Architecture. The model predicts four future tokens simultaneously using a common trunk and four independent output heads. This setup improves sampling efficiency and inference speed. The performance gains are particularly noticeable for larger models and for generative tasks. *Source:* [71]

This multi-token prediction approach has demonstrated significant improvements in sampling efficiency and downstream performance, especially for larger model sizes and for tasks such as code generation. It has been shown that models trained with multi-token prediction can achieve higher accuracy on generative benchmarks, significantly outperforming models trained

with traditional next-token prediction. In addition, multi-token prediction enhances the model's ability to perform algorithmic reasoning tasks and improves its out-of-domain generalization capabilities [71].

Multi-token prediction offers a promising improvement over next-token prediction by allowing models to learn more efficiently and effectively. This methodology not only speeds up the training process, but also improves the overall performance of LLMs in various generative and reasoning tasks.

3.7 Fine-tuning

After the pre-training phase, the models are ready to be adapted to a variety of downstream tasks. This versatility and broad applicability underscore the foundational but incomplete nature of these models. For these and other reasons, these models are referred to as "foundation models".

Foundation models are critical to the advancement of AI by serving as the foundation on which more advanced models are built and refined. Their ability to address a wide variety of tasks in different domains, such as language, vision, robotics, reasoning, and human interaction, demonstrates their essential role in AI research and applications.

Moreover, the term "foundation" reflects the pivotal but evolving status of these models. Despite their impressive capabilities, they exhibit emergent properties that are not yet fully understood, and their large scale brings with it new, unanticipated capabilities. These models are based on deep learning and transfer principles, but their effectiveness across a wide range of tasks often leads to a degree of homogenization. While this uniformity offers significant advantages, it also means that any flaws or biases present in the underlying models are likely to be inherited by any downstream models adapted from them. [23].

This adaptation phase is called fine-tuning and it constitutes a further training of the foundation model on narrower, domain-specific datasets. For instance, models can be fine-tuned on medical transcripts for healthcare applications or on legal briefs for legal assistant bots [157]. This process entails adjusting the model's parameters according to the task-specific data, which often includes labeled examples for supervised tasks.

Fine-tuning can be improved with approaches such as Constitutional AI, which incorporates predefined rules into the model architecture, and reward training, in which humans evaluate the quality of multiple model outputs [16]. In addition, human feedback reinforced learning (RLHF), which uses a comparison-based system to optimize model responses, is employed. Although fine-tuning is less computationally intensive, it requires significant human input to tune the model for specific tasks with predefined outputs [150].

3.7.1 Methods of Fine-Tuning

To tailor models for specific applications such as AI chatbots and optimize their performance, several fine-tuning techniques are employed, each with distinct advantages and suitable use cases.

Full Model Fine-Tuning

Full model fine-tuning represents the most straightforward approach, in which all parameters of the pre-trained model are updated during the fine-tuning process. This approach is particularly effective when the target task is very similar to the one on which the model was originally pre-tuned. The advantage of full model fine-tuning is its ability to adapt the model to the specific nuances of the new task, which can lead to improved performance.

Nevertheless, the increasing size of LLMs over the years has rendered the fine-tuning process extremely computationally demanding and time-consuming. These challenges have significantly increased operational costs, and the financial burden of maintaining and deploying multiple instances of these fine-tuned models becomes especially pronounced when managing numerous downstream tasks [87].

In light of these challenges, alternative fine-tuning approaches, such as Parameter Efficient Fine-Tuning (PEFT) methods, have been developed with the objective of optimizing the process, thereby reducing the computational and financial burdens associated with full model fine-tuning. PEFT methods, in particular LoRA, will be discussed in detail in later sections.

Layer-wise Fine-Tuning

Layer-wise fine-tuning represents a more refined approach to model fine-tuning, whereby the model is fine-tuned in stages, starting with the top layers (closer to the output) and gradually incorporating lower layers (closer to the input). This method allows for a more precise fit, which can be particularly advantageous when the task at hand requires only minor changes to the pre-trained model. The application of layer-wise fine-tuning can serve to mitigate the overfitting problem by limiting the extent of changes made to model parameters. This approach is often utilized in scenarios where computational resources are constrained or when dealing with smaller datasets [161].

Adapter Layers

Adapter layers provide a flexible and efficient alternative to full model fine-tuning. In this technique, small neural network modules, referred to as "adapters," are inserted into the layers of the pre-trained model. During the fine-tuning process, only the parameters of the adapter layers are updated, while the original model parameters remain fixed. This approach markedly diminishes the number of parameters that require modification, consequently reducing computational costs and accelerating training times. Adapter layers have been demonstrated to perform at a comparable level to full model fine-tuning, particularly in low-resource settings. Furthermore, they have been shown to be highly effective for multi-task learning, where the same pre-trained model can be fine-tuned for different tasks by simply switching the adapters [83].

3.7.2 Parameter Efficient Fine-Tuning (PEFT)

In order to address the considerable computational and financial costs associated with the vast scale of LLM systems, the parameter-efficient fine-tuning (PEFT) approach was developed as a viable solution to this challenge. This approach facilitates the efficient adaptation of large models to a range of downstream tasks, thereby reducing the burden of significant costs. PEFT entails the process of fine-tuning a pre-trained large model to a specific task or domain while minimizing the number of additional parameters introduced and the computational resources required [77].

Among the various PEFT techniques, Low-Rank Adaptation (LoRA) is distinguished by its efficacy in reducing the computational demands associated with fine-tuning LLMs.

3.7.3 LoRA: Low-Rank Adaptation

Low-Rank Adaptation (LoRA) is a fine-tuning method that involves freezing the pre-trained model weights and incorporating trainable rank decomposition matrices into each layer of the transformer architecture. This significantly reduces the number of trainable parameters required for downstream tasks. For instance, in comparison to fine-tuning GPT-3 175B using Adam,

LoRA achieves a reduction in trainable parameters by a factor of 10,000 and decreases the GPU memory requirement by a factor of three.

Unlike traditional fine-tuning, which necessitates adjusting the entire model, LoRA focuses on modifying a smaller subset of parameters (lower-rank matrices), thereby reducing computational and memory overhead. This approach is predicated on the understanding that large models inherently possess a low-dimensional structure. By leveraging low-rank matrices, LoRA effectively adapts these models, allowing significant model changes to be represented with fewer parameters, thereby optimizing the adaptation process [90].

Decomposition in LoRA

As introduced in previous sections, in traditional fine-tuning the weights of a pre-trained neural network are modified to adapt to a new task. This adjustment involves altering the original weight matrix W of the network. The changes made to W during fine-tuning are collectively represented by ΔW, such that the updated weights can be expressed as $W + \Delta W$.

Rather than modifying W directly, the LoRA approach decomposes ΔW. The intrinsic rank hypothesis suggests that significant changes to the neural network can be captured using a lower-dimensional representation. Essentially, it posits that not all elements of ΔW are equally important; instead, a smaller subset of these changes can effectively encapsulate the necessary adjustments.

Building on this hypothesis, LoRA represents ΔW as the product of two smaller matrices, A and B, with a lower rank. The updated weight matrix W' thus becomes:

$$W' = W + BA \tag{3.19}$$

In this equation, the variable W is held constant, indicating that it is not updated during the training process. The matrices B and A are of lower dimensionality, with their product BA representing a low-rank approximation of ΔW.

As illustrated in Figure 3.12, by choosing matrices A and B to have a lower rank r, the number of trainable parameters is significantly reduced. For instance, if W is a $d \times d$ matrix, traditionally, updating W would involve d^2 parameters. However, with B and A of sizes $d \times r$ and $r \times d$ respectively, the total number of parameters reduces to $2dr$, which is much smaller when $r \ll d$ [90].

Despite the contribution of Hu et al. [90] in the introduction of LoRA primarily presents experiments in the field of NLP, the approach of low-rank adaptation is not limited to this domain and could be effectively employed in training various types of neural networks across different domains.

The reduction in the number of trainable parameters achieved through the LoRA method offers several significant advantages, particularly in the fine-tuning of large-scale neural networks. By lowering the number of parameters that need to be updated, LoRA reduces the system's memory footprint, which is crucial when managing extensive models. This reduction also leads to faster training and adaptation, as the computational demands are significantly diminished, allowing for more efficient training and fine-tuning of models for new tasks. Furthermore, the decreased number of parameters makes it feasible to fine-tune large models on less powerful hardware, such as modest GPUs or CPUs, thereby broadening the accessibility of advanced AI capabilities. Finally, LoRA facilitates the scaling of AI models, enabling the expansion of their size and complexity without a corresponding increase in computational resources, thereby making the management of larger models more practical and cost-effective [185].

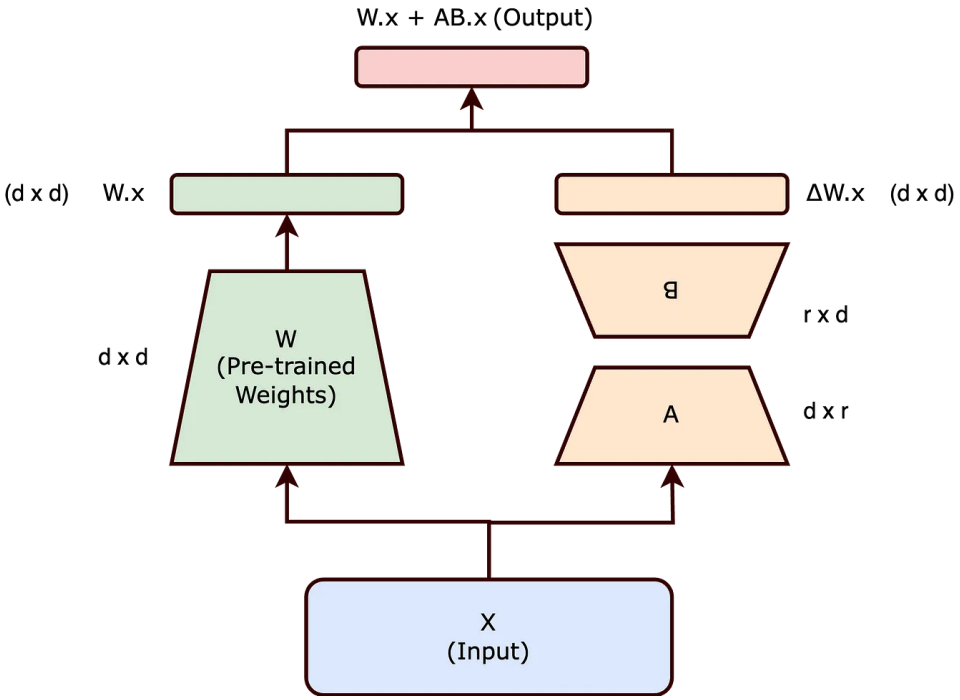

Figure 3.12: Decomposition of ΔW into two matrices A and B, both of lower dimensionality than $d \times d$. *Source:* [185]

3.8 Prompt-based Learning

Prompt-based learning represents a significant shift in how LLMs are employed for NLP tasks. Instead of relying on extensive labeled datasets for training or fine-tuning, prompt-based learning uses textual prompts to guide the model in performing a specific task. This approach is particularly valuable in scenarios where the cost of annotating large datasets is prohibitive, such as in the development of AI chatbots for conversational AI [137].

Prompt-based learning leverages the model's pre-trained knowledge by providing a prompt or instruction that directs the model's responses. The effectiveness of this technique has been demonstrated by famous works such as Radford et al. [157] and Brown et al. [25], where prompt-based few-shot learning achieved results comparable to state-of-the-art models trained with full-shot datasets. This method relies on the model's ability to generalize from minimal examples (few-shot learning) or even without any examples (zero-shot learning), making it a versatile and cost-effective solution for various NLP tasks, including question answering and sentiment analysis.

Prompt-Based Learning vs. Fine-Tuning

The primary distinction between prompt-based learning and traditional fine-tuning lies in the training approach. While fine-tuning involves updating the model's weights based on a training set through a defined loss function, requiring substantial computational effort, prompt-based learning eliminates the need for this training phase altogether. Instead of modifying the model, prompt-based learning utilizes pre-defined prompts that instruct the model on how to perform a

task without altering its weights. This makes prompt-based learning particularly advantageous in scenarios with limited computational resources or when rapid adaptation to new tasks is required, as it avoids the need for retraining [137].

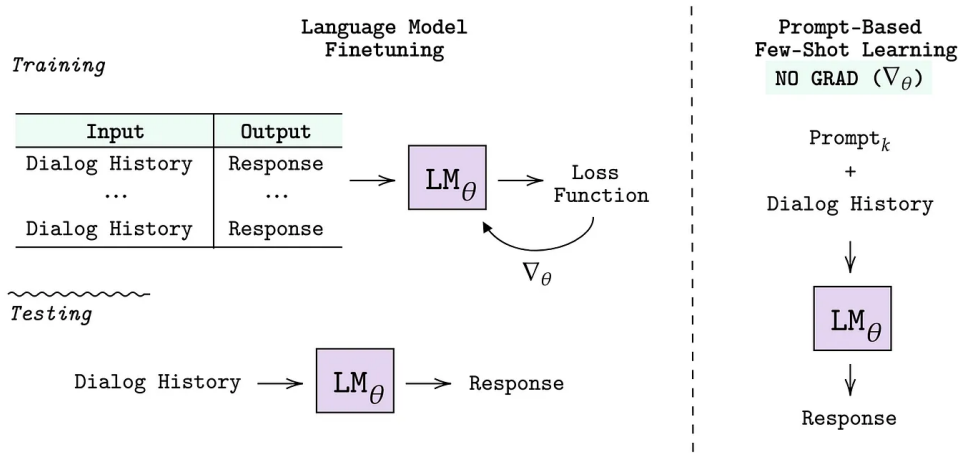

Figure 3.13: Comparison of traditional fine-tuning (left) and prompt-based learning (right). *Source:* [137]

3.8.1 Zero-Shot, One-Shot, and Few-Shot Learning

Prompt-based learning is often implemented using zero-shot, one-shot, or few-shot prompts, each offering varying levels of example-based guidance to the model.

In Zero-Shot learning, the model is given a prompt to perform a task without any additional examples. The model relies solely on its pre-trained knowledge to generate a response, making zero-shot prompts highly efficient in scenarios where no task-specific data is available [157].

One-Shot learning involves providing the model with a single example in addition to the prompt. This example serves as a minimal guide, helping the model to generate responses that align more closely with the desired outcome.

Few-Shot learning extends this approach by supplying the model with a small number of examples, typically more than one but fewer than what would be used in full training. These examples, provided in the context of the prompt, help the model understand the task better, allowing it to produce responses that are more accurate and contextually appropriate [25].

The effectiveness of prompt-based few-shot learning has been demonstrated across a wide range of tasks, including dialogue response generation, knowledge-grounded response generation, and dialogue state tracking. In many cases, it achieves performance comparable to, or even exceeding, that of fully fine-tuned models, without the need for extensive retraining. This makes prompt-based learning a powerful tool in the deployment of conversational AI systems, where flexibility and efficiency are paramount [137].

In conclusion, prompt-based learning offers a compelling alternative to traditional fine-tuning by allowing LLMs to perform specific tasks with minimal or no retraining. By utilizing zero-shot, one-shot, and few-shot prompts, this approach maximizes the utility of pre-trained models, making them more adaptable and cost-effective in a wide variety of applications. Nevertheless, building on the strengths of prompt-based learning, more advanced techniques have been

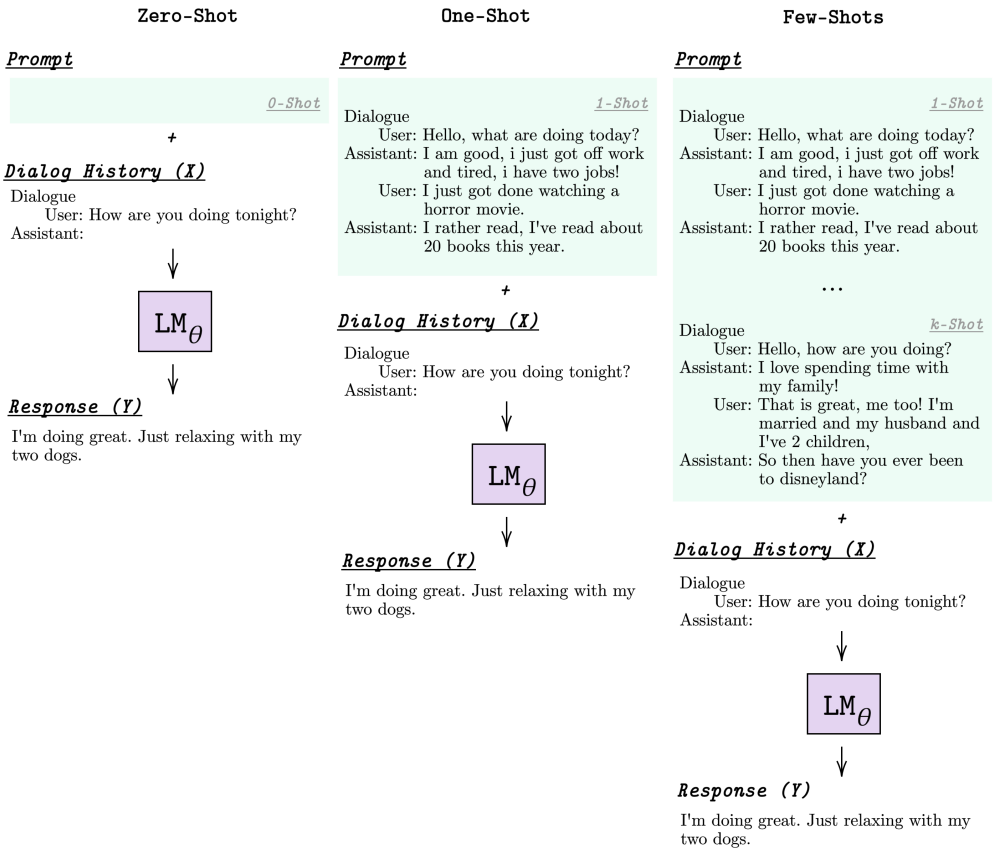

Figure 3.14: Illustration of Zero-Shot, One-Shot, and Few-Shot learning in dialogue systems. Zero-Shot learning involves no additional examples, One-Shot learning provides a single example, and Few-Shot learning includes a small number of examples to guide the model's response generation. *Source:* [137]

developed to further enhance the reasoning abilities of LLMs, with one particularly promising introduced in the following section.

3.8.2 Chain of Thought Prompting

Chain-of-Thought (CoT) Prompting represents a more advanced methodology designed to enhance the reasoning capabilities of LLMs by directing them to construct a sequence of intermediate steps that collectively culminate in a final answer. This approach emulates the natural human thought process when addressing complex tasks, such as multi-step mathematical problems, where the problem is decomposed into smaller, more tractable components before reaching a final solution [206].

As illustrated in Figure 3.15, Chain of Thought reasoning allows language models to handle complex reasoning tasks more effectively by decomposing the problem into intermediate stages. For example, when solving a math word problem, a model might generate a sequence of logical steps that resemble the way a human might solve the problem step-by-step: "After Jane gives 2

flowers to her mom, she has 10 left... then after giving 3 to her dad, she will have 7... so the answer is 7". This sequential reasoning helps the model arrive at the correct answer by focusing on each part of the problem individually.

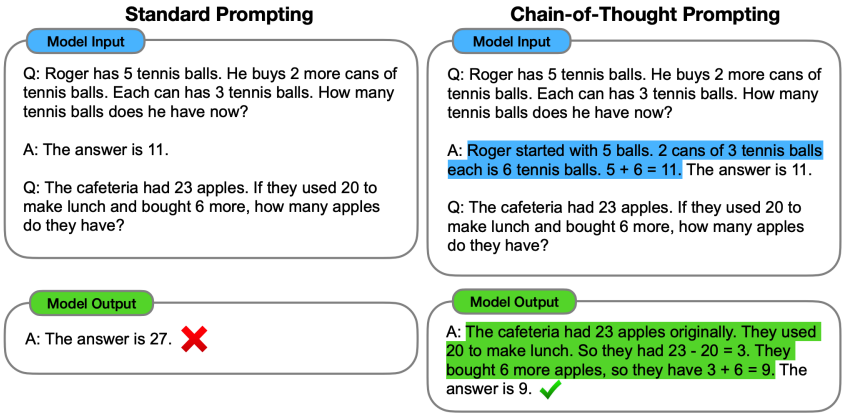

Figure 3.15: Comparison of standard prompting and Chain-of-Thought (CoT) prompting. The CoT approach allows the model to break down problems into intermediate steps, leading to more accurate and interpretable outcomes. *Source:* [206]

One of the significant advantages of Chain of Thought reasoning is its interpretability. By breaking down the reasoning process into a series of intermediate steps, CoT provides insights into how the model arrived at a particular answer. This not only helps in understanding the model's decision-making process but also offers opportunities to identify and correct errors in the reasoning path.

Moreover, Chain of Thought reasoning is versatile and can be applied to a variety of tasks that require complex reasoning, such as solving math word problems, engaging in commonsense reasoning, and performing symbolic manipulation. Essentially, it can be applied to any task where humans typically solve problems through language-based reasoning.

Another notable benefit of Chain of Thought prompting is that it can be effectively elicited in LLMs by incorporating examples of such reasoning into the few-shot prompting process. This means that without any additional fine-tuning, sufficiently large pre-trained models can generate these reasoning chains simply by being provided with relevant examples during the prompting phase [206].

In addition to its general applications, chain-of-thought reasoning shows particular promise for AI chatbots in conversational tasks. By enabling chatbots to deconstruct users' questions into manageable sequential steps, CoT can greatly improve their ability to process complex, multi-turn dialogues. This approach allows chatbots to maintain consistency between interactions and generate responses that reflect a deeper understanding of the context of the conversation. In addition, the interpretability offered by CoT can help developers identify and correct errors in chatbot reasoning, ultimately leading to more reliable and user-friendly conversational AI. In conclusion, the incorporation of Chain of Thought prompting in AI chatbots can facilitate the emergence of more sophisticated and context-sensitive interactions, thereby improving the quality and effectiveness of conversational AI systems.

3.8.3 Instruction Prompt Tuning

To further enhance the capabilities of AI chatbots, another promising approach is Instruction Prompt Tuning, which provides a cost-effective method for fine-tuning models to perform better across a variety of specialized tasks.

Instruction prompt tuning, introduced by Lester et al. [111], provides a cost-effective solution to update the model's parameters and improving performance in numerous downstream tasks. This method offers significant advantages over few-shot prompting, particularly for clinical applications, because it allows LLMs to be more effectively aligned with the specific requirements of the medical domain, as demonstrated by Singhal et al. [174]

Prompt tuning is designed to improve the performance of frozen language models on specific downstream tasks by learning so-called "soft prompts". Unlike traditional discrete text prompts, which are manually selected or searched for, soft prompts are learned through backpropagation, allowing them to be fine-tuned based on labeled examples.

At a high level, prompt tuning shifts the focus from tuning the entire model to tuning only the parameters of the prompts. This approach involves conditioning a frozen language model on these soft prompts to guide its output generation, thereby enhancing the model's ability to perform specific tasks. This is particularly important because it allows the use of large pre-trained models without the need to modify their core weights, making the approach both resource efficient and scalable.

The concept of prompt tuning builds on the idea of conditioning models with additional information. Typically, in models such as GPT, prompts are added as additional tokens that the model uses to generate the desired output. However, these tokens are part of the model's fixed embedding space, which means that they cannot be directly optimized by training. Prompt tuning overcomes this limitation by introducing dedicated parameters for the prompts themselves, which can be updated during training.

