What this book covers

Chapter 1, *What Is Generative AI?,* explains how generative AI has revolutionized the processing of text, images, and video, with deep learning at its core. This chapter introduces generative models such as LLMs, detailing their technical underpinnings and transformative potential across various sectors. This chapter covers the theory behind these models, highlighting neural networks and training approaches, and the creation of human-like content. The chapter outlines the evolution of AI, Transformer architecture, text-to-image models like Stable Diffusion, and touches on sound and video applications.

Chapter 2, *LangChain for LLM Apps,* uncovers the need to expand beyond the stochastic parrots of LLMs–models that mimic language without true understanding–by harnessing LangChain's framework. Addressing limitations like outdated knowledge, action limitations, and hallucination risks, the chapter highlights how LangChain integrates external data and interventions for more coherent AI applications. The chapter critically engages with the concept of stochastic parrots, revealing the deficiencies in models that produce fluent but meaningless language, and explicates how prompting, chain-of-thought reasoning, and retrieval grounding augment LLMs to address issues of contextuality, bias, and intransparency.

Chapter 3, *Getting Started with LangChain,* provides foundational knowledge for you to set up your environment to run all examples in the book. It begins with installation guidance for Docker, Conda, Pip, and Poetry. The chapter then details integrating models from various providers like OpenAI's ChatGPT and Hugging Face, including obtaining necessary API keys. It also deals with running open-source models locally. The chapter culminates in constructing an LLM app to assist customer service agents, exemplifying how LangChain can streamline operations and enhance the accuracy of responses.

Chapter 4, *Building Capable Assistants,* tackles turning LLMs into reliable assistants by weaving in fact-checking to reduce misinformation, employing sophisticated prompting strategies for summarization, and integrating external tools for enhanced knowledge. It explores the Chain of Density for information extraction and discusses LangChain decorators and expression language for customizing behavior. The chapter introduces map-reduce in LangChain for handling long documents and discusses token monitoring to manage API usage costs.

It looks at implementing a Streamlit application to create interactive LLM applications and using function calling and tool usage to transcend basic text generation. Two distinct agent paradigms, plan-and-solve and zero-shot, are implemented to demonstrate decision-making strategies.

Chapter 5, *Building a Chatbot like ChatGPT*, delves into enhancing chatbot capabilities with **retrieval-augmented generation** (**RAG**), a method that provides LLMs with access to external knowledge, improving their accuracy and domain-specific proficiency. This chapter discusses document vectorization, efficient indexing, and the use of vector databases like Milvus and Pinecone for semantic search. We implement a chatbot, incorporating moderation chains to ensure responsible communication. The chatbot, available on GitHub, serves as a basis for exploring advanced topics like dialogue memory and context management.

Chapter 6, *Developing Software with Generative AI*, examines the burgeoning role of LLMs in software development, highlighting the potential for AI to automate coding tasks and serve as dynamic coding assistants. It explores the current state of AI-driven software development, experiments with models to generate code snippets, and introduces a design for an automated software development agent using LangChain. Critical reflections on the agent's performance emphasize

the importance of human oversight for error mitigation and high-level design, setting the stage for a future where AI and human developers work symbiotically.

Chapter 7, *LLMs for Data Science*, explores the intersection of generative AI and data science, spotlighting LLMs' potential to amplify productivity and drive scientific discovery. The chapter outlines the current scope of automation in data science through AutoML and extends this notion with the integration of LLMs for advanced tasks like augmenting datasets and generating executable code. It covers practical methods for LLMs to conduct exploratory data analysis, run SQL queries, and visualize statistical data. Finally, the use of agents and tools demonstrates how LLMs can address complex data-centric questions.

Chapter 8, *Customizing LLMs and Their Output*, delves into conditioning techniques like fine-tuning and prompting, essential for tailoring LLM performance to complex reasoning and specialized tasks. We unpack fine-tuning, where an LLM is further trained on task-specific data, and prompt engineering, which strategically guides the LLM to generate desired outputs. Advanced prompting strategies such as few-shot learning and chain-of-thought are implemented, enhancing the reasoning capabilities of LLMs. The chapter not only provides concrete examples of fine-tuning and prompting but also discusses the future of LLM advancements and their applications in the field.

Chapter 9, *Generative AI in Production*, addresses the complexities of deploying LLMs within real-world applications, covering best practices for ensuring performance, meeting regulatory requirements, robustness at scale, and effective monitoring. It underscores the importance of evaluation, observability, and systematic operation to make generative AI beneficial in customer engagement and decision-making with financial consequences. It also outlines practical strategies for deployment and ongoing monitoring of LLM apps using tools like Fast API, Ray, and newcomers such as LangServe and LangSmith. These tools can provide automated evaluation and metrics that support the responsible adoption of generative AI across sectors.

Chapter 10, *The Future of Generative Models*, ventures into the potential advancements and socio-technical challenges of generative AI. It examines the economic and societal impacts of these technologies, debating job displacement, misinformation, and ethical concerns like human value alignment. As various sectors brace for disruptive AI-induced changes, it reflects on the responsibility of corporations, lawmakers, and technologists to forge effective governance frameworks.
This final chapter emphasizes the importance of steering AI development toward augmenting human potential while addressing risks such as deepfakes, bias, and the weaponization of AI. It highlights the urgency for transparency, ethical deployment, and equitable access to guide the generative AI revolution positively.