Transforming Conversational AI

Exploring the Power of Large Language Models in Interactive Conversational Agents

Michael McTear
Marina Ashurkina

Apress®

TRANSFORMING CONVERSATIONAL AI

EXPLORING THE POWER OF LARGE LANGUAGE MODELS IN INTERACTIVE CONVERSATIONAL AGENTS

Michael McTear
Marina Ashurkina

Apress®

Transforming Conversational AI: Exploring the Power of Large Language Models in Interactive Conversational Agents

Michael McTear
Belfast, Northern Ireland, UK

Marina Ashurkina
London, UK

ISBN-13 (pbk): 979-8-8688-0109-9
ISBN-13 (electronic): 979-8-8688-0110-5
https://doi.org/10.1007/979-8-8688-0110-5

Managing Director, Apress Media LLC: Welmoed Spahr
Acquisitions Editor: Shivangi Ramachandran
Development Editor: James Markham
Project Manager: Gryffin Winkler

Cover designed by eStudioCalamar

Distributed to the book trade worldwide by Springer Science+Business Media New York, 1 New York Plaza, Suite 4600, New York, NY 10004-1562, USA. Phone 1-800-SPRINGER, fax (201) 348-4505, e-mail orders-ny@springer-sbm.com, or visit www.springeronline.com. Apress Media, LLC is a California LLC and the sole member (owner) is Springer Science + Business Media Finance Inc (SSBM Finance Inc). SSBM Finance Inc is a **Delaware** corporation.

For information on translations, please e-mail booktranslations@springernature.com; for reprint, paperback, or audio rights, please e-mail bookpermissions@springernature.com.

Apress titles may be purchased in bulk for academic, corporate, or promotional use. eBook versions and licenses are also available for most titles. For more information, reference our Print and eBook Bulk Sales web page at http://www.apress.com/bulk-sales.

Any source code or other supplementary material referenced by the author in this book is available to readers on GitHub. For more detailed information, please visit https://www.apress.com/gp/services/source-code.

Paper in this product is recyclable

Contents

About the Authors

Michael McTear is an emeritus professor of Ulster University who has worked in spoken dialogue technologies and conversational AI for more than 20 years. He is the author of several books, including *Spoken Dialogue Technology* (Springer, 2004), *The Conversational Interface* (Springer, 2016), and *Conversational AI* (Springer, 2020). Currently Michael is involved in several research and development projects investigating the use of conversational agents in socially relevant projects such as mental health monitoring and home monitoring of older adults. Michael's main motivation for writing this book is to bring new developments in conversational AI to the attention of conversation designers and other professionals in a clear and accessible manner.

Marina Ashurkina studied linguistics and translation studies. She has over eight years of experience working with dialogue systems, including working in the company api.ai before it was acquired by Google and became Dialogflow. Also she had her own consultancy Cherry.ai helping companies build smart assistants and worked on building a multilingual voice assistant platform for Huawei. In 2020, Marina created and lectured on a conversation design course to 60 students with a focus on building skills for smart speakers. She was also a Product Manager in Generative Assistants Inc., a US-based startup striving to streamline the creation of generative AI assistants. Besides that, Marina is a certified Project Manager Professional, Scrum Master, and Product Owner, which helps her to set up and drive complex Conversational AI projects.

About the Technical Reviewer

Tom Taulli (@ttaulli) is an advisor and board member of various AI companies. He is also the author of books like *Generative AI: How ChatGPT and Other AI Tools Will Revolutionize Business* and *Artificial Intelligence Basics: A Non-Technical Introduction*.

Acknowledgments

In writing this book, we have been guided by an efficient and supportive team from Apress, including Shobana Srinivasan (Production Editor), Tom Taulli (Technical Reviewer), Jim Markham (Development Editor), Gryffin Winkler (Editorial Project Manager), and Linthaa Muralidharan, (Production Supervisor).

Thank you for all your constructive comments and guidance that have helped us improve our book.

We received useful feedback and suggestions from several friends and colleagues, including Mikhail Burtsev (Landau AI Fellow at London Institute for Mathematical Sciences) and Muskaan Singh (Ulster University) for helpful comments on Chapters 3 and 4 and Arseny Fitilbam (founder and CTO of JIQ.ai) for providing anonymized examples of conversations and statistical data for Chapter 8.

Writing a book requires a lot of time and effort, and during the months of drafting, editing, and rewriting chapters, we have been encouraged by our partners who have been patient and supportive during the long hours in which we were working on the book. Michael would like to thank and acknowledge the support of his wife Sandra; Marina is grateful for the support and encouragement she received from her husband Adam.