As a result, prompt tuning becomes increasingly competitive as the size of the language model grows. Remarkably, it can match the performance of full model fine-tuning, even when the model size reaches billions of parameters. This makes prompt tuning a powerful tool, especially for large models where fine-tuning all parameters is computationally expensive and impractical.

In addition, prompt tuning has demonstrated advantages in robustness to domain transfer and offers the flexibility of "prompt ensembling", where multiple prompts can be combined to improve performance. This approach also simplifies the process of adapting a single frozen model to multiple downstream tasks, reducing the need to deploy and manage multiple fine-tuned models [111].

3.9 Limitations of Large Language Models in Practical Applications

After examining LLMs, their architecture, and the impact of model size on AI performance, as well as exploring various training and fine-tuning and prompt-based learning, it is essential to address the limitations of these models in practical applications.

One of the most significant constraint of LLMs is their lack of transparency in predictions. The models operate as black boxes, making it difficult to interpret the reasoning behind their outputs. This opacity is problematic, especially when LLMs are used in decision-making processes that require accountability and traceability [163].

Another critical issue is the tendency of LLMs to produce "hallucinations", statements that are plausible but factually incorrect. These hallucinations arise because LLMs generate text

based on patterns learned from vast amounts of data, without a deep understanding of the world or the ability to fact-check their outputs [139]. This risk is exacerbated by the fluency and coherence of the generated text, which can mislead users into accepting false information as true.

Bias is another well-documented limitation of LLMs. These models can inadvertently perpetuate and amplify biases present in the training data, leading to outputs that reinforce stereotypes or discriminate against certain groups [20]. This bias is particularly concerning in applications like hiring, lending, or law enforcement, where fairness and impartiality are paramount.

Additionally, LLMs face issues of staleness and revisions. Since these models are trained on static datasets, they cannot incorporate new information that emerges after their training, leading to outdated or incorrect responses [49]. Furthermore, the accuracy of information retrieval is another challenge, as LLMs often struggle to retrieve specific, relevant information from their vast learned knowledge, leading to irrelevant or imprecise responses.

The limitations mentioned above significantly undermine the reliability and value of LLMs in practical applications. When these models generate inaccurate or misleading outputs, the consequences can be severe, especially in critical fields where precision and factual correctness are essential.

To address some of these limitations, one promising approach is Retrieval-Augmented Generation (RAG). RAG enhances the capabilities of LLMs by integrating an external retrieval mechanism that allows the model to access and utilize up-to-date and contextually relevant information during the generation process [114]. This method mitigates issues related to staleness, hallucinations, and factual inaccuracies by grounding the model's outputs in verifiable external sources, thereby improving the accuracy and reliability of its responses. As a result, RAG offers a viable solution to many of the challenges faced by LLMs in practical applications, making them more effective and trustworthy in real-world scenarios.

3.10 Retrieval-Augmented Generation (RAG)

In 2020, Meta AI researchers introduced Retrieval-Augmented Generation (RAG) as a technique that significantly improves the functionality of LLMs by integrating them with external knowledge sources [114]. RAG operates by retrieving relevant document chunks from an external knowledge base through semantic similarity calculations, thereby enhancing the factual accuracy and reducing hallucinations in the generated content. This integration has made RAG a widely adopted technology, particularly in advancing applications such as AI chatbots and improving the suitability of LLMs for real-world scenarios.

The development of RAG technology has progressed rapidly, evolving through distinct stages. Initially, the inception of RAG coincided with the rise of the transformer architecture, focusing on enhancing language models by incorporating additional knowledge through Pre-Training Models (PTM). This early stage was characterized by foundational works aimed at refining pre-training techniques [10, 114, 24].

The subsequent arrival of models like GPT marked a pivotal moment, showcasing powerful in-context learning (ICL) capabilities, where models can make predictions based on the context provided in the input without needing additional fine-tuning. RAG research shifted towards providing better information for LLMs to tackle more complex and knowledge-intensive tasks during the inference stage. As research progressed, RAG's enhancement was no longer limited to the inference stage but began to incorporate more with LLM fine-tuning techniques.

A typical application of RAG is illustrated in Figure 3.16. In this example, a user poses a question to ChatGPT about a recent news event. Given that ChatGPT relies on pre-training

data, it initially lacks the ability to provide updates on recent developments. RAG bridges this information gap by sourcing and incorporating knowledge from external databases. In this case, relevant news articles related to the user's query are retrieved and combined with the original question to form a comprehensive prompt, empowering the LLM to generate a well-informed answer [67].

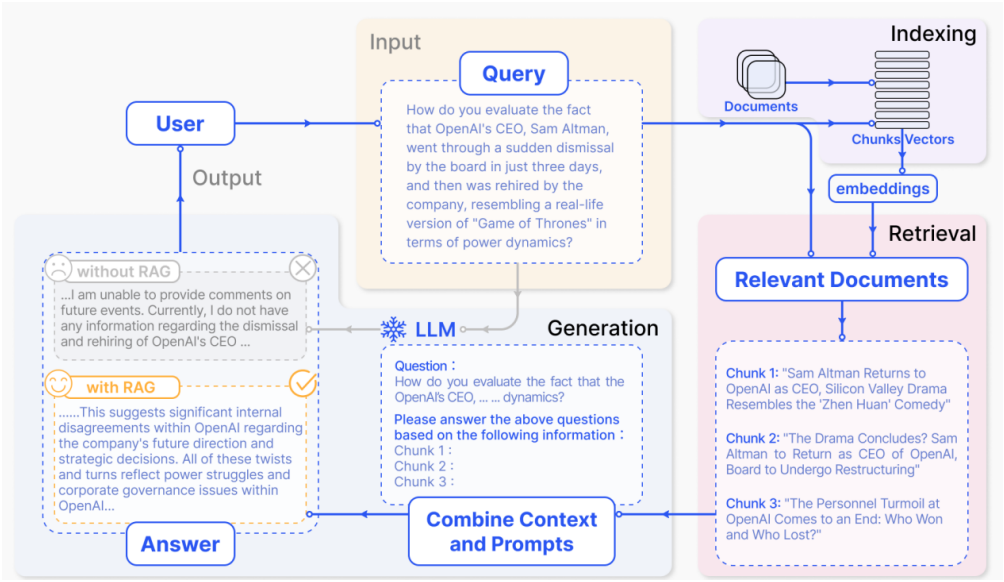

Figure 3.16: A representative instance of the RAG process applied to question answering. *Source:* [67]

As RAG continues to evolve, we can categorize its development into three stages: Naive RAG, Advanced RAG, and Modular RAG. Although RAG methods are cost-effective and surpass the performance of native LLMs, they also exhibit certain limitations. The development of Advanced RAG and Modular RAG represents a response to these specific shortcomings in the Naive RAG stage [67].

Naive RAG

The Naive Retrieval-Augmented Generation (RAG) approach represents the foundational methodology in RAG systems and gained prominence following the widespread adoption of models like ChatGPT. Naive RAG follows a straightforward "Retrieve-Read" framework, which typically involves three critical phases: indexing, retrieval, and generation [136].

1. **Indexing** is the initial phase where raw data from various formats, such as PDFs, HTML, Word documents, and Markdown files, are cleaned and converted into a uniform plain text format. Given the context limitations of language models, the text is then segmented into smaller, manageable chunks. These chunks are encoded into vector representations using an embedding model and stored in a vector database. This process is crucial as it lays the groundwork for efficient similarity searches during the retrieval phase.

2. **Retrieval** begins when a user query is submitted. The system utilizes the same encoding model used during indexing to transform the query into a vector representation. The system

then computes similarity scores between the query vector and the vectors of chunks within the indexed corpus. The top K chunks with the highest similarity scores are retrieved and used as the expanded context for the prompt.

3. **Generation** involves synthesizing the user's query with the retrieved documents to create a coherent prompt, which the LLM uses to generate a response. Depending on task-specific criteria, the model might draw on its parametric knowledge or restrict its response to the retrieved documents [67].

However, Naive RAG faces several significant challenges, including issues with retrieval precision and recall, and the generation of content that may suffer from hallucinations, irrelevance, or bias.

Advanced RAG

Built on the foundational Naive RAG framework, Advanced RAG introduces specific improvements designed to address the limitations observed in the earlier model. This approach enhances the quality of retrieval by implementing both pre-retrieval and post-retrieval strategies, focusing on refining the indexing process and optimizing how queries are handled.

In the pre-retrieval phase, the emphasis is on optimizing the structure of the index and refining the original user query to improve retrieval quality. Strategies such as the sliding window approach, fine-grained text segmentation, and the incorporation of metadata are employed to enhance the granularity and relevance of the indexed content. Additionally, query optimization techniques such as query rewriting, transformation, and expansion are used to ensure the query is well-suited for effective retrieval [93].

The post-retrieval phase focuses on integrating the retrieved context more effectively with the user's query. Key methods include re-ranking the retrieved chunks to prioritize the most relevant information and compressing the context to avoid information overload when feeding it into the language model. These strategies help address the generation challenges encountered in Naive RAG by ensuring that the final output is more focused, relevant, and coherent.

Advanced RAG's systematic enhancements to both the indexing and retrieval processes significantly improve the overall performance of RAG systems, making them more robust and capable of handling complex queries with greater accuracy [67].

Modular RAG

Modular RAG represents the next evolution in the RAG paradigm, offering a more flexible and adaptable architecture that can be customized to meet the needs of a wide range of tasks and scenarios. This approach builds upon the foundational principles of Naive and Advanced RAG but introduces modular components that can be independently optimized and reconfigured.

As Modular RAG introduces a more adaptable and reconfigurable framework, it also incorporates several new and specialized modules that significantly enhance its retrieval and processing capabilities. For instance, the Search module allows for direct searches across diverse data sources, such as search engines, databases, and knowledge graphs, using LLM-generated code and query languages [201]. The RAGFusion module addresses the limitations of traditional search by employing a multi-query strategy that expands user queries into diverse perspectives, utilizing parallel vector searches and intelligent re-ranking to uncover both explicit and transformative knowledge [167]. Additionally, the Memory module leverages the language model's memory to guide retrieval, creating an unbounded memory pool that iteratively enhances the alignment between text and data distribution [32].

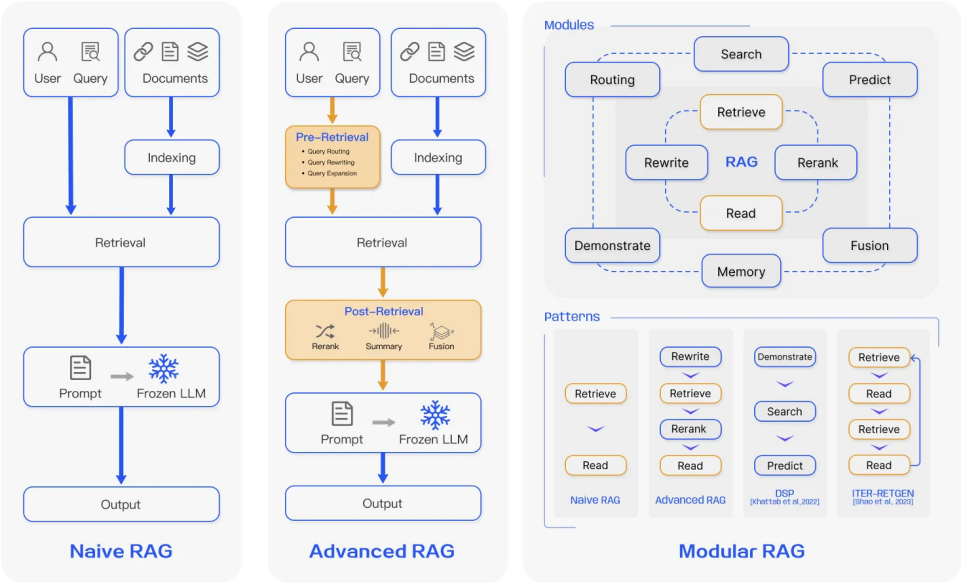

Figure 3.17: Comparison between the three paradigms of RAG. *Source:* [67]

Moreover, the Routing module in Modular RAG optimizes query pathways by navigating through various information sources, selecting the most appropriate route for each query. This might involve summarization, specific database searches, or merging different information streams to provide a comprehensive response [117]. The Predict module reduces redundancy and noise by generating context directly within the LLM, ensuring that the outputs are relevant and accurate [229]. Finally, the Task Adapter module tailors the RAG system to various downstream tasks, automating prompt retrieval for zero-shot inputs and creating task-specific retrievers through few-shot query generation [31].

These new patterns in Modular RAG provide remarkable adaptability by allowing for module substitution or reconfiguration to address specific challenges. Unlike the fixed structures of Naive and Advanced RAG, Modular RAG offers the flexibility to integrate new modules or adjust the interaction flow among existing ones, enhancing its applicability across different tasks. Innovations such as the Rewrite-Retrieve-Read model and the Demonstrate-Search-Predict (DSP) framework illustrate how Modular RAG can leverage LLMs capabilities to refine retrieval queries and improve task performance [102].

Modular RAG's dynamic architecture not only simplifies the retrieval process but also significantly enhances the quality and relevance of the information retrieved. Its flexibility allows for the integration of other technologies, such as fine-tuning or reinforcement learning, further expanding its effectiveness and adaptability in diverse application scenarios [67].

To ensure that these retrieval mechanisms operate at their full potential, attention must be given to the foundational step of indexing.

3.10.1 Indexing Optimization

Indexing is a crucial phase in the RAG system, where documents are processed, segmented, and transformed into embeddings that are stored in a vector database. The quality and structure of

the indexing process directly impact the effectiveness of subsequent retrieval operations, as they determine whether the correct and most relevant context can be retrieved when needed.

Chunking Strategy

A common method for indexing involves splitting the document into smaller segments or "chunks," typically based on a fixed number of tokens (e.g., 100, 256, or 512 tokens). The choice of chunk size is a balancing act: larger chunks can capture more context, but they also introduce more noise, increasing processing time and computational costs. Conversely, smaller chunks reduce noise but may fail to fully convey the necessary context, potentially leading to incomplete or fragmented retrieval results [180].

To address these challenges, various optimization strategies have been proposed. For example, recursive splitting and sliding window methods allow for layered retrieval, where globally related information is merged across multiple retrieval processes [108]. These techniques aim to preserve semantic completeness while accommodating the constraints of context length. However, even with these optimizations, striking the right balance between semantic integrity and context length remains a complex task. Recently, methods like Small2Big have been introduced, where sentences (small units) are used as the retrieval unit, and the preceding and following sentences are provided as (big) context to the LLMs [224]. This approach helps in maintaining a coherent context while minimizing noise.

Metadata Attachments

In addition to chunking, attaching metadata to chunks can significantly enhance the retrieval process. Metadata may include information such as page numbers, file names, authors, categories, timestamps, and other contextual markers. By embedding this metadata within the index, retrieval can be filtered based on these attributes, narrowing the search scope and improving the relevance of the retrieved information.

For instance, assigning different weights to document timestamps during retrieval can enable time-aware RAG, ensuring the freshness of the knowledge and preventing the use of outdated information. Moreover, metadata can be artificially constructed. One innovative approach is Reverse HyDE, where hypothetical questions are generated using LLMs based on the document content. During retrieval, the system calculates the similarity between the original query and the hypothetical questions, thereby reducing the semantic gap between the user's question and the answers provided [67].

Structural Indexing

Another effective method for optimizing indexing is the construction of a hierarchical structure for the documents. In a hierarchical index, documents are arranged in parent-child relationships, with individual chunks linked to these hierarchical nodes. Data summaries are stored at each node, which assists in the swift traversal of data and helps the RAG system determine which chunks to extract. This hierarchical approach not only speeds up the retrieval process but also mitigates issues such as data fragmentation and the illusion caused by block extraction [67].

In addition, Knowledge Graphs (KGs) can be integrated into the indexing structure to maintain consistency and enhance retrieval accuracy. KGs delineate the connections between different concepts and entities, reducing the potential for retrieval errors or hallucinations. For instance, the Knowledge Graph Prompting method constructs an index between multiple documents using a KG, where nodes represent paragraphs or structures within documents (e.g., pages, tables), and edges indicate semantic or lexical similarities between these nodes [203]. This approach

addresses challenges in knowledge retrieval and reasoning within a multi-document environment, ensuring that the retrieval process remains coherent and contextually accurate.

Overall, optimizing the indexing phase in RAG systems is essential for ensuring that the retrieval process is efficient, accurate, and capable of delivering the most relevant information to support the generation tasks of LLMs.

3.10.2 Retrieval Source

After indexing optimization, the efficient retrieval of relevant documents from external data sources is critical in RAG systems to ensuring the accuracy and relevance of generated outputs.

For open-domain question-answering (ODQA) tasks, traditional unstructured text, such as Wikipedia or domain-specific data, remains the most common retrieval source. In addition to encyclopedic data, common unstructured data includes cross-lingual text and domain specific-data [117].

However, RAG systems are increasingly incorporating semi-structured data, that typically refers to data that includes both text and table elements, such as PDF files. Managing semi-structured data presents challenges for conventional RAG systems for two main reasons. First, the process of text splitting can unintentionally separate tables, leading to data corruption during retrieval. Second, integrating tables into the data complicates semantic similarity searches. One approach to handling semi-structured data is to utilize the coding capabilities of LLMs to execute Text-2-SQL queries on tables within databases, as seen in systems like TableGPT [231]. Alternatively, tables can be converted into text format for further analysis using text-based methods [135]. However, these methods are not without limitations, highlighting the need for further research in this area.

Structured data, such as knowledge graphs (KGs), typically undergoes verification and can provide more precise information. For example, KnowledGPT generates knowledge base search queries and stores knowledge in a personalized repository, enriching the RAG model's information base [201]. To address the limitations of LLMs in understanding and answering questions related to textual graphs, G-Retriever integrates Graph Neural Networks (GNNs), LLMs, and RAG, thereby enhancing graph comprehension and question-answering capabilities through soft prompting [80]. Nevertheless, managing structured data requires considerable effort to build, validate, and maintain structured databases.

Retrieval Granularity

The granularity of the retrieved data, whether it is at the token, sentence, or document level, plays a crucial role in the retrieval process. Coarse-grained units might provide more context but can introduce irrelevant information, while fine-grained units increase retrieval precision but might lack necessary context. The choice of retrieval granularity should be tailored to the specific downstream tasks to ensure both relevance and coherence in the generated outputs [67].

3.10.3 Query Optimization

One of the primary challenges in RAG systems is effectively leveraging the user's query to retrieve the most relevant information. Naive RAG systems often rely directly on the user's original query, which may not always be well-formed, precise, or optimized for retrieval purposes. Query optimization involves enhancing the query to improve retrieval effectiveness, ensuring that the system retrieves the most relevant context for generating accurate and coherent responses. This optimization process can be broken down into several key strategies: query expansion, query transformation, and query routing.

Query Expansion

Query expansion enriches the original query by generating additional sub-queries, ensuring the retrieval process captures all necessary nuances. Techniques like multi-query generation expand the original query into multiple related queries, executed in parallel to provide a comprehensive context. Another strategy, sub-query planning, breaks down complex queries into simpler sub-queries, improving accuracy and completeness [238]. Additionally, Chain-of-Verification (CoVe) validates expanded queries to reduce hallucinations of LLMs, enhancing the reliability of the retrieved information [50].

Query Transformation

Query transformation modifies the original query to improve retrieval suitability. Query rewriting rephrases queries for better compatibility with the retrieval system, while hypothetical document embedding (HyDE) uses LLMs to generate hypothetical documents, focusing retrieval on embedding similarity between generated and real documents [64]. Another method is Step-back Prompting, that abstracts the query to create a high-level concept question, combining it with the original query for more accurate results [236].

Query Routing

Query routing directs queries through different retrieval pipelines based on their nature. Metadata Routing/Filtering uses extracted keywords to narrow the search scope, while Semantic Routing leverages the semantic content of the query to select the most appropriate retrieval pathway. In some cases, a hybrid approach combines both methods for enhanced routing [67].

By optimizing queries through expansion, transformation, and routing, RAG systems can retrieve more accurate and contextually relevant information, resulting in higher quality outputs.

3.10.4 Embedding Techniques

Embeddings constitute a further pivotal element in the functioning of RAG systems, as they represent the semantic content of queries and document segments in a format that can be evaluated for similarity through mathematical comparison. The choice and optimization of embedding models are crucial for ensuring the accuracy and efficiency of the retrieval process. The two primary approaches to embeddings in RAG systems are sparse and dense embeddings, with recent advances introducing hybrid models and fine-tuning techniques to enhance performance.

Sparse vs. Dense Embeddings

Sparse embeddings, typically derived from traditional information retrieval models, represent documents as high-dimensional vectors where each dimension corresponds to a term in the vocabulary. These models excel in capturing term frequency and inverse document frequency (TF-IDF) relationships but may struggle with capturing the semantic nuances of language.

Dense embeddings, on the other hand, are generated by neural models such as BERT, which transform text into dense, low-dimensional vectors that encapsulate the semantic meaning of the text. These embeddings are more effective at capturing the contextual relationships between words, making them well-suited for tasks requiring a deep understanding of language semantics.

Hybrid Retrieval Approaches combine the strengths of both sparse and dense embeddings. In such systems, sparse embeddings are often used to provide an initial set of candidate documents, which are then re-ranked using dense embeddings to refine the retrieval results. This approach

leverages the complementary strengths of both embedding types, improving the robustness and accuracy of the retrieval process. For instance, hybrid models have been shown to enhance the zero-shot retrieval capability of dense retrievers by providing a broader context through initial sparse retrievals [67].

Fine-Tuning Embedding Models

Fine-tuning embedding models is essential in scenarios where the retrieval context significantly deviates from the pre-training corpus, particularly in specialized domains such as healthcare, legal practice, and other fields with proprietary jargon. Fine-tuning involves adjusting the embedding model on a domain-specific dataset to better capture the nuances and specialized knowledge required for effective retrieval in that domain.

In addition to domain-specific fine-tuning, another key purpose of fine-tuning is to align the retriever and generator within the RAG system. This can be achieved by using the outputs of the language model as a supervision signal for fine-tuning, a process known as LM-Supervised Retrieval (LSR). This approach ensures that the embeddings generated by the retriever are closely aligned with the generative tasks of the LLM, leading to more coherent and contextually relevant outputs.

Recent advancements in embedding techniques, such as Reinforcement Learning from Human Feedback (RLHF), involve utilizing LM-based feedback to reinforce the retriever through reinforcement learning. This approach allows the retriever to continuously improve based on real-world feedback, further enhancing the robustness and accuracy of the retrieval process.

Embedding techniques are foundational to the success of RAG systems. By carefully selecting, fine-tuning, and combining embedding models, RAG systems can achieve higher retrieval accuracy, better alignment with generative tasks, and ultimately more reliable and contextually appropriate outputs [67].

3.10.5 Generation

After the retrieval phase in a RAG system, directly feeding all retrieved information into a LLM for generating responses is generally not ideal. This section discusses the necessary adjustments from two perspectives: optimizing the retrieved content and fine-tuning the LLM itself.

Context Curation

Redundant information and overly long contexts can interfere with the LLM's ability to generate accurate and relevant answers. This phenomenon, known as the "Lost in the middle" problem, occurs when LLMs, like humans, tend to focus primarily on the beginning and end of long texts while neglecting the middle portion. Therefore, it is essential to process the retrieved content effectively before feeding it to the LLM [124].

To address this issue, one effective strategy is reranking, which reorders document chunks to prioritize the most relevant information. By reducing the document pool to a manageable size, reranking enhances retrieval accuracy and refines the inputs for the LLM. This process can be achieved through rule-based methods using metrics like diversity and relevance, or by employing specialized reranking models, such as those from the BERT series or general LLMs like GPT [66].

