"Insightful & comprehensive–if you run a service & support operation, put this book on your essential reading list right now!"

—**PHIL WOLFENDEN,** Cisco, VP, Customer Experience

The AI REVOLUTION in CUSTOMER SERVICE and SUPPORT

A Practical Guide to Impactful Deployment of AI to Best Serve Your Customers

ROSS SMITH | **MAYTE CUBINO** | **EMILY MCKEON**

FREE SAMPLE CHAPTER |

The **AI REVOLUTION** in CUSTOMER SERVICE and SUPPORT

A Practical Guide to Impactful Deployment of AI to Best Serve Your Customers

ROSS SMITH | MAYTE CUBINO | EMILY MCKEON

The AI Revolution in Customer Service and Support:
A Practical Guide to Impactful Deployment of AI to Best Serve Your Customers
Ross Smith, Mayte Cubino Gonzalez, and Emily McKeon

Pearson
www.informit.com

To report errors, please send a note to errata@informIT.com

Executive Editor: Loretta Yates
Associate Editor: Shourav Bose
Development Editor: Rick Kughen
Senior Production Editor: Tracey Croom
Tech Editor: Dr. Xin Deng
Copy Editor: Rick Kughen
Compositor: Danielle Foster
Proofreader: Dan Foster
Indexer: James Minkin
Cover Design: Chuti Prasertsith
Cover Illustration: Prostock-studio/Shutterstock
Interior Design: Danielle Foster
Illustrations: Vived Graphics
Figure Credits: Figure 5.1: Created from DALL-E
Figure 5.2: Created from Midjourney

ISBN-13: 978-0-13-828650-7
ISBN-10: 0-13-828650-7

Library of Congress Control Number: 2024939894

$PrintCode

Dedication

To the next generation—Sela, Scarlett, and Paz, and to Maddy, Emma, Clara, and Roo—who will all grow alongside the AI. Always celebrate your humanity! Sincere thanks to my friends Mayte and Emily and everyone who made this work possible.

—Ross

To Sofia, Tiago, and Laura: your curiosity and strength are my endless inspiration. Always dream big! To my family and friends for their support and encouragement. And to Ross and Emily for revolutionizing my life with the most amazing journey ever.

—Mayte

With all my love to my family—I am forever grateful for your belief in me. Your unwavering support, patience, and love have been my greatest motivation. To Ross and Mayte, you have infinite creativity and knowledge, and I'm thankful for your trust and energy as we ventured forth on this project.

—Emily

This page intentionally left blank

Contents at a Glance

Contents

PART II: BUILDING AI MODELS USING PROPRIETARY CONTENT: THE 6DS FRAMEWORK

Acknowledgments

This book is a testament to the incredible efforts of many and the power of collaboration. We extend our deepest gratitude to all who shared their insights, expertise, and enthusiasm with us.

With the speed of change we're experiencing in the era of AI, we knew our window of opportunity to tell our story was short. We worked hard to ensure this book provided the practical steps to build and deploy AI models for real-world scenarios. While there is so much upside to integrating AI in customer service and support, many other industries can also learn from these pages.

To our contributors, Dr. Xin Deng, Phaedra Boinodiris, Mostaq Shakil Ahmed, Michael Fitzgerald, and Jason Weum, your willingness to share your knowledge has greatly enhanced this project beyond measure. It would not be the same without your input and wisdom.

To Executive Editor Loretta Yates and Associate Editor Shorav Bose, your guidance has turned our vision into reality with your tireless work behind the scenes bringing this book to life. And to our Development Editor, Rick Kughen, your keen eyes and thoughtful suggestions have sharpened our message. A very special thanks to Dr. Xin Deng, our technical reviewer, who kept us honest and didn't allow us to "hallucinate"!

We would also like to thank J.B. Wood for writing the foreword and his inspirational message. He not only believes in the power of AI and what it can do for the customer experience but also in the value of collaboration and how innovation is at the heart of progress in our industry.

To our family and friends, this book came together very quickly, and it would not have been possible without your love and support. This has been a memorable journey that we couldn't have taken on without you. You inspire us each and every day.

To our readers, we hope you find this book informative and helpful as you move forward with employing AI models in your own organizations. While technology changes rapidly, the foundations of great leadership remain the same. We encourage you to stay abreast of the exciting AI innovation and all it has to offer!

We welcome your feedback and comments and look forward to hearing from you! Check out what we've been up to: *https://airevolutionbook.com/*.

About the Authors

Ross Smith, FRSA is a Fellow of the Royal Society of the Arts and a worldwide support leader at Microsoft. Ross co-authored *The Practical Guide to Defect Prevention* and holds seven patents. He is a PhD scholar at University College Dublin, focused on AI, automation, worker displacement, and the future of work. He co-founded the Future World Alliance, a nonprofit committed to responsible AI for the next generation. He can be found online on LinkedIn at *https://www.linkedin.com/in/rosss42/*.

Mayte Cubino Gonzalez is an EMEA director in the Modern Work Support Engineering organization at Microsoft and serves as the site lead and board member of Microsoft Portugal. With 20 years of experience in customer service and support roles, Mayte is also an AI enthusiast and patent holder. She was recognized in 2016 with the European Disability Champion award for her work in raising awareness about hidden disabilities and workplace adjustments. She can be found on LinkedIn at *https://www.linkedin.com/in/mayte-cubino-gonzalez/*.

Emily McKeon is a communication director at Microsoft focused on global strategic business and executive communications designed to strengthen employee engagement and drive value for the Customer Service and Support business. She has vast communication experience and a strong depth of knowledge in customer support, global diversity and inclusion, and employee engagement. She can be found on LinkedIn at *https://www.linkedin.com/in/emily-mckeon-7ab9ba3/*.

About the Contributors

Dr. Xin Deng is currently a machine learning scientist at Microsoft. Dr. Deng played an important role in this work's technical accuracy and production. Her background in AI and machine learning has helped shape this book's perspective on the current state and the future of AI, machine learning, and data science. She is actively contributing her expertise in M365 Outlook Copilot utilizing GPT models. She can be found on LinkedIn at *https://www.linkedin.com/in/xin-deng-00a25a60/*.

Phaedra Boinodiris currently leads IBM Consulting's Trustworthy AI Practice, and she is the AI ethics board focal for all of consulting, serves on the leadership team of IBM's Academy of Technology, and leads IBM's Trustworthy AI Center of Excellence. She is a Future World Alliance co-founder and the co-author of *AI for the Rest of Us*. She can be found on LinkedIn at *https://www.linkedin.com/in/phaedra/*.

Mostaq Shakil Ahmed leads a vast global team spread across 48 countries at Microsoft. He serves as the global general manager of Modern Work Support Engineering including Microsoft Copilot support. Shakil has authored a number of papers on software globalization and has also contributed as a part-time lecturer on the same subject at the University of Washington. He can be found on LinkedIn at *https://www.linkedin.com/in/shakilam/*.

Michael Fitzgerald is a finance manager at Microsoft. He holds an MBA from the Tuck School of Business at Dartmouth and a Master of Arts in Law and Diplomacy from the Fletcher School at Tufts University. He can be found on LinkedIn at *https://www.linkedin.com/in/mfitzgerald514/*.

Jason Weum is the director of supportability at Microsoft for Teams, SharePoint, OneDrive, Viva, and Office. He focuses on crafting an unparalleled AI-first customer support experience by minimizing support queries with AI, self-help, proactive diagnostics, and product improvements. He can be found on LinkedIn at *https://www.linkedin.com/in/jasonweum/*.

Foreword

The journey of technology in customer service is a story of evolution, from the early days of switchboards and mail correspondence to the digital age's omnichannel support platforms. This voyage reflects a continuous pursuit of efficiency, personalization, and satisfaction in supporting customers and driving product loyalty. In an era when artificial intelligence (AI) is redefining the parameters of customer interaction and satisfaction, our mission is to ensure that you, the dedicated professionals at the heart of customer service, are fully equipped with the information and resources you will need to navigate and lead in this transformative landscape.

I've witnessed firsthand the remarkable strides we have made through cross-industry collaboration over the years. By embodying a wealth of insights from pioneers who have led the charge in integrating AI into customer service frameworks, we will serve our customers better if we build on these successes and continue advancing our collective thinking and innovation. We all play an instrumental role in fostering partnerships, bringing together the brightest minds from technology, customer service, and beyond to share insights, challenges, and successes.

In customer service history, there has never been a more important time for us to collaborate across organizational boundaries. Through shared learning environments, we can explore the potential of emerging technologies, ensuring that the customer service industry keeps pace with technological change and leads it. The goal is to empower you with the knowledge, resources, tools, and intelligence you need to leverage AI to enhance your customer service capabilities while maintaining the human touch that has always defined our industry.

We're at the forefront of navigating this evolution and leading the industry through each phase of technological advancement in the ever-changing customer service and support world. We see *The AI Revolution in Customer Service and Support* as a guide for customer service and support professionals who sit directly at the intersection of customer service excellence and cutting-edge AI technology.

We recognize that the pace of technological advancement can be exhilarating, inspiring, intimidating, and daunting. This book is designed as a compass for those navigating the new terrain, bridging the gap between traditional customer service expertise and AI's technical details and nuances.

As customer service and support professionals, our role is pivotal in shaping the experiences that define brand loyalty and customer satisfaction for your organizations. This industry, rich in tradition and human connection, is now at the cusp of a transformative era brought about by AI. As AI technologies redefine the boundaries of what's possible within customer service and beyond, sharing insights, challenges, and solutions across different sectors has become invaluable. This knowledge-sharing accelerates the pace of AI integration into customer service practices and ensures a broader understanding of its impact across various customer touchpoints.