Introduction

We were motivated to write this book by the launch in November 2022 of ChatGPT and by the ensuing excitement and disruption across the world of Conversational AI. The book is written for a broad audience who are already working, starting to work, or simply interested in Conversational AI. This will include conversation designers, for whom these new technologies are bringing challenges as well as new opportunities; product owners, project managers, software developers, and data scientists who wish to learn about these new methods and technologies; and final year undergraduates and graduates of computer science who are keen to learn about Conversational AI. The book will also be of interest to professionals involved in content generation and discovery across diverse fields, including marketing, law, medicine, and education, as well as members of the general public eager to find out more about this revolutionary new technology.

In writing this book, we have been guided by two primary objectives. Firstly, we want to provide a practical guide for those who wish to explore Conversational AI and its associated technologies. A focal point of the book is the intricate art of prompt engineering. We illustrate with detailed examples the role of Prompt Engineers, nowadays much sought after specialists who can skillfully develop and optimize prompts to enhance the performance of systems powered by LLMs.

Our second aim is to enable you to understand and appreciate the complexities of the technologies of Conversational AI in a relatively non-technical way. Modern Conversational and Generative AI differ considerably from technologies that most readers will have encountered previously, and so we believe strongly that it is important to have a basic understanding and appreciation of how these new systems work. Also, given the controversies that surround the whole area of modern AI, we feel that it is important to consider risks and various ethical considerations that have featured prominently in media discussions.

Conversational AI is a dynamic and rapidly evolving field, with new advancements being reported almost weekly. Our aim in this book is to provide a comprehensive overview of the core concepts and principles of conversational AI, equipping you with a solid understanding of this ground-breaking technology.

In our final chapter, we will delve into the latest developments in conversational AI, highlighting the most significant breakthroughs and emerging trends up to the time of publication. To ensure you stay at the forefront of this exciting field, we encourage you to explore the list of resources provided, which will guide you to continue learning and staying informed about future advancements.

Overview of the Book

There are ten chapters in the book. Here is a brief summary of what we cover in each chapter.

Chapter 1, "A New Era in Conversational AI," introduces groundbreaking developments in Conversational AI since the launch of ChatGPT in November 2022. Key terms in Conversational AI are explained along with illustrative examples of interactions with ChatGPT and similar chatbots and an overview of how AI-powered chatbots are revolutionizing diverse application areas, transforming the way we interact with technology.

Chapter 2, "Designing Conversational Systems," reviews current approaches to conversation design and assesses the impact of recent developments, showing how Large Language Models can be leveraged to help designers brainstorm user intents, system responses, and conversation flows. The chapter also describes what is involved in leading a Conversational AI project, outlining the roles and responsibilities within a cross-functional team to ensure successful project execution.

Chapter 3, "The Rise of Neural Conversational Systems," introduces the encoder–decoder architecture which provides a foundation for neural conversational systems. We explore transformers and the attention mechanism which have become state-of-the-art and revolutionized the field of Conversational AI. We conclude by outlining the advantages and disadvantages of the neural conversational approach compared to the traditional rule-based approach described in Chapter 2.

Chapter 4, "Large Language Models," introduces Large Language Models (LLMs) and explains how they have transformed Conversational AI. We delve into the intricate mechanisms of LLMs and explore their fundamental differences from traditional search engines, how they can be augmented with external knowledge, and what is involved in fine-tuning. We also address the challenges and limitations of LLMs.

Chapter 5, "Introduction to Prompt Engineering," introduces the essential terminology and concepts central to prompt engineering. It explores web interfaces for famous LLMs and examines different use cases. The chapter demonstrates practical examples of crafting effective prompts, common design techniques, and patterns. It also presents actionable examples for

Conversation Designers, illustrating methods to significantly reduce the time and effort required to develop intent-based virtual agents through prompt engineering. This chapter will help readers learn how to craft prompts for many scenarios. Moreover, this chapter lays the foundation for the advanced prompt engineering topics in Chapter 6.

Chapter 6, "Advanced Prompt Engineering," offers an extensive overview of advanced tools and examples to develop prompt engineering skills further. It is written for those who want to go beyond basic LLM interfaces and acquire hands-on experience configuring and setting up the optimal combination of LLM parameters, chaining prompts, and creating LLM applications. This chapter covers system prompts and prompt settings, playgrounds, and APIs and discusses prompt hacking. It also reviews several sophisticated prompt patterns with reasoning elements, such as Chain-of-Thought, ReAct, and Self-Consistency.