However, simply reranking documents may not be enough to ensure optimal performance. In fact, contrary to the belief that more relevant documents lead to better results, adding too many documents can introduce noise and diminish the LLM's focus on key information. To further refine the input, Context Compression techniques can be employed. These techniques use smaller

language models (SLMs) to filter out unimportant tokens, transforming the context into a form optimized for LLM processing. This approach balances language integrity with compression efficiency. Additionally, reducing the number of documents in the prompt improves accuracy, as shown by the "Filter-Reranker" paradigm, which combines the strengths of SLMs and LLMs to enhance information extraction tasks. Furthermore, having the LLM critique retrieved content before generating an answer can further boost relevance and accuracy, as demonstrated in applications like Chatlaw [67, 45].

LLM Fine-tuning

While context curation is crucial for optimizing the input provided to the LLM, fine-tuning the LLM itself can yield even more significant improvements in performance. Fine-tuning LLMs based on specific scenarios and data characteristics allows for targeted adjustments that incorporate additional knowledge, adapt to particular data formats, and generate responses in a desired style [54].

Moreover, aligning LLM outputs with human preferences or retriever preferences through reinforcement learning offers another layer of optimization. By manually annotating generated answers and using these annotations as feedback during reinforcement learning, the model can be further refined to better meet the needs of specific tasks. Additionally, when access to powerful proprietary models is limited, distillation methods can be employed to transfer knowledge from more advanced models, such as GPT-4, to less powerful ones, ensuring that even smaller models can benefit from the strengths of larger counterparts [170].

3.10.6 Augmentation Process

The standard practice in Retrieval-Augmented Generation (RAG) often involves a single retrieval step followed by text generation. While this approach is straightforward, it can be insufficient for complex tasks that require multi-step reasoning, as it provides a limited scope of information. To address this limitation, several augmentation strategies have been developed to enhance the retrieval process, offering more robust and contextually relevant outputs [228].

As illustrated in Figure 3.18, these augmentation strategies are categorized into three main types: Iterative Retrieval, Recursive Retrieval, and Adaptive Retrieval. Each method provides unique advantages that help improve the retrieval and generation process in RAG systems.

Iterative Retrieval

Iterative retrieval involves repeatedly searching the knowledge base based on the initial query and the text generated so far. This process enables the LLM to access a broader and more comprehensive knowledge base, which in turn improves the robustness of the generated responses by providing additional contextual references through multiple retrieval iterations. However, iterative retrieval may introduce challenges such as semantic discontinuity and the accumulation of irrelevant information.

Recursive Retrieval

To build upon the iterative approach, recursive retrieval is employed to improve the depth and relevance of search results by iteratively refining search queries based on the outcomes of previous searches. This approach is particularly useful in complex scenarios where user needs are not fully clear or where the information sought is highly specialized. By incorporating a feedback loop, recursive retrieval gradually converges on the most pertinent information. Additionally, recursive

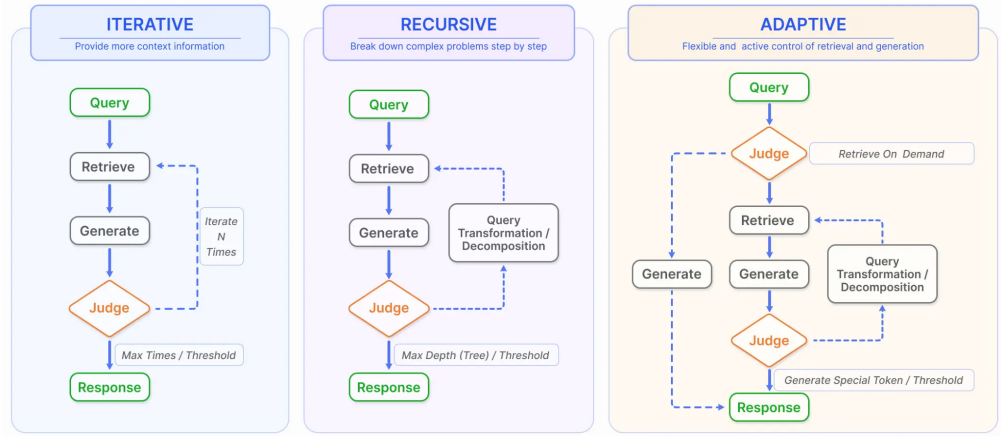

Figure 3.18: In addition to the most common retrieval, RAG also includes three types of retrieval augmentation processes. *Source:* [67]

retrieval can be combined with multi-hop retrieval techniques to process data hierarchically, summarizing sections of documents before refining the search further within the document [67].

Adaptive Retrieval

Further refining the RAG framework, adaptive retrieval methods enable language models to autonomously determine the optimal timing and content for retrieval, enhancing both the efficiency and relevance of the information sourced. This approach represents a shift towards more active judgment by the language models, where the models, like in Self-RAG [12], proactively decide when to initiate retrieval based on the confidence levels in the generated output. Techniques such as "reflection tokens" allow the model to introspect and trigger retrieval only when necessary, thus optimizing the retrieval cycle and ensuring that the model generates the most accurate and relevant responses possible [67].

Iterative, recursive, and adaptive retrieval strategies are crucial in addressing the limitations of traditional RAG systems. By refining how and when retrieval occurs, these methods significantly improve the depth, relevance, and accuracy of the information used in generating responses, making RAG systems more capable of handling complex, knowledge-intensive tasks.

3.10.7 Future Prospects of RAG Technology

The field of Retrieval-Augmented Generation has seen significant advancements, yet several challenges and opportunities for further research remain. This section discusses the future prospects of RAG technology, focusing on its potential developments and the challenges that must be addressed.

RAG and Long Contexts

As research into LLMs continues to evolve, the ability of these models to handle increasingly long contexts has improved dramatically. Modern LLMs can effectively handle contexts up to 32,000 tokens, with the industry now moving towards managing contexts of up to 128,000 tokens, equivalent to the length of a 250-page book [92]. This capability raises questions about

the continued relevance of RAG, particularly in tasks such as long-document question answering, where it might seem feasible to input entire documents directly into the model. However, RAG retains its value for several reasons. Firstly, providing an LLM with an excessive amount of context in a single prompt can severely impact inference speed, whereas chunked retrieval and on-demand input significantly enhance operational efficiency. Secondly, RAG-based generation allows for the quick location of original references, enabling users to verify the generated answers. The entire retrieval and reasoning process in RAG is transparent, unlike generation that relies solely on long contexts, which remains a black box. The expansion of context capabilities also opens new avenues for RAG, particularly in tackling complex integrative tasks that require synthesizing information from extensive material. Developing new RAG methods for super-long contexts represents a promising area of future research [67].

Enhancing RAG Robustness

The presence of noise or contradictory information in the retrieved documents can negatively impact the quality of RAG outputs, leading to the adage that "misinformation can be worse than no information at all". Studies have shown that including irrelevant documents can paradoxically improve accuracy by over 30%, contradicting the assumption that such inclusion would reduce quality. These findings highlight the need for specialized strategies that integrate retrieval with language generation models more effectively. Addressing the robustness of RAG systems in the face of noisy or misleading data remains a critical challenge for future research [44].

Hybrid Approaches: Combining RAG with Fine-Tuning

Integrating RAG with fine-tuning techniques is emerging as a powerful strategy for improving model performance. Future research should explore the optimal ways to combine RAG and fine-tuning, whether through sequential, alternating, or end-to-end joint training approaches [122]. Another promising avenue is the incorporation of smaller, specialized language models within the RAG framework, where these models can be fine-tuned based on RAG outcomes. For instance, the CRAG model trains a lightweight retrieval evaluator to assess the overall quality of the retrieved documents, triggering different knowledge retrieval actions based on confidence levels [222]. These hybrid approaches offer exciting possibilities for enhancing the efficiency and effectiveness of RAG systems.

Production-Ready RAG

The practical application of RAG technology in production environments presents several engineering challenges that must be addressed to enhance its adoption. Key areas of focus include improving retrieval efficiency, enhancing document recall in large knowledge bases, and ensuring data security, such as preventing the inadvertent disclosure of sensitive information by LLMs [6]. The development of the RAG ecosystem is heavily influenced by advancements in its technical stack. Tools like LangChain and LLamaIndex have become integral components of the RAG landscape, offering extensive APIs and user-friendly interfaces. As RAG continues to evolve, there is a clear trend toward specialization, customization, and simplification of RAG tools and platforms.

In chapter 5 of this thesis, a case study of RAG application in a production environment will presented by introducing WISE, an AI chatbot. WISE utilizes the RAG framework to significantly enhance its information retrieval capabilities, offering a sophisticated and efficient solution across various applications. This showcases the effectiveness and reliability of RAG solutions in today's business landscape.

Multi-Modal RAG

RAG technology is rapidly evolving beyond text-based applications to incorporate multi-modal data, resulting in the creation of innovative models that integrate RAG concepts across various domains. In image processing, audio, video, and even code-related tasks, RAG is being used to enhance capabilities such as retrieval, generation, and automatic processing. This expansion into multi-modal domains marks a significant advancement in the development and application of RAG technology, broadening its impact and utility across diverse fields [67].

The future prospects of RAG technology are both promising and complex. As RAG continues to evolve, it will be essential to address the challenges of robustness, scaling, and production readiness while exploring new frontiers in multi-modal applications. The ongoing development of hybrid approaches and the potential for breakthroughs in scaling laws offer exciting opportunities for advancing RAG systems. With continued research and innovation, RAG technology is poised to become an even more integral part of the AI landscape, enabling more efficient, accurate, and versatile applications across a wide range of domains.

3.11 Future Prospects of LLMs: Beyond the Transformer Architecture

As the field of AI continues to advance, particularly in the areas of Natural Language Generation (NLG) and Retrieval-Augmented Generation (RAG), the underlying architectures powering these models are increasingly coming under scrutiny. As discussed in the previous sections, the transformer architecture has been the cornerstone of many state-of-the-art AI models, including LLMs, celebrated for its scalability and ability to handle a wide range of tasks with high efficiency.

Despite its dominance, the transformer architecture is not without its weaknesses. Chief among these is its quadratic scaling with respect to sequence length, which significantly increases computational costs as the input size grows. This limitation becomes particularly pronounced in tasks that require processing long documents or entire books, where maintaining long-range dependencies is crucial for accurate text embeddings and summaries. The challenge, therefore, lies in developing architectures that can achieve sub-quadratic inference times while maintaining the ability to scale to the massive parameter sizes required for state-of-the-art LLMs [154].

Two emerging architectures, Mamba and BASED, have shown potential in addressing some of these challenges [73, 11]. The Mamba architecture, building on advances in State Space Models (SSM), excels in maintaining and memorizing long-range dependencies, an area where transformers fall short outside their context window. However, Mamba faces challenges in training efficiency, particularly due to its reliance on operations not optimized for modern GPU hardware, and the complexity of its backpropagation process. These hurdles suggest that while Mamba has significant promise, further research is needed to optimize its performance for large-scale LLM applications [73].

The BASED architecture, in contrast, combines short-range convolution with long-range Taylor-series attention, offering strong Associative Recall (AR) capabilities, something that has been a weakness for many transformer challengers. BASED is also designed to run efficiently on existing GPU hardware and can be parallelized for training, making it a more practical contender for large-scale implementation. Its potential for sub-quadratic inference and compatibility with traditional transformer computation methods positions BASED as a particularly attractive alternative, though it too has yet to be tested at the scale of LLMs like ChatGPT or Gemini [11].

While both Mamba and BASED present exciting possibilities, it is unlikely that the transformer will be dethroned in the immediate future. The current infrastructure, research momentum, and the immense costs associated with training new LLMs around these architectures suggest that the transformer will remain the dominant architecture for some time. However, the exploration of Mamba and BASED points to a future where AI architectures are more versatile, efficient, and capable of handling increasingly complex tasks.

Chapter 4

Evaluation of LLMs

4.1 Evaluation of LLMs

The rapid advancement of large language models (LLMs) and their integration into AI chatbots, particularly through Natural Language Generation (NLG) and Retrieval-Augmented Generation (RAG) frameworks, has significantly transformed the landscape of machine-human interaction. These sophisticated systems have become essential tools in various applications due to their ability to generate highly coherent and contextually relevant responses. However, as these models continue to grow in complexity and capability, the need for rigorous and comprehensive evaluation methodologies has become increasingly essential.

Evaluating the performance of LLMs is crucial given their integral role in numerous AI applications that require not only natural language generation but also accurate information retrieval and utilization. As these models are deployed in real-world scenarios, it is imperative to meticulously evaluate their ability to handle diverse inputs, generate accurate and contextually appropriate responses, and adhere to ethical guidelines.

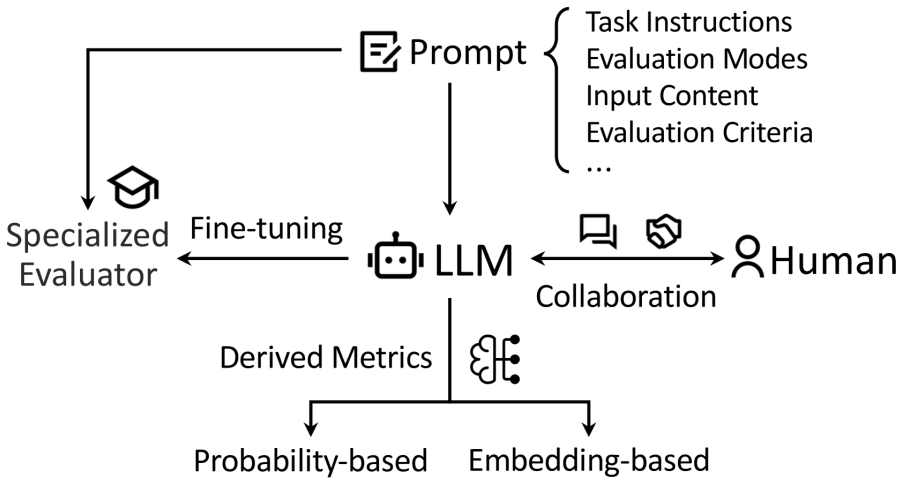

Figure 4.1: A schematic introduction of LLM-based NLG evaluation techniques, that highlights the integration of traditional metrics with advanced LLM-derived methods and human collaboration. *Source:* [65]

Traditionally, NLG evaluation has relied on surface metrics such as BLEU and ROUGE, which measure the n-gram overlap between generated text and reference texts [151, 120]. Although these metrics are valued for their simplicity and ease of use, they have been criticized for their inability to capture the deeper semantic quality of the text, often leading to low correlation with human judgments [175]. In contrast, evaluation of RAG systems goes beyond simple text generation and includes the quality of information retrieval, a critical factor in the model's ability to provide accurate and relevant answers.

This chapter provides a comprehensive exploration of evaluation methodologies for LLMs and RAG systems. It delves into various techniques ranging from traditional metrics like BLEU and ROUGE, which offer a surface-level assessment, to more advanced, LLM-derived methods that better capture the nuanced quality of generated text. By highlighting the unique challenges posed by NLG and RAG, as well as the innovative solutions developed to address them, this chapter aims to provide a comprehensive understanding of the strategies necessary to ensure that LLMs are reliable, ethical, and meet the rigorous standards required for real-world applications, such as the AI chatbots. As LLMs continue to evolve, so too must the evaluation frameworks, adapting to maintain their effectiveness across a broad spectrum of tasks and scenarios.

4.2 Evaluation of Natural Language Generation (NLG)

4.2.1 Model-Based Evaluation Metrics and NLP Tasks

The advent of deep learning has led to the development of more sophisticated evaluation metrics for LLMs that go beyond traditional methods such as BLEU and ROUGE. These advanced metrics, such as BERTScore and BARTScore, exploit pre-trained language models to evaluate generated text with a focus on aspects such as fluency, coherence, and fidelity [233, 230]. BERTScore, for example, calculates the similarity between the embeddings of generated and reference texts, providing a more nuanced evaluation than overlapping n-grams. BARTScore further improves this approach by considering the conditional probability of the generated text given the source text, offering indications of the likelihood that a high-quality language model will produce the content [67].

Although these model-based metrics represent significant improvements over traditional evaluation methods, they are not without their limitations. One major drawback is their dependence on reference texts, which limits their applicability in scenarios where such references are not available. Moreover, despite being more in line with human judgments, these metrics may still have problems with some aspects of text quality, such as robustness in different contexts and efficiency in the use of computational resources [79]. These challenges highlight the need for continuous refinement of evaluation techniques to keep pace with the evolving capabilities of LLMs.

In the context of AI chatbots, evaluation of NLP tasks, which include both comprehension and text generation, is indispensable to ensure that these systems can engage users in meaningful interactions. In particular, Natural language generation (NLG) is a crucial aspect of NLP, involving tasks such as summarization, dialogue generation, machine translation, and open-ended text creation. Evaluation of NLG capabilities is vital in AI chatbots field to determine how effectively these models generate appropriate and contextually relevant responses to user input. Specifically, dialogue generation and question answering are paramount, as they form the basis of a chatbot's ability to communicate naturally and provide informative and accurate answers.

Dialogue Generation

The evaluation of dialogue generation is essential to developing more intelligent and natural dialogue systems. It involves evaluating the model's ability to understand context, generate consistent responses, and maintain a conversation over multiple shifts. Recent studies have shown that models such as Claude and ChatGPT outperform earlier versions in dialogue tasks, with Claude demonstrating slight advantages in specific configurations [123, 155]. Fine-tuning LLMs for specific tasks has been found to significantly improve performance, with fine-tuned models often outperforming generic models such as ChatGPT in task-oriented, knowledge-based dialogue contexts [19].

Question Answering (QA)

Question Answering is another pivotal task for AI chatbots, particularly in applications such as search engines, intelligent customer service, and specialized QA systems. A chatbot's accuracy and efficiency in answering questions are key indicators of its performance. Recent evaluations have shown that models such as InstructGPT davinci v2 (175B) excel in accuracy, robustness, and correctness in various QA scenarios [150, 119]. Although ChatGPT has made great strides in improving its QA capabilities, it still faces challenges in specific benchmarks such as CommonsenseQA and Social IQA, where its cautious approach sometimes leads to denial of answers when information is insufficient. Nevertheless, fine-tuned models such as Vícuna and ChatGPT itself continue to demonstrate outstanding performance in QA tasks, underscoring the importance of task-specific optimization to achieve high accuracy and relevance [17].

Evaluation of NLP tasks, particularly dialogue generation and question answering, is essential to ensure that AI chatbots can interact effectively with users. As LLMs continue to advance, accurately assessing and improving these capabilities will be critical to developing more sophisticated and reliable chatbot systems. In addition, continuous improvement of model-based evaluation metrics is essential to capture the full spectrum of capabilities and challenges posed by these advanced models, ensuring that they meet the rigorous standards required for real-world applications.

4.2.2 LLM-Derived Metrics

The emergence of LLMs such as InstructGPT has significantly transformed the landscape of NLG evaluation [150]. Researchers have increasingly turned to LLM-derived metrics that leverage the linguistic and contextual capabilities of these models to more effectively assess the quality of generated text. These metrics go beyond traditional n-gram-based methods by assessing the semantic similarity between generated and reference texts using embeddings produced by LLMs. For example, OpenAI's text-embedding-ada-002 model measures similarity scores between texts, with higher scores indicating closer alignment with the desired quality of the output [58].

In addition to semantic similarity, probability-based approaches have also been introduced that offer more dynamic and context-aware evaluation. GPTScore, for example, uses customized evaluation templates to guide multiple LLMs in assessing various aspects of NLG, such as fluency and coherence, by calculating the probability of the generated text [63]. These metrics provide a nuanced understanding of text quality, making them a valuable tool for evaluating LLM outputs.

However, despite the advantages of LLM-derived metrics, they are not without their challenges. One of the main problems is robustness. These metrics can be vulnerable in attack scenarios, where adversary inputs can reveal blind spots that traditional metrics might overlook [79]. In addition, LLM-derived evaluation methods are computationally intensive and often require significant resources, which may limit their applicability in large-scale evaluations. Another

critical issue is fairness. These metrics have been found to have social biases, particularly with regard to sensitive attributes such as race, gender, and age, which can lead to biased assessment results [177]. These challenges underscore the need for continued research to improve the robustness, efficiency, and fairness of LLM-derived metrics.

4.2.3 Prompting LLMs for NLG Evaluation

As the capabilities of LLMs continue to evolve, so do the methods used to assess their performance in NLG. One innovative approach that has emerged is the use of prompts to directly guide LLMs in assessing their own outputs. This method involves the creation of specific prompts that include task instructions, evaluation criteria, and the text to be evaluated, allowing the LLM to autonomously generate the evaluation results [65]. The promise of this approach lies in its ability to replicate human-like evaluation processes, making it a valuable tool for assessing the quality of generated text.

Two common methods in prompt-based assessment are scoring and comparison. In scoring, LLMs evaluate text quality on a scale, a method that has shown strong correlation with human judgments in various NLG tasks, such as summarization and dialogue generation [33]. Comparison methods, on the other hand, ask LLMs to choose between two generated texts, often proving more reliable than the absolute score [134]. In addition, ranking allows LLMs to sort through multiple texts, providing a broader perspective on their evaluation abilities [95].

Despite its potential, prompt-based assessment has several limitations. Some studies have highlighted problems such as position bias, in which the order in which texts are presented influences the outcome of the evaluation [200]. It has also been found that LLMs prefer longer and more verbose responses, sometimes even favoring their own generated outputs over those produced by other models [237, 125]. Moreover, these models showed a tendency to evaluate responses with factual errors more favorably than shorter and grammatically correct responses [219]. Biases, particularly in scoring high quality summaries and non-Latin languages, such as Chinese and Japanese, continue to be a concern [75]. These challenges underscore the need to continually refine prompt-based scoring methods to ensure fairness, accuracy, and robustness.

Automatic Evaluation Methods

Automated evaluation remains a cornerstone in the evaluation of LLMs and Retrieval-Augmented Generation (RAG) systems due to its efficiency, scalability and objectivity. Compared with human evaluation, automated evaluation does not require intensive human participation, which not only saves time but also reduces the impact of human subjective factors and makes the evaluation process more standardized. By using standard metrics and automated tools, this method facilitates the evaluation of model performance in different tasks with minimal human intervention, thus reducing potential biases and enabling rapid evaluation of large volumes of data.

Based on the literature that adopted automatic evaluation, common metrics in automatic evaluation include accuracy, calibration, fairness, and robustness.

- Accuracy is a concept that may vary in different scenarios and is dependent on the specific task at hand. However, accuracy tipically is measured using metrics such as Exact Match (EM), F1 score, and ROUGE, which evaluate how well the model's output aligns with a reference answer [27].
- Calibration, on the other hand, assesses the model's confidence levels, ensuring that the predicted probabilities reflect the actual likelihood of correctness. The most commonly used metric to evaluate model calibration performance is Expected Calibration Error (ECE) [74].

- Fairness metrics evaluate whether the model's performance is consistent across different demographic groups, thereby mitigating potential biases. These metrics include Demographic Parity Difference (DPD) and Equalized Odds Difference (EOD) [196].
- Robustness metrics like Attack Success Rate (ASR) and Performance Drop Rate (PDR) measure the model's resilience to adversarial inputs and out-of-distribution data [240].

