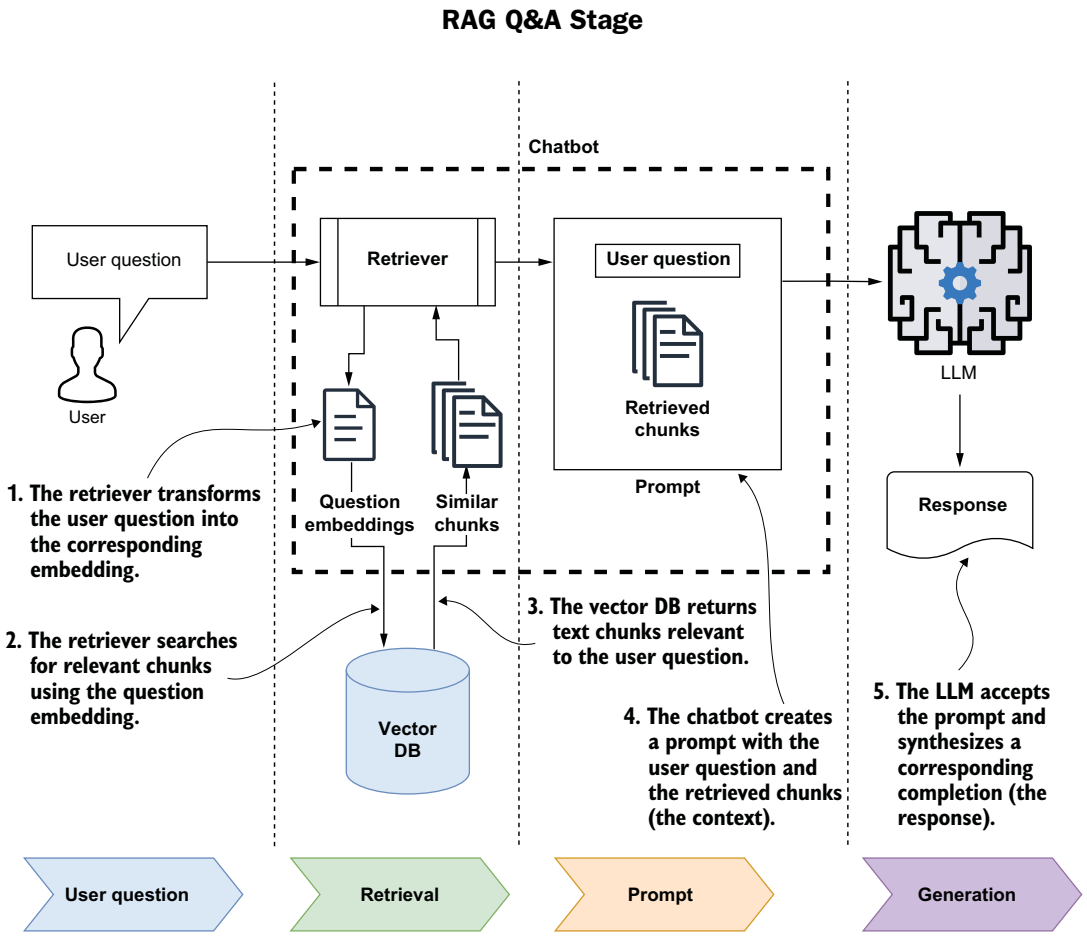
AI Agents and Applications

With LangChain, LangGraph and MCP

Roberto Infante

MANNING



Retrieval-Augmented Generation (RAG) Q&A stage: retrieval and generation

AI Agents and Applications

WITH LANGCHAIN, LANGGRAPH AND MCP

ROBERTO INFANTE

MANNING
SHELTER ISLAND

For online information and ordering of this and other Manning books, please visit www.manning.com. The publisher offers discounts on this book when ordered in quantity. For more information, please contact

Special Sales Department
Manning Publications Co.
20 Baldwin Road
PO Box 761
Shelter Island, NY 11964
Email: orders@manning.com

Manning Publications Co.
20 Baldwin Road
PO Box 761
Shelter Island, NY 11964

Development editor:	Dustin Archibald
Technical editors:	Keerthivasan Santhanakrishnan and Antowan Malik Batts
Review editor:	Kishor Rit
Production editor:	Kathy Rossland
Copy editor:	Julie McNamee
Proofreader:	Melody Dolab
Technical proofreader:	Andrew R. Freed
Typesetter and cover designer:	Marija Tudor

ISBN 9781633436541
Printed in the United States of America

To my mother and father

brief contents

contents

preface

In late 2022, something changed. Large language models (LLMs) stopped feeling like experimental demonstrations and started becoming genuinely useful. A quick attempt to summarize a paragraph evolved into a chatbot capable of answering questions, and a small script turned into a service that other teams wanted to try. Before long, LLMs were no longer a curiosity—they had become an essential part of the software development toolkit.