Chapter 7, "Conversational AI Platforms," reviews the transformation of conversational AI platforms from traditional to hybrid and ultimately to new LLM-based platforms. This chapter lists the most important components of classic platforms and how they are influenced by the rise of LLMs. Generative AI features become a new norm in hybrid platforms to automate the process of creating conversational systems and to enrich the end-user experience with live text generation and dynamic reasoning inside the application.

Chapter 8, "Evaluation Metrics," explores various approaches for the evaluation of conversational systems. We begin by examining metrics employed in the assessment of traditional intent-based conversation systems. Next we provide a comprehensive overview of different frameworks for evaluating LLMs. Following this, we discuss the essential product metrics for evaluating conversational systems as a whole. Finally, we introduce the innovative concept of employing LLMs as a tool for assessing the quality of conversations.

Chapter 9, "AI Safety and Ethics," delves into ethical considerations, including the handling of bias, toxic content, misinformation, privacy, and data protection. We examine how these critical issues are currently being tackled through regulatory measures and the establishment of standards aimed at fostering trustworthy and responsible AI.

Chapter 10, "Final Words," reviews recent advancements in Conversational AI and the role of LLMs. We also explore the exciting possibilities that lie ahead in this rapidly evolving and captivating field.

The Appendix contains a list of LLM-powered chatbots that you can use to test the examples in the book.

The Notebook is a web-based resource accessible through `https://github.com/Apress/Transforming-Conversational-AI` to copy and paste the examples of prompts provided in the book.

CHAPTER

1

A New Era in Conversational AI

On November 30, 2022, OpenAI, a prominent US company with headquarters in San Francisco, released a publicly available version of a chatbot called ChatGPT that transformed the world of Conversational Artificial Intelligence (AI) and ignited what has come to be known as "The Conversational AI Arms Race." Within just five days of its launch, ChatGPT had acquired a million users, and within two months, it was estimated to have 100 million active users. In February 2023, Microsoft, having invested heavily in OpenAI, launched a version of its Bing search engine powered by the technology behind ChatGPT. Google responded in March 2023 by releasing its own AI-powered chatbot called Bard. Others followed, including Anthropic, funded initially by Google and subsequently by Amazon, with a chatbot called Claude, as well as major Chinese tech firms, such as Baidu and Alibaba.

Approximately a year after the launch of ChatGPT, on November 6, 2023, OpenAI unveiled a host of enhancements, innovative products, and tools at

M. McTear and M. Ashurkina, *Transforming Conversational AI*,
https://doi.org/10.1007/979-8-8688-0110-5_1

its inaugural developer conference, OpenAI DevDay.[1] These developments are likely to provide new opportunities for those involved in the creation of AI applications. At the same time, they present formidable challenges for competitor companies and we can expect another upsurge of activity as the industry responds to and addresses these challenges.

So what is this all about and why does it matter? Our aim in this chapter is to introduce you to the world of ChatGPT and similar chatbots. We will begin by defining some of the commonly used terms in this area of Artificial Intelligence, such as Conversational AI, Generative AI, and Large Language Models. Following this, we will introduce some examples of how you can engage in natural and meaningful dialogues with ChatGPT and other chatbots. More detailed examples and explanations of how they work and how they can be used will be provided in later chapters. Next we will examine some areas in which these chatbots are being used, looking at the benefits as well as some of the concerns around their use. The chapter concludes with a list of useful resources for you to consult if you wish to delve further into this fascinating field.

By the end of the chapter, you will have gained a good understanding of the main concepts in the fields of Conversational and Generative AI, insights into the diverse types of applications leveraging these technologies, and an awareness of how these applications are likely to impinge on many aspects of our daily lives.

Understanding Key Terms in Conversational AI

Before we go any further, it will be useful to explain some of the terms that we will be using throughout this book.

Conversational AI is a fairly recent term that describes an area of Natural Language Processing (NLP) and Artificial Intelligence (AI) concerned with developing systems that can process human language and interact with humans in a natural way that mimics human conversation. These systems are known by various names, including **conversational agents or assistants**, **chatbots**, and **digital personal assistants**. The term **(spoken) dialogue system** is used widely in academic and industrial research laboratories, while in commercial applications such as automated customer service, they are known as **voice user interfaces**. **Embodied Conversational Agents (ECAs)** are another type of application that features computer-generated animated characters and social robots that can display emotions, gestures, and facial expressions. In some cases, they can also recognize and interpret these cues when displayed by the humans they interact with, thus providing a more human-like and engaging form of interaction. Recently, Meta has been

[1] https://devday.openai.com/

developing Conversational AI characters with unique interests and personalities (see further Chapter 10).[2]