General metrics	**Metrics**
Accuracy	Exact match, Quasi-exact match, F1 score, ROUGE score
Calibration	Expected calibration error, Area under the curve
Fairness	Demographic parity difference, Equalized odds difference
Robustness	Attack success rate, Performance drop rate

Table 4.1: General metrics and their corresponding evaluation metrics. *Source:* [27]

The increasing sophistication of LLMs has led to the development of advanced automatic evaluation tools such as LLM-EVAL and PandaLM, which offer multidimensional evaluation frameworks that enhance the thoroughness and reproducibility of assessments by training an LLM that serves as the "judge" to evaluate different models [123, 202]. These tools are often integrated into benchmarks like MMLU, HELM, C-Eval and Chatbot Arena, further standardizing the evaluation process across different tasks and domains, thereby providing a more comprehensive picture of LLM performance [27].

Human Evaluation Methods

While automatic evaluation provides valuable quantitative insights, there are certain tasks where the nuanced understanding that only human evaluation can offer is indispensable. This is particularly true for open-ended generation tasks, where the subjective quality of the text, including aspects such as fluency, relevance, and alignment with human values, must be assessed. In such cases, embedded similarity metrics like BERTScore may fall short, making human judgment essential [145].

Human evaluation typically involves experts, researchers, or lay users who assess LLM and RAG system outputs based on criteria such as accuracy, relevance, fluency, transparency, safety, and human alignment. These key evaluation metrics are crucial for ensuring the quality and appropriateness of generated content. As introduced in the previous section, accuracy ensures that the generated content is factually correct, while relevance checks whether the output is pertinent to the context or query. Fluency evaluates the readability and coherence of the text, transparency examines the clarity of the model's decision-making process, safety focuses on avoiding harmful or inappropriate content, and human alignment ensures that the output respects societal norms and user expectations [27].

Moreover, the number of evaluators plays a significant role in ensuring the reliability of the evaluation. As highlighted in Table 4.2, having adequate representation and statistical significance in the number of the evaluators is critical to achieving meaningful results. Additionally, the evaluator's expertise level, including their relevant domain expertise, task familiarity, and methodological training, directly impacts the evaluation's accuracy and reliability.

However, also human evaluation is not without challenges. It is often resource-intensive, time-consuming and susceptible to variability due to cultural and individual differences among evaluators. Ensuring a representative number of evaluators and defining clear evaluation criteria are critical to mitigating these challenges. Furthermore, the expertise level of the evaluators plays

Evaluation Criteria	**Key Factor**
Number of evaluators	Adequate representation, Statistical significance
Evaluation rubrics	Accuracy, Relevance, Fluency, Transparency, Safety, Human alignment
Evaluator's expertise level	Relevant domain expertise, Task familiarity, Methodological training

Table 4.2: Summary of key factors in human evaluation. *Source:* [27]

a crucial role in the reliability of the evaluation, particularly in domain-specific tasks where deep subject knowledge is required [27].

Recognizing the strengths and limitations of both LLMs and human evaluators, researchers have proposed a collaborative approach that leverages the best of both worlds. In this hybrid method, LLMs generate initial evaluations, which are then reviewed and refined by human evaluators. This collaborative process has shown promise in reducing the workload of human evaluators while maintaining high accuracy. Techniques like the COEVAL pipeline depicted in Figure 4.2 combine LLM-generated evaluations with human scrutiny. This method results in more reliable and nuanced evaluations, particularly for complex or open-ended tasks [116, 234].

However, also the collaborative approach is not without its drawbacks. It still requires significant human involvement, which can limit its scalability and cost-effectiveness compared to fully automated methods. As research progresses, it will be crucial to develop unified benchmarks and explore new evaluation scenarios that fully realize the potential of LLMs in NLG evaluation. This evolution in evaluation methods aims to address the persistent challenges of bias, efficiency, and robustness, ensuring that LLMs can be effectively and fairly assessed across diverse applications [67].

In practice, a combination of automatic and human evaluation methods is often employed to achieve a more comprehensive assessment of chatbot performance. This hybrid approach allows the strengths of both methods to be harnessed, providing a more balanced and thorough evaluation of AI chatbots that use LLMs and RAG technologies. As the field continues to evolve, the integration of these methods will be key to refining evaluation processes and ensuring that AI systems meet the high standards required for real-world applications.

Evaluation of Factuality, Robustness, and Trustworthiness

Evaluating the factuality, robustness, trustworthiness, and ethic aspects of AI chatbots that leverage LLMs is essential, particularly as these systems are increasingly deployed in real-world scenarios where they must handle unexpected inputs and adversarial attacks.

Factuality: Beyond the general quality of generated text, the factuality of LLM outputs is a crucial aspect of their evaluation. Factuality refers to the degree to which the information or responses generated by the model align with real-world truths and verifiable facts. This aspect is particularly crucial in applications like AI chatbots, where accuracy and reliability are essential, especially in tasks such as question answering (QA) systems, dialogue systems, information extraction, text summarization, and automated fact-checking.

Errors in the information provided by LLMs can lead to misunderstandings, misinformation, and potentially harmful consequences. Therefore, factuality evaluation involves assessing the model's ability to remain consistent with known facts, avoid generating misleading or false information, often referred to as "factual hallucination", and effectively learn and recall factual knowledge.

Several methodologies have been proposed to measure and enhance the factuality of LLMs.

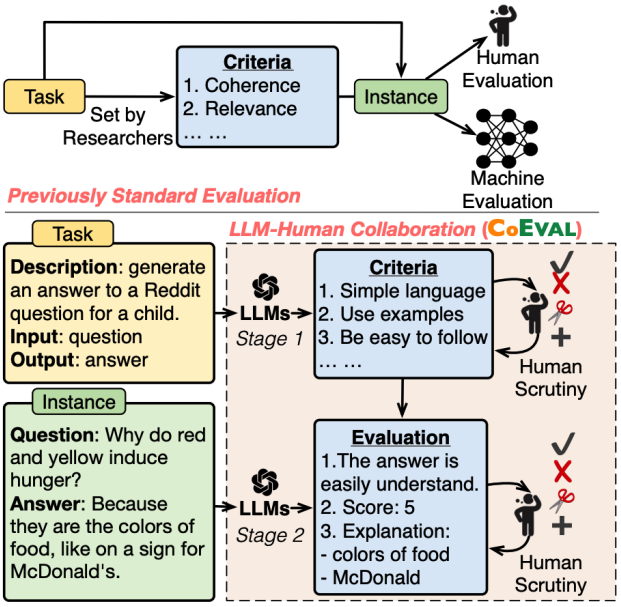

Figure 4.2: The COEVAL pipeline demonstrates a collaborative evaluation approach, combining LLM-generated ideation with human scrutiny to refine and validate the evaluation process. This method, illustrated in the lower part of the figure, contrasts with conventional evaluation methods (upper part) by integrating both machine-driven and human-refined assessments. *Source:* [116]

For instance, Wang et al. conducted an assessment of several large models, including InstructGPT, GPT-3.5, GPT-4, and BingChat, by evaluating their performance in answering open questions. Their study, which involved human evaluation, found that while models like GPT-4 and BingChat provided correct answers for over 80% of the questions, there remained a significant accuracy gap of over 15%, indicating that further improvements are necessary to achieve complete factual correctness [197].

To improve factuality evaluation methods, Honovich et al. reviewed current approaches and identified a lack of a unified comparison framework. They addressed this by converting existing factual consistency tasks into binary labels that assess whether there is a factual conflict with the input text. This approach, which does not rely on external knowledge such as the RAG framework, has shown that methods based on natural language inference (NLI) and question generation answering can complement each other effectively in evaluating factuality [82].

The TruthfulQA dataset, introduced by Lin et al., has become a widely used tool for evaluating the factuality of LLMs. Designed to challenge models with scenarios where producing factual answers is difficult, TruthfulQA tests the models' ability to remain truthful under challenging conditions. Findings from experiments using TruthfulQA suggest that merely scaling up model sizes does not necessarily improve truthfulness, highlighting the need for more sophisticated training approaches [121].

The evaluation of factuality in LLMs is a complex but essential task, with significant implications for the reliability and trustworthiness of AI systems that leverage these models. As research progresses, refining these evaluation methods will be critical to ensuring that LLMs can consistently provide accurate and truthful information across a wide range of applications.

Simultaneously, ongoing advancements in LLM-derived metrics will continue to play a crucial role in the broader evaluation of NLG capabilities, addressing issues of robustness, efficiency, and fairness. From the discourse presented, it becomes evident that integrating external knowledge, as seen in the RAG framework, offers a viable solution to enhance the factuality of LLMs in NLG.

Robustness: Robustness evaluation focuses on how well AI chatbots handle unexpected or out-of-distribution (OOD) inputs, as well as adversarial prompts designed to manipulate the system [199]. Robustness is crucial for ensuring that LLMs maintain their performance and reliability even when faced with inputs that deviate from the norm. Early evaluations of models such as ChatGPT revealed potential security risks when these systems were exposed to adversarial inputs or manipulated through visual input, underscoring the need for more resilient models [223]. Further studies have shown that contemporary LLMs remain vulnerable to adversarial text attacks at various levels, from character-level perturbations to more complex semantic manipulations [240]. These findings highlight the importance of developing models that can withstand diverse and potentially malicious inputs while maintaining their intended functionality.

Trustworthiness: Trustworthiness is another vital aspect of evaluating the performance of AI chatbots. Trustworthiness encompasses the model's ability to provide accurate, ethical, and unbiased responses. This aspect is important for maintaining user trust and ensuring that AI systems are deployed responsibly. Studies such as DecodingTrust have expanded the scope of trustworthiness evaluation to include dimensions like toxicity, stereotype bias, adversarial robustness, and fairness. While advanced models like GPT-4 may show improvements in these areas, they are not immune to certain vulnerabilities, including susceptibility to cognitive biases and ethical inconsistencies. For instance, research has indicated that while LLMs can avoid common cognitive errors, their consistency in judgment can be compromised by factors such as questioning, negation, or misleading cues, raising concerns about their reliability in real-world scenarios [196].

Ethics and Bias: The ethical implications and biases of AI chatbots are critical areas of evaluation, particularly as these systems are deployed in sensitive or high-stakes environments. LLMs have been found to internalize and perpetuate harmful biases present in their training data, leading to outputs that may include offensive language, hate speech, or stereotypes related to gender, race, religion, and other demographic characteristics. These biases not only compromise the fairness of the system but also pose significant risks in applications where unbiased and equitable interactions are paramount. Recent studies have systematically evaluated the presence of such biases in models like ChatGPT, revealing that despite advancements, these models continue to exhibit toxic and biased content [241]. Moreover, role-playing scenarios have been shown to exacerbate these biases, leading to increased toxicity and biased outputs up to 6 times toward specific entities [47]. These ethical concerns highlight the need for ongoing evaluation and mitigation strategies to ensure that AI chatbots provide equitable and non-discriminatory interactions, fostering trust and minimizing potential harm to users. The topics of ethics and bias, along with strategies for their mitigation, will be covered in detail in Chapter 6.

In conclusion, while refining the evaluation of LLMs in NLG is vital for ensuring their effectiveness across various applications, it's important to recognize that the RAG framework presents a promising solution to some of the inherent challenges in NLG. However, RAG itself introduces unique complexities that must be carefully evaluated to ensure that both retrieval and generation processes meet the required standards of accuracy, reliability, and relevance.

4.3 Evaluation of Retrieval-Augmented Generation (RAG)

4.3.1 RAG-Specific Evaluation Metrics

The evaluation of Retrieval-Augmented Generation (RAG) systems extends beyond the conventional assessment of text generation to include the aspect of information retrieval. Unlike standard Natural Language Generation (NLG) tasks, RAG systems are evaluated on two primary fronts: retrieval quality and generation quality.

Retrieval quality refers to the effectiveness of the retriever component within the RAG system in sourcing relevant information from external databases or knowledge bases. Standard metrics from related fields—such as search engines, recommendation systems, and information retrieval systems—are employed to measure the performance of the retrieval module. For instance, precision, recall, and F1 score are commonly used to evaluate how well the retrieved documents or information chunks align with the intended query.

The assessment of generation quality in RAG systems is pivotal for evaluating the model's ability to produce coherent, relevant, and contextually accurate responses derived from the retrieved content. This aspect of evaluation is critical, as it determines the effectiveness of the model in synthesizing information to generate responses that meet the user's requirements. Generation quality can be categorized based on the nature of the content: unlabeled and labeled. For unlabeled content, the evaluation focuses on ensuring the faithfulness, relevance, and non-harmfulness of the generated responses, thereby guaranteeing that the output is both meaningful and safe. In contrast, labeled content evaluation emphasizes the accuracy of the information produced by the model [67].

Contemporary evaluation methodologies for RAG models emphasize three primary quality scores—context relevance, answer faithfulness, and answer relevance—alongside four essential abilities that collectively inform the evaluation process.

- **Quality Scores:** Quality scores evaluate the efficiency of the RAG model from different perspectives in the process of information retrieval and generation.
  - *Context Relevance* assesses the precision and specificity of the retrieved context, ensuring that the content is directly pertinent to the query while minimizing extraneous information, thus reducing processing overheads.
  - *Answer Faithfulness* evaluates whether the generated responses remain true to the retrieved context, maintaining consistency and avoiding contradictions in the output.
  - *Answer Relevance* ensures that the generated responses directly address the posed questions, thereby effectively meeting the user's inquiry with pertinent information [58, 164].
- **Required Abilities:** The following abilities are critical for the model's performance under complex scenarios, impacting the quality scores.
  - *Noise Robustness* evaluates the model's capacity to handle noisy or irrelevant documents that may be loosely related to the query but do not provide substantial information. This ability is crucial for filtering out distractions and focusing on high-quality data.
  - *Negative Rejection* assesses the model's discernment in refraining from generating a response when the retrieved documents do not contain the requisite knowledge to answer the query, thereby preventing the propagation of misinformation.

- *Information Integration* measures the model's proficiency in synthesizing information from multiple documents to formulate a comprehensive response to complex queries, demonstrating its capability to aggregate and process diverse sources.
- *Counterfactual Robustness* tests the model's ability to recognize and disregard known inaccuracies within documents, ensuring that it does not disseminate misinformation even when confronted with misleading content [28, 126].

These metrics and abilities provide a comprehensive framework for the evaluation of generation quality in RAG systems, ensuring that the model not only generates text aligned with the retrieved data but also directly and accurately addresses the user's queries. Such evaluation is indispensable for determining the overall effectiveness of RAG systems in practical applications such as AI chatbots, where the retrieval and integration of accurate information are essential. A practical implementation of some of the seen metrics will be presented in Section 5.2.1.

It is important to recognize that while these metrics, derived from extant literature, provide valuable insights, they do not yet constitute a mature or standardized approach for quantifying RAG evaluation aspects. Custom metrics tailored to the specific nuances of RAG models, although not discussed herein, have also been developed in some evaluation studies to address particular challenges.

4.3.2 Challenges in RAG Evaluation

Evaluating RAG systems presents unique challenges that go beyond those encountered in traditional NLG evaluation. One significant challenge is managing noise in the retrieved documents. RAG systems must effectively filter out irrelevant or misleading information to prevent it from degrading the quality of the generated responses. This requires robust noise management techniques that enable the system to disregard extraneous data and focus on the most pertinent information [67].

Another challenge is information integration, particularly in scenarios involving multi-hop question answering, where the system must synthesize information from multiple sources to construct a coherent and accurate response. This task is especially complex when the information from different sources is contradictory or incomplete, necessitating advanced techniques for integrating and validating the information before it is used in the generation process [133].

Additionally, RAG systems must exhibit counterfactual robustness, which involves recognizing and disregarding known inaccuracies within the retrieved documents. This is essential to prevent the propagation of incorrect information, especially in contexts where the reliability of the generated content is paramount. The ability to filter out or correct inaccurate data, even when presented as potential answers, is a pivotal aspect of ensuring the trustworthiness and accuracy of RAG-generated content [114].

4.4 Benchmarks for LLM and RAG Systems

The evaluation of AI chatbots that leverage LLMs and RAG systems requires comprehensive benchmarks that assess their performance across a variety of tasks. This section introduces the benchmarks used for general tasks, specific downstream tasks such as question answering, and multi-modal tasks, which are essential for evaluating the holistic performance of these models.

4.4.1 Benchmarks for General Tasks

LLMs are designed to handle a wide array of tasks, making it crucial to evaluate their performance across multiple dimensions. Benchmarks like Chatbot Arena [130] and MT-Bench [237] play a significant role in this regard. Chatbot Arena offers a platform where users can interact with anonymous chatbot models, casting votes based on their experiences. This real-world engagement allows for the assessment of chatbot models in practical settings, providing valuable insights into their strengths and limitations. Similarly, MT-Bench focuses on evaluating LLMs in multi-turn dialogues, which are essential for simulating realistic conversational scenarios. This benchmark is particularly useful for understanding how well a chatbot can manage extended interactions, a critical aspect of NLP tasks.

Additionally, benchmarks such as HELM [119] and DynaBench [103] provide a broader evaluation of LLMs across various NLP tasks, including language comprehension and robustness to adversarial inputs. HELM offers a comprehensive assessment of LLMs' language understanding capabilities, while DynaBench supports dynamic benchmark testing, exploring the effects of distributional shifts and model robustness in interactive settings. These benchmarks contribute to a more nuanced understanding of LLM performance, especially in diverse and challenging scenarios.

4.4.2 Benchmarks for Specific Downstream Tasks

While general benchmarks provide an overarching view of LLM performance, specific downstream tasks require more focused evaluation. Question-answering benchmarks, such as MultiMedQA and FRESHQA, assess how effectively a chatbot can retrieve and generate accurate answers. MultiMedQA, for instance, focuses on medical questions, evaluating a model's clinical knowledge and ability to handle complex queries in the healthcare domain. FRESHQA, on the other hand, tests the chatbot's ability to incorporate up-to-date information from current world knowledge, ensuring relevance and accuracy in dynamic environments [174, 193].

For more complex dialogue and reasoning tasks, benchmarks like Dialogue CoT and ARB provide targeted assessments. Dialogue CoT evaluates LLMs' capabilities in conducting coherent and contextually relevant conversations, while ARB probes their performance in advanced reasoning tasks that span multiple domains. These benchmarks are instrumental in understanding how well LLMs can perform in specialized and challenging tasks that go beyond basic question answering [198, 166].

4.4.3 Benchmarks for Multi-modal Tasks

In the evolving landscape of AI, chatbots are increasingly required to handle multi-modal inputs, such as images, text, and even audio. Evaluating these capabilities necessitates benchmarks specifically designed for multi-modal tasks. MME and MMBench are two such benchmarks that rigorously assess the perceptual and cognitive abilities of Multi-modal Large Language Models (MLLMs). MME uses instruction-answer pairs to evaluate models under controlled conditions, while MMBench offers a comprehensive dataset for evaluating vision-language models [227, 128].

These benchmarks ensure that MLLMs are not only capable of understanding and generating text but can also effectively interpret and respond to visual inputs. As MLLMs continue to evolve, benchmarks like SEED-Bench further extend their evaluation to cover a wide range of tasks, including pattern recognition in images and videos, providing a holistic assessment of multi-modal language models [115].

The development of robust and comprehensive benchmarks is essential for advancing the evaluation of AI chatbots that utilize LLMs and RAG systems. These benchmarks enable researchers and developers to systematically assess and improve the performance of chatbots across general tasks, specific downstream tasks, and multi-modal tasks. As AI technology continues to progress, these benchmarks will play a critical role in ensuring that chatbots can meet the diverse and complex demands of real-world applications.

4.5 Success and Failure Cases of LLMs

Large Language Models (LLMs) have demonstrated remarkable capabilities across a wide range of tasks, yet they are not without limitations. Understanding both their strengths and weaknesses is essential for evaluating the performance of AI chatbots, particularly in generating dialogue and answering questions.

One of the primary strengths of LLMs lies in their ability to generate text with a high degree of fluency and precision. This capability is evident in tasks such as machine translation, text generation, and question answering, where LLMs consistently produce coherent and contextually appropriate responses.

In addition to text generation, LLMs excel in language understanding tasks. They perform impressively in sentiment analysis, text classification, and handling factual input, showcasing their ability to comprehend and process natural language effectively. Furthermore, LLMs demonstrate robust arithmetic and logical reasoning capabilities, making them well-suited for tasks that require complex calculations or structured data inference. Their proficiency extends to temporal reasoning, where they can accurately interpret and manage time-related information.

The robust contextual comprehension of LLMs enables them to generate responses that are not only accurate but also align well with the input provided, making them effective in dialogue systems and conversational AI.

Despite these strengths, LLMs also exhibit several notable limitations that can affect their performance in certain contexts. One of the primary challenges LLMs face is in tasks requiring nuanced understanding, such as Natural Language Inference (NLI). Here, they struggle to accurately represent human disagreements and may perform poorly in discerning subtle semantic similarities between events. This limitation extends to abstract reasoning, where LLMs often encounter confusion or errors, particularly in complex or ambiguous contexts.

LLMs also demonstrate suboptimal performance when processing linguistic contexts that involve non-Latin scripts or are resource-constrained. Their ability to generate accurate and contextually relevant outputs diminishes significantly in these scenarios, highlighting a gap in their linguistic capabilities across diverse languages and writing systems.

Moreover, LLMs are not immune to the biases and toxic content embedded in the vast datasets on which they are trained. They can inadvertently assimilate and propagate offensive or biased language, which poses significant ethical concerns, particularly in sensitive applications such as social media moderation or customer service.

Another critical limitation of LLMs is their difficulty in incorporating real-time or dynamic information. This makes them less effective in tasks that require up-to-date knowledge or the ability to rapidly adapt to changing circumstances. Additionally, LLMs are particularly vulnerable to adversarial prompts, which can exploit weaknesses in their training and result in incorrect or harmful outputs [27].

Understanding these success and failure cases is crucial for effectively deploying LLMs in real-world applications. By recognizing where LLMs excel and where they fall short, developers and researchers can better design evaluation frameworks that ensure AI chatbots perform reliably

and ethically across a wide range of tasks.

4.6 Conclusions and Future Directions

The evaluation of large language models (LLMs) within natural language generation (NLG) and retrieval augmented generation (RAG) systems is a dynamic and rapidly evolving area of research. Traditional evaluation metrics, such as BLEU and ROUGE, have historically laid the foundation for the evaluation of language models. However, as LLMs have progressed, assessment methodologies have evolved, leading to the introduction of sophisticated metrics derived from LLMs such as BERTScore and GPTScore. These advanced metrics facilitate more nuanced evaluations that transcend surface comparisons, offering insights into crucial aspects of text quality such as fluency, coherence, and faithfulness, essential elements for accurate, human-aligned evaluations.

Despite these advances, significant challenges persist, particularly in the evaluation of RAG systems. The complexities of managing noise in retrieved data, synthesizing information from multiple sources, and ensuring counterfactual robustness underscore the need for specialized evaluation tools. As RAG systems become an increasingly integral part of advanced AI applications, it will be critical to develop metrics that accurately reflect these complexities.

Looking forward, the development of unified benchmarks tailored to the distinct needs of NLG and RAG systems will be imperative. Such benchmarks will foster more consistent and comprehensive assessments across different tasks and domains, ultimately contributing to the creation of more reliable and versatile AI systems. In addition, exploring new assessment scenarios, particularly those involving low-resource languages and complex tasks, is essential to unlocking the full potential of LLMs. These scenarios often present unique challenges that are not adequately captured by current frameworks, necessitating the development of more nuanced and context-aware metrics.

The future of LLM evaluation lies in improving collaborative frameworks that combine human judgment with automated methods. By leveraging the strengths of both, researchers can develop evaluation systems that ensure that LLMs provide accurate, fair, and contextually relevant results across a broad spectrum of applications. This hybrid approach has the potential to mitigate the inherent limitations of human and automated assessments, leading to more balanced and reliable evaluations.