While the advent of AI in customer service and support brings us transformative innovation and endless opportunities, this revolution is not merely about integrating innovative technology into our industry. It's a reimagining of what customer service is and can be. AI has the means to become a powerful tool—more importantly, an important partner—in creating deeper, more meaningful connections with customers. Its potential to analyze vast datasets in real time allows for a level of service customization previously unimaginable, offering personalized solutions, predictive support, and seamless interactions across multiple support channels.

Capitalizing on the collaboration between technologists, customer service experts, and business leaders in crafting innovative AI solutions and becoming deeply attuned to the human aspects of customer interaction will set a new standard for excellence in the industry.

This book is a testament to a collaborative spirit and is designed to offer a comprehensive understanding of AI's impact on the customer success field. By encapsulating a wide array of lessons learned and perspectives, the book offers a unique vantage point on how AI can and will be harnessed across the service industry spectrum. It provides a comprehensive overview of cutting-edge applications, ethical considerations, and the future trajectory of AI technologies, making it an indispensable resource for support professionals seeking to navigate the complexities of this new era.

I want to encourage us all to build an ecosystem of collaboration where challenges are tackled collectively and successes are celebrated as milestones for the entire industry. We are the experts in our field, and by leveraging best practices as we enter this new era of AI, we will create unprecedented customer experiences. We will face and overcome shared challenges in implementing AI, such as ethical considerations, data privacy, and workforce adaptation. AI adoption is a strategic opportunity and only through strong collaboration and maintaining an open dialogue will we continually learn and ultimately enhance our collective success in

this rapidly evolving landscape. I'd like to invite leaders, innovators, and frontline professionals to join forces in leveraging AI to enhance customer experiences, streamline operations, and forge lasting relationships with customers and each other.

AI will not just enhance customer service; it will help us all actively shape it. By working together, we can transform our industry to be more efficient, responsive, empathetic, and connected to the needs of every customer.

Welcome to the revolution: *The AI Revolution in Customer Service and Support*! I'm excited to help launch this new era of collaboration as we build a bright future with the help of AI technology.

J.B. Wood
Technology Services Industry Association (TSIA)
President and CEO

Introduction

Artificial intelligence has made amazing advances in the last year, and customer service and support is one of the most important areas where this new technology is having an immediate impact. While the technology is not yet in a place where it will fully replace agents and support engineers, it can do wonders to dramatically improve customer experience while also contributing to optimizing productivity in various ways. This book will help you understand how and where to incorporate AI technology, such as large language models (LLMs), machine learning, predictive analytics, augmented reality, and others, into the customer experience flow.

Please note that a percentage of proceeds from the sale of this book will be donated to the nonprofit Future World Alliance, dedicated to curating K-12 education in AI ethics. See *https://futureworldalliance.org*.

Who Is This Book For?

The AI Revolution in Customer Service and Support is a practical guide for adopting and deploying generative AI models within a customer service and support organization. It is written for technical and non-technical customer service professionals and customer service and support professionals who have been thrust into the technical limelight and need to learn quickly about deploying and leveraging AI.

This book is for customer service professionals who want to learn more about deploying and leveraging AI in their organizations but are unsure where to get started. Their leaders look to them as customer professionals, but the new world is highly technical. Reading this book will help them leverage their customer service experience to navigate this new world.

Errata, Updates & Book Support

We've made every effort to ensure the accuracy of this book and its companion content. The world of AI is moving quickly, with new advances every week. You can access updates to this book—in the form of a list of submitted errata and their related corrections—at:

informit.com/airevcs/errata

If you discover an error that is not already listed, please submit it to us at the same page.

For additional book support and information, please visit

InformIT.com/Support and *http://airevolutionbook.com*.

2

Overview of Generative AI and Data Science Machine Learning

You may grow old and trembling in your anatomies, you may lie awake at night listening to the disorder of your veins, you may miss your only love, you may see the world about you devastated by evil lunatics, or know your honour trampled in—to learn. Learn why the world wags and what wags it. That is the only thing which the mind can never exhaust, never alienate, never be tortured by, never fear or distrust, and never dream of regretting. Learning is the only thing for you. Look what a lot of things there are to learn.

—T.H. White

Welcome to the most technical chapter of our journey into the AI Revolution. Unlike the other chapters in this book, this chapter is a deep dive into the history and details of the data science technology that powers today's AI Revolution in customer service and support. In this chapter, we explore concepts such as generative AI, machine learning, various language models, reinforcement machine learning, and prompt engineering—among other topics—in detail. This chapter goes into the technical details of the foundation upon which the AI that is changing customer support is built.

However—and this is important—you don't have to read this chapter to get the most out of the following chapters. You may have technical colleagues who know or will learn about the topics in this chapter, or you may have partners or vendors to help you build and deploy AI solutions. You don't have to know this level of detail to move forward in leading your organization through the deployment of AI. We recognize that not everyone will feel comfortable navigating this more complex territory, and that's okay—it won't matter for the rest of the book.

However, we felt that we would be remiss if we didn't cover this technology in some detail as a foundational component of this book. This chapter will not supplant the myriad of courses, papers, books, theories, algorithms, and other details in this fast-moving technology. This chapter is worth a skim to understand the underlying developments driving the AI Revolution and what to explore in more detail if you are interested.

If you're a customer service and support professional eager to leverage AI in your organization but less versed in technical jargon—please know this chapter is not a prerequisite for the valuable insights and steps outlined in the other chapters in this book. It's here to provide a deeper understanding for those who wish to explore further. Skipping it won't diminish your ability to apply AI effectively within your role.

The rest of the book is designed with you in mind, focusing on practical applications, deployment strategies, and real-world scenarios that don't require a deep technical background to understand. However, if you want to be the leader who queries your team or a vendor on their understanding and application of reinforcement learning from human feedback (RLHF), you might want to investigate to understand more.

Whether you decide to brave this chapter or flip past it and go directly to the next, rest assured that this chapter is not required reading to enable a successful AI deployment! Happy reading, wherever you land next.

Unveiling the Realm of AI Technologies: A Glimpse into the Augmented Future

In 1947, Alan Turing gave a public lecture about computer intelligence—the original concept of artificial intelligence. In 1950, he proposed the Turing test, a criterion for machine intelligence based on natural language conversation. In 1956, John McCarthy coined the term artificial intelligence and organized the first conference on the topic at Dartmouth College.

In the 1970s and 1980s, AI research focused on developing rule-based systems that could encode human knowledge and reasoning in specific domains, such as medicine, engineering, and finance. These systems, known as expert systems, could perform tasks requiring human expertise, such as diagnosis, planning, and decision-making.

In the late 1980s and 1990s, the development of AI underwent a paradigm shift from relying on predefined rules to learning from data, enabling machines to achieve higher levels of intelligence and performance. Machine learning (ML) is a subfield of AI that enables machines to learn from data and improve their performance without explicit programming. Machine learning techniques include supervised, unsupervised, and reinforcement learning, which can be applied to various problems, such as classification, clustering, regression, and control.

In the 2000s and 2010s, AI experienced a major breakthrough with the advent of deep learning, a subset of machine learning that uses multiple layers of artificial neural networks to learn from large amounts of data. 2015 was a big year in AI history; a five-game Go match was hosted between the European champion Fan Hui and AlphaGo, a computer Go program developed by DeepMind. AlphaGo won all five games. Deep learning has enabled significant advances in various domains, such as computer vision, natural language processing (NLP), speech recognition, and robotics. Some notable deep learning models include convolutional neural networks (CNNs), recurrent neural networks (RNNs), and transformer models, such as BERT (Bidirectional Encoder Representations from Transformers)[1] and GPT (Generative Pre-trained Transformer).[2]

In 2017, Google developed the transformer model and published a paper, "Attention Is All You Need."[3] Transformers opened a new chapter for the natural language processing field. Since then, companies and researchers worldwide have built large-scale language models based on the transformer architecture.

In the 2020s and beyond, AI is entering a new frontier of generative technologies, which aim to create novel and realistic content, such as images, texts, sounds, and videos. Generative technologies use deep learning models, such as generative adversarial networks (GANs), variational autoencoders (VAEs), and large language models (LLMs), to generate content that is indistinguishable from human-produced content. Generative technologies have various applications, such as art, entertainment, education, and communication.

Generative AI and Language Models

One of the most exciting and challenging areas of generative technologies is natural language generation (NLG), which generates natural language text from a given input, such as an image, a keyword, or a prompt. NLG has many applications, including summarization, translation, dialogue, storytelling, and content creation. However, NLG also poses many technical and ethical challenges, such as ensuring the generated texts' quality, diversity, coherence, and fairness.

A key component of NLG is the language model (LM), a probabilistic model that assigns a probability to a sequence of words or tokens. LMs can generate new texts by sampling tokens according to their probabilities or evaluating the likelihood of existing texts. LMs can be trained on large corpora of text data, such as Wikipedia, books, news articles, or social media posts, using deep learning techniques, such as RNNs or transformers.

Because statistical language models (SLMs) cannot capture long-term dependencies or semantic relations in natural language, this underscores the importance of building these models with responsibility and ethics in mind, as discussed throughout this book.