Natural Language Processing (NLP) is a branch of Artificial Intelligence that is concerned with giving computers the ability to process, understand, and generate natural language. NLP has its roots in the 1950s with early attempts at machine translation, and has passed through several stages:

- Symbolic NLP (from the 1950s to early 1990s): in which hand-crafted rules were developed to understand and generate natural language texts
- Statistical NLP (from the 1990s to 2010s): in which machine learning algorithms were used in tasks such as classifying texts and user inputs
- Neural NLP (from around 2010 to the present): in which deep learning methods have been applied to NLP tasks

NLP can be broken down into **Natural Language Understanding (NLU)** and **Natural Language Generation (NLG)**. Interactive NLP systems, such as Dialogue Systems, also include a **Dialogue Management (DM)** component that processes inputs and determines the system's actions and responses. Voice-based (or spoken) dialogue systems also include an **Automated Speech Recognition (ASR)** component that converts spoken input into text and a **Text-to-Speech (TTS)** component that converts text output to speech.

Note Recently, the term NLU has come to be used to describe chatbots and conversational systems that have been developed using traditional technologies involving intents, entities, and pre-defined system responses and conversational flows, as described in Chapter 2, as opposed to systems developed using neural technologies, as described in Chapters 3 and 4.

Generative AI is a new and rapidly emerging area of AI that is concerned with generating new data. This data can be in the form of textual content, such as responses to prompts, summarizations, and text transformations, such as translation to different formats or different languages. More recently, Generative AI is being used to generate images, 3D models, videos, and music. Generative AI leverages the capabilities of Large Language Models (LLMs) to create this new content and has a wide range of potential applications in fields such as art, music, gaming, entertainment, and scientific research. In the

[2] https://about.fb.com/news/2023/09/introducing-ai-powered-assistants-characters-and-creative-tools/

commercial arena, Generative AI is being deployed to enhance productivity in repetitive tasks such as the creation of marketing content, legal documents, and more.

Large Language Models (or LLMs) represent a breakthrough in recent AI. Large Language Models can understand and generate human-like language and are used to perform many tasks in NLP, including translation, summarization, question-answering, and content generation. We provide a fairly non-technical overview of LLMs, how they are trained, and how they are used, in Chapter 4.

AI-powered chatbots. This term refers to chatbots and conversational agents that make use of new technologies such as Large Language Models in contrast to earlier systems based on hand-crafted rules. We show examples of several AI-powered chatbots in the book. As well as ChatGPT, we also provide examples generated by Google's Bard, Anthropic's Claude, and others.

■ **Note** For many of these systems, there are free as well as subscription-based versions. You can find a list of these in the Appendix.

ChatGPT is a conversational interface to various LLMs developed by OpenAI. The interface allows users to insert a prompt to which ChatGPT generates a response, or more precisely, a completion, as the prompt provides a completion to the words of the user's input. The latest version of ChatGPT can also generate images from textual prompts and search the Internet, and there are also speech-to-text and text-to-speech capabilities. We describe how the completion is generated in more detail in Chapter 3, while the creation of effective prompts to ensure useful output is explained with multiple examples in Chapters 5 and 6.

GPT. The GPT in ChatGPT refers to the **Generative Pre-trained Transformer Architecture** that is the basis for AI-powered chatbots. The Transformer architecture is described in Chapter 3, while Chapter 4 provides an overview of pre-trained (or foundational) Large Language Models that make use of the architecture.

■ **Note** The term GPT is now being used to refer to applications in which ChatGPT can be customized by anyone wishing to develop chatbots for their own specific purposes using non-coding methods. It is planned to create a GPT Store where these GPTs can be stored and made accessible.[3]

---

[3] https://openai.com/blog/introducing-gpts

Interacting with ChatGPT and Similar Chatbots

While the underlying technologies powering ChatGPT and similar AI-powered chatbots have been in existence for several years, it wasn't until its launch in November 2022 that ChatGPT became the fastest-growing computing technology in history. This was largely due to its simple user-friendly chat interface that allowed anyone with Internet access to engage in open-ended conversations on any topic with an AI entity that could provide detailed answers to questions, execute tasks such as document summarization, generation of emails and other content, language translation, computer code production, and much more. Figure 1-1 depicts the intuitive ChatGPT chat interface.