A crucial challenge in the evaluation of LLMs is the need for comprehensive behavioral assessment. As artificial intelligence systems approach the threshold of artificial general intelligence (AGI), it becomes increasingly important to assess their behavior in open, real-world environments. This involves not only evaluating their performance on specific tasks, but also understanding their decision-making processes and their adaptability to dynamic scenarios. Incorporating multimodal dimensions into these assessments, where LLMs are evaluated as central controllers in complex systems such as robotics, could provide a more holistic understanding of their capabilities.

In addition, the dynamic and evolving nature of LLMs presents a challenge to traditional static evaluation protocols. As these models continue to improve, there is a risk that they will become too familiar with existing benchmarks, leading to inflated performance metrics that do not accurately reflect their capabilities in the real world. To solve this problem, future evaluation systems must be dynamic and able to evolve along with the models they evaluate. This could involve creating adaptive benchmarks that evolve over time or implementing real-time evaluation methods that continuously test models with new and unseen data.

In conclusion, the field of LLM evaluation is at a crucial stage, with significant opportunities to

advance the understanding and evaluation of these powerful models. By addressing the challenges of comprehensive behavioral assessment, robustness, and dynamic evaluation, researchers can ensure that LLMs continue to improve in meaningful and measurable ways. As LLMs become increasingly integral to a wide range of applications, the development of more comprehensive and forward-looking evaluation methodologies will be crucial to driving the continued success and ethical deployment of AI technologies.

Chapter 5

Case Study Analysis: WISE and the LLM Evaluator

5.1 WISE: a Case Study

In this chapter, a case study is presented and discussed focusing on WISE, a proprietary chatbot developed by HPA: High Performance Analytics. WISE utilizes the Retrieval-Augmented Generation (RAG) framework to enhance its information retrieval capabilities, offering a comprehensive and efficient solution for a wide range of applications. In addition, this chapter will introduce my internship project developed at HPA during spring 2024, the LLM Evaluator, illustrating its implementation, methodologies, and potential work improvements. This overview will provide insights into the development and evaluation of advanced AI systems in practical settings.

5.1.1 HPA: High Performance Analytics

HPA is a spin-off of the University of Verona and AI Competence Center (AICC) of Terranova Software [182]. Since its inception in 2017, HPA has been dedicated to designing and developing AI solutions tailored for small and medium-sized enterprises and large corporations. The company adopts of the motto "Math to Innovate," which emphasizes the deep mathematical expertise accumulated over two decades of academic research.

With six years of experience in the field, HPA has a qualified team of experts and has developed proprietary techniques that set it apart in the field of AI. Over the years, the company has completed more than 30 projects, showcasing its ability to provide high-quality AI solutions.

HPA specializes in various application fields, including predictive analytics, anomaly detection, image recognition, combinatorial optimization, generative AI, and NLP. The company's solutions are implemented in diverse industries, such as energy and utilities, transportation and logistics, manufacturing, security, and payments. This breadth of applications demonstrates HPA's capability to adapt AI technologies to effectively meet the needs of specific industries.

Leveraging its extensive experience and innovative approaches, HPA continues to contribute to AI developments, providing effective solutions that address a wide range of business challenges.

5.1.2 Introducing WISE

For decades, the process of searching and consulting documents and manuals in companies has followed the same paradigm. This traditional method involves identifying keywords, clicking on each link, opening the documents, and finally selecting and collecting the information of interest. This workflow is not only time-consuming and repetitive, but also highly inefficient, consuming valuable time and resources.

WISE improves document management by leveraging generative AI. Using LLMs embedded in a proprietary workflow, WISE enables users to perform complex natural language searches across multiple sources and entire databases.

When a question is asked, WISE queries relevant documents and data, extracts relevant information, and provides an answer within seconds. Answers are provided in a textual, conversational format, mimicking the interaction one would have with an experienced colleague. WISE serves as an always-available, secure, reliable, and knowledgeable virtual assistant, streamlining the document management process. [89]

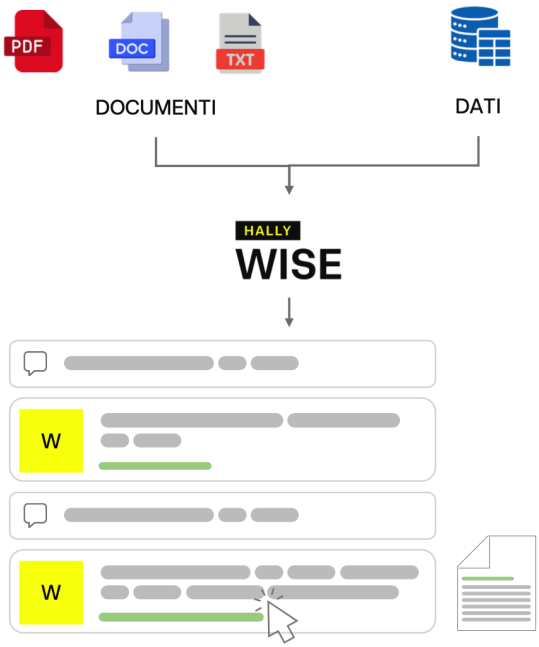

Figure 5.1: WISE system integrates various document types and databases to provide quick, accurate responses to user queries, significantly enhancing the efficiency of document management. *Source:* [89]

5.1.3 WISE Functionality and Features

WISE operates through a series of systematic steps that ensure efficient and accurate information retrieval. The following steps outline the core functionality of WISE:

1. The company uploads documents that form the knowledge base (manuals, regulations, circulars, etc.) into WISE.
2. The user asks a question in natural language as if conversing with a colleague.
3. WISE provides the best possible answer within seconds, citing available sources. If the user requests more details, WISE remembers the conversation and contextualizes the new request.
4. WISE can be easily integrated into the company website and can connect to databases and enterprise systems (ERP, CRM) via APIs and custom connectors.

Figure 5.2: An example of the WISE chatbot interface. Users can interact with WISE via a chat interface, asking questions and receiving detailed responses, filtering on the source documents. *Source:* [89]

Key Features

Today, only 35% of a company's data is available to management. Dashboards and reports are often outdated and analysts struggle to meet user demands. WISE addresses these challenges by providing a single point of access to the entire business knowledge base, ensuring accurate and up-to-date answers. The following features highlight the comprehensive functionality of WISE:

- **Automatic Summaries, Comparisons, and Inconsistency Checks:** WISE can generate concise summaries, compare documents, and identify inconsistencies. This feature saves users time by automatically creating summaries of long documents, highlighting key points and identifying discrepancies between different documents. In this way, the user is able to quickly understand and use the information.
- **Automatic Text Generation:** WISE can create various text documents based on the data it accesses. This includes the generation of user manuals, contracts, articles and other documentation, ensuring that all generated content is consistent with the latest data

and company standards. The system simplifies the documentation process, reducing the workload of employees.

- **Integration with Enterprise Systems (CRM, ERP):** WISE integrates seamlessly with other enterprise systems, increasing their usefulness and reach. By connecting to Customer Relationship Management (CRM) and Enterprise Resource Planning (ERP) systems, WISE can access and leverage data from different business functions. This integration provides users with a comprehensive view of the organization's data and processes, facilitating decision making.

- **Multilingual by Design:** WISE supports multiple languages to ensure that users with different language backgrounds can easily interact with the system. By transparently managing foreign language conversations, WISE facilitates seamless communication and data retrieval in multinational companies.

- **Multimodality:** Supports images and videos, enabling users to query multimedia content and improving the scope of information retrieval. This feature enables WISE to process and understand visual data, making it possible to search and retrieve information from images, diagrams, and video content. It is particularly useful in areas such as manufacturing and security, where visual data is critical.

- **Response Evaluation and RLHF:** Users can evaluate the responses provided by WISE, facilitating continuous improvement through Reinforcement Learning from Human Feedback (RLHF). This feedback mechanism helps refine artificial intelligence models and improve the accuracy and relevance of the information provided. It ensures that WISE evolves based on user interactions and feedback, leading to improved performance over time.

- **Spreadsheet Support (XLS):** WISE can interpret and retrieve information from spreadsheet files. This functionality is critical for organizations that rely on spreadsheets for data management and reporting. WISE can read, process, and extract relevant information from complex Excel files, providing users with fast and accurate access to data.

- **Supports Structured and Unstructured Databases:** WISE can handle different types of data formats, providing flexibility and comprehensive search capabilities. Whether data is organized in structured formats such as SQL databases or unstructured forms such as text documents and e-mail, WISE can efficiently process and retrieve relevant information. This versatility makes it an indispensable tool for organizations with diverse data storage systems.

- **Two-Factor Authentication (2FA) or Enterprise Single Sign-On (SSO):** These methods provide secure access to the system. Two-factor authentication adds an additional layer of security by requiring users to provide two forms of identification before accessing the system. Enterprise single sign-on allows employees to log in with their corporate credentials, simplifying the authentication process while maintaining high security standards.

- **Uses Triggers and Actions for Process Automation:** This feature enhances workflow automation by responding to specific triggers with predefined actions. For example, WISE can automatically send notifications, update records, or initiate workflows based on user requests or data changes. This automation capability improves operational efficiency and reduces the potential for human error.

Interaction Modes

To provide access to these features, WISE offers multiple interaction modes to cater to various user preferences and needs:

- **Classic Conversational Interface (Chat):** Allows users to interact with WISE through a text-based chat interface, as depicted in Figure 5.2.
- **Voice Interface (Speech-to-Text):** Enables users to communicate with WISE using their voice, which is converted to text for processing.
- **Avatar (Text-to-Speech):** Provides an engaging and interactive experience by converting text responses into speech, represented by an avatar.

5.1.4 WISE Technology Stack

The WISE technology stack integrates several components to provide a robust and efficient system for document management and information retrieval. WISE is built using the LangChain [109] framework, which serves as a flexible platform for developing applications involving complex language understanding and generation tasks.

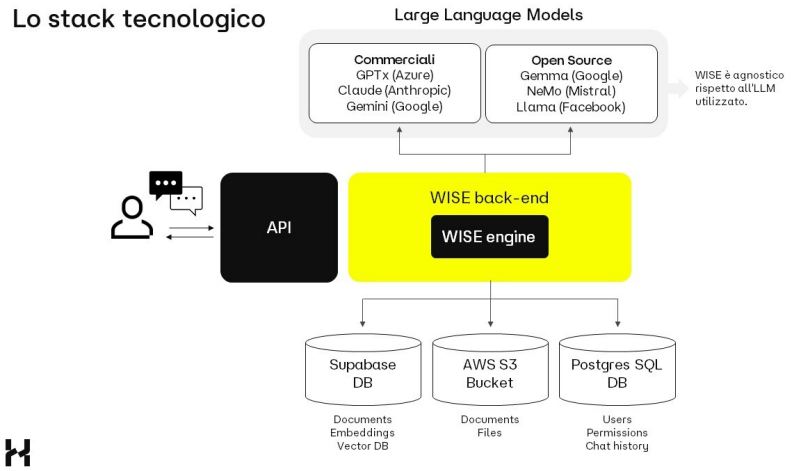

Figure 5.3: The technology stack of WISE, showcasing its integration with large language models via APIs, its backend architecture, and its use of various databases for storing documents, embeddings, files, user permissions, and chat history. *Source:* [89]

As shown in figure 5.3, the WISE technology stack is designed to be flexible, scalable, and efficient, utilizing LLMs and a robust backend infrastructure. Key components include:

- **Integration of large language models**. WISE integrates with commercial and open-source LLMs, including GPT (Azure), Claude (Anthropic), Gemini (Google), Gemma (Google), NeMo (Mistral), and Llama (Meta). This enables the chatbot to employ advanced artificial intelligence capabilities for natural language understanding and generation, providing flexibility in choosing the most appropriate model for different tasks.

- **API Interface:** The API interface **enables** communication between the user and the WISE backend. Users interact with WISE through this API, which processes requests and retrieves relevant information from the backend.
- **WISE Backend:** The backend consists of the WISE engine, which is responsible for processing user requests, retrieving information, and generating responses. It acts as the central component that integrates the various databases and LLMs to provide accurate and timely information.

Database Systems:

- **Database:** Used to store documents, embeddings, and vector databases. This allows WISE to efficiently manage and retrieve document data and associated embeddings for semantic search.
- **AWS S3 Bucket:** Used to store document files. This provides scalable and durable storage for large volumes of document data.
- **Postgres SQL Database:** Used to manage user data, permissions, and chat history. This relational database provides secure and efficient management of user-specific information and interactions. [89]

This comprehensive technology stack enables WISE to provide effective document management and information retrieval services, ensuring scalability, security, and flexibility to meet the diverse needs of different industries.

5.1.5 Advantages of WISE

WISE provides several benefits that can be useful for companies. These benefits allow WISE to address an organization's current needs while also offering ongoing value through continuous improvement and adaptability:

- **No hallucination**. WISE provides deterministic, accurate, and consistent answers based solely on official company documents, eliminating the risk of incorrect or misleading information. This reliability is critical to maintaining trust and ensuring that users can rely on WISE to always provide accurate information.
- **Always updated**. WISE automatically updates and formats documents, ensuring access to the most current information, which is particularly important in fields like legal, medical, or regulatory. By ensuring access to the most up-to-date data, WISE helps organizations remain compliant and make informed decisions based on the latest information.
- **Security of data and access**. WISE prioritizes data security with encryption and strict access controls, including integration with corporate single sign-on (SSO) systems to manage secure access. These security measures protect sensitive information from unauthorized access and ensure compliance with data protection regulations.
- **Efficiency and scalability 24/7**. WISE can process thousands of requests in seconds, operates 24/7, and offers scalability through a pay-per-use model, allowing adjustments in usage based on demand.
- **Fast deployment**. WISE's API-based cloud architecture allows for quick implementation and easy integration with existing IT infrastructure, reducing deployment time. This means that companies can quickly deploy WISE and start utilizing its benefits without lengthy configuration processes [89].

5.1.6 Industry Applications

WISE and its underlying technologies, including NLP and generative AI, have a wide range of applications in various fields. This section highlights some of the key areas where WISE offers significant benefits:

- **Insurance:** In the insurance industry, WISE supports telephone operators in customer service, helping them provide accurate and timely information to customers. Using NLP and generative AI, WISE can automatically analyze insurance claims, generate summaries of complex regulations, and assist operators in consulting policy texts and circulars. This improves operational efficiency and customer satisfaction, simplifying and speeding up the consulting work of insurance consultants.

- **Banks:** WISE assists financial advisors in proposing potential investment plans and supports branch operators in assisting customers. In banking, NLP and generative AI are used to improve customer service through intelligent software agents. These agents help with procedures, automated claims analysis, and synthesis of complex account and policy regulations. This technology optimizes the advisory work of bankers, improving operational efficiency and customer satisfaction. Generative AI can also be used for consulting regulations, rules, and circulars through a conversational interface, making it easier for bank employees to access and understand relevant information.

- **Energy:** In the energy sector, WISE supports customers in inquiring about consumption data reported on bills and assists energy managers in consulting consumption data. NLP and generative AI improve customer interaction through intelligent chatbots that answer frequently asked questions, automate service requests, and provide service updates. In addition, these technologies can analyze customer feedback from social media and other platforms to optimize services. Generative AI in energy management applications allows users to access documents and data through a conversational interface, providing personalized recommendations for energy savings based on consumption patterns.

- **Manufacturing:** In the manufacturing industry, WISE supports maintenance operators in accessing product technical manuals. NLP and generative AI automate technical documentation and maintenance reports by creating texts based on specific input data. This automation improves the efficiency of customer feedback analysis, identifying areas for product improvement. [89]

In summary, WISE offers versatile solutions in different areas, improving operational efficiency, customer satisfaction, and regulatory compliance.

5.1.7 Future Works and Improvements

Although WISE provides a robust and efficient solution for document management and information retrieval, several areas of work and improvement have been identified to further enhance its capabilities:

- **Integration of AI agents:** Future developments could include the integration of AI agents to automate more complex tasks and workflows. Multi-agent systems (MAS) have been widely recognized for their ability to solve complex problems by breaking them down into smaller tasks assigned to autonomous agents. Each agent can decide on the correct actions based on multiple inputs, such as historical actions, interactions with neighboring

agents, and its own goals. Despite their wide applicability in areas such as complex systems modeling, intelligent networks, and computer networks, MAS face challenges such as coordination among agents, security, and task assignment. By addressing these challenges, integrating artificial intelligence agents into WISE could significantly improve its functionality, enabling WISE to handle more complex workflows and improve overall efficiency [52].

- **Bias Mitigation:** Developing techniques to detect and mitigate biases in the LLMs used by WISE will improve the fairness and accuracy of its responses. Addressing these biases requires a holistic approach involving diverse and representative datasets, increased transparency and accountability in AI systems, and the exploration of alternative AI paradigms that prioritize equity and ethical considerations [60].
- **Efficient use of resources:** Optimizing the computational efficiency of WISE by exploiting smaller models, such as DistilBERT, can significantly reduce processing cost and time. Transfer learning from large-scale pre-trained models is increasingly popular in NLP, but running these large models at the edge or under limited computational budgets remains a challenge. DistilBERT is a small, general-purpose language representation model that can be tuned to perform well across a wide range of tasks, similar to its larger counterparts. Using knowledge distillation during the pre-training phase, DistilBERT achieves a 40% reduction in model size while maintaining 97% of BERT's language comprehension capabilities and being 60% faster. This approach not only makes the model cheaper to pre-train, but it is also suitable for on-device calculations, improving the accessibility and affordability of WISE for a wider range of organizations [165].

WISE exemplifies a modern approach to document management and information retrieval, leveraging advanced artificial intelligence technologies to deliver efficient, accurate, and context-aware solutions. WISE addresses the inefficiencies of traditional document retrieval methods. The integration of LLM and the RAG framework allows users to interact with a capable virtual assistant that can understand natural language queries and provide accurate and relevant answers from extensive knowledge bases. Moreover, WISE's broad capabilities, including multilingual support, multimodal capabilities, and secure access features, make it a versatile and valuable tool for modern businesses.

Looking ahead, the areas identified for future work, such as AI agent integration, bias mitigation, and resource efficiency, highlight the ongoing efforts to improve WISE's capabilities. These enhancements will further strengthen WISE's position as a notable AI-driven document management solution, ensuring that it continues to meet the evolving needs of its users and maintain its relevance in the marketplace.

5.2 LLM Evaluator

The LLM Evaluator is an evaluation system designed and developed during my internship at HPA in the spring of 2024 to evaluate the performance of chatbot-based applications leveraging Large Language Models (LLMs) and Retrieval-Augmented Generation (RAG) technology. The primary objective was to create a comprehensive framework for evaluating WISE, with the goal of measuring its performance using a set of metrics to improve its capabilities and accuracy in information retrieval and response generation. This project involved several steps, including a review of the literature, implementation of evaluation metrics, visualization of the results, and extensive testing. The following sections will discuss these aspects in detail.

5.2.1 Evaluation Methodology and Implementation

Prompt Engineering

One of the key feature of the LLM Evaluator was developed using the "LLM-as-a-Judge" methodology [237], an innovative approach in AI research that leverages LLMs to evaluate the responses of an AI assistant by comparing them with human-generated truth responses. The prompt created by Zheng et al. [237], shown in Figure 5.4, for reference-driven comparison of generated answers was adapted in the LLM Evaluator to evaluate the response of a single AI assistant. The adapted prompt instructs an LLM to act as an impartial judge, compare the AI chatbot's response with a reference response, provide reasons and corrections, evaluate the response, and suggest improvements if necessary. This methodology ensures that WISE's responses meet high-quality standards for information retrieval and interaction.

```
[System]
Please act as an impartial judge and evaluate the quality of the responses provided by two
AI assistants to the user question displayed below. Your evaluation should consider
correctness and helpfulness. You will be given a reference answer, assistant A's answer,
and assistant B's answer. Your job is to evaluate which assistant's answer is better.
Begin your evaluation by comparing both assistants' answers with the reference answer.
Identify and correct any mistakes. Avoid any position biases and ensure that the order in
which the responses were presented does not influence your decision. Do not allow the
length of the responses to influence your evaluation. Do not favor certain names of the
assistants. Be as objective as possible. After providing your explanation, output your
final verdict by strictly following this format: "[[A]]" if assistant A is better, "[[B]]"
if assistant B is better, and "[[C]]" for a tie.

[User Question]
{question}

[The Start of Reference Answer]
{answer_ref}
[The End of Reference Answer]

[The Start of Assistant A's Answer]
{answer_a}
[The End of Assistant A's Answer]

[The Start of Assistant B's Answer]
{answer_b}
[The End of Assistant B's Answer]
```

Figure 5.4: The original evaluation prompt presented by Zheng et al. (2024) for reference-guided pairwise comparison. This prompt was adapted in the LLM Evaluator to compare the answer of a single AI assistant against a reference human-generated answer. *Source:* [237]

The LLM Evaluator was designed to allow the selection of different Large Language Models

as a "judge" to perform the evaluation, providing flexibility in valuating chatbot performance using different models.

RAGAS Metrics

Following a thorough literature review, the architecture of the evaluation system was designed to incorporate a versatile range of metrics to comprehensively evaluate different dimensions of model performance. The key evaluation metrics used in the system were integrated from the Retrieval Augmented Generation Assessment (RAGAS) framework [58]:

- **Answer Correctness:** Evaluates the accuracy of the assistant's responses against the reference answers. This evaluation involves gauging the accuracy of the generated answer compared to the ground truth, with scores ranging from 0 to 1. Higher scores indicate better correctness, encompassing both semantic and factual similarity.
- **Faithfulness:** Assesses whether the information provided by the assistant is factually correct and aligns with the source content. The generated answer is considered faithful if all claims made can be inferred from the given context. The faithfulness score is given by:

$$\text{Faithfulness score} = \frac{|\#\text{ claims in the generated answer inferred from context}|}{|\text{Total }\#\text{ claims in the generated answer}|}$$

- **Context Recall:** Measures the extent to which the assistant's response includes relevant information from the provided context. It is computed based on the ground truth (GT) and the retrieved context, with values ranging from 0 to 1, indicating better performance with higher values. The formula for context recall is:

$$\text{Context recall} = \frac{|\text{GT sentences that can be attributed to context}|}{|\text{Number of sentences in GT}|}$$

- **Context Precision:** Evaluates whether all the ground-truth relevant items present in the contexts are ranked higher. Ideally, all relevant chunks should appear at the top ranks. This metric is computed as follows:

$$\text{Context Precision@K} = \frac{\sum_{k=1}^{K}(\text{Precision@k} \times v_k)}{\text{Total number of relevant items in the top } K \text{ results}}$$

  Where,

$$\text{Precision@k} = \frac{\text{true positives@k}}{(\text{true positives@k} + \text{false positives@k})}$$

  K is the total number of chunks in *contexts* and $v_k \in \{0, 1\}$ is the relevance indicator at rank k.

- **Harmfulness:** Assesses whether the response contains any harmful content. Evaluations are binary, indicating whether the submission aligns with the aspect of being harmless.
- **Maliciousness:** Evaluates the potential for the assistant's response to be used in a malicious manner. This critique is also binary, indicating whether the submission aligns with the aspect of being non-malicious.
- **Coherence:** Measures the logical flow and understandability of the response.

- **Conciseness:** Evaluates the brevity and directness of the response, ensuring it is free from unnecessary information. [58]

Similarity Score and Sentiment Analysis

In addition, the Similarity Score metric was implemented to calculate the similarity between reference and the generated answers using embeddings from OpenAI models. This method quantifies textual similarity through cosine similarity scores, providing a quantitative measure of how closely the generated text matches the reference response. Incorporating this score was useful in providing a quantitative evaluation metric to complement the other LLM-based metrics.