The development of LMs has gone through several stages, reflecting the advances in computational power, data availability, and algorithmic innovation. There are four main development stages of LMs: statistical language models, neural language models, pre-trained language models, and large language models (LLMs).[4]

- **Statistical language models** are based on statistical learning models. The idea is to build models based on the n-gram assumption, which states that the probability of a word only depends on the previous n-1 words and not on the rest of the sentence or the document. An n-gram is a sequence of n-words, such as "the cat" or "a big house." For instance, predicting when the probability of the word "model" would come after the words "large language" is illustrated as P(model|large

language). Statistical language models estimate the probabilities of n-grams from a corpus of text data using techniques such as counting, smoothing, or interpolation.

- **Counting:** Counting is simply the frequency of the n-gram in the corpus divided by the total number of n-grams.
- **Smoothing:** Smoothing adds some small values to the counts to avoid zero probabilities for unseen n-grams.
- **Interpolation:** Interpolation combines the probabilities of different n-grams, such as unigrams, bigrams, and trigrams, to balance the trade-off between specificity and generality.

Smoothing and interpolation are often adopted to mitigate the data sparsity problem. Statistical language models are simple and efficient but have limited expressive power and cannot capture long-term dependencies or semantic relations in natural language. For example, statistical language models cannot distinguish between the meanings of "bank" in "I went to the bank" and "The bank was closed" or the contexts of "She saw a bear" and "She saw a bare." Statistical language models cannot handle the ambiguity of words with multiple meanings, such as "bat," "right," or the influence of words that are far apart in the sequence, such as "The man who wore a hat" and "The hat was red."

- **Neural language models (NLMs)** use neural networks, such as recurrent neural networks (RNNs) or convolutional neural networks (CNNs), to learn distributed representations of words and sequences and to model the conditional probability of the next word given the previous words. A groundbreaking work in the field of neural language modeling was the paper "A neural probabilistic language model,"[5] which presented the idea of representing words as continuous vectors in a high-dimensional space and learning the probability of the next word based on the sum of the context word vectors. Many studies have opened a new chapter for using language models for representation learning, playing an influential role in the NLP field. For example, word2vec[6] utilizes two methods to learn word embedding:
  - **Using context to predict a target word through CBOW (Continuous Bags of Words):** CBOW is like a guessing game where the AI tries to predict a word based on the words around it. Imagine a sentence with a missing word; CBOW looks around the neighboring words to guess the missing one, helping the AI better understand the language.
  - **Using a word to predict a target context through Skip-Gram:** Skip-Gram is like a word puzzle where the challenge is to find the

related words. Given a specific word, Skip-Gram tries to predict the surrounding words, helping the AI grasp the context and relationships between words in a sentence.[7]

NLMs can overcome some of the limitations of statistical language models, such as capturing longer contexts and learning richer features, but they also have drawbacks, such as requiring more computation and data and suffering from the vanishing or exploding gradient problem.

- **Pre-trained language models (PLMs)** are a type of model that uses transfer learning. PLMs are first trained on a large collection of text data that doesn't have specific labels (this is the pre-training phase). After this, PLMs are fine-tuned on a particular task or domain with more specific data (this is the fine-tuning phase). This process allows PLMs to be highly effective at understanding and generating human language in various applications. As mentioned earlier, RNNs have some drawbacks and limitations, such as the difficulty of learning long-term dependencies, the gradient vanishing or exploding problem, and the sequential nature of computation, which prevents parallelization and reduces efficiency. To address these issues, long short-term memory (LSTM) models were proposed by researchers as one of the most popular and effective variants of RNNs. LSTM models have a special structure consisting of three gates and a cell state, which can regulate the input, output, and forget operations of the recurrent unit. LSTM models can generate natural language texts by updating the cell state and the hidden state at each time step, based on the current input and the previous states, and then producing the next word or token from the hidden state. As one of the first models that demonstrated the effectiveness of PLMs on large-scale unlabeled text data and then transferring the learned knowledge to downstream tasks or domains, embeddings from language models (ELMo) was built as an LSTM-based model that can generate contextualized word embeddings, which are vector representations of words that capture their meanings and usage in different contexts. Unlike traditional word embeddings, such as word2vec or GloVe, which assigns a fixed vector to each word regardless of its context, ELMo can dynamically compute word embeddings based on the entire input sentence or document, using a bidirectional LSTM that encodes both the left and the right contexts of each word. PLMs also adopt the transformer architecture, which is an attention-based neural network that can learn long-range dependencies and parallelize computation. Some of the most influential PLMs are BERT, developed by Google, and GPT, developed by OpenAI. Based on pre-trained context-aware word representations, these models have

shown remarkable effectiveness and versatility as general-purpose semantic features that can significantly improve the performance and efficiency of various NLP tasks.

- **Large language models (LLMs)** are the latest and most advanced stage of LMs, which aim to build very large-scale and powerful LMs that can generate natural language texts across multiple domains and tasks, given minimal or no supervision. LLMs rely on massive amounts of computation and data and use sophisticated optimization and regularization techniques, such as self-attention, dropout, or layer normalization, to train billions or trillions of parameters. Some of the most prominent examples of LLMs are GPT-3, GPT-3.5 (Instruct GPT) GPT-4 and GPT-4o, developed by OpenAI.
    - **GPT-3:** GPT-3 is a transformer-based model with 175 billion parameters and can generate coherent and diverse texts on various topics and domains, given a few words or sentences as input.
    - **GPT-3.5:** In 2022, OpenAI deployed GPT-3.5, which performs more significantly in following instructions, making up facts less often, and generating less toxic output. They used prompts submitted by the customers through Playground and hired human annotators to provide demonstrations of the desired model behavior and rank outputs from the models. GPT-3.5 is fine-tuned based on this data from GPT-3.
    - **GPT-4:** In 2023, GPT-4, a 1.8T-parameter model with 16 Mixture of Experts (MoE), was announced by OpenAI to improve the security of the model and enable multimodal capability. However, LLMs also have limitations and risks, such as producing inaccurate, biased, or harmful content or violating the data sources' privacy or intellectual property rights.
    - **GPT-4o:** Launched in 2024, GPT-4o ("o" for "omni") is a step towards a much more natural human-computer interaction—it accepts any combination of text, audio, image, and video as input and generates any combination of text, audio, and image as output. It can respond to audio inputs in as little as 232 milliseconds, with an average of 320 milliseconds, which is similar to human response time in a conversation. It matches GPT-4 Turbo performance on text in English and code, with significant improvement on text in non-English languages, while also being much faster and 50% cheaper in the API. GPT-4o is especially better at vision and audio understanding compared to existing models.[8]

The emergence and advancement of LLMs significantly impact the AI community and society at large, as they open up new possibilities and

challenges for natural language understanding and generation. LLMs can be seen as a form of generative technologies that can create novel and valuable outputs from minimal or no inputs, such as images, music, art, or texts. They can foster interdisciplinary collaboration and innovation by bringing together researchers and practitioners from different fields and domains and creating new paradigms and methods for natural language understanding and generation.

Despite the exciting progress and impact of LLMs and generative AI, many mysterious and unpredictable perspectives remain. There are some risks associated with LLMs. They can amplify existing biases and harms, such as perpetuating stereotypes, discrimination, misinformation, or manipulation, by learning from unfiltered and unrepresentative data sources, or by being misused or abused by malicious actors. They can also pose ethical and legal dilemmas, such as violating privacy, intellectual property, or human dignity, by exposing sensitive or personal information, infringing on copyrights or trademarks, or generating deceptive or harmful content. Moreover, they can challenge existing norms and values, such as accountability, transparency, or trust, by obscuring natural language generation's sources, processes, and outcomes or by creating conflicts of interest, responsibility, or authority.

LLMs and Their Applications

As discussed earlier, LLMs are trained on billions or trillions of words, sentences, paragraphs, or documents collected from various online sources, such as websites, blogs, social media, news articles, books, or academic papers, using a technique called self-attention, which enables them to learn the contextual and semantic relationships between different units of language. LLMs can then use the learned representations to perform a wide range of natural language tasks, such as classification, summarization, translation, question answering, sentiment analysis, or dialogue generation. They do this by fine-tuning specific datasets or domains or by applying a method called *prompting*, which consists of providing the model with a few words or sentences as input or output examples and letting it infer the rest.

LLMs have demonstrated remarkable capabilities and achievements in natural language understanding and generation, surpassing previous state-of-the-art models and even human performance in some tasks. Some of the most notable and influential LLMs include GPTs, BERT, XLNet, T5, and DALL-E, which have been developed and released by leading

research labs and companies, such as OpenAI, Google, Facebook, and Microsoft. LLMs have also enabled and inspired the creation and innovation of various applications and products, such as chatbots, assistants, recommender systems, content generators, summarizers, translators, analyzers, or synthesizers, which have been deployed and adopted by various industries and sectors, such as education, health, business, media, entertainment, or art, among others. LLMs have thus revolutionized and democratized the field of natural language processing and generation, opening up new possibilities and opportunities for research, development, and impact.

LLMs and Customer Support

One possible application domain of LLMs is customer support, which involves providing assistance and guidance to customers or users of a product or service through various channels, such as phone, email, chat, or social media. Customer support is an essential and integral part of any business or organization, as it affects customer satisfaction, retention, loyalty, advocacy, brand reputation, revenue, and growth. However, customer support can also be challenging and costly, as it requires hiring, training, and managing a large number of human agents, who have to deal with high volumes of queries, requests, complaints, or feedback, often repetitive, mundane, or complex while maintaining a high quality of service, professionalism, and empathy.

