Galileo

Mastering RAG

A comprehensive guide for building enterprise-grade RAG systems

PREFACE

It's fascinating how quickly we've gotten accustomed to "prompt and you shall get" wizardry, haven't we? What was once far-fetched, a bold idea in a sci-fi novel, has already found widespread popularity, so much so that we run to answering engines for quick recipes, lesson plans, travel itineraries, homework help, and a medley of other things—life advice, even!

Large language models (LLMs), a term sometimes interchangeably used with OpenAI's ChatGPT, have become mainstream—ranking in the top 5% of all news coverage topics, just in the year 2023. As they become increasingly used across all industries, LLMs are poised to augment creative and technical tasks alike.

Ultimately, LLMs aren't magic. They've been trained on huge amounts of data and these models have learned how to apply information about one context to another. This has made them smart autocomplete bots—generating coherent and relevant responses in most situations.

But setting aside the discussion and debate on whether LLMs can truly understand, interpret, and communicate, we, engineers, scientists, and users, must look at LLMs as smart assistants—tools—that will provide us with a gentle footing in all our tasks.

That said, this ebook assumes that you already have a basic understanding of how LLMs work and can build simple LLM applications. In the scope of this ebook, we're more interested in an architectural approach called Retrieval Augmented Generation (RAG), which helps provide additional context to enhance LLM responses by pulling in information from external databases or documents the user provides. This means each response now is more specific, contextual, and in-depth—instead of just relying on an LLM›s pre-learned information. It also addresses the problem of "hallucinations" to a great extent—along with enabling real-time context, in addition to user-provided information, and factuality of responses.

However, implementing an enterprise-level RAG system is rife with challenges. Firstly, there's no "go-to" framework that developers can use as a reference before they journey into this space. Then, there's very little research into productionizing these complex systems, including the scenarios to consider before and during this step. Lastly, how does one monitor and refine the system continuously after deployment?

This "ebook" aims to be your go-to guide for all things RAG-related. If you're a machine learning engineer, a data scientist, an AI researcher, or a technical product manager looking to educate, experiment with, and build enterprise-level RAG-powered LLM applications, this ebook can be a great guide for you to refer to. Having said that, if you're a grad student or a computer scientist enthusiast looking for a comprehensive resource to understand the nuances of an RAG system, this ebook can serve as a great starting point. The book is divided into six chapters:

Chapter 1 briefly introduces LLMs and RAG systems. The assumption here is that you're already familiar with the basics of generative models, how they differ from discriminative models, and how they work.

Chapter 2 details the challenges or pain points associated with RAG systems and some practical tips for addressing them.

Chapter 3 covers different prompting techniques that you can use to reduce hallucinations in your RAG applications.

Chapter 4 – consisting of many subchapters – explores chunking for RAGs, discusses vector embeddings and re-ranking techniques to improve retrieval, and provides tips on choosing the best vector databases for your RAG system. In the end, it offers a practical guide to starting your journey in building an enterprise-RAG system through architectural considerations.

Chapter 5 prepares you for productionizing your RAG system through a detailed walkthrough of 8 test case scenarios.

Chapter 6 concludes with different methods to observe and manage your RAG system after deployment.

Chapter 7 explores ways to improve RAG performance after deployment, ensuring your system is always effective.

We're confident that going through this comprehensive resource will better position you to experiment with LLMs and RAGs and appreciate the intricacies of such systems. Some of these concepts are relatively new, and something better and more interesting may emerge tomorrow. That said, the topics we've covered in the ebook are structured to build a foundation—a gentle footing—upon which you can confidently work towards building enterprise-level RAG systems. The concepts and ideas that you'll carry with you from here will remain evergreen. During this exercise, you'll also explore different ways in which the AI systems you build are safe, transparent, and secure—the linchpin of a good business—and be someone who customers can trust.

Written by Pratik Bhavsar