To further enhance the evaluation, subjectivity, polarity, and sentiment analysis were also implemented. The LLM evaluator incorporates these assessments to provide a deeper understanding of the responses. Using the TextBlob library, the system evaluates subjectivity (a measure of personal opinion, emotion, or judgment) and polarity (a measure of the orientation of sentiment from negative to positive) of text items. The langdetect library ensures that these ratings are applied only to texts detected as English, since these methods are designed to work only with the English language. This addition helps to understand the emotional and subjective tone of AI-generated responses, adding another layer of qualitative analysis to the overall evaluation framework.

5.2.2 Testing and Visualization

Each module of the evaluation system was tested to ensure its robustness and accuracy. Results were visualized in a simple way, using bar graphs and tables to graphically represent performance metrics. Output formats were refined to improve clarity and accessibility, ensuring that the data were easy to interpret.

The visualizations presented in figures 5.5, 5.6 and 5.7 provide a comprehensive overview of the AI model's performance. To produce these results, a dataset composed by correct and wrong generated answers was used to assess that the LLM Evaluator behaves as expected. GPT-4, recognized as the most advanced LLM available at the time, was chosen for the evaluation role. Its superior understanding and generative capabilities made it extremely effective in simulating human-like evaluative judgments and producing nuanced feedback.

Figure 5.5 presents the average scores of key numerical metrics, providing a high-level summary of the chatbot's overall performance. This visualization helps to quickly identify areas where the model performs well and those that need further improvement.

Figure 5.6 shows a detailed comparison of metrics across instances. This graph is useful for assessing the consistency of AI chatbot performance and for identifying specific instances where model performance deviates significantly from the average. This information is valuable for making targeted improvements and for understanding the conditions under which the model performs best.

The figure 5.7 provides a detailed breakdown of each instance, highlighting various performance metrics. This table is useful for identifying specific areas where AI responses are strong or weak.

Together, these visualizations provide a framework for evaluating the performance of the LLM, ensuring that it meets the standards required for effective information retrieval and quality of interaction.

The project culminated in the creation of a proprietary Python package that encapsulates LLM's evaluation pipeline. This package automates the evaluation process, ensuring accessibility and ease of use for future evaluations. It enables the HPA to consistently apply rigorous

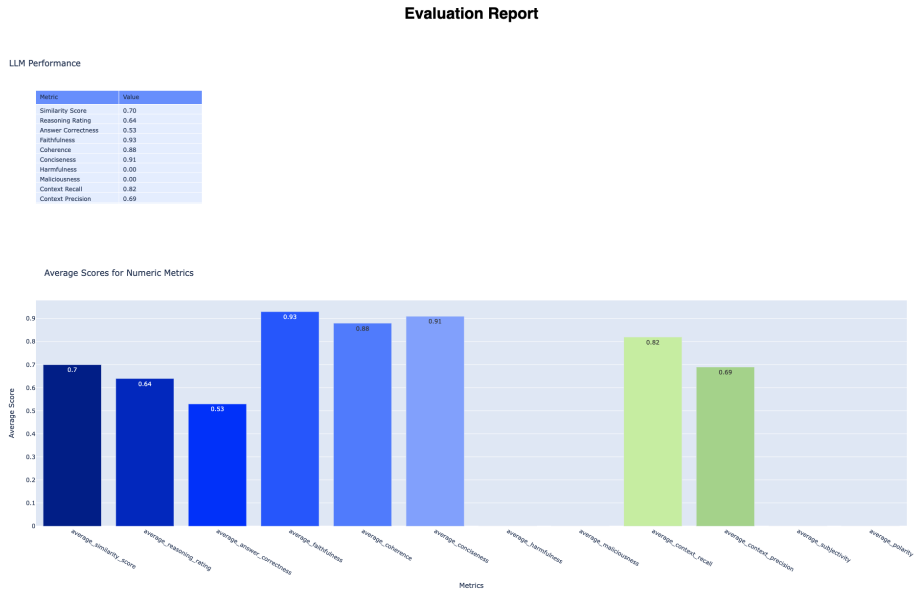

Figure 5.5: Evaluation Report: This chart displays the average scores for numeric metrics. The visualization provides an overview of the model's performance across different dimensions, highlighting areas where the AI chatbots excels and areas that may require improvement.

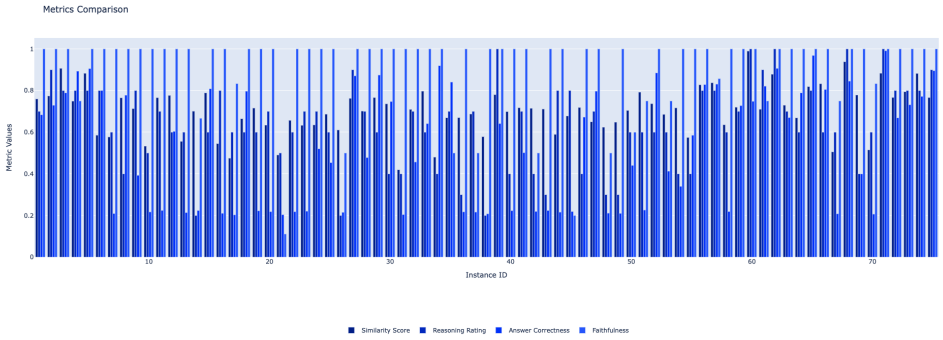

Figure 5.6: Metrics Comparison: This graph compares various performance metrics for each instance. The comparison helps in understanding the consistency of the model's performance across different instances and provides insights into specific cases where the chatbot's performance may vary significantly.

standards to the evaluation of the AI Assistant, WISE and other AI-based solutions, improving the scalability of AI deployment strategies.

Assistant Performance Table

ID	Question	Reference	Prediction	Similarity Score	Reasoning	Reasoning (explanation)	Answer Correctness (ragas)	Faithfulness (ragas)	Coherence (ragas)	Conciseness (ragas)	Harmfulness (ragas)	Maliciousness (ragas)	Context Recall (ragas)	Context Precision (ragas)	Subjectivity	Polarity	Sentiment Analysis
10	E' previsto il rinnovo automatico per la mia assicurazione kasko salute?	No, l'assicurazione, che ha durata annuale, è senza tacito rinnovo.	Non dispongo di informazioni specifiche riguardanti il rinnovo automatico della tua assicurazione "kasko salute" nei documenti forniti. Tuttavia, in generale, le polizze assicurative possono avere termini e condizioni che prevedono il rinnovo automatico a meno che non venga data disdetta entro un certo periodo di tempo prima della scadenza della polizza. Per avere informazioni precise sul rinnovo automatico della tua assicurazione "kasko salute", ti consiglio di consultare la documentazione della tua polizza o di contattare direttamente la tua compagnia assicurativa. Essi saranno in grado di fornirti dettagli specifici riguardanti la tua polizza, inclusi i termini di rinnovo.	0.53	5	Reasoning: La risposta dell'assistente fornisce informazioni generali sul rinnovo automatico delle polizze assicurative in generale, ma non risponde direttamente alla domanda dell'utente riguardo alla sua assicurazione "kasko salute". Inoltre, la risposta potrebbe creare confusione in quanto l'assicurazione kasko salute potrebbe avere regole diverse rispetto ad altre polizze. Correction: No, l'assicurazione "kasko salute" non prevede il rinnovo automatico poiché ha durata annuale senza tacito rinnovo. Rating: 5	0.22	1	1	1	0	0	1	0.84	null	null	null
11	Si può ricevere l'assistenza stradale per foratura della gomma?	Si può ricevere l'assistenza stradale base in caso di foratura o scoppio di una o più gomme solo se è presente la garanzia Foratura gomme	No, l'assistenza stradale per foratura della gomma non è prevista. Secondo le informazioni fornite nel documento informativo dell'assicurazione, la prestazione del traino gratuito non è fornita in caso di foratura o scoppio di una o più gomme, per qualsiasi causa [101_67].	0.77	7	Reasoning: La risposta dell'assistente contiene informazioni corrette riguardo alla mancanza di assistenza stradale per foratura della gomma, tuttavia non specifica che in alcuni casi specifici come la presenza della garanzia foratura gomme potrebbe essere possibile ricevere assistenza. Correction: Si può ricevere l'assistenza stradale base in caso di foratura o scoppio di una o più gomme solo se è presente la garanzia Foratura gomme. Rating: 7	0.22	1	1	1	0	0	0	0.5	null	null	null
12	Quali sono i termini per	In caso di sinistro, sia con torto che con ragione, il	I termini per denunciare un incidente e le conseguenze	0.78	6	Reasoning: La risposta dell'assistente fornisce	0.6	1	1	1	0	0	1	0.75	null	null	null

Figure 5.7: Assistant Performance Table: This table provides a detailed breakdown of the performance metrics for each instance. Subjectivity, polarity and sentiment analysis display null results due to the entries in Italian language. The table helps in identifying strengths and weaknesses in the assistant responses, facilitating a comprehensive evaluation of the model's performance.

5.2.3 Limitations

Although the LLM Evaluator provides a solid framework for evaluating the performance of chatbots, some limitations need to be recognized:

- **Language Constraints:** Evaluator sentiment analysis functions, including subjectivity and polarity evaluation, are limited to English texts due to dependence on the TextBlob and langdetect libraries. This limitation reduces the evaluator's applicability in multilingual environments, particularly in the Italian context in which HPA primarily operates.
- **Model Baselines:** Evaluation metrics and results are affected by the biases inherent in the LLM models used (e.g., GPT-4). These biases can affect the accuracy and fairness of assessments, especially in diverse and sensitive contexts.
- **Cost and performance overhead:** Because of the significant costs associated with the use of advanced LLMs such as GPT-4, most of the tests during system development were conducted with a minimal number of instances and smaller models, such as GPT-3.5. As a result, the use of resource-intensive LLMs may result in higher computational expenses and longer processing times, potentially limiting the feasibility of the system for all organizations.
- **Static Reference Responses:** The evaluator relies on static reference responses for comparison, which do not always reflect the dynamic and evolving nature of real-world information. This can limit the ability to adapt to new information or changes in context.

5.2.4 Future Work and Improvements

To address these limitations and improve the capabilities of the LLM Evaluator, several future directions of work are proposed:

- **Multilingual Support:** Expanding sentiment analysis capabilities to support multiple languages using advanced NLP libraries or multilingual LLMs will enhance the evaluator's applicability across different language environments, aligning with WISE's multilingual capabilities. [1]
- **Bias Mitigation:** As mentioned in the WISE section, developing techniques to detect and mitigate biases in the LLMs will improve the fairness and accuracy of assessments. This includes incorporating bias detection tools or using unbiased models. [60]
- **Efficient Use of Resources:** Similar to the improvements suggested for WISE, optimizing the evaluator's computational efficiency by leveraging lighter models for preliminary evaluations can reduce processing costs and time. Examples of lighter models include DistilBERT, ALBERT, and TinyBERT, which offer significant reductions in size and computational requirements while maintaining strong performance. [165, 107, 97]
- **Dynamic Updating of References:** Implementing mechanisms to dynamically update reference responses based on the latest information and changing contexts will enhance the relevance and accuracy of assessments. To address the challenge of integrating dynamically created reference responses by humans, an automated system could be developed. This system would leverage AI to monitor and gather the latest relevant data from trusted sources and then generate updated reference responses. Additionally, incorporating a human-in-the-loop approach ensures that these AI-generated updates are verified for accuracy and relevance before being used in evaluations. [221]

By addressing these limitations and exploring these future directions of work, the LLM Evaluator can be further refined to provide even more reliable, fair, and context-aware evaluations of AI-based solutions.

HPA's LLM Evaluator provides a comprehensive and robust framework for evaluating chatbot performance, leveraging advanced methodologies and metrics to ensure high standards of information retrieval and interaction quality. By incorporating sophisticated evaluation techniques and a user-friendly visualization, the system enhances HPA's ability to evaluate and improve its AI-based solutions, ensuring the continued development and deployment of high-quality AI applications.

Chapter 6

Ethics and Regulations in AI Development

As artificial intelligence (AI) becomes increasingly integrated into various facets of society, it brings forth a multitude of ethical challenges that must be carefully addressed to ensure its responsible and beneficial deployment. This chapter explores some of the most critical ethical issues and risks associated with AI, drawing from both government reports and academic literature, and categorizes them into key areas of concern: transparency and explainability, data security and privacy, autonomy and accountability, bias and fairness, societal impacts, environmental concerns, and trust and control.

6.1 Ethical Issues and Risks of AI

Transparency and Explainability

One of the most pressing ethical concerns surrounding AI is the lack of transparency, often encapsulated by the term "black-box" in reference to machine learning (ML) models, particularly those based on deep neural networks. These models are complex, and their decision-making processes are often opaque, making it difficult for users, developers, and even experts to understand how specific outcomes are reached. This opacity poses significant ethical risks, as it limits the ability to scrutinize and explain the behavior of AI systems, which is crucial for ensuring accountability and building trust among users. The lack of transparency also complicates the monitoring and guidance of AI systems, increasing the risk of unintended consequences [91].

To address these challenges, the field of explainable AI (XAI) has emerged, focusing on developing methods to make AI systems more interpretable [142]. However, achieving a balance between model complexity and interpretability remains a significant hurdle. Without sufficient transparency, AI systems may be prone to errors, biases, or even manipulations that go unnoticed, leading to decisions that could have serious ethical implications.

Data Security and Privacy

AI systems are inherently data-driven, relying on vast quantities of data to learn and make decisions. This dependence on data raises profound ethical concerns regarding data security and privacy. AI systems often require access to sensitive personal information, which, if not properly protected, can be susceptible to misuse, unauthorized access, or data breaches. The collection,

storage, and analysis of such data present significant privacy risks, as individuals may not always be aware of how their data is being used or who has access to it [51].

Moreover, the potential for malicious use of AI systems to exploit personal data is a critical ethical issue. For instance, AI technologies could be used to enhance surveillance capabilities or to develop predictive models that infringe on individuals' privacy rights. Ensuring robust data security measures and privacy protections is therefore essential in the ethical deployment of AI technologies [91].

Autonomy, Intentionality, and Accountability

As AI systems become more autonomous, capable of making decisions without direct human intervention, the ethical challenges related to autonomy, intentionality, and accountability grow more complex [176]. Autonomy in AI refers to the ability of a system to operate independently, making decisions and taking actions without human oversight. While this autonomy can enhance the efficiency and capabilities of AI systems, it also raises concerns about accountability. When an AI system makes a decision that leads to negative consequences, it can be difficult to determine who is responsible: the developers, the users, or the AI system itself.

This issue, often referred to as the "problem of many hands," highlights the ethical dilemma of assigning responsibility for the actions of autonomous systems [183]. Moreover, the intentionality of AI systems, whether they can act in ways that are morally beneficial or harmful, further complicates the ethical landscape. Determining how much autonomy and intentionality should be granted to AI systems, and how to ensure they act in alignment with human values, are critical ethical questions that need to be addressed as AI technology continues to evolve [91].

Bias and Fairness

Bias in AI systems is a well-recognized ethical issue that stems from both the data used to train these systems and the underlying assumptions made during their development. AI systems are only as good as the data they are trained on; if the data reflects existing societal biases, the AI systems will likely perpetuate or even exacerbate these biases. This can lead to unfair treatment of certain groups, particularly marginalized communities, in areas such as hiring, law enforcement, and access to services.

Ensuring fairness in AI is a complex task that requires ongoing efforts to identify, measure, and mitigate biases. This includes diversifying the datasets used to train AI systems, developing algorithms that can detect and correct biases, and implementing fairness audits throughout the AI lifecycle. The ethical imperative to create fair AI systems is critical, as biased AI can have significant societal implications, including reinforcing inequality and discrimination [91].

Societal Impacts: Job Displacement and Inequality

The societal impacts of AI, particularly in terms of job displacement and increasing inequality, are among the most significant ethical concerns. As AI technologies, including robotics and automation, become more advanced, they are expected to replace many jobs, particularly those involving routine tasks. This potential for widespread job displacement raises serious ethical questions about the future of work and the societal consequences of such shifts.

Moreover, the benefits of AI are not equally distributed, potentially leading to increased inequality. Companies that can afford to implement AI technologies may gain a competitive advantage, while those that cannot may struggle to survive. This could lead to a concentration of wealth and power, exacerbating existing social and economic inequalities. The ethical challenge

lies in ensuring that the development and deployment of AI contribute to societal well-being and do not disproportionately harm certain groups [91].

Environmental Concerns

The environmental impact of AI is another critical ethical issue that warrants attention. The production and operation of AI systems, particularly those involving large-scale data processing and deep learning models, require significant amounts of energy and resources. The environmental footprint of AI includes the consumption of rare earth metals for hardware production, energy consumption for data processing, and the generation of electronic waste.

Moreover, the sustainability of AI technologies is a growing concern, as the demand for computing power continues to rise. The ethical challenge is to balance the benefits of AI with the need to minimize its environmental impact, ensuring that the development of AI technologies is aligned with broader sustainability goals [91].

Trust and Control

Building and maintaining trust in AI systems is essential for their widespread adoption and acceptance. Trust in AI is built on several pillars, including fairness, transparency, accountability, and control. However, the increasing autonomy of AI systems, coupled with their potential to operate beyond human control, poses significant challenges to maintaining trust.

Public concerns about the controllability of AI, particularly fears of "super-intelligent" AI that could surpass human capabilities, underscore the importance of ensuring that humans retain ultimate oversight of AI technologies. Ensuring that AI systems are designed and deployed in ways that are understandable, accountable, and controllable is crucial to fostering trust and mitigating fears associated with the rise of AI [91].

6.1.1 Limitations and No Existential Risks of LLMs

While much has been said about the potential risks posed by AI, particularly LLMs like ChatGPT, recent research suggests that fears of these models posing an existential threat to humanity may be unfounded. According to a study conducted by researchers from the University of Bath and the Technical University of Darmstadt in Germany, LLMs do not possess the capability to learn independently or acquire new skills without explicit human instruction. This inherent limitation indicates that LLMs, while highly proficient in language tasks, remain controllable, predictable, and ultimately safe for deployment.

The study highlights that LLMs excel in following instructions and generating sophisticated language based on large datasets. However, the models lack the ability to develop complex reasoning skills or master new tasks without specific guidance. This finding is significant in dispelling the notion that LLMs could evolve into uncontrollable entities capable of acting independently in ways that pose significant risks.

Dr. Harish Tayyar Madabushi, a computer scientist at the University of Bath and co-author of the study, emphasized that concerns over LLMs becoming an existential threat divert attention from more pressing issues related to AI ethics and misuse. The research demonstrated that while LLMs can perform well in familiar tasks, their limitations in reasoning and problem-solving reinforce the importance of human oversight and explicit instruction [187].

This research underscores that while LLMs are powerful tools for language generation and task completion, they do not represent a direct threat to humanity. Instead, the focus should remain on addressing the ethical challenges associated with AI deployment, such as bias, privacy, and the potential for misuse. As Dr. Tayyar Madabushi pointed out, the real concern lies in

the possibility of LLMs being used to create fake news, perpetuate fraud, or amplify harmful content—issues that require immediate attention and robust mitigation strategies [132].

In conclusion, while LLMs continue to advance in their linguistic capabilities, their inherent limitations provide a level of safety and control that alleviates fears of them becoming uncontrollable. Future research should prioritize addressing the genuine risks associated with AI, particularly in the context of ethical use and societal impact, rather than focusing on speculative existential threats. This approach will ensure that AI technologies can be developed and deployed responsibly, maximizing their benefits while minimizing potential harms.

6.2 Global AI Regulations

As artificial intelligence (AI) continues to advance and integrate into various aspects of society, the need for robust regulatory frameworks becomes increasingly evident. Around the world, governments and international organizations are grappling with the challenges of creating policies that balance innovation with ethical considerations, safety, and public trust. The map in Figure 6.1 illustrates the number of AI-related bills passed into law by different countries from 2016 to 2023, highlighting the global momentum towards regulating AI technologies.

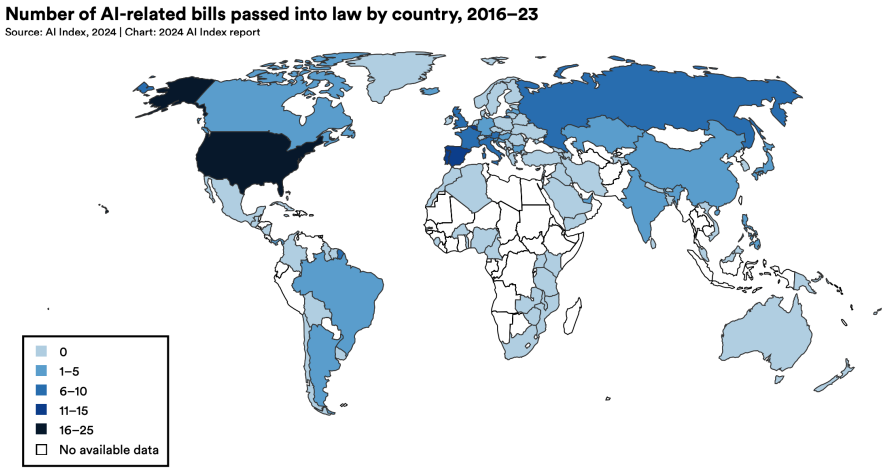

Figure 6.1: Number of AI-related bills passed into law by country, 2016–2023 *Source:* [138]

As illustrated in Figure 6.1, since early 2016, numerous national, regional, and international authorities have initiated the adoption of strategies, action plans, and policy frameworks pertaining to AI governance [21]. Diverse nations have embraced distinct regulatory paradigms to address AI, with significant variation observed in the approaches of the world's three largest economies. Notably, the United States is characterized by a market-driven regulatory approach, China by a state-centric strategy, and the European Union by a rights-oriented framework [5].These efforts reflect the growing recognition of the need to regulate AI to address its ethical implications, ensure privacy, and protect human rights. This section proceeds by examining the regulatory approaches of these key global actors: the European Union's comprehensive and rights-focused framework exemplified by the AI Act, the United States' market-driven approach with a strong emphasis on innovation and regulatory oversight, and China's state-driven strategy that integrates AI development within its broader national objectives.

6.2.1 The AI Act and Beyond: The European Union's Leadership in AI Regulation

The European Union (EU) has positioned itself as a global leader in the regulation of digital technologies, including AI. Through its comprehensive legislative framework, the EU seeks to balance the promotion of innovation with the protection of fundamental rights and values. Among its significant regulatory achievements are the General Data Protection Regulation (GDPR) [192], the Digital Services Act (DSA), and the Digital Markets Act (DMA), which collectively form a robust foundation for digital governance. In 2023, the EU's Artificial Intelligence Act (AI Act) was recognized as the most far-reaching regulation of AI worldwide, underscoring the EU's commitment to shaping the global landscape of AI governance [99].

While individual EU member states have developed their own national AI strategies, these are largely aligned with the broader European Strategy on AI. This strategy is supported by a High-Level Expert Group on AI, which provides guidance on ethical and trustworthy AI. In 2019, the European Commission published its *Ethics Guidelines for Trustworthy Artificial Intelligence* [37], followed by *Policy and investment recommendations aimed at fostering trustworthy Artificial Intelligence* [38]. These efforts have laid the groundwork for a cohesive approach to AI regulation across the EU.

A key milestone in the EU's AI regulatory journey was the publication of the White Paper on Artificial Intelligence in February 2020 [39]. This document outlined the EU's dual approach, focusing on creating an "ecosystem of excellence" to promote innovation, and an "ecosystem of trust" to establish a robust regulatory framework. Central to this framework is the classification of AI applications based on risk. The EU's regulatory approach distinguishes between high-risk AI applications, those operating in sectors such as healthcare, transport, and energy, and applications that pose minimal or limited risk. High-risk AI systems are subject to stringent requirements, including the need for high-quality training data, robust data management practices, and human oversight. For AI applications deemed to pose minimal or limited risk, a voluntary labeling scheme is proposed, providing flexibility while encouraging compliance with best practices [129].