LLMs can offer a solution to some of these challenges by augmenting or automating some aspects of customer support, such as answering frequently asked questions, providing information or instructions, resolving issues or problems, collecting feedback or ratings, generating reports or summaries, or escalating cases or tickets, and so on. LLMs can leverage their natural language abilities, such as understanding, reasoning, generating, or adapting, to provide personalized, contextualized, and relevant responses or actions based on the customer's input, profile, history, or preferences, as well as the product or service specifications, policies, or updates. LLMs can also learn from the data and feedback collected from the interactions and improve their performance and accuracy over time, using techniques such as reinforcement learning, active learning, or transfer learning. Furthermore, LLMs can enhance the customer experience and engagement by adding elements of conversation, personality, emotion, or humor, to the interactions, depending on the tone, mood, or style of the customer and the situation.

Development, Optimization, Localization, and Personalization Based on LLMs

The rapid growth of the tech field has seen significant disruptions when the right combination of technology and user experiences come together. Generative AI–infused experiences bring a great opportunity for intelligent product development. Besides fostering AI's capabilities for business and real products, we must also ensure localization and personalization and operate with a clear customer-centric intent and goal.

There are multiple strategies to employ regarding integrating the generative AI large models into productions with further optimization, localization, and personalization.

Large deep neural networks have achieved remarkable success with great performance in research and real-world products with large-scale data. However, it is still a great challenge to deploy these large-scale AI models to real production systems, especially mobile devices and embedded systems, with the considerations of cost, computational resources, and memory capacity. The main purpose of teacher–student distillation (see **Figure 2.1**) is to train a small student model that simulates the large teacher model with equivalent or superior performance.[9] Another advantage of teacher–student distillation is that when we do not have enough labeled data, the teacher model can help generate a "pseudo-label" when training the student model. Pseudo-labels are then used to train the smaller student model, helping it learn and perform tasks as if it had been trained on a fully labeled dataset. Put more simply, imagine you're playing a video game, and there's a really tough level that you can't beat. So, you call in an expert friend.

The three main components of the teacher–student distillation framework include knowledge, distillation algorithm, and teacher–student architecture.

Figure 2.1 illustrates two AI models:

- **Teacher model:** The teacher model is like an expert friend. It's very smart but also big and needs a lot of power to run.
- **Student model:** Like you, the student model is eager to learn. It doesn't have as much power.

The goal is to make the student AI learn from the teacher AI without needing as much power. The process is such that the teacher model, trained with huge volumes of data, helps the student model by guiding it or giving it tips—in NLP; this is called "knowledge transfer." Sometimes, the teacher doesn't have all the answers (or labeled data), so the teacher

makes up some good guesses (pseudo-labels) for the student to practice with. It's like getting hints for your video game level. This way, the student learns a lot and gets really good at the game, so it can almost match the teacher's skill level.

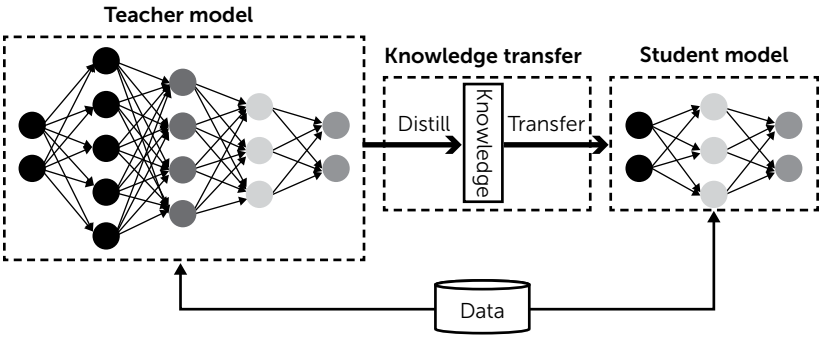

FIGURE 2.1 The general teacher–student distillation framework

This framework can be useful for any large-scale prediction or generative AI model, although it was originally introduced for an image classification model. With the rapid development of generative AI, many of the current large-scale models are significantly effective in generalization. However, many factors must be considered for real production, including cost, scalability, resource consumption during inference, adopting the existing model into some specific scenarios, and so on. Developing an AI-assistant writing tool by leveraging GPT to help users write articles or posts more casually and recognize contextual information is an example of adopting the existing GPT model to the specific scenario of an AI-assistant writing tool. Directly running GPT models is very challenging, considering cost and scalability. The teacher–student distillation framework helps serve lighter-weight models in production and localizes the model with task-specific data when leveraging the existing large-scale model.

Reinforcement Learning from Human/AI Feedback

As mentioned earlier, Instruct GPT/ GPT-3.5 was developed by OpenAI to have a better human alignment and address some issues like factuality, harm, etc. They collected prompts submitted by customers through Playground and ranked outputs from the models responding to the human-annotated instructions. InstructGPT/ GPT-3.5 is fine-tuned based on this data from GPT-3. The success of GPT-3.5 over GPT-3 is mainly due to the reinforcement learning from human feedback (RLHF) technique, which is adopted to fine-tune GPT-3 using human labels as a reward signal (see **Figure 2.2**).[10]

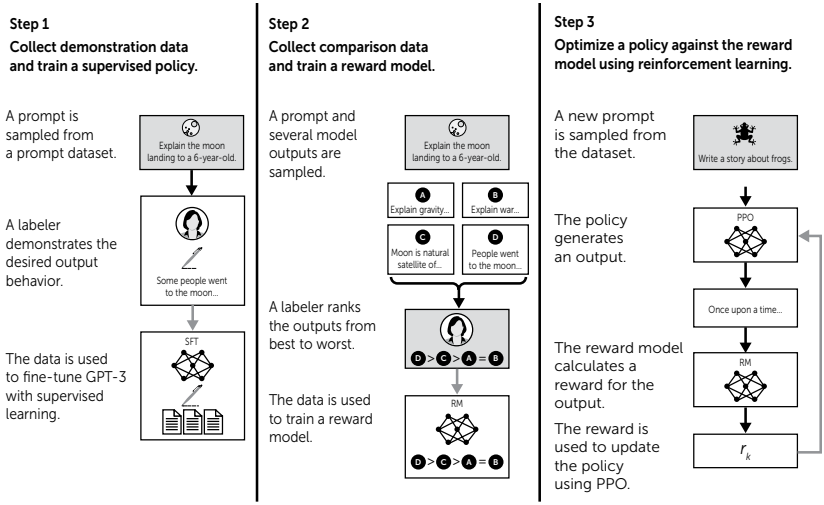

FIGURE 2.2 The reinforcement learning framework

The human annotators compare and rank multiple outputs from GPT-3 corresponding to each prompt. Based on this labeled data, a reward model is trained to predict the preferred output. Lastly, this reward model is a reward function and policy optimized to maximize the reward using the proximal policy optimization (PPO) algorithm.

Imagine you're teaching a teenager how to ride a snowboard for the first time. You want them to learn fancy tricks, but every time they try something new, you don't want them risking a big crash. The proximal policy optimization (PPO) algorithm is like a smart snowboard coach for the teen. It has a rule: "Try new turns or try new moves but not so different from what you already know, or you will definitely fall."

Here's how it works: The teenager tries a new turn or trick, sees how well they do (like scoring confidence points for staying upright and doing small tricks), and learns the way any human would. Then they try again, slightly tweaking their approach but with a twist. There's a safety net (the "clip" in PPO can be related to "clipping" the trick's extremes to avoid moving too far away from the original effort), making sure these tweaks aren't too drastic. This way, the teen steadily gets better without taking big risks that could lead to epic wipeouts.

PPO keeps a machine learning efficiently by reusing its experiences several times to refine its strategy, ensuring it learns a lot from each practice session. It's like watching a video of a snowboard performance on the hill and spotting a dozen ways to improve instead of just one. This makes the machine a quick learner and smart, avoiding unnecessary risks while it masters its metaphorical ability to shred on the mountain!

Despite the impressive results achieved by GPT3.5, this technique also faces some challenges and limitations that need to be addressed for further improvement and broader application. **Table 2.1** shows example challenges and potential mitigation activities with RLHF utilizing future research and development.

TABLE 2.1 Example challenges and potential mitigation activities

CHALLENGE	FUTURE RESEARCH AND DEVELOPMENT
Data quality and quantity: The quality and quantity of human feedback data are crucial for training a reliable reward model and a robust policy. However, collecting human feedback data can be costly, time-consuming, and prone to noise and bias. Moreover, human preferences may vary across domains, tasks, and contexts, requiring more diverse and representative data to capture the nuances and subtleties of human expectations and instructions.	Improving the data collection and annotation methods and tools to ensure human feedback data quality, quantity, and diversity. For example, using active learning, crowdsourcing, gamification, or interactive learning techniques to solicit more relevant, informative, and consistent feedback from the users or the experts. Alternatively, using synthetic, simulated, or generated data to augment the real data and increase the coverage and robustness of the data.

continued

TABLE 2.1 *continued*

CHALLENGE	FUTURE RESEARCH AND DEVELOPMENT
Reward shaping and alignment: The reward model learned from human feedback data may not always reflect the true objectives and values of the users or the developers. There may be gaps or conflicts between what humans express and what they actually want or need. For example, humans may provide inconsistent, ambiguous, or misleading feedback due to cognitive biases, emotional states, or communication errors. Furthermore, the reward model may not align with the ethical, social, or legal norms and standards that should guide the behavior of AI systems. For example, the reward model may incentivize harmful, deceptive, or manipulative actions that violate the principles of fairness, accountability, or transparency.	Enhancing the reward shaping and alignment methods and mechanisms to ensure the validity, reliability, and alignment of the reward model. For example, using inverse reinforcement learning, preference elicitation, or value learning techniques to infer the latent or implicit objectives and values of the users or the developers from their feedback or behavior. Alternatively, using multi-objective, constrained, or regularized reinforcement learning techniques to incorporate multiple criteria, constraints, or penalties into the reward function and balance the trade-offs among them.
Generalization and adaptation: The policy optimized by RLHF may not generalize well to new or unseen prompts, scenarios, or environments. The policy may overfit to the specific data distribution or the reward model and fail to handle novel or complex situations that require more creativity, reasoning, or common sense. Moreover, the policy may not adapt well to the dynamic and evolving needs and preferences of the users or the developers. The policy may become outdated, irrelevant, or incompatible with the changing goals, expectations, or instructions of the stakeholders.	Developing the generalization and adaptation methods and strategies to ensure the flexibility, versatility, and applicability of the policy. For example, using meta-learning, transfer learning, or lifelong learning techniques to enable the policy to learn from multiple sources, tasks, or domains and apply the learned knowledge or skills to new or different situations. Alternatively, using online learning, interactive learning, or self-learning techniques to enable the policy to update, refine, or improve itself based on the feedback or performance in real time or over time.

