I POCKET GUIDES

Beginning Spring AI

A Quick Guide to AI Engineering in Spring

Andrew Lombardi · Joseph Ottinger

Apress®

Apress Pocket Guides

Apress Pocket Guides present concise summaries of cutting-edge developments and working practices throughout the tech industry. Shorter in length, books in this series aims to deliver quick-to-read guides that are easy to absorb, perfect for the time-poor professional.

This series covers the full spectrum of topics relevant to the modern industry, from security, AI, machine learning, cloud computing, web development, product design, to programming techniques and business topics too.

Typical topics might include:

- A concise guide to a particular topic, method, function or framework
- Professional best practices and industry trends
- A snapshot of a hot or emerging topic
- Industry case studies
- Concise presentations of core concepts suited for students and those interested in entering the tech industry
- Short reference guides outlining 'need-to-know' concepts and practices.

More information about this series at https://link.springer.com/bookseries/17385.

Beginning Spring AI

A Quick Guide to AI Engineering in Spring

Andrew Lombardi
Joseph Ottinger

Apress®

Beginning Spring AI: A Quick Guide to AI Engineering in Spring

Andrew Lombardi
Laguna Beach, CA, USA

Joseph Ottinger
Youngsville, NC, USA

ISBN-13 (pbk): 979-8-8688-1290-3 ISBN-13 (electronic): 979-8-8688-1291-0
https://doi.org/10.1007/979-8-8688-1291-0

Managing Director, Apress Media LLC: Welmoed Spahr
Acquisitions Editor: Melissa Duffy
Development Editor: Laura Berendson
Coordinating Editor: Gryffin Winkler

Cover designed by eStudioCalamar

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

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

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

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

If disposing of this product, please recycle the paper

For my wife, my artistic love, and our kids:
the philosophers, adventurers, and creators who
make life an endless adventure.

Table of Contents

About the Authors

Andrew Lombardi is a veteran entrepreneur and software engineer. He's been running the consulting firm Mystic Coders for 25 years, authored multiple outstanding books on Spring for Apress and WebSocket for O'Reilly, coding, speaking internationally, and offering technical guidance to companies as large as Airbus and companies as controversial and unique as Twitter 1.0. He firmly believes that the best thing he's done so far is being a great dad.

Joseph B. Ottinger is a distributed systems architect with experience in many cloud platforms. He was the editor-in-chief of both the *Java Developer Journal* and TheServerSide.com and has also contributed to many, many publications, open source projects, and commercial projects over the years, using many different languages (but primarily Java, Kotlin, Python, and JavaScript). He's also a previously published author online (with too many publications to note individually) and in print, through Apress.

About the Technical Reviewer

Manuel Jordan Elera is an autodidactic developer and researcher who enjoys learning new technologies for his own experiments and creating new integrations. Manuel won the Springy Award 2013 Community Champion and Spring Champion. In his little free time, he reads the Bible and composes music on his guitar. Manuel is known as dr_pompeii. He has tech-reviewed numerous books, including *Pro Spring MVC with WebFlux* (Apress, 2020), *Pro Spring Boot 2* (Apress, 2019), *Rapid Java Persistence and Microservices* (Apress, 2019), *Java Language Features* (Apress, 2018), *Spring Boot 2 Recipes* (Apress, 2018), and *Java APIs, Extensions and Libraries* (Apress, 2018). You can read his detailed tutorials on Spring technologies and contact him through his blog at `www.manueljordanelera.blogspot.com`. You can follow Manuel on his Twitter account, `@dr_pompeii`.

Acknowledgments

I would like to thank my family for supporting me during the writing of this book for the last few months along with crazy work running the company. I'd like to thank Joe for writing another book together which has turned out some pretty useful and interesting text as a team, our friends who remind me that there is also life outside of a screen and sunshine, rainbows, and the blinking cursor which is a prompt to keep going.

—Andrew Lombardi

I would like to thank whoever came up with the idea of thanking people in the front matter for books. Apparently, it's been done for pretty much the entire history of literature, originally used to credit deities and other patrons, so maybe we get to thank the purveyors of bad copper somehow.... This has not been a good acknowledgments paragraph, so let me try to rescue it somehow: I'd like to thank my family for their constant and undeserved encouragement, Andrew for putting up with my oddball ideas, the various small animals I keep throwing seeds to in my backyard for keeping me amused and appreciative, my wife, along with eden Hudson and Jess Astra for reminding me sometimes what a joy it is to read as well as write, and most of all, *you*, dear reader, for being willing to learn and grow with us. And spaghetti. I almost forgot to thank spaghetti.

—Joseph Ottinger

Introduction

Welcome to *Beginning Spring AI!*

Artificial intelligence has rapidly evolved over the last several years, and its influence now reaches into virtually every corner of modern software development. From generating text and images on-demand to analyzing audio content and extracting meaningful insights, AI is no longer the next frontier—it's a set of powerful tools ready to be applied right now. The Spring ecosystem, known for its robust and developer-friendly frameworks, has embraced this new era through Spring AI, a suite of libraries that serve as a gateway into the world of large language models (LLMs) and other advanced AI services.

What Is Spring AI?