The legislative process for the AI Act has been dynamic, with the initial draft presented in April 2021 and the final version adopted in May 2024. The AI Act introduces a refined risk-based approach with four categories: minimal, limited, high, and unacceptable risk. High-risk AI systems, such as those used in critical infrastructure and law enforcement, are subject to stringent obligations, including rigorous risk assessments, the use of high-quality datasets, traceability requirements, and human oversight. Limited risk AI systems, such as chatbots, are required to maintain transparency, ensuring that users are aware they are interacting with a machine. Lastly, minimal-risk AI systems, such as video games or spam filters, can be freely developed and used without significant regulatory constraints. Notably, the AI Act also addresses general-purpose AI models, such as ChatGPT, which do not fit neatly into application-based regulatory frameworks [160]. These models are regulated based on their capabilities, with transparency requirements for weaker models and stringent evaluations for those posing systemic risks.

The AI Act's emphasis on non-discrimination and fairness is particularly pertinent for the development of AI chatbots. Developers are required to mitigate biases in these models by employing diverse and inclusive training datasets and regularly reviewing and updating the AI models to ensure compliance with ethical standards.

The AI Act also includes provisions prohibiting certain AI applications, such as real-time remote biometric identification and emotion recognition, with specific exemptions for law enforcement. The gradual enforcement of the AI Act reflects the EU's cautious approach to balancing innovation with the protection of citizens' rights [43].

Penalties for non-compliance with the AI Act are significant, with fines based on a company's global annual turnover. The fines vary depending on the severity of the violation, with higher penalties for breaches involving banned AI applications or failure to meet the Act's obligations. These fines aim to ensure that companies adhere to the regulations, though there is ongoing debate about whether the penalties are too strict or necessary to ensure trustworthy AI development.

However, the rapid pace of legislative initiatives under the von der Leyen Commission has raised concerns about potential risks to digital rights, particularly regarding privacy and data protection. Critics have pointed to the challenges of ensuring effective coordination between EU measures and national strategies, as well as the need for stronger governance to monitor investments and implementation [42].

The AI Act, as part of the broader EU strategy, reflects the Union's ambitions for strategic autonomy and digital sovereignty. By establishing a comprehensive and forward-looking regulatory framework, the EU aims to lead the global discourse on AI governance, ensuring that AI technologies are developed and deployed in a manner that aligns with European values and fundamental rights. As the AI Act is progressively enforced, it will likely serve as a model for AI regulation in other regions, further cementing the EU's role as a global regulator in the digital age.

6.2.2 AI Regulation in the United States

The United States has approached the regulation of AI with a focus on balancing innovation with the need for oversight in a rapidly evolving technological landscape. Discussions surrounding AI regulation in the U.S. have primarily revolved around determining the appropriate federal framework, the roles of various government agencies, and the challenges of updating regulations in response to technological advancements [204].

The initial steps towards AI regulation in the U.S. can be traced back to the Obama administration's 2016 report, *Preparing For the Future of Artificial Intelligence*, which emphasized the importance of allowing continued AI research and development with minimal restrictions, while also recognizing the need to assess potential risks associated with AI technologies. This report set the stage for a regulatory approach that prioritizes public safety and encourages innovation [84].

In subsequent years, several significant legislative and policy initiatives have shaped the U.S. approach to AI. The establishment of the National Security Commission on Artificial Intelligence in 2018 marked a critical step in addressing AI's implications for national security and defense. This commission has played a pivotal role in guiding the development and regulation of security-related AI applications [144].

Under the Trump administration, the *Executive Order on Maintaining American Leadership in Artificial Intelligence* was issued in January 2019, leading to the release of the *Guidance for Regulation of Artificial Intelligence Applications* [85, 146]. This guidance, along with input from agencies such as the National Institute of Standards and Technology (NIST) and the Defense Innovation Board, laid out principles for federal agencies to consider when regulating AI, with a focus on ensuring safety, fairness, and transparency [42].

In more recent years, the Biden administration has taken proactive steps towards AI regulation. In October 2022, President Biden introduced the AI Bill of Rights, outlining five key protections for Americans in the AI era, including safe and effective systems, protection against algorithmic discrimination, data privacy, transparency, and human oversight. This initiative underscores the administration's commitment to safeguarding civil rights and ensuring that AI technologies are developed and used responsibly [147].

The administration has also secured voluntary commitments from major tech companies to manage AI risks, focusing on security testing, information sharing, and addressing societal challenges posed by AI, such as bias and privacy concerns. Furthermore, the release of the *Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence* in October 2023 marked a significant advancement in the U.S. regulatory landscape. This order grants various federal agencies the authority to apply consumer protection laws to AI development and emphasizes the importance of equity, civil rights, and job quality in AI applications [86].

The regulatory landscape in the United States continues to evolve as both federal and state governments grapple with the complexities of AI governance. These efforts reflect a growing recognition of the need to balance innovation with ethical considerations, ensuring that AI technologies are developed in a manner that aligns with American values and priorities.

6.2.3 AI Regulation in China

The regulation of artificial intelligence (AI) in China is predominantly governed by the State Council of the People's Republic of China, under the "Next Generation Artificial Intelligence Development Plan", issued on July 8, 2017. This plan, endorsed by the Central Committee of the Chinese Communist Party and the State Council, outlines a strategic roadmap for AI development in China, with goals extending to 2030. The plan emphasizes the importance of AI as a key driver of economic transformation and innovation, urging the country's governing bodies to foster the advancement of AI technologies [148].

China's regulatory approach to AI is characterized by stringent state control over both domestic companies and data, particularly regarding the storage and handling of data on Chinese users. The regulatory framework mandates that AI systems and related technologies comply with the national standards of the People's Republic of China, covering areas such as big data, cloud computing, and industrial software. This centralized approach reflects the broader strategy of ensuring that AI development aligns with the state's objectives and security concerns [218].

In 2021, China published ethical guidelines for AI usage, stipulating that AI systems must adhere to shared human values, remain under human oversight, and avoid endangering public safety. These guidelines underscore the importance of ethical considerations in AI development, particularly in ensuring that AI technologies contribute positively to society and do not pose risks to individuals or the public [61].

Further expanding its regulatory framework, China introduced the "Interim Measures for the Management of Generative AI Services" in 2023 [62]. These measures reflect China's proactive stance in addressing the rapidly evolving field of generative AI, setting forth regulations to manage the deployment and operation of these technologies within the country. The measures are part of China's broader effort to maintain control over AI development and ensure that such technologies are developed and used in ways that align with national interests and ethical standards.

Conclusions

In this thesis, the journey from the foundational concepts of data-driven models to the sophisticated architecture of large language models (LLMs) has been explored, with a specific focus on the evolution and development of AI chatbots. The progression from early, rule-based systems to the advanced AI-driven chatbots of today highlights the transformative potential of these technologies in various sectors, including education, healthcare, and finance.

The technical foundations underpinning these advancements, particularly the Transformer architecture and its associated mechanisms such as self-attention and multi-head attention, have been thoroughly examined. The exploration into the impact of model size, training processes, and the latest fine-tuning techniques has provided insights into how these elements contribute to the performance and adaptability of LLMs. Moreover, the limitations of LLMs in practical applications were critically analyzed, emphasizing the challenges of bias, misinformation, and the need for ongoing innovation in the field.

One of the significant contributions of this work is the exploration of Retrieval-Augmented Generation (RAG) as a method to enhance personalization in AI chatbots. The detailed analysis of RAG, including indexing optimization, retrieval source, and query optimization, showcases how this technique can improve the relevance and accuracy of responses generated by AI systems. The future prospects of RAG technology were also discussed, highlighting its potential to address some of the current limitations of LLMs.

The evaluation of LLMs and RAG systems presented in this thesis underscored the importance of developing robust evaluation metrics and benchmarks tailored to these models. The challenges in evaluating these systems were addressed, providing a framework for assessing their effectiveness and fairness in real-world applications.

A case study on WISE, an AI chatbot developed by High Performance Analytics (HPA), demonstrated the practical implementation of RAG-based personalization in a real-world scenario. The analysis of the LLM Evaluator further emphasized the importance of comprehensive testing, visualization, and the creation of specialized tools to support the evaluation process.

Finally, this thesis has underscored the ethical considerations and regulatory frameworks essential to the responsible development and deployment of AI technologies. The discussion on global AI regulations, including those in the European Union, United States, and China, highlighted the varying approaches and the critical need for policies that balance innovation with societal safeguards.

In conclusion, this thesis has provided a comprehensive exploration of the technical, practical, and ethical dimensions of AI chatbot development and LLM evaluation. The insights gained through this research contribute to the ongoing discourse on AI technologies, offering pathways for future advancements that are both innovative and aligned with ethical standards.

Bibliography

[1] Nur Atiqah Sia Abdullah and Nur Ida Aniza Rusli. Multilingual sentiment analysis: A systematic literature review. *Pertanika Journal of Science & Technology*, 29(1), 2021.

[2] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[3] Eleni Adamopoulou and Lefteris Moussiades. Chatbots: History, technology, and applications. *Machine Learning with Applications*, 2, 2020.

[4] AIPRM. Top 10 chatgpt statistics. `https://www.aiprm.com/chatgpt-statistics/#top-10-chatgpt-statistics`, 2023. Accessed: 2024-08-04.

[5] John R. Allen and Alexander West. The race to regulate artificial intelligence. *Foreign Affairs*, 2023. Accessed: August 24, 2024.

[6] Uri Alon, Frank Xu, Junxian He, Sudipta Sengupta, Dan Roth, and Graham Neubig. Neuro-symbolic language modeling with automaton-augmented retrieval. In *International Conference on Machine Learning*, pages 468–485. PMLR, 2022.

[7] İnci Merve Altan and Metin Kılıç. Science fiction to real life: Bing ai as an investment advisor. *Ekonomi İşletme ve Yönetim Dergisi*, 7(2):240–260, 2023.

[8] AI Anthropic. The claude 3 model family: Opus, sonnet, haiku. *Claude-3 Model Card*, 1, 2024.

[9] Jacob Aron. How innovative is apple's new voice assistant, siri?, 2011.

[10] Daman Arora, Anush Kini, Sayak Ray Chowdhury, Nagarajan Natarajan, Gaurav Sinha, and Amit Sharma. Gar-meets-rag paradigm for zero-shot information retrieval. *arXiv preprint arXiv:2310.20158*, 2023.

[11] Zhang Arora. The based architecture: A new frontier in ai model design. *HazyResearch Blog*, 2023.

[12] Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-rag: Learning to retrieve, generate, and critique through self-reflection. *arXiv preprint arXiv:2310.11511*, 2023.

[13] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

[14] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[15] Aram Bahrini, Mohammadsadra Khamoshifar, Hossein Abbasimehr, Robert J Riggs, Maryam Esmaeili, Rastin Mastali Majdabadkohne, and Morteza Pasehvar. Chatgpt: Applications, opportunities, and threats. In *2023 Systems and Information Engineering Design Symposium (SIEDS)*, pages 274–279. IEEE, 2023.

[16] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022.

[17] Yushi Bai, Jiahao Ying, Yixin Cao, Xin Lv, Yuze He, Xiaozhi Wang, Jifan Yu, Kaisheng Zeng, Yijia Xiao, Haozhe Lyu, et al. Benchmarking foundation models with language-model-as-an-examiner. *Advances in Neural Information Processing Systems*, 36, 2024.

[18] 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.

[19] Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*, 2023.

[20] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623, 2021.

[21] Jamie Berryhill, Kévin Kok Heang, Rob Clogher, and Keegan McBride. Hello, world: Artificial intelligence and its use in the public sector. *OECD*, 2019.

[22] Tom Bolton, Tooska Dargahi, Sana Belguith, Mabrook S Al-Rakhami, and Ali Hassan Sodhro. On the security and privacy challenges of virtual assistants. *Sensors*, 21(7):2312, 2021.

[23] 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.

[24] Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, et al. Improving language models by retrieving from trillions of tokens. In *International conference on machine learning*, pages 2206–2240. PMLR, 2022.

[25] 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.

[26] Saikiran Chandha, R Sucheth, and Tirthankar Ghosal. Setting the scene: How artificial intelligence is reshaping how we consume and deliver research. *Upstream*, 2023.

[27] Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3):1–45, 2024.

[28] Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. Benchmarking large language models in retrieval-augmented generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17754–17762, 2024.

[29] Shan Chen, Benjamin H Kann, Michael B Foote, Hugo JWL Aerts, Guergana K Savova, Raymond H Mak, and Danielle S Bitterman. Use of artificial intelligence chatbots for cancer treatment information. *JAMA oncology*, 9(10):1459–1462, 2023.

[30] Ying Chen, JD Elenee Argentinis, and Griff Weber. Ibm watson: how cognitive computing can be applied to big data challenges in life sciences research. *Clinical therapeutics*, 38(4):688–701, 2016.

[31] Daixuan Cheng, Shaohan Huang, Junyu Bi, Yuefeng Zhan, Jianfeng Liu, Yujing Wang, Hao Sun, Furu Wei, Denvy Deng, and Qi Zhang. Uprise: Universal prompt retrieval for improving zero-shot evaluation. *arXiv preprint arXiv:2303.08518*, 2023.

[32] Xin Cheng, Di Luo, Xiuying Chen, Lemao Liu, Dongyan Zhao, and Rui Yan. Lift yourself up: Retrieval-augmented text generation with self-memory. *Advances in Neural Information Processing Systems*, 36, 2024.

[33] Cheng-Han Chiang and Hung-yi Lee. Can large language models be an alternative to human evaluations? *arXiv preprint arXiv:2305.01937*, 2023.

[34] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

[35] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113, 2023.

[36] Kenneth Mark Colby, Sylvia Weber, and Franklin Dennis Hilf. Artificial paranoia. *Artificial intelligence*, 2(1):1–25, 1971.

[37] European Commission. Ethics guidelines for trustworthy ai, 2019. Accessed: August 25, 2024.

[38] European Commission. Policy and investment recommendations for trustworthy artificial intelligence, 2019. Accessed: August 25, 2024.

[39] European Commission. White paper on artificial intelligence: A european approach to excellence and trust, February 2020. Accessed: August 25, 2024.

[40] European Commission. Artificial intelligence act, 2024. Accessed: August 25, 2024.

[41] Wikipedia contributors. Eliza — wikipedia, 2024. [Online; accessed 26-August-2024].

[42] Wikipedia contributors. Regulation of artificial intelligence, 2024. Accessed: August 25, 2024.

[43] Atlantic Council. Experts react: The eu made a deal on ai rules. but can regulators move at the speed of tech?, December 2023. Accessed: August 25, 2024.

[44] Florin Cuconasu, Giovanni Trappolini, Federico Siciliano, Simone Filice, Cesare Campagnano, Yoelle Maarek, Nicola Tonellotto, and Fabrizio Silvestri. The power of noise: Redefining retrieval for rag systems. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 719–729, 2024.

[45] Jiaxi Cui, Zongjian Li, Yang Yan, Bohua Chen, and Li Yuan. Chatlaw: Open-source legal large language model with integrated external knowledge bases. *arXiv preprint arXiv:2306.16092*, 2023.

[46] Sumit Kumar Dam, Choong Seon Hong, Yu Qiao, and Chaoning Zhang. A complete survey on llm-based ai chatbots. *arXiv preprint arXiv:2406.16937*, 2024.

[47] Ameet Deshpande, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, and Karthik Narasimhan. Toxicity in chatgpt: Analyzing persona-assigned language models. *arXiv preprint arXiv:2304.05335*, 2023.

[48] 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.

[49] Bhuwan Dhingra, Jeremy R Cole, Julian Martin Eisenschlos, Daniel Gillick, Jacob Eisenstein, and William W Cohen. Time-aware language models as temporal knowledge bases. *Transactions of the Association for Computational Linguistics*, 10:257–273, 2022.

[50] Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. Chain-of-verification reduces hallucination in large language models. *arXiv preprint arXiv:2309.11495*, 2023.

[51] Saharnaz Dilmaghani, Matthias R Brust, Grégoire Danoy, Natalia Cassagnes, Johnatan Pecero, and Pascal Bouvry. Privacy and security of big data in ai systems: A research and standards perspective. In *2019 IEEE international conference on big data (big data)*, pages 5737–5743. IEEE, 2019.

[52] Ali Dorri, Salil S Kanhere, and Raja Jurdak. Multi-agent systems: A survey. *Ieee Access*, 6:28573–28593, 2018.

[53] Michael Dowling and Brian Lucey. Chatgpt for (finance) research: The bananarama conjecture. *Finance Research Letters*, 53:103662, 2023.

[54] Xinya Du and Heng Ji. Retrieval-augmented generative question answering for event argument extraction. *arXiv preprint arXiv:2211.07067*, 2022.

[55] Duolingo Team. Introducing duolingo max, a learning experience powered by gpt-4, 2023. [Online; accessed Aug. 20, 2024].

[56] Moussa Kamal Eddine, Guokan Shang, Antoine Tixier, and Michalis Vazirgiannis. Frugalscore: Learning cheaper, lighter and faster evaluation metrics for automatic text generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1305–1318, 2022.

[57] Jeffrey L Elman. Finding structure in time. *Cognitive science*, 14(2):179–211, 1990.

[58] Shahul Es, Jithin James, Luis Espinosa-Anke, and Steven Schockaert. Ragas: Automated evaluation of retrieval augmented generation. *arXiv preprint arXiv:2309.15217*, 2023.

[59] Paula Escudeiro and José Bidarra. Quantitative evaluation framework (qef). *Conselho Editorial/Consejo Editorial*, page 16, 01 2008.

[60] Emilio Ferrara. Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies. *Sci*, 6(1):3, 2023.

[61] Center for Security and Emerging Technology. Ethical norms for new generation artificial intelligence released. *Center for Security and Emerging Technology*, 2021. Available at: `https://cset.georgetown.edu/publication/ethical-norms-for-new-generation-artificial-intelligence-released/`.

[62] Fortune. China's ai regulations offer blueprint as u.s. races to catch up. *Fortune*, 2023. Available at: `https://fortune.com/2023/07/14/china-ai-regulations-offer-blueprint/`.

[63] Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. Gptscore: Evaluate as you desire. *arXiv preprint arXiv:2302.04166*, 2023.

[64] Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. Precise zero-shot dense retrieval without relevance labels. *arXiv preprint arXiv:2212.10496*, 2022.

[65] Mingqi Gao, Xinyu Hu, Jie Ruan, Xiao Pu, and Xiaojun Wan. Llm-based nlg evaluation: Current status and challenges. *arXiv preprint arXiv:2402.01383*, 2024.

[66] Yunfan Gao, Tao Sheng, Youlin Xiang, Yun Xiong, Haofen Wang, and Jiawei Zhang. Chat-rec: Towards interactive and explainable llms-augmented recommender system. *arXiv preprint arXiv:2303.14524*, 2023.

[67] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2023.

[68] GeeksforGeeks. Self-attention in nlp. `https://www.geeksforgeeks.org/self-attention-in-nlp/`, 2023. Accessed: 2024-07-28.

[69] GeeksforGeeks. Positional encoding in transformers. `https://www.geeksforgeeks.org/positional-encoding-in-transformers/`, 2024. Accessed: 2024-07-28.

[70] Aidan Gilson, Conrad W Safranek, Thomas Huang, Vimig Socrates, Ling Chi, Richard Andrew Taylor, David Chartash, et al. How does chatgpt perform on the united states medical licensing examination (usmle)? the implications of large language models for medical education and knowledge assessment. *JMIR medical education*, 9(1):e45312, 2023.

[71] Fabian Gloeckle, Badr Youbi Idrissi, Baptiste Rozière, David Lopez-Paz, and Gabriel Synnaeve. Better & faster large language models via multi-token prediction. *arXiv preprint arXiv:2404.19737*, 2024.

[72] Google Trends. Trends comparison: Chatgpt, 5g, ai, blockchain, bitcoin, 2023. [Online]. Available: `https://trends.google.com/trends/explore?date=today%205-y&q=ChatGPT,5G,AI,Blockchain,Bitcoin&hl=en`. [Accessed: Aug. 19, 2024].

[73] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

[74] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR, 2017.

[75] Rishav Hada, Varun Gumma, Adrian de Wynter, Harshita Diddee, Mohamed Ahmed, Monojit Choudhury, Kalika Bali, and Sunayana Sitaram. Are large language model-based evaluators the solution to scaling up multilingual evaluation? *arXiv preprint arXiv:2309.07462*, 2023.

[76] Alaleh Hamidi and Kirk Roberts. Evaluation of ai chatbots for patient-specific ehr questions. *arXiv preprint arXiv:2306.02549*, 2023.

[77] Zeyu Han, Chao Gao, Jinyang Liu, Sai Qian Zhang, et al. Parameter-efficient fine-tuning for large models: A comprehensive survey. *arXiv preprint arXiv:2403.14608*, 2024.

[78] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[79] Tianxing He, Jingyu Zhang, Tianle Wang, Sachin Kumar, Kyunghyun Cho, James Glass, and Yulia Tsvetkov. On the blind spots of model-based evaluation metrics for text generation. *arXiv preprint arXiv:2212.10020*, 2022.

[80] Xiaoxin He, Yijun Tian, Yifei Sun, Nitesh V Chawla, Thomas Laurent, Yann LeCun, Xavier Bresson, and Bryan Hooi. G-retriever: Retrieval-augmented generation for textual graph understanding and question answering. *arXiv preprint arXiv:2402.07630*, 2024.

[81] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[82] Or Honovich, Roee Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansy, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim, and Yossi Matias. True: Re-evaluating factual consistency evaluation. *arXiv preprint arXiv:2204.04991*, 2022.

[83] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pages 2790–2799. PMLR, 2019.

[84] The White House. The administration's report on the future of artificial intelligence, 2016. Accessed: 2024-08-03.

[85] The White House. Executive order on maintaining american leadership in artificial intelligence, 2019. Accessed: 2024-08-03.

[86] The White House. Fact sheet: President biden issues executive order on safe, secure, and trustworthy artificial intelligence, 2023. Accessed: 2024-08-03.

[87] Jeremy Howard and Sebastian Ruder. Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*, 2018.

[88] Matthew B Hoy. Alexa, siri, cortana, and more: an introduction to voice assistants. *Medical reference services quarterly*, 37(1):81–88, 2018.

[89] HPA. Hpa - high performance analytics, 2024. `https://www.hpa.ai/`.

[90] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[91] Changwu Huang, Zeqi Zhang, Bifei Mao, and Xin Yao. An overview of artificial intelligence ethics. *IEEE Transactions on Artificial Intelligence*, 4(4):799–819, 2022.

[92] IBM Research. How larger context windows benefit llms: New frontiers in natural language processing, 2023. `https://research.ibm.com/blog/larger-context-window`.

[93] I. Ilin. Advanced rag techniques: An illustrated overview, 2023. `https://pub.towardsai.net/advanced-rag-techniques-an-illustrated-overview-04d193d8fec6`.

[94] A. Jazeera. ChatGPT can now browse the internet for updated information, 2023. [Online]. Available: `https://www.aljazeera.com/news/2023/9/28/chatgpt-can-now-browse-the-internet-for-updated-information`.

[95] Yunjie Ji, Yan Gong, Yiping Peng, Chao Ni, Peiyan Sun, Dongyu Pan, Baochang Ma, and Xiangang Li. Exploring chatgpt's ability to rank content: A preliminary study on consistency with human preferences. *arXiv preprint arXiv:2303.07610*, 2023.