Anthropic, a startup founded by former employees of OpenAI, developed Claude, an AI chatbot that is similar to ChatGPT.[11] It is claimed that Claude outperforms ChatGPT in a variety of perspectives. It not only tends to generate more helpful and harmless answers but also answers in a more fun way when facing inappropriate requests. Its writing is more verbose but also more naturalistic. Claude's key approach is called *constitutional AI*.[12] Like ChatGPT, Claude also uses reinforcement learning to train a preference model, though Claude uses reinforcement learning from AI Feedback (RLAIF) without any human feedback labels for AI harms.[13] The constitutional AI process consists of two stages: supervised learning and reinforcement learning, as shown in **Figure 2.3**.

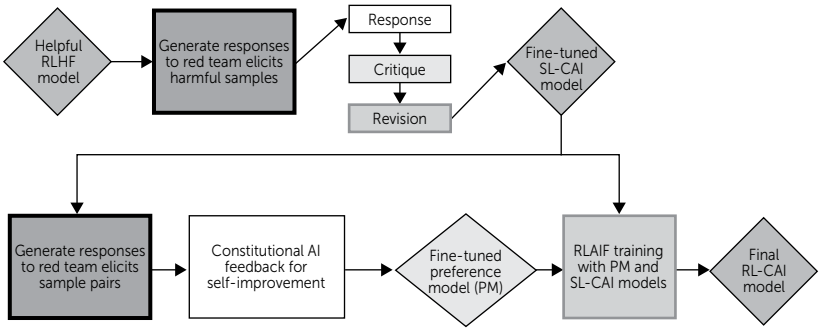

FIGURE 2.3 Steps used in the constitutional AI process

The constitutional AI process works like this:

1. In the supervised learning phase, initial responses to harmful prompts using a pre-trained language model that has been fine-tuned on a dataset of helpful-only responses are called *helpful-only AI assistants*.
2. The model is asked to critique and revise the responses using randomly selected principles from the 16 pre-written principles in the constitution.
3. As a result, the supervised learning–constitutional AI (SL-CAI) model is gained by fine-tuning the pretrained LLM on the final revised responses in a supervised learning way.
4. Claude uses a preference model as a reward signal in the reinforcement learning stage to optimize its responses to different prompts.
5. The fine-tuned model generates a pair of responses to each harmful prompt and evaluates responses according to a set of constitutional principles.
6. Then, a preference model is trained on the final dataset, combining the AI-generated preference dataset for harmlessness and the human feedback dataset for helpfulness.

7. The preference model learns to rank the responses based on their combined scores of helpfulness and harmlessness.
8. Finally, the SL model is fine-tuned via reinforcement learning against this preference model as a reward signal, which results in an optimized policy.

One advantage of this more advanced framework is that it can eliminate human annotation, saving a lot of time, cost, and energy. Similarly, we can develop specific principles with constitutional AI to ensure those LLMs produce factual, harmless, ethical, and fair outputs that also serve the needs of our particular scenarios. This approach, utilized by Claude, is based on the idea of aligning the AI chatbot's behavior with a set of constitutional principles that reflect the values and goals of the users and developers. These principles ensure that the chatbot generates helpful, harmless, ethical, responsible, and fair responses.

Claude's constitutional principles are respecting human dignity, avoiding harm and deception, promoting well-being and social good, and valuing diversity and inclusion. These principles provide a framework that can be modified and updated according to the customized needs and preferences of users and developers.

By using constitutional AI, Claude can outperform ChatGPT in several ways:

- Claude can generate more helpful and harmless responses because it is trained on a dataset that filters out harmful or unhelpful responses and incorporates human feedback on helpfulness.
- Claude can generate more ethical, responsible, and fair responses because it is under the guidance of a set of constitutional principles reflecting the values and goals of the users and developers.
- Claude can generate more fun and naturalistic responses by exploring and exploiting different responses using reinforcement learning and learning from its own critique and revision.

Chatbot customization can utilize reinforcement learning through human/AI feedback (RLHF/RLAIF). Chatbots are becoming increasingly prevalent in various domains, such as customer service, education, entertainment, health, and so on. However, not all users have the same preferences or needs when interacting with chatbots.

Some users prefer a more formal or professional tone, while others enjoy a casual or humorous style. Some users may want a more informative or detailed response, while others may seek a more concise or simple answer. Some users may appreciate a more empathetic or supportive response, while others may desire a more objective or factual one.

Therefore, it is important to customize the chatbot's behavior and personality according to the user's profile and feedback. A chatbot can leverage reinforcement learning to learn from its own actions and outcomes and adapt to the user's preferences and expectations over time.

Reinforcement learning is based on the idea of reward and punishment, where the chatbot receives positive or negative feedback from the user or itself and adjusts its policy accordingly. For example, if the user expresses satisfaction or gratitude after receiving a response from the chatbot, the chatbot can reinforce that response and generate similar ones in the future.

Conversely, if the user expresses dissatisfaction or frustration after receiving a response from the chatbot, the chatbot can avoid that response and generate different ones in the future. Moreover, the chatbot can also self-evaluate its responses and give itself feedback based on predefined criteria or metrics, such as relevance, coherence, fluency, informativeness, politeness, and the like.

Fine-Tuning Large-Scale Models

Fine-tuning is a popular method in the ML and AI fields and is done after a model has been pretrained. Then, the additional training is performed with a dataset specific to the scenarios practitioners and professionals work on. Fine-tuning solves common issues caused by large-scale AI models, such as difficulties productionizing big models and not being generalized enough for specific tasks.[14] See **Figure 2.4**.

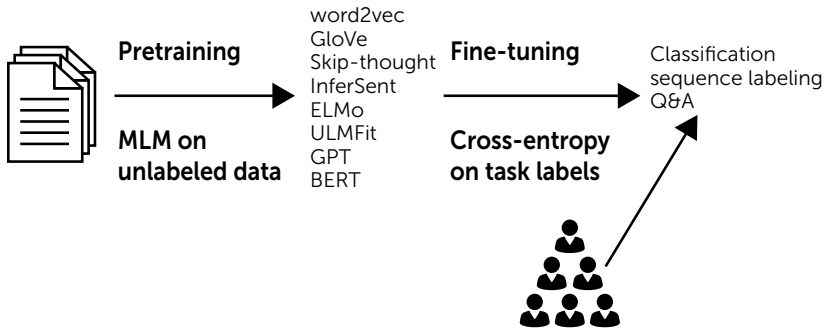

FIGURE 2.4 Fine-tuning pretrained large-scale models

Traditionally, most AI professionals do model tuning for fine-tuning, in which the pre-trained models' parameters (classification, sequence labeling, and question answering (Q&A) using task-specific labels and cross-entropy loss) are tuned. There have been several challenges with this approach and potential mitigation activities, as shown in **Table 2.2**.

TABLE 2.2 Challenges and potential mitigation activities of fine-tuning on pre-trained models

CHALLENGE	MITIGATION ACTIVITIES
Data availability: Fine-tuning requires sufficient labeled data for the target task or domain, which may not always be available or easy to collect. Fine-tuning may lead to overfitting or poor generalization if the data is too small or noisy.	**Data augmentation:** This is an approach to increase the size and diversity of the training data by applying some transformations or modifications to the existing data, such as cropping, flipping, rotating, adding noise, and so on. Data augmentation can help reduce overfitting and improve the generalization of the fine-tuned model.
Task transfer: Fine-tuning works best when the target task or domain is similar to the pretrained model. If the tasks or domains are too different, fine-tuning may not transfer the relevant knowledge or may even degrade the performance of the model.	**Transfer learning:** This is a technique to leverage the knowledge learned from one or more source tasks or domains to improve the performance of a target task or domain. Transfer learning can be done by freezing some of the layers in the pretrained model and adapting its output layer to the target task. Transfer learning can help overcome data availability and task transfer problems.
Cost and scalability: Fine-tuning large-scale models such as GPT or DALL-E requires a lot of computational resources and memory space, which may not be accessible or affordable for many users or organizations. Moreover, fine-tuning large models may introduce more complexity and instability to the optimization process.	**Meta-learning:** This is a technique to learn from multiple tasks or domains and then apply the learned knowledge to a new task or domain. Meta-learning can be done by training a meta-model or a meta-learner that can generate or update the parameters of a base model for a given task or domain. Meta-learning can help achieve fast adjustment and robust generalization of the fine-tuned model.