Here's why that's so exciting: LLMs allow software to "speak human." They can revise a contract, turn logs into meaningful answers, draft code, and plan the next step—and then invoke the right tools and data to actually accomplish the task. Combined with retrieval and tool use, an application stops feeling like a rigid machine and begins to feel more like a collaborative partner. The potential is significant, but turning that potential into production systems isn't simple. It still requires careful work to integrate data flows, design effective prompts, ground answers with retrieval, orchestrate multi-step workflows, and monitor how the system behaves once it's deployed.

My own journey into this field followed a similar path to many developers. I began by experimenting in Jupyter Notebooks, exploring APIs, and learning where the models succeeded and where they struggled. Those early explorations gradually evolved into small, agent-based side projects at work, built for a handful of early adopters seeking productivity gains. As the technology advanced—through improvements in OpenAI's APIs, the growing capabilities of LangChain, the orchestration power of LangGraph, and emerging techniques such as advanced Retrieval-Augmented Generation (RAG) and ReAct—those prototypes became more sophisticated. At the same time, I began writing this book. More than once, a chapter I had just completed felt outdated only a few weeks later. It was both exhilarating and, at times, exhausting.

That experience strongly influenced the way this book is written. Rather than chase every new parameter or short-lived "best practice," the focus here is on the concepts, architectures, and design patterns that have proven stable and that underpin reliable LLM applications. We'll build concrete systems—engines, chatbots, and agents—but the aim is to provide reusable foundations: how to structure retrieval, design prompts, compose chains, evaluate behavior, and orchestrate multi-step workflows with clarity and confidence.

Frameworks play a crucial role in this process. LangChain standardizes the essential components—loaders, splitters, embeddings, retrievers, vector stores, and prompts—so that you don't need to reinvent the plumbing for every project. LangGraph extends this by structuring workflows as graphs and coordinating agent loops, while LangSmith adds visibility for debugging and evaluation. Together, they enable developers to focus on the application itself rather than the underlying infrastructure.

Why this book, and why now? The foundations are finally stable enough to teach effectively, and the need is greater than ever. Development teams want practical guidance that bridges the gap between ideas and implementation without tying them to a single vendor or a fleeting technique. If this book succeeds, you'll come away with a clear mental model for building LLM-powered systems, a set of proven patterns you can depend on, and the confidence to keep building—even as the landscape continues to evolve beneath our feet.

acknowledgments

Writing a book like this is never a solo effort, and I feel profoundly grateful to have had so many talented and generous people alongside me on this journey. First, I would like to thank Juan-Mauro Bozzano, Romulus Corneanu, Carolyn Kao, and Ivan Lin for their thoughtful feedback at the very beginning, when the first chapters were still little more than outlines. Their comments and conversations helped shape the direction of the book in its early stages. I'm also indebted to David Bujan and Diego López-de-Ipiña from Deusto University, who encouraged me to bring greater precision and rigor to the work—especially in those formative drafts when many of the ideas were still taking shape.

A special word of thanks goes to Michael Stephens at Manning, who believed in this project from the very first conversation. At the time, LLMs were only beginning to capture the world's attention, and the road ahead was far from clear. His support, together with timely guidance and encouragement to adapt as the technology evolved, helped keep the book relevant and focused.

I'm particularly grateful to my editor, Dustin Archibald, whose steady guidance and thoughtful advice carried me through every stage of the writing process, and to Andrew Freed, whose sharp technical insights made the explanations clearer and more accurate. Many thanks also to technical editor Antowan Batts. Antowan is a financial systems analyst and applied economist with expertise in finance, data analysis, and global supply chains. He teaches economics and focuses on connecting analytical insights to real-world business challenges.

To the reviewers—Abdullah Al Imran, Abhishek Guha, Andres Mariscal, Antowan Malik Batts, Artur Guja, Aryan Jadon, Balaji Venkateswaran, Borko Djurkovic, Byron Galbraith, Carnell Greenfield, David Caswell, David Jacobs, Derek Morgan, Felipe P.