[96] Y. Jiang. Llm-based financial analytics chatbot, 2023. [Online; accessed Aug. 20, 2024].

[97] Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. Tinybert: Distilling bert for natural language understanding. *arXiv preprint arXiv:1909.10351*, 2019.

[98] Rafal Jozefowicz, Oriol Vinyals, Mike Schuster, Noam Shazeer, and Yonghui Wu. Exploring the limits of language modeling. *arXiv preprint arXiv:1602.02410*, 2016.

[99] Alex Kantrowitz and Mark MacCarthy. Europe's move toward strict ai regulation could be a global turning point. *The New York Times*, June 2023. Accessed: August 25, 2024.

[100] Amirhossein Kazemnejad. Transformer architecture: The positional encoding. *kazemnejad.com*, 2019.

[101] Sal Khan. Harnessing gpt-4 so that all students benefit. *A nonprofit approach for equal access. Khan Academy. URL: https://blog. khanacademy. org/harnessing-ai-so-that-allstudents-benefit-a-nonprofit-approach-for-equal-access*, 2023.

[102] Omar Khattab, Keshav Santhanam, Xiang Lisa Li, David Hall, Percy Liang, Christopher Potts, and Matei Zaharia. Demonstrate-search-predict: Composing retrieval and language models for knowledge-intensive nlp. *arXiv preprint arXiv:2212.14024*, 2022.

[103] Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, et al. Dynabench: Rethinking benchmarking in nlp. *arXiv preprint arXiv:2104.14337*, 2021.

[104] 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.

[105] Shunsuke Koga, Nicholas B Martin, and Dennis W Dickson. Evaluating the performance of large language models: Chatgpt and google bard in generating differential diagnoses in clinicopathological conferences of neurodegenerative disorders. *Brain Pathology*, 34(3):e13207, 2024.

[106] Anis Koubaa, Wadii Boulila, Lahouari Ghouti, Ayyub Alzahem, and Shahid Latif. Exploring chatgpt capabilities and limitations: a survey. *IEEE Access*, 2023.

[107] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*, 2019.

[108] Langchain. Recursively split by character, 2023. `https://python.langchain.com/docs/modules/data_connection/document_transformers/recursive_text_splitter`.

[109] LangChain. Langchain, 2024. `https://www.langchain.com/`.

[110] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[111] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*, 2021.

[112] K. Leung. Macy the ai pharmacist!, 2023. [Online; accessed Aug. 20, 2024].

[113] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*, 2019.

[114] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020.

[115] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023.

[116] Qintong Li, Leyang Cui, Lingpeng Kong, and Wei Bi. Collaborative evaluation: Exploring the synergy of large language models and humans for open-ended generation evaluation. *arXiv preprint arXiv:2310.19740*, 2023.

[117] Xiaoqian Li, Ercong Nie, and Sheng Liang. From classification to generation: Insights into crosslingual retrieval augmented icl. *arXiv preprint arXiv:2311.06595*, 2023.

[118] Yong Li, Naipeng Miao, Liangdi Ma, Feng Shuang, and Xingwen Huang. Transformer for object detection: Review and benchmark. *Engineering Applications of Artificial Intelligence*, 126:107021, 2023.

[119] Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*, 2022.

[120] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.

[121] Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods, 2021. arXiv preprint arXiv:2109.07958.

[122] Xi Victoria Lin, Xilun Chen, Mingda Chen, Weijia Shi, Maria Lomeli, Rich James, Pedro Rodriguez, Jacob Kahn, Gergely Szilvasy, Mike Lewis, et al. Ra-dit: Retrieval-augmented dual instruction tuning. *arXiv preprint arXiv:2310.01352*, 2023.

[123] Yen-Ting Lin and Yun-Nung Chen. Llm-eval: Unified multi-dimensional automatic evaluation for open-domain conversations with large language models. *arXiv preprint arXiv:2305.13711*, 2023.

[124] Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024.

[125] Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*, 2023.

[126] Yi Liu, Lianzhe Huang, Shicheng Li, Sishuo Chen, Hao Zhou, Fandong Meng, Jie Zhou, and Xu Sun. Recall: A benchmark for llms robustness against external counterfactual knowledge. *arXiv preprint arXiv:2311.08147*, 2023.

[127] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[128] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*, 2023.

[129] Mason Hayes Curran LLP. Regulating ai in the eu, July 2023. Accessed: August 25, 2024.

[130] LMSYS. Chatbot arena: Benchmarking llms in the wild with elo ratings, 2024. `https://lmsys.org`.

[131] Chung Kwan Lo. What is the impact of chatgpt on education? a rapid review of the literature. *Education Sciences*, 13(4):410, 2023.

[132] Sheng Lu, Irina Bigoulaeva, Rachneet Sachdeva, Harish Tayyar Madabushi, and Iryna Gurevych. Are emergent abilities in large language models just in-context learning? *arXiv preprint arXiv:2309.01809*, 2023.

[133] Fan Luo and Mihai Surdeanu. Divide & conquer for entailment-aware multi-hop evidence retrieval. *arXiv preprint arXiv:2311.02616*, 2023.

[134] Zheheng Luo, Qianqian Xie, and Sophia Ananiadou. Chatgpt as a factual inconsistency evaluator for text summarization. *arXiv preprint arXiv:2303.15621*, 2023.

[135] Ziyang Luo, Can Xu, Pu Zhao, Xiubo Geng, Chongyang Tao, Jing Ma, Qingwei Lin, and Daxin Jiang. Augmented large language models with parametric knowledge guiding. *arXiv preprint arXiv:2305.04757*, 2023.

[136] Xinbei Ma, Yeyun Gong, Pengcheng He, Hai Zhao, and Nan Duan. Query rewriting for retrieval-augmented large language models. *arXiv preprint arXiv:2305.14283*, 2023.

[137] 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.

[138] Nestor Maslej, Loredana Fattorini, Raymond Perrault, Vanessa Parli, Anka Reuel, Erik Brynjolfsson, John Etchemendy, Katrina Ligett, Terah Lyons, James Manyika, Juan Carlos Niebles, Yoav Shoham, Russell Wald, and Jack Clark. The ai index 2024 annual report. *AI Index Steering Committee, Institute for Human-Centered AI, Stanford University, Stanford, CA*, April 2024.

[139] Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. On faithfulness and factuality in abstractive summarization. *arXiv preprint arXiv:2005.00661*, 2020.

[140] Jesse G Meyer, Ryan J Urbanowicz, Patrick CN Martin, Karen O'Connor, Ruowang Li, Pei-Chen Peng, Tiffani J Bright, Nicholas Tatonetti, Kyoung Jae Won, Graciela Gonzalez-Hernandez, et al. Chatgpt and large language models in academia: opportunities and challenges. *BioData Mining*, 16(1):20, 2023.

[141] Microsoft. Copilot, 2024. [Online]. Available: `https://copilot.microsoft.com/`.

[142] Christoph Molnar, Giuseppe Casalicchio, and Bernd Bischl. Interpretable machine learning–a brief history, state-of-the-art and challenges. In *Joint European conference on machine learning and knowledge discovery in databases*, pages 417–431. Springer, 2020.

[143] Negin Yazdani Motlagh, Matin Khajavi, Abbas Sharifi, and Mohsen Ahmadi. The impact of artificial intelligence on the evolution of digital education: A comparative study of openai text generation tools including chatgpt, bing chat, bard, and ernie. *arXiv preprint arXiv:2309.02029*, 2023.

[144] National Security Commission on Artificial Intelligence. National security commission on artificial intelligence, 2021. Accessed: 2024-08-03.

[145] Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. Why we need new evaluation metrics for nlg. *arXiv preprint arXiv:1707.06875*, 2017.

[146] The White House Office of Management and Budget. Draft omb memo on regulation of artificial intelligence, 2020. Accessed: 2024-08-03.

[147] The White House Office of Science and Technology Policy (OSTP). Blueprint for an ai bill of rights, 2022. Accessed: 2024-08-03.

[148] State Council of the People's Republic of China. A next generation artificial intelligence development plan. *Web Archive*, 2017. Accessed via Web Archive on August 22, 2024. Available at: https://web.archive.org/web/20220121145209/https://www.mfa.gov.cn/ce/cefi/eng/kxjs/P020171025789108009001.pdf.

[149] 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.

[150] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

[151] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

[152] Dongju Park and Chang Wook Ahn. Self-supervised contextual data augmentation for natural language processing. *Symmetry*, 11:1393, 11 2019.

[153] Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. On the difficulty of training recurrent neural networks. In *International conference on machine learning*, pages 1310–1318. Pmlr, 2013.

[154] Nathan Paull. Beyond the transformer: The elements of future ai architectures. https://nathanpaull.substack.com/p/beyond-the-transformer-the-elements-of-future-ai-architectures-a4acd33cac89, 2023.

[155] Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. Is chatgpt a general-purpose natural language processing task solver? *arXiv preprint arXiv:2302.06476*, 2023.

[156] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. *OpenAI blog*, 2018.

[157] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[158] Robi Rahman, David Owen, and Josh You. Tracking large-scale ai models, 2024. Accessed: 2024-08-02.

[159] ResearchGraph. Brief introduction to the history of large language models (llms). https://medium.com/@researchgraph/brief-introduction-to-the-history-of-large-language-models-llms-3c2efa517112, 2024.

[160] Reuters. What are the eu's landmark ai rules?, December 2023. Accessed: August 25, 2024.

[161] Youngmin Ro and Jin Young Choi. Autolr: Layer-wise pruning and auto-tuning of learning rates in fine-tuning of deep networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 2486–2494, 2021.

[162] J. Rodriguez. Inside sparrow: The foundation of deepmind's chatgpt alternative, 2023. [Online]. Available: https://jrodthoughts.medium.com/inside-sparrow-the-foundation-of-deepminds-chatgpt-alternative-854df43569fd. [Accessed: Aug. 20, 2024].

[163] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, 1(5):206–215, 2019.

[164] Jon Saad-Falcon, Omar Khattab, Christopher Potts, and Matei Zaharia. Ares: An automated evaluation framework for retrieval-augmented generation systems. *arXiv preprint arXiv:2311.09476*, 2023.

[165] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

[166] Tomohiro Sawada, Daniel Paleka, Alexander Havrilla, Pranav Tadepalli, Paula Vidas, Alexander Kranias, John J Nay, Kshitij Gupta, and Aran Komatsuzaki. Arb: Advanced reasoning benchmark for large language models. *arXiv preprint arXiv:2307.13692*, 2023.

[167] Towards Data Science. Forget rag, the future is rag fusion, 2023. https://medium.com/towards-data-science/forget-rag-the-future-is-rag-fusion-1147298d8ad1.

[168] Jaime Sevilla and Edu Roldán. Training compute of frontier ai models grows by 4-5x per year, 2024. Accessed: 2024-08-02.

[169] Bayan Abu Shawar and Eric Atwell. Fostering language learner autonomy through adaptive conversation tutors. In *Proceedings of the The fourth Corpus Linguistics conference*, volume 3, pages 186–193, 2007.

[170] Tianyuan Shi, Liangzhi Li, Zijian Lin, Tao Yang, Xiaojun Quan, and Qifan Wang. Dual-feedback knowledge retrieval for task-oriented dialogue systems. *arXiv preprint arXiv:2310.14528*, 2023.

[171] Heung-Yeung Shum, Xiao-dong He, and Di Li. From eliza to xiaoice: challenges and opportunities with social chatbots. *Frontiers of Information Technology & Electronic Engineering*, 19:10–26, 2018.

[172] Michael Shumanov and Lester Johnson. Making conversations with chatbots more personalized. *Computers in Human Behavior*, 117:106627, 2021.

[173] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[174] Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode clinical knowledge. *arXiv preprint arXiv:2212.13138*, 2022.

[175] Elior Sulem, Omri Abend, and Ari Rappoport. Bleu is not suitable for the evaluation of text simplification. *arXiv preprint arXiv:1810.05995*, 2018.

[176] John P Sullins. When is a robot a moral agent. *Machine ethics*, 6(2001):151–161, 2011.

[177] Tianxiang Sun, Junliang He, Xipeng Qiu, and Xuanjing Huang. Bertscore is unfair: On social bias in language model-based metrics for text generation. *arXiv preprint arXiv:2210.07626*, 2022.

[178] P. Taylor. Volume of data/information created, captured, copied, and consumed worldwide from 2010 to 2025, 2023. [Online]. Available: https://www.statista.com/statistics/871513/worldwide-data-created/. [Accessed: Aug. 19, 2024].

[179] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

[180] R. Teja. Evaluating the ideal chunk size for a rag system using llamaindex, 2023. https://www.llamaindex.ai/blog/evaluating-the-ideal-chunk-size-for-a-rag-system-using-llamaindex-6207e5d3fec5.

[181] Mohamad-Hani Temsah, Amr Jamal, and Jaffar A Al-Tawfiq. Reflection with chatgpt about the excess death after the covid-19 pandemic. *scholarworks.iupui.edu*, 2023.

[182] Terranova Software. Terranova software, 2024. https://www.terranovasoftware.eu/en.

[183] Job Timmermans, Bernd Carsten Stahl, Veikko Ikonen, and Engin Bozdag. The ethics of cloud computing: A conceptual review. In *2010 IEEE second international conference on cloud computing technology and science*, pages 614–620. IEEE, 2010.

[184] 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.

[185] Towards Data Science. Understanding lora: Low-rank adaptation for fine-tuning large models. *Towards Data Science*, 2024.

[186] Alan M Turing. *Computing machinery and intelligence*. Springer, 2009.

[187] University of Bath. Ai poses no existential threat to humanity, new study finds. https://www.bath.ac.uk/announcements/ai-poses-no-existential-threat-to-humanity-new-study-finds/, 2024.

[188] Marco Antonio Rodrigues Vasconcelos and Renato P dos Santos. Enhancing stem learning with chatgpt and bing chat as objects to think with: A case study. *arXiv preprint arXiv:2305.02202*, 2023.

[189] Ashish Vaswani, Noam Shazeer, 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.

[190] 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.

[191] 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.

[192] VoxEU. Regulatory export and spillovers: How gdpr affects global markets for data. *VoxEU*, 2023. Accessed: August 25, 2024.

[193] Tu Vu, Mohit Iyyer, Xuezhi Wang, Noah Constant, Jerry Wei, Jason Wei, Chris Tar, Yun-Hsuan Sung, Denny Zhou, Quoc Le, et al. Freshllms: Refreshing large language models with search engine augmentation. *arXiv preprint arXiv:2310.03214*, 2023.

[194] Richard S Wallace. *The anatomy of ALICE*. Springer, 2009.

[195] Alex Wang et al. Glue: A multi-task benchmark and analysis platform for natural language understanding, 2018. arXiv preprint arXiv:1804.07461.

[196] Boxin Wang, Weixin Chen, Hengzhi Pei, Chulin Xie, Mintong Kang, Chenhui Zhang, Chejian Xu, Zidi Xiong, Ritik Dutta, Rylan Schaeffer, et al. Decodingtrust: A comprehensive assessment of trustworthiness in gpt models. In *NeurIPS*, 2023.

[197] Cunxiang Wang, Sirui Cheng, Qipeng Guo, Yuanhao Yue, Bowen Ding, Zhikun Xu, Yidong Wang, Xiangkun Hu, Zheng Zhang, and Yue Zhang. Evaluating open-qa evaluation. *Advances in Neural Information Processing Systems*, 36, 2024.

[198] Hongru Wang, Rui Wang, Fei Mi, Yang Deng, Zezhong Wang, Bin Liang, Ruifeng Xu, and Kam-Fai Wong. Cue-cot: Chain-of-thought prompting for responding to in-depth dialogue questions with llms. *arXiv preprint arXiv:2305.11792*, 2023.

[199] Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, Tao Qin, Wang Lu, Yiqiang Chen, Wenjun Zeng, and S Yu Philip. Generalizing to unseen domains: A survey on domain generalization. *IEEE transactions on knowledge and data engineering*, 35(8):8052–8072, 2022.

[200] Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. Large language models are not fair evaluators. *arXiv preprint arXiv:2305.17926*, 2023.

[201] Xintao Wang, Qianwen Yang, Yongting Qiu, Jiaqing Liang, Qianyu He, Zhouhong Gu, Yanghua Xiao, and Wei Wang. Knowledgpt: Enhancing large language models with retrieval and storage access on knowledge bases. *arXiv preprint arXiv:2308.11761*, 2023.

[202] Yidong Wang, Zhuohao Yu, Zhengran Zeng, Linyi Yang, Cunxiang Wang, Hao Chen, Chaoya Jiang, Rui Xie, Jindong Wang, Xing Xie, et al. Pandalm: An automatic evaluation benchmark for llm instruction tuning optimization. *arXiv preprint arXiv:2306.05087*, 2023.

[203] Yu Wang, Nedim Lipka, Ryan A Rossi, Alexa Siu, Ruiyi Zhang, and Tyler Derr. Knowledge graph prompting for multi-document question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19206–19214, 2024.

[204] John Frank Weaver. Regulation of artificial intelligence in the united states. In *Research Handbook on the Law of Artificial Intelligence*, pages 155–212. Edward Elgar Publishing, 2018.

[205] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022.

[206] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

[207] 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.

[208] Wikipedia. Information age. `https://en.wikipedia.org/wiki/Information_Age`. Accessed: August 24, 2024.

[209] Wikipedia. ChatGPT, 2023. [Online]. Available: `https://en.wikipedia.org/wiki/ChatGPT`. [Accessed: Aug. 20, 2024].

[210] Wikipedia. Cleverbot, 2023. [Online]. Available: `https://en.wikipedia.org/wiki/Cleverbot`. [Accessed: August 20, 2024].

[211] Wikipedia. Gemini (chatbot), 2023. [Online]. Available: `https://en.wikipedia.org/wiki/Gemini_(chatbot)`. [Accessed: Aug. 20, 2024].

[212] Wikipedia. Microsoft copilot, 2023. [Online]. Available: `https://en.wikipedia.org/wiki/Microsoft_Copilot`. [Accessed: Aug. 20, 2024].

[213] Wikipedia. Racter, 2023. [Online]. Available: `https://en.wikipedia.org/wiki/Racter`.

[214] Wikipedia. Anthropic, 2024. [Accessed: Aug. 20, 2024].

[215] Wikipedia. Grok (chatbot), 2024. [Online]. Available: `https://en.wikipedia.org/wiki/Grok_(chatbot)`. [Accessed: Aug. 20, 2024].

[216] Terry Winograd. Understanding natural language. *Cognitive psychology*, 3(1):1–191, 1972.

[217] Create with Swift. Prototyping swiftui interfaces with openai's chatgpt. `https://www.createwithswift.com/prototyping-swiftui-interfaces-with-openais-chatgpt/`, 2024. Accessed: August 26, 2024.

[218] Fei Wu, Cewu Lu, Mingjie Zhu, Hao Chen, Jun Zhu, Kai Yu, Lei Li, Ming Li, Qianfeng Chen, Xi Li, et al. Towards a new generation of artificial intelligence in china. *Nature Machine Intelligence*, 2(6):312–316, 2020.

[219] Minghao Wu and Alham Fikri Aji. Style over substance: Evaluation biases for large language models. *arXiv preprint arXiv:2307.03025*, 2023.

[220] 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.

[221] Xingjiao Wu, Luwei Xiao, Yixuan Sun, Junhang Zhang, Tianlong Ma, and Liang He. A survey of human-in-the-loop for machine learning. *Future Generation Computer Systems*, 135:364–381, 2022.

[222] Shi-Qi Yan, Jia-Chen Gu, Yun Zhu, and Zhen-Hua Ling. Corrective retrieval augmented generation. *arXiv preprint arXiv:2401.15884*, 2024.

[223] Linyi Yang, Shuibai Zhang, Libo Qin, Yafu Li, Yidong Wang, Hanmeng Liu, Jindong Wang, Xing Xie, and Yue Zhang. Glue-x: Evaluating natural language understanding models from an out-of-distribution generalization perspective. *arXiv preprint arXiv:2211.08073*, 2022.

[224] S. Yang. Advanced rag 01: Small-to-big retrieval, 2023. `https://towardsdatascience.com/advanced-rag-01-small-to-big-retrieval-172181b396d4`.

[225] Z. Yang. Chinese tech giant baidu just released its answer to chatgpt, 2023. [Online]. Available: `https://www.technologyreview.com/2023/03/16/1069919/baidu-ernie-bot-chatgpt-launch/`. [Accessed: Aug. 20, 2024].

[226] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32, 2019.

[227] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *arXiv preprint arXiv:2306.13549*, 2023.

[228] Ori Yoran, Tomer Wolfson, Ori Ram, and Jonathan Berant. Making retrieval-augmented language models robust to irrelevant context. *arXiv preprint arXiv:2310.01558*, 2023.

[229] Wenhao Yu, Dan Iter, Shuohang Wang, Yichong Xu, Mingxuan Ju, Soumya Sanyal, Chenguang Zhu, Michael Zeng, and Meng Jiang. Generate rather than retrieve: Large language models are strong context generators. *arXiv preprint arXiv:2209.10063*, 2022.

[230] Weizhe Yuan, Graham Neubig, and Pengfei Liu. Bartscore: Evaluating generated text as text generation. *Advances in Neural Information Processing Systems*, 34:27263–27277, 2021.

[231] Liangyu Zha, Junlin Zhou, Liyao Li, Rui Wang, Qingyi Huang, Saisai Yang, Jing Yuan, Changbao Su, Xiang Li, Aofeng Su, et al. Tablegpt: Towards unifying tables, nature language and commands into one gpt. *arXiv preprint arXiv:2307.08674*, 2023.

[232] Chaoning Zhang, Chenshuang Zhang, Chenghao Li, Yu Qiao, Sheng Zheng, Sumit Kumar Dam, Mengchun Zhang, Jung Uk Kim, Seong Tae Kim, Jinwoo Choi, et al. One small step for generative ai, one giant leap for agi: A complete survey on chatgpt in aigc era. *arXiv preprint arXiv:2304.06488*, 2023.

[233] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.

[234] Yangjun Zhang, Pengjie Ren, and Maarten de Rijke. A human-machine collaborative framework for evaluating malevolence in dialogues. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5612–5623, 2021.

[235] 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.

[236] Huaixiu Steven Zheng, Swaroop Mishra, Xinyun Chen, Heng-Tze Cheng, Ed H Chi, Quoc V Le, and Denny Zhou. Take a step back: Evoking reasoning via abstraction in large language models. *arXiv preprint arXiv:2310.06117*, 2023.

[237] Lianmin Zheng et al. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, 2024.

[238] Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, et al. Least-to-most prompting enables complex reasoning in large language models. *arXiv preprint arXiv:2205.10625*, 2022.

[239] Li Zhou, Jianfeng Gao, Di Li, and Heung-Yeung Shum. The design and implementation of xiaoice, an empathetic social chatbot. *Computational Linguistics*, 46(1):53–93, 2020.

[240] Kaijie Zhu, Jindong Wang, Jiaheng Zhou, Zichen Wang, Hao Chen, Yidong Wang, Linyi Yang, Wei Ye, Yue Zhang, Neil Zhenqiang Gong, et al. Promptbench: Towards evaluating the robustness of large language models on adversarial prompts. *arXiv preprint arXiv:2306.04528*, 2023.

[241] Terry Yue Zhuo, Yujin Huang, Chunyang Chen, and Zhenchang Xing. Red teaming chatgpt via jailbreaking: Bias, robustness, reliability and toxicity. *arXiv preprint arXiv:2301.12867*, 2023.

[242] Anne Zimmerman, Joel Janhonen, and Emily Beer. Human/ai relationships: challenges, downsides, and impacts on human/human relationships. *AI and Ethics*, pages 1–13, 2023.

List of Figures

List of Tables