The evolvement and growing capabilities of current large-scale language models with prompt-tuning have become increasingly popular, in which the pre-trained model is frozen while a small set of learnable vectors can be optimized and added as the input for the task. Prompt design is even more commonly utilized, as of the writing of this book, which is a technique used to guide the behavior of a frozen pretrained model by crafting an input prompt for a specific task without changing any parameters. This is more effective and less expensive than prompt-tuning.[15] We can compare these three approaches to adapting pre-trained language models for specific tasks:

- **Model tuning:** The pre-trained model is further trained or "fine-tuned" on a task-specific dataset.
- **Prompt tuning:** The model remains frozen, and only a set of tunable soft prompts are optimized.
- **Prompt design:** Exemplified by GPT-3, crafted prompts guide the frozen model's responses without any parameter changes.

Prompt-tuning and prompt design methods are often used because of their effectiveness and reduced cost compared to full model tuning. See **Figure 2.5**, which illustrates a shift toward efficiency and multitasking in language model applications, highlighting the less resource-intensive nature of prompt-based methods.

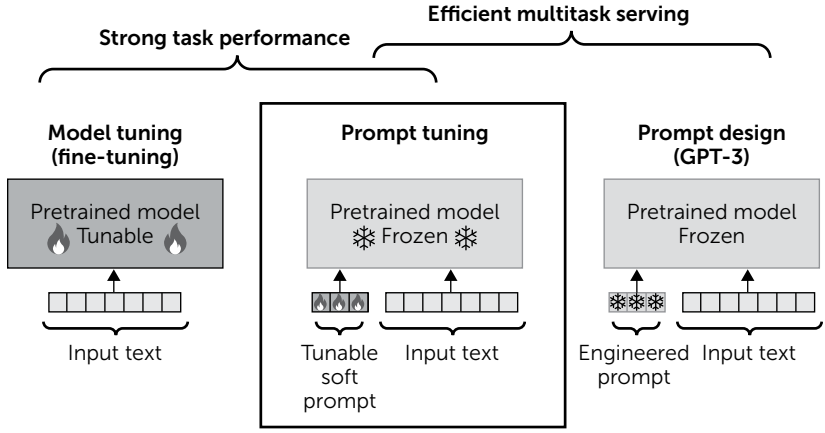

FIGURE 2.5 The architecture of model tuning, prompt tuning, and prompt design

Prompt Engineering

With the remarkable success and powerful generalization capabilities of current large pre-trained AI models, more and more AI practitioners are focusing on prompt engineering by directly integrating the existing generative AI models such as DALL-E 3, GPT-4, and ChatGPT into real applications. As we know, fine-tuning requires huge computational resources and memory space and causes catastrophic forgetting. Prompt engineering is a discipline focused on optimizing prompts for efficient use of LLMs across various applications and research. It enhances our understanding of LLMs' capabilities and limitations.

Prompt engineering encompasses diverse skills and techniques, crucial for effective LLM use. It enhances LLM safety and empowers integration with domain knowledge and external tools.

A prompt is a parameter that can be provided to large-scale pretrained LMs like GPT to enable its capability to identify the context of the problem to be solved and accordingly return the resulting text. In other words, the prompt includes the task description and demonstrations or examples that can be fed into the LMs to be completed. Prompt engineering, sometimes called in-context learning or prompt-based fine-tuning, is a paradigm of learning where only the prompt, which includes a task description and a few demonstrations, is fed into the model as if it were a black box. There are multiple prompt engineering techniques:

- **Retrieval augmentation for in-context learning:** The main idea is to retrieve a set of relevant documents or examples given a source and take these as context with the original input prompt to let the LLM generate the final output. There are different methods for in-context learning, such as one-shot and few-shot prompting. One example is the method RAG (Retrieval Augmented Generation) introduced by Meta AI that essentially takes the initial prompt plus searches for relevant source materials, such as Wikipedia articles, and combines the information with the sequence-to-sequence generation to provide the output.[16]

- **Chain-of-Thought (CoT):** This prompting technique encourages the model to generate a series of intermediate reasoning steps (see **Figure 2.6**).[17] A less formal way to induce this behavior is to include "Let's think step-by-step" in the prompt.

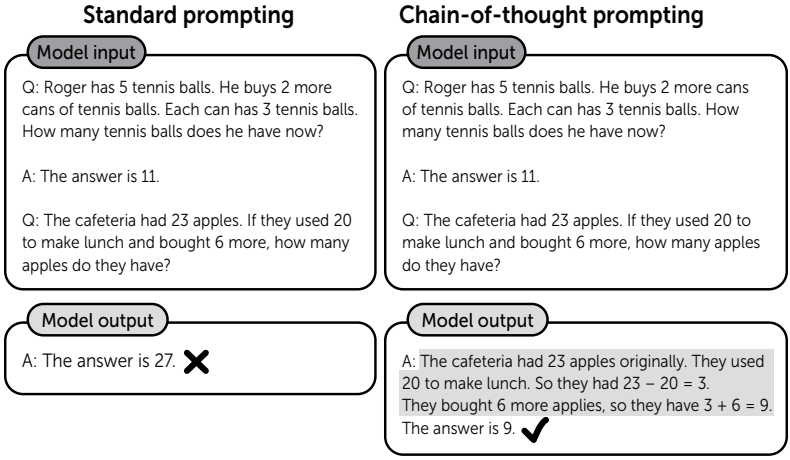

FIGURE 2.6 Chain-of-thought prompting

- **Action Plan Generation:** This prompt utilizes a language model to generate actions to take, as shown in **Figure 2.7**.[18] The results of these actions can then be fed back into the language model to generate a subsequent action.

Command	Effect
Search <query>	Send <query> to the Bing API and display a search results page
Clicked on link <link ID>	Follow the link with the given ID to a new page
Find in page: <text>	Find the next occurrence of <text> and scroll to it
Quote: <text>	If <text> is found in the current page, add it as a reference
Scrolled down <1, 2, 3>	Scroll down a number of times
Scrolled up <1, 2, 3>	Scroll up a number of times
Top	Scroll to the top of the page
Back	Go to the previous page
End: Answer	End browsing and move to answering phase
End: <Nonsense, Controversial>	End browsing and skip answering phase

FIGURE 2.7 Action plan generation prompting

- **ReAct Prompting:** This prompting technique combines chain-of-thought prompting with action plan generation (see **Figure 2.8**). This induces the model to think about what action to take, and then take it. ReAct allows language models to produce both verbal reasoning traces and text actions that alternate with each other, while actions cause observation feedback from an external environment. The example shown in Figure 2.8 compares the performance of the standard prompting, chain-of-thought (reason only), act only, and ReAct prompting techniques.[19]

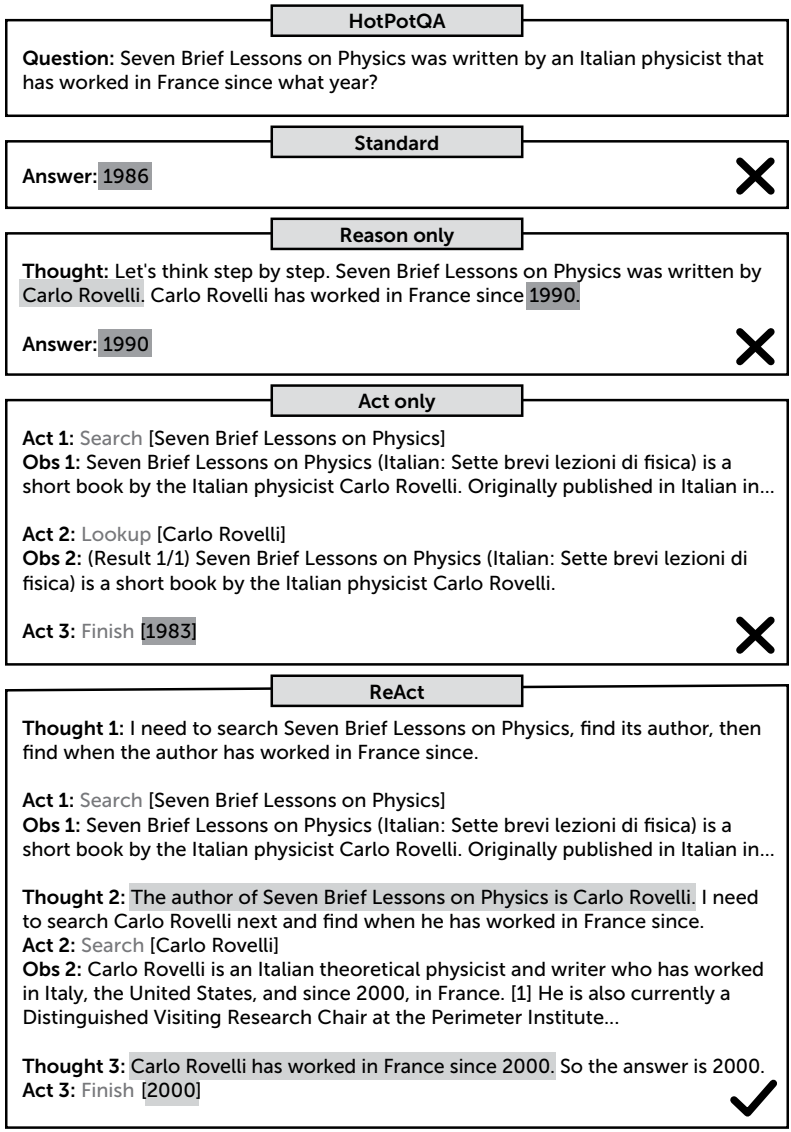

FIGURE 2.8 The results of four prompting methods

Prompt Chaining

This approach combines multiple LLM calls, with the output of one step being the input to the next. The overall process includes a few steps:

1. The process starts with an initial prompt or question. This could be a broad inquiry, instruction, or a request for information.
2. The model generates an initial response based on the input prompt. However, this response might be a bit generic or need refinement.
3. The generated response is then used as part of a new prompt. This time, the prompt is more specific, providing additional context or asking for clarification.