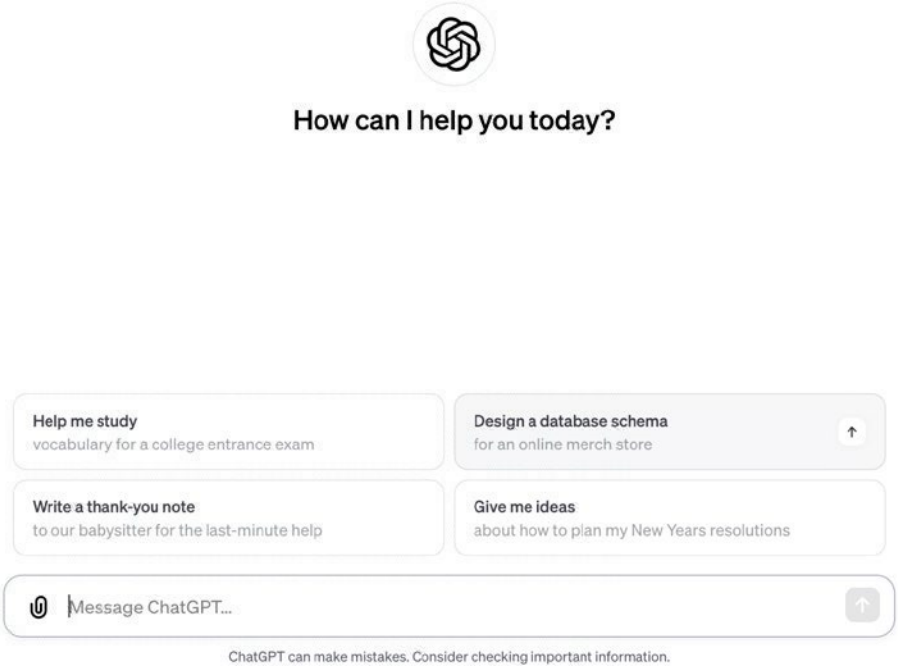

Figure 1-1. The ChatGPT chat interface[4]

In addition to offering interaction tips and a disclaimer about potential errors, users are presented with a message box for entering textual prompts. The latest version, ChatGPT-4, expands the input options to include images as well as documents. Similarly, other advanced chatbots, such as Anthropic's Claude, also accept documents as input.

[4] https://chat.openai.com/

What differentiates a chat interface from a simple request to an LLM is the feature of keeping the entire conversational context. The user can freely switch from one topic to another, and the bot will retain and remember all the information to support the conversation as a human would. This makes the conversation genuinely remarkable. It's possible due to the large context window. At the time of writing this book, Claude 2.1 (Anthropic) has an industry-leading context window of 200K tokens, which was released on November 21st.[5] It can remember and track text as large as 150,000 words or 500 pages.

We will introduce the concepts of tokens, context window, prompt parameters, and techniques for prompt engineering in Chapters 5 and 6. You can copy the provided prompt examples from the notebook and try them out in different chat interfaces. Chapters 6 and 7 will introduce playgrounds and Conversational AI platforms where you can build your own LLM application similar to ChatGPT. This section offers a few examples to illustrate the power and versatility of this ground-breaking technology.

We have used search engines such as Google to access information for several decades already. That's why when we interact with ChatGPT or similar chat interfaces, we unconsciously use them as search engines. We need to start thinking differently about them. To get a better result, we need to improve our query. Instead of asking a simple question, we can provide context, ask them to follow instructions, and give detailed descriptions of what we need. We can ask the LLM to take on a specific role, such as a teacher, lawyer, financial advisor, detective, and many more.

Let's provide a simple example of how a differently formulated query can improve the conversation with ChatGPT. If we want to learn about the history of the UK's landmark the London Tower Bridge, instead of just asking, "How was London Tower Bridge built?" we can provide a detailed prompt, as shown in Figure 1-2.

[5] www.anthropic.com/index/claude-2-1

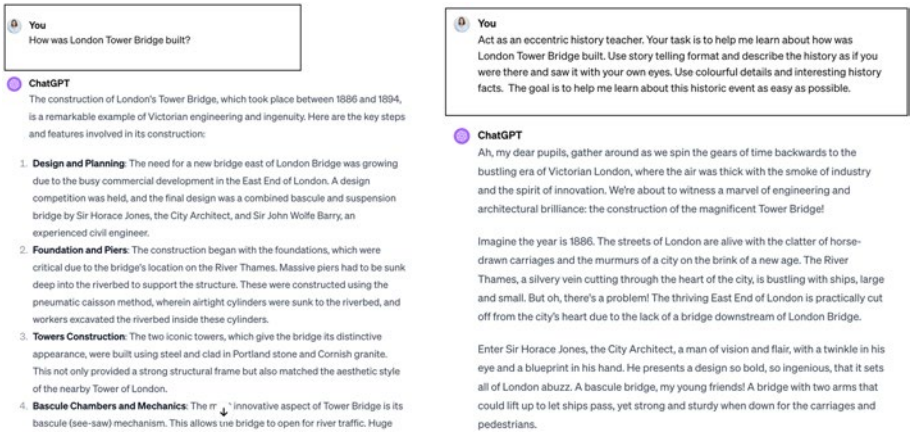

Figure 1-2. Interacting with ChatGPT is different from interacting with a search engine. Prompt engineering techniques make the generated text unique and creative

Using different techniques for constructing prompts, we can create better experiences and generate unique text. Prompt engineering is as much a technical task as it is creative.