Spring AI represents a cohesive set of abstractions and utilities that bridge your Spring-based applications with leading AI platforms. By simplifying complex integrations, Spring AI empowers you to connect to popular text, image, and audio models using a standardized approach—eliminating the need to learn multiple proprietary APIs or wrestle with inconsistent data formats. Instead, you can leverage a consistent, Spring-friendly programming model to interact with models like ChatGPT, stable diffusion image models, and speech-to-text engines, all within the familiar boundaries of your existing Spring projects.

What You'll Learn in This Book

In this straightforward yet comprehensive guide, you will gain the practical knowledge and hands-on experience necessary to start building AI-enhanced applications right away. We begin by walking through project setup and configuration, ensuring you have a solid foundation for adding AI functionalities to your Spring-based environment.

1. **Connecting to ChatGPT and Other Large Language Models**
   Learn how to integrate large language models into your workflow. From the straightforward task of handling simple text queries, you'll progress toward generating structured content suitable for predictable, repeatable outputs. You'll also see how LLMs can be granted controlled access to your proprietary data, enabling sophisticated chatbots and assistants capable of interacting with real-world systems—from inventory databases to IoT devices—based on user input.

2. **Audio Generation and Analysis**
   Move beyond text-based content and tap into the world of audio. Discover how to instruct an LLM to create audio outputs—from synthetic speech to sonic branding elements—and then learn how to analyze spoken words. This capability opens up possibilities such as creating voice interfaces, automated transcription services, and real-time sentiment analysis for customer support calls.

3. **Visual Content Creation and Interpretation**
   Unleash the power of image-based AI models. You'll see how to use LLMs to generate entirely new images or enhance existing ones. On the flip side, you'll learn how to have an AI model interpret and describe visual content, making it possible for your applications to "see" and understand the world around them, from recognizing products in a catalog to summarizing complex diagrams for accessibility.

4. **Ethical, Legal, and Cost Considerations**
   While the potential of AI is vast, it's essential to understand its implications. We'll address common concerns around the ethical use of AI, highlight privacy and compliance considerations, and discuss cost management strategies. You'll gain the insights needed to responsibly build and deploy AI solutions that respect user data, operate transparently, and control expenses.

Why Spring AI?

By the end of this book, you'll have a toolkit that allows you to interact with multiple AI services through a single, coherent approach. Spring AI's abstraction layers and its deep integration with the broader Spring ecosystem significantly reduce the learning curve. You'll be able to plug into cutting-edge AI models using a codebase that is both approachable and maintainable—letting you focus on creating value instead of wrestling with technical minutiae.

Your Journey Begins

As you turn the pages, you'll progress from foundational setup tasks to crafting sophisticated, AI-driven features. You'll discover not only how to leverage the capabilities of large language models but also how to integrate them seamlessly into real-world applications, from text-based assistants to voice-driven interfaces and vision-based analyzers.

Whether you're a seasoned Spring developer looking to add AI to your skill set, or you're new to the Spring ecosystem and want to start your journey at the cutting edge, this book is your guide to building smarter, more responsive, and more dynamic applications. Let's dive in and uncover the power of Spring AI together!

CHAPTER 1

Introduction

Welcome to *Beginning Spring AI*!

"Spring AI" refers to a suite of libraries within the Spring Framework designed to help programmers harness some of the most popular artificial intelligence (AI) technologies available today.

In this book, we'll take you on a guided tour of these libraries and their features. Our goal is not to provide a comprehensive reference on every feature but to offer enough information so you can see the potential of these technologies and use the majority of what you find most valuable.

Note This chapter does not contain code. It lays a foundational understanding of AI, providing key definitions and concepts, as well as an overview of the technologies covered in this book. If you prefer diving right into the code, feel free to skip ahead to Chapter 2. However, we suggest revisiting this chapter later on, as it contains valuable insights that will enhance your understanding.

AI Is Everywhere

It's nearly impossible to browse the Internet today without encountering AI in some form. Whether it's a simple search for information, or summarizing content, possibly writing prompts on Quora, or creating

A. Lombardi and J. Ottinger, *Beginning Spring AI*, Apress Pocket Guides,
https://doi.org/10.1007/979-8-8688-1291-0_1

visual arts or music, AI is being applied across numerous domains. There are even entire songs generated by AI that some may find quite respectable.

For programmers, most IDEs now come with AI integrations, suggesting code completions and improvements. With the right prompts, it's possible to have an AI generate substantial portions of a working application, causing some managers to wonder if they even need human engineers anymore.[1]

Authors also feel the impact of AI "writing." Many tools now suggest grammar and spelling corrections, as well as more nuanced changes, to evoke a specific tone or style. AI can even write stories, and when given detailed prompts, the results can sometimes pass for human-written content.

Note It's worth noting for the record: This book has indeed been impacted by AI, and we'll point out why and how later. We promise. At least, that's what the AI said to say. Or did it write it? Sometimes it's hard to tell.[2]

This raises an important question: Are humans being replaced by machines? If AI can create applications, music, or stories that rival human quality, then why do we need humans at all? There are several compelling answers to this, and some less convincing ones as well.

[1] Spoiler alert: Yes, the managers do need human engineers. We'll get to why soon, some in this chapter and some in Chapter 6.

[2] This is intended to be humor. If an AI generates any actual content of note in this book, we'll be pointing it out, even if it's not obvious. With that said, when we say AI is used in many writing tools, we mean it; a lot of our grammar was checked and occasionally fixed by AI. It's also worth noting that not even *one* of the footnotes was suggested by an AI—or, indeed, any human.