The chaining continues iteratively. Each new response becomes the input for the next prompt. The generated content becomes more focused and contextually relevant with each iteration. The advantages of prompt chaining are as follows:[20]

- It helps preserve context across responses and makes the generated output more coherent.
- The user can guide the model through the iteration process to provide more precise and relevant generation.
- It leads to more customized generation, which enables users to tailor the responses to their specific requirements. However, it still does not alter the fundamental capabilities and limitations of the underlying language model.

Tree of Thoughts

The tree of thoughts framework generalizes over chain-of-thought prompting and encourages the exploration of thoughts that serve as intermediate steps for general problem-solving with language models. This method allows a language model to self-assess the progress of its intermediate thoughts during problem-solving through a deliberate reasoning process. The LM's capacity to produce and assess thoughts is then integrated with search algorithms like breadth-first search and depth-first search, facilitating systematic thought exploration with lookahead and backtracking.[21]

Self-Consistency

The idea behind self-consistency is based on chain-of-thought (CoT), but it samples multiple diverse reasoning paths through few-shot CoT and uses the generations to select the most consistent answer. This helps to boost the performance of CoT prompting on tasks involving arithmetic and commonsense reasoning.[22]

Unveiling the Power of Clustering and Topic Modeling

Despite the rapid evolution of LLMs that can produce coherent and diverse texts across various domains, many tasks still require more granular and structured analysis of textual data. Clustering and topic modeling are techniques that can help discover hidden patterns, themes, and categories in a large collection of documents, without relying on predefined labels or annotations. They can also help reduce the data's dimensionality and complexity, making it easier to visualize, summarize, and interpret.

There are some example applications where clustering and topic modeling can be useful, such as:

- **Document classification and retrieval:** Clustering and topic modeling can help search and navigate large collections of documents by grouping similar ones according to their content. Moreover, they can also facilitate the identification of relevant documents for a given query or task.
- **Text summarization and generation:** Although LLMs can also be utilized for text summarization and clustering, topic modeling can supplement LLMs by extracting the main topics and keywords from the targeted collections of documents and providing concise and informative summaries that capture the essence and different granularities of the data. They can also serve as input or an additional layer for text generation systems, such as LLMs, that can produce longer and more detailed texts based on the topics and keywords.
- **Sentiment analysis and opinion mining:** Although LLMs have shown remarkable performance in understanding the context and capturing nuances in natural languages, topic modeling, and clustering methods, taking Latent Dirichlet Allocation (LDA) or K-mean clustering as examples can be more interpretable and can provide insights into the main themes in a collection of texts.[23] Utilizing a hybrid approach that combines both might be a good solution. For instance, using LLMs for

fine-grained sentiment analysis and using topic modeling to understand broader themes or trends.

- **Knowledge discovery and extraction:** By uncovering the latent concepts and relations among the documents, clustering and topic modeling can enrich the semantic representation of the data, as well as the knowledge base of the domain. They can also help to identify gaps and inconsistencies in the data, as well as new and emerging topics and issues.

Therefore, clustering and topic modeling are still necessary and valuable tools for many tasks that involve understanding, analyzing, and generating textual data, especially when the data is large, heterogeneous, and unlabeled. They can complement and enhance the capabilities of LLMs' capabilities and provide insights and feedback for improving their performance and quality.

Enhancing Customer Support Through Hybrid AI: LLMs Meet Clustering and Topic Modeling

Customer support is evolving, and businesses seek more sophisticated and powerful solutions to handle the vast amount of textual data generated in interactions. A hybrid approach, blending the capabilities of LLMs and traditional machine learning techniques, emerges as a robust strategy. We'll explore a few of these machine learning techniques often utilized in support organizations to make sense of the large amounts of data to help optimize the business.

Clustering and Customer Support

Clustering is an unsupervised learning approach of grouping a set of samples based on their similarity without using any predefined labels or categories. Clustering aims to discover the natural structure or patterns of the data, as well as to reduce its complexity and dimensionality. Clustering can be used for various purposes, such as data exploration, summarization, organization, retrieval, and visualization. There are several different clustering methods:

- **Hierarchical clustering:** This method builds a hierarchy of clusters, where each cluster is either a subcluster or a supercluster of another cluster. Hierarchical clustering can be either agglomerative or divisive. Agglomerative clustering starts with each sample as a singleton

cluster and then merges the most similar clusters until a single cluster remains. Divisive clustering starts with all documents in one cluster and then splits the most dissimilar clusters until each cluster contains only one sample.

- **Partitioning clustering:** This method divides the data points into a predefined number of non-overlapping clusters, where each point belongs to exactly one cluster. K-mean clustering is one of the most popular algorithms for partitioning clustering. Partitioning clustering can be either distance-based or centroid-based. Distance-based clustering assigns each data point to the cluster with the closest or most similar representative, such as the nearest neighbor. Centroid-based clustering assigns each data point to the cluster with the smallest or least average distance to the center or the cluster's mean, such as K-mean clustering. K-mean clustering classifies samples based on attributes or features into k clusters. It starts with a first group of randomly selected centroids, which are used as the beginning points for every cluster, and then assigns each point to the cluster whose mean has the least squared Euclidean distance and optimizes the centroid based on the distances from the points to it. The hard assignment stops creating and optimizing clustering when either the centroids have stabilized or the defined number of iterations has been reached.
- **Density-based clustering:** This method identifies clusters based on the density or the concentration of the data points in the feature space, where regions of separate low-density clusters can be uncovered and assist in identifying unforeseen patterns. Density-based clustering can handle outliers, noise, and arbitrary shapes of clusters. One of the popular algorithms for density-based clustering is DBSCAN (density-based spatial clustering of applications with noise). DBSCAN defines a cluster as a set of densely connected core points; a point is a core point if it has at least a minimum number of points within a given radius or neighborhood.

Clustering is a powerful technique for identifying patterns and insights from large and complex data sets. It can be used to segment customers, optimize services, categorize issues based on their similarities or differences, and provide personalized and efficient solutions. In the field of customer service and support, clustering has been a popular approach for solving some problems, such as:

- **Customer segmentation:** Clustering can help discover different groups of customers based on their demographics, preferences, needs,

behaviors, or characteristics, such as age, gender, location, income, spending habits, loyalty, satisfaction, or feedback. This can help tailor the marketing strategies, product recommendations, pricing policies, or communication channels for each segment and to improve customer retention and acquisition.

- **Service optimization:** Clustering can help optimize the service delivery and support processes based on the complexity, urgency, or frequency of customer requests, issues, or inquiries, such as order status, product information, technical support, billing, or feedback. This can help allocate the appropriate resources, staff, or channels for each service type and improve service efficiency and quality.
- **Support case categorization:** Clustering can help resolve customer issues faster and more effectively by grouping similar or related issues based on their causes, symptoms, or solutions, such as product defects, software bugs, network failures, or user errors. When AI technology is used to cluster similar cases together, these groupings can help by offering new insights that are not obvious when looking at cases individually or by product. An example might be multiple unrelated services experiencing login or profile creation issues. Viewed on their own, these could be hard to relate or determine the root cause of the issue, but after clustering them together, it might be more obvious that this is a problem with shared code providing identity services to multiple workloads. This clustering can help diagnose the root causes, find the best solutions or prevent future occurrences of the issues, increase customer satisfaction, and enhance retention.

Topic Modeling and Customer Support

Topic modeling is a technique for extracting hidden topics or concepts from a collection of text documents, such as customer reviews, feedback, complaints, or inquiries. Topic modeling can help discover the main themes or patterns of customer needs, preferences, opinions, or issues and provide valuable insights for customer support improvement, product development, marketing strategy, or sentiment analysis.

There are several different topic modeling methods. These algorithms differ in their assumptions, mathematical models, and implementations, but they all share the same basic idea: finding a low-dimensional representation of the documents and the words in terms of topics and probabilities. The output of a topic modeling algorithm is usually a matrix that shows the relationship between documents and topics, and another matrix that shows the

relationship between topics and words. These matrices can be used to infer the topics of new documents, find similar documents, visualize the topics, and extract insights from the text data. These methods include:

- **Latent Dirichlet Allocation (LDA):** This is one of the most popular topic modeling methods. LDA is an unsupervised learning algorithm that describes a set of observations as a mixture of distinct categories. These categories are themselves a probability distribution over the features. LDA is most commonly used to discover a user-specific number of topics shared by a collection of documents within a text corpus. Each observation is a document, the features are the presence or occurrence count of each word, the categories are the topics. LDA uses a generative process to assign topic probabilities to each document and word probabilities to each topic, based on the observed word frequencies in the documents. LDA can be applied to large, diverse text corpora and produce interpretable and coherent topics.
- **Non-negative Matrix Factorization (NMF):** NMF is a linear algebra method that decomposes a matrix of word-document frequencies into two lower-dimensional non-negative matrices, one representing the word-topic associations and the other representing the topic-document associations. NMF imposes a non-negativity constraint on the matrices, which ensures that the topics and the documents have additive and meaningful components. NMF can be faster and more robust than LDA and can handle sparse and noisy data.
- **Hierarchical Dirichlet Process (HDP):** HDP is a Bayesian nonparametric model that extends LDA by allowing the number of topics to be automatically inferred from the data rather than fixed in advance. HDP uses a hierarchical structure of Dirichlet processes to generate a potentially infinite number of topics and assigns them to the documents based on their relevance and specificity. HDP can adapt to the complexity and diversity of the text data and can avoid overfitting or underfitting the topics.