OpenAI models have a knowledge cut-off; if used as-is, they can't provide real-time information. GPT-4, OpenAI's most recent model, has world knowledge up to April 2023. To solve this issue, OpenAI first introduced ChatGPT plugins, which were able to make requests to third-party applications to get relevant data. In November 2023, OpenAI rolled out GPTs, custom versions of ChatGPT, which can be connected to the real world via a function called 'actions'. Microsoft took an early stance on the issue of knowledge cut-off with Bing by integrating search, browsing, and chat into a unified experience in February 2023. Google also integrated search into its conversational interface in Bard. We will talk about GPTs in Chapter 10 of this book.

Throughout this book, we will provide numerous examples of how to interact with ChatGPT and similar chat interfaces. For a list of the most prominent AI-powered conversational systems, refer to the Appendix.

Using AI-Powered Chatbots: Examples of Some Relevant Application Areas

Chatbots have been used for a number of years in many diverse application areas, including customer service, education, healthcare, and as social companions. Traditionally, these applications were developed using conventional design and development methods, as detailed in Chapter 2. Now the field of Conversational AI has been revolutionized by the emergence

of Large Language Models (LLMs) and deep neural network architectures. While these advancements have opened up exciting new possibilities, they have also presented a unique set of challenges. In this section, we delve into some key application areas where these emerging and innovative technologies are making a significant impact.

Customer Service

Customer service is one of the most popular use cases for Conversational AI. According to Gartner, Inc., chatbots will become a primary customer service channel for roughly a quarter of organizations by 2027.[6] Businesses were eager to automate customer service long before Generative AI. They used more conventional technology and deployed it to multiple channels such as the web, phone, emails, and messengers.

Generative AI uncovers new opportunities for customer support. With its ever-growing context window size, it can remember the entire conversation history with a specific customer across multiple channels and provide better customer support. Customer support agents can benefit from using live AI assistants which can create drafts and suggestions during live calls with customers. LLM-powered conversational agents sound more fluent and human-like. Customer service can benefit from using Generative AI for summarizing customers' cases, identifying the sentiment of the conversation, and gaining insights from data. Also, automated internal employee support is a growing business case as a large amount of the company's data can be used as a knowledge base for conversational agents.

However, implementing Generative AI in customer-facing applications may come with risks. Traditional tools which used machine learning classifiers to identify pre-defined intents and follow specific scenarios offered more control over the technology. LLMs come with incredible opportunities as well as certain risks, such as hallucinations, where the LLM fabricates information, bias, privacy, latency, copyright, and other issues which we aim to address comprehensively in this book.

With the right approach and required skills within the team, it's possible to bring Generative AI to customer support. Let's take as an example South Korea's leading mobile operator KT, which has trained its own LLM in the Korean language. GiGA Genie has become the most popular AI voice assistant in South Korea and has had conversations with over 8 million customers as of

[6] www.gartner.com/en/newsroom/press-releases/2022-07-27-gartner-predicts-chatbots-will-become-a-primary-customer-service-channel-within-five-years

September 2022. By leveraging LLMs, the company has achieved significant quality improvements, better language understanding, and more human-sounding sentences.[7]

Education

Shortly after the launch of ChatGPT in November 2022, concerns were voiced in educational circles warning of the dangers arising from the potential misuse of this new technology. There was a fear that students would be able to have essays generated by ChatGPT that would be indistinguishable from their own work, making the detection of plagiarism almost impossible. Furthermore, given the propensity for LLMs to output inaccurate information, students could be fed content that they would be unable to critically evaluate.

These are legitimate concerns that warrant careful consideration. However, it is important to recognize that alongside these challenges, the emergence of new technologies like ChatGPT brings forth many exciting opportunities for students as well as educators. Embracing these technologies responsibly can empower students to explore innovative learning methodologies, while educators can foster a dynamic and enriched learning environment, tailoring their approaches to cater to individual needs and inspiring a new era of educational excellence.

LLMs provide a versatile learning tool for pupils and students at all levels of education, from elementary school through to university and beyond, in tasks such as writing essays, translating texts, summarizing documents, and generating computer code. The challenge is to treat this generated content not as a final product but as an initial suggestion that can be refined based on specific criteria. Assessment of the student's work should extend beyond the text output produced by the LLM to a focus on how the student iteratively refined and re-designed prompts to the LLM throughout the learning process. In this way, the LLM becomes a facilitator in the content production process.

LLMs also have the potential to serve as tools for improving the student's writing skills and critical thinking abilities, while also supporting other tasks such as the development of reading comprehension or the learning of foreign languages. Each student can work individually with their own chat-based LLM interface, thus benefiting from a personalized learning experience in which they receive individualized constructive feedback.

There are also many benefits for teachers. For example, LLMs can be used to produce lesson plans or to brainstorm the topics to be covered in a lecture. These outputs could be tailored to cater to different levels of student

[7] https://blogs.nvidia.com/blog/kt-large-language-models/

proficiency, creating personalized lesson plans that align with individual learning needs.

LLMs can also be beneficial in semi-automated grading processes where the teacher can input the student's work into the LLM and obtain a concise summary highlighting the strengths and weaknesses of the work. LLMs can also be used as a powerful tool for plagiarism detection.

Balancing the concerns mentioned earlier with the potential benefits requires a concerted effort to establish robust guidelines, ethical frameworks, and educational practices that harness the transformative power of ChatGPT while mitigating its risks.

For more detailed discussion of the benefits as well as the challenges of LLMs in education, you can check out the following papers: "Practical and Ethical Challenges of Large Language Models in Education: A systematic review"[8] and "ChatGPT for Good? On Opportunities and Challenges of Large Language Models for Education."[9]

Healthcare

Healthcare is a domain where LLMs have demonstrated enormous potential, but also where there are significant concerns. ChatGPT, for instance, has proved capable of passing medical exams (e.g., the U.S. Medical Licensing Exam), and there are already several specialized LLMs tailored for medical applications, including BERT for Biomedical Text Mining (BioBERT), ClinicalBERT, GatorTron, Med-PALM, and many more. At the same time, the critical nature of healthcare requires careful consideration of issues related to misinformation, bias, potential breaches of patient privacy, and others.

In this section, we will look at how LLMs can enhance the work and educational experience of healthcare professionals and medical students. Additionally, we explore the positive impact LLMs can have on the lives of patients. Following this, we will outline some of the challenges associated with the use of LLMs in healthcare and propose some solutions to mitigate these concerns.

LLMs can alleviate the burdens faced by healthcare professionals in various time-consuming and repetitive tasks. For example, LLMs can drastically reduce the time and effort required for creating summaries of medical interviews with patients, composing standardized reports and discharge summaries, and even translating documents into other languages.

---

[8] https://bera-journals.onlinelibrary.wiley.com/doi/full/10.1111/bjet.13370#:~:text=Large%20language%20models%20have%20been,question%20generation%20and%20essay%20scoring

[9] www.sciencedirect.com/science/article/abs/pii/S1041608023000195

LLMs can also provide efficient access to medical research, delivering summaries and responses tailored to individual patients. Furthermore, they can also act as a basis for conversational assistants, capable of examining and explaining medical images and other test results, assisting in diagnosis, and supporting clinical decision-making.

With the integration of frameworks like Retrieval Augmented Generation (RAG), which we will describe later in Chapters 4 and 7, LLMs can analyze relevant documents such as electronic health records, radiology reports, and other medical documentation to predict diagnoses, recommend treatment options, and offer clinical decision support to healthcare professionals.

Medical students can benefit from the use of LLMs in several ways. In addition to providing summaries of relevant research papers, LLMs can enable students to create learning simulations in which the students can engage in realistic interactions with simulated patients and develop skills for taking patient histories, assessing diagnosis, and formulating treatment plans.

For patients, in a healthcare environment facing increasing resource constraints, Conversational AI-powered virtual nurses can serve as complementary tools for patients, offering preliminary guidance and triage until a healthcare professional becomes available.

Given the paramount importance of patient safety in healthcare, there are several ethical issues to consider. *Fairness* is concerned with the data used to train the LLM and the need to prevent bias and ensure accurate predictions. However, obtaining suitable datasets for LLM training poses a challenge due to data privacy concerns and the general reluctance of individuals to share their personal data for LLM training.

In healthcare applications, the explainability of LLM predictions and decisions is crucial for maintaining *transparency*. Robust regulatory frameworks must be established to oversee the usage of LLMs in healthcare applications, ensuring *accountability* and adherence to ethical principles.