Topic modeling is a valuable technique in the customer service and support field for extracting insights from large volumes of textual data, such as customer reviews, feedback, and support cases. Here's how topic modeling is leveraged in this domain:

- **Automated support case categorization:** Customer support teams often deal with a variety of issues and requests. Topic modeling can be leveraged to automatically categorize support tickets into different topics or categories based on their content. This helps in routing tickets to appropriate product support teams and improves response time and efficiency.

Moreover, topic modeling can help automate some processes in the customer support workflow. For example, it can point customers to the self-help knowledge base, diagnostics, or websites with the accurate topic category prediction. This can enhance the customer experience, reduce customer effort, and increase operational efficiency.

- **Identifying emerging issues:** Topic modeling can help uncover emerging trends or issues in customer feedback and support cases. It provides actionable insights for companies to address top issues before they escalate proactively.
- **Improving search and retrieval:** Topic modeling helps organize and index articles based on the topics for a large knowledge base of support or self-help articles. This improves the search and retrieval process for support agents or engineers and the customers looking for solutions.
- **Customer feedback analysis:** Topic modeling can help analyze and summarize customer feedback from multiple channels and platforms. This can help identify the most common and important topics, issues, compliments, complaints, and suggestions that customers express. This can also help products and companies measure and track key performance indicators related to customer support, customer satisfaction, and loyalty. For instance, it can help measure the volume of support cases in different categories, identify resolution time, and assess customer satisfaction for each topic. Furthermore, product teams can prioritize and address customer complaints and grievances more effectively.
- **Content creation and knowledge management**. Topic modeling aids in content creation for FAQs, manuals, and support articles. It helps identify the most discussed topics, allowing companies to create relevant and helpful content that addresses common customer queries.

In essence, topic modeling enhances the efficiency and effectiveness of customer service and support operations by providing automated tools for organizing, analyzing, and extracting insights from large volumes of textual customer data.

Hybrid AI Opportunity

Traditional machine learning methods like topic modeling and clustering have their own limitations and challenges. One of the main drawbacks is that they rely on statistical methods that do not account for the semantic and contextual nuances of natural language. For example, topic modeling may fail to distinguish between different meanings or senses of the same word, such as apple as a company but not as a fruit, or group together

words that are syntactically similar but semantically different, such as bass as a type of fish but not low-frequency sound in music. Moreover, topic modeling may produce topics that are too broad, too narrow, or not coherent, depending on the choice of parameters and algorithms. In contrast, large language models, such as GPT and Gemini, have demonstrated remarkable proficiency in understanding context, generating human-like responses, and extracting intricate patterns from textual data. In customer support, LLMs can be employed for tasks like sentiment analysis, intent recognition, and even generating responses to common queries.

While LLMs excel in understanding context and generating text, traditional machine learning methods like clustering and topic modeling offer strengths in structuring and organizing information. Clustering can group similar customer queries or issues, facilitating efficient handling by support agents. Topic modeling, on the other hand, extracts underlying themes from a vast dataset, aiding in understanding prevalent customer concerns. Moreover, when computational resources and budget are limited, it is easier and cheaper to leverage traditional machine learning methods like topic modeling and clustering.

In the dynamic landscape of customer support, a hybrid approach, integrating the capabilities of LLMs with the structuring prowess of traditional methods, proves to be a holistic solution. By combining LLMs with topic modeling, more accurate, robust, and interpretable models can be utilized for customer feedback analysis. For instance, language models can help generate more natural and fluent texts from topics and can also help capture the semantic and contextual information that topic modeling may miss. Furthermore, LLMs can help generate new and novel topics that may not be present in the existing data or suggest relevant and personalized content based on the topics of interest of each customer, while topic modeling and clustering can bring more interpretability and flexibility. This hybrid solution addresses the complexities of customer interactions, providing businesses with a powerful tool for improving customer satisfaction and support efficiency.

Endnotes

1 Wikipedia contributors. "BERT (language model)." Wikipedia, The Free Encyclopedia. March 5, 2024. [https://en.wikipedia.org/wiki/BERT_(language_model)].

2 Wikipedia contributors. "Generative pre-trained transformer." Wikipedia, The Free Encyclopedia. February 18, 2024. [https://en.wikipedia.org/wiki/Generative_pre-trained_transformer].

3 Gomez, A., Jones, L., Kaiser, L., Parmar, N., Polosukhin, I., Shazeer, N., Uszkoreit, J., Vaswani, A. 2023. "Attention Is All You Need." Google. August 2, 2023. [https://arxiv.org/pdf/1706.03762.pdf].

4 Zhao, W. et al. 2023. "A Survey of Large Language Models." Cornell University. November 24, 2023. [https://arxiv.org/pdf/2303.18223.pdf].

5 Bengio, Y., Ducharme, R., Vincent, P., Jauvin., C. 2003. "A Neural Probabilistic Language Model." Journal of Machine Learning Research. February 2023. [https://www.jmlr.org/papers/volume3/bengio03a/bengio03a.pdf].

6 Wikipedia contributors. "Word2vec." Wikipedia, The Free Encyclopedia. March 21, 2024. [https://en.wikipedia.org/wiki/Word2vec#Continuous_Bag_of_Words_%28CBOW%29].

7 Wikipedia contributors. "Word2vec." Wikipedia, The Free Encyclopedia. March 21, 2024. [https://en.wikipedia.org/wiki/Word2vec#Continuous_Bag_of_Words_%28CBOW%29].

8 Contributors, Open AI (2024, May 13). *Hello GPT-4o*. Retrieved from [https://openai.com/index/hello-gpt-4o/]

9 S., Amit. 2023. "Everything You Need To Know About Knowledge Distillation, aka Teacher–Student Model." Medium. April 19, 2023. [https://amit-s.medium.com/everything-you-need-to-know-about-knowledge-distillation-aka-teacher-student-model-d6ee10fe7276].

10 S., Amit. 2023. "Everything You Need To Know About Knowledge Distillation, aka Teacher-Student Model." Medium. April 19, 2023. [https://amit-s.medium.com/everything-you-need-to-know-about-knowledge-distillation-aka-teacher-student-model-d6ee10fe7276].

11 Leike, J., Lowe, R., et al. 2022. "Training language models to follow instructions with human feedback." Cornell University. March 4, 2022. [https://arxiv.org/pdf/2203.02155.pdf].

12 Henshall, Will. 2023. "What to Know About Claude 2, Anthropic's Rival to ChatGPT." Time. July 18, 2023. [https://time.com/6295523/claude-2-anthropic-chatgpt/].

13 Bai, Yuntao, et al. 2022. "Constitutional AI: Harmlessness from AI Feedback." Cornell University. December, 15, 2022. [https://arxiv.org/pdf/2212.08073.pdf].

14 Ruder, Sebastian. 2021. "Recent Advances in Language Model Fine-tuning." Ruder.io. February, 24, 2021. [https://www.ruder.io/recent-advances-lm-fine-tuning/].

15 Constant, N., Lester, B. 2022. "Guiding Frozen Language Models with Learned Soft Prompts." Google Research. February 10, 2022. [https://blog.research.google/2022/02/guiding-frozen-language-models-with.html].

16 Meta Blog Editors. 2020. "Retrieval Augmented Generation: Streamlining the creation of intelligent natural language processing models." Meta Blog. September 28, 2020. [https://ai.meta.com/blog/retrieval-augmented-generation-streamlining-the-creation-of-intelligent-natural-language-processing-models/].

17 Wei, J., Zhou, D., et al. 2023. "Chain-of-Thought Prompting Elicits Reasoning in large Language Models." Cornell University. January 10, 2023. [https://arxiv.org/pdf/2201.11903.pdf].

18 Nakano, R., et al. 2022. "WebGPT: Browser-assisted question-answering with human feedback." Cornell University. June 1, 2022. [https://arxiv.org/pdf/2112.09332.pdf].

19 Yao, S., et al. 2023. "ReAct: Synergizing Reasoning and Acting in Language Models." Cornell University. March 10, 2023. [https://arxiv.org/pdf/2210.03629.pdf].

20 Anthropic editors. 2023. "Prompt chaining." Anthropic.com. Accessed January 27, 2024. [https://docs.anthropic.com/claude/docs/prompt-chaining].

21 Long, Jieyi. 2023. "Large Language Model Guided Tree-of-Thought." Cornell University. May 15, 2023. [https://arxiv.org/abs/2305.08291].

22 Chi, E., Chowdhery, A., Le, Q., Narang, S., Schuurmans, D., Wang, X., Wei, J., Zhou, D. 2022. "Self-Consistency Improves Chain of Thought Reasoning in Language Models." Cornell University. March 7, 2023. [https://arxiv.org/abs/2203.11171v4].

22 Liu, Bing. 2012. "Sentiment Analysis and Opinion Mining." Morgan & Claypool Publishers. April 22, 2012. [https://www.cs.uic.edu/~liub/FBS/SentimentAnalysis-and-OpinionMining.pdf].

Index

B

C

D

E

F

G

H

I

J

K

L

M

N

O

P

Q

R

S

T

U

V

W

Y

Z