Finally, given the relative novelty of LLMs in healthcare, there is a need for comprehensive training and education in programs for healthcare professionals, emphasizing the capabilities, limitations, and potential risks associated with LLM technology.

There is an extensive literature on LLMs in healthcare. This article, "Large Language Models in Health Care: Development, Applications, and Challenges"[10] provides a readable overview, with a particular emphasis on the challenges involved. See also "Embracing Large Language Models for Medical Applications: Opportunities and Challenges."[11]

[10] https://onlinelibrary.wiley.com/doi/10.1002/hcs2.61
[11] www.ncbi.nlm.nih.gov/pmc/articles/PMC10292051/

Social Companions

Originally the term *chatbot* was used to characterize a conversational system that engaged primarily in casual chit-chat with users for the purposes of entertainment in contrast to task-oriented systems with more "serious" purposes such as responding to queries or helping users complete a task. Nowadays the term has broadened to encompass all types of conversational systems.

One particularly compelling application of modern chatbots is to act as virtual social companions for individuals like older adults or those living alone who may struggle with the challenges of depression and related disorders. In this context, a social companion can play a crucial role in enhancing the overall well-being of these individuals, providing assistance with activities of daily living, identifying potential risks, and offering both practical support and companionship. The deployment of chatbots as social companions can contribute to the automation of the previously highlighted issue of the scarcity of public health and care workers in many contemporary communities. In the following paragraphs, we provide brief descriptions of two instances where chatbots are actively employed in social care applications.

CLOVA CareCall Service

The CLOVA CareCall Service, developed by NAVER, South Korea's leading platform company, was deployed initially to monitor the health symptoms of users during the COVID pandemic. The service has since been re-purposed to provide support to elderly individuals with simulated LLM-powered conversations on a range of topics based on a large-scale conversational dataset. Using LLMs has enabled the system to provide open-domain conversations on a range of topics, including general health of users as well as their hobbies and interests.

The service has been evaluated through focus group observations and interviews and has generally received positive support. However, on occasions, it was found that users expected the system to be able to support social services that were beyond the system's scope. Users also felt that the system was impersonal as it was unable to follow up on past conversations due to the lack of long-term memory in LLM-powered chatbots. Attempts to address these problems include the use of in-context learning in which prompts are augmented with additional information. We describe in-context learning and prompt augmentation in more detail in Chapters 4 and 6.

This posting from the European AI Alliance provides a brief description of the CareCall system.[12] You can find more detail in this paper from the CHI '23 conference.[13]

---

[12] https://futurium.ec.europa.eu/en/european-ai-alliance/best-practices/ai-people-clova-carecall-service-naver

[13] https://dl.acm.org/doi/fullHtml/10.1145/3544548.3581503

The e-VITA project

e-VITA, a three-year collaborative European and Japanese research project, has developed a virtual coach aimed at empowering older adults to effectively manage their health, well-being, and daily routines.[14] The virtual coach provides personalized support and motivation across a range of critical areas, including cognition, physical activity, mobility, mood enhancement, social interaction, leisure, and spiritual well-being.

During the initial phase of the project, dialogues with the virtual coach were developed using Rasa's open source Conversational AI platform. Developing these dialogues involved creating training examples to enable the classification of intents based on user inputs across the various domains covered by the coach; specifying the system's responses; and designing conversation flows (known in Rasa as stories).

Following the launch of ChatGPT in November 2022, there was a growing demand from users to be able to access the latest Conversational AI technologies. As a result, LLM-powered dialogues based on the OpenAI API were integrated into the system. These dialogues were employed in two different ways:

1. Fallback intent: When the system was unable to classify a user's utterance using its predefined intent classification capabilities, the LLM was invoked to recognize the intent and enable the dialogue to continue. This mechanism ensured that users could seamlessly engage with the system even if their input did not fit neatly into predefined categories.
2. Casual dialogue: Users could explicitly request to continue a dialogue with the LLM when a predetermined story had reached a conclusion. This resulted in a more open-ended conversation that did not need to follow the constraints of a predefined script. This approach allowed users to engage in a more spontaneous way with the virtual coach, thus providing a more natural and personalized conversational experience.

The use of the LLM-based approach in the project was subject to certain constraints, particularly for the European Union (EU) on account of regulations regarding the use of AI systems (for more detail see Chapter 9). For this reason, the scope of topics for the LLM-powered conversations was restricted to information contained in documents provided by the project's Content Group. These documents were fed into the API to provide a contextual basis

[14] www.e-vita.coach/