Spring AI: Building Intelligent Applications - TT3340

In-depth exploration of foundational AI principles, cutting-edge AI models, and practical implementation techniques, all within the robust Spring ecosystem.

Duration: 3 Days
Skill Level: Intermediate
Available Format: Instructor-Led Online ; On Public Schedule

What You'll Learn

Overview

Unlock the full potential of AI and Spring technologies with our comprehensive course designed for developers eager to master the latest advancements in artificial intelligence and application development. This course offers an in-depth exploration of foundational AI principles, cutting-edge AI models, and practical implementation techniques, all within the robust Spring ecosystem.

You will begin by delving into the core concepts of artificial intelligence, understanding how AI thinks, processes information, and generates human-like text. Through hands-on labs and real-world examples, you'll learn how to create effective prompts, harness the power of vector databases, and integrate Large Language Models (LLMs) to enhance AI capabilities.

The course covers practical skills in deploying AI solutions with Spring, from using the AI Model API to configuring Spring Boot for seamless integration of AI models and

vector stores. Additionally, you'll explore advanced features like function calling, chat client APIs, and multimodal AI models, ensuring you can build dynamic, interactive, and responsive applications.

By the end of this course, you'll have the knowledge and skills to implement, optimize, and evaluate AI-driven applications, leveraging the full power of Spring AI to drive innovation and efficiency in your projects. Whether you're looking to enhance existing applications or develop new AI-powered solutions, this course provides the tools and insights to help you succeed.

Objectives

This 'skills-centric' course is about 50% hands-on lab and 50% lecture, designed to train attendees in Java and Spring development skills, coupling the most current, effective techniques with the soundest industry practices. Working within in an engaging, hands-on learning environment, guided by our expert team, attendees will learn:

- Grasp the core principles of artificial intelligence and how AI processes information.
- Learn about vector databases and their significance in AI applications.
- Understand Large Language Models (LLMs) and their applications in generating human-like text.
- Create effective prompts to optimize AI model performance.
- Use prompt templates to streamline the prompt creation process.
- Test and refine prompts to achieve desired AI responses.
- Utilize the Spring AI Model API for integrating various AI models.
- Configure Spring Boot to support AI model integration and vector stores.
- Understand the role of ETL (Extract, Transform, Load) processes in data preparation for AI models.
- Create and configure a ChatClient using Spring's Chat Client API.
- Implement different response types and streaming responses for real-time interactions.
- Use system text, advisors, and chat memory to enhance the flexibility and personalization of chat responses.
- Explore the capabilities of multimodal LLMs, integrating text, images, audio, and video.
- Implement multimodal models using Spring AI's tools and APIs.
- Understand the benefits of multimodal approaches over traditional single-modality models.

- Evaluate the content generated by AI models to ensure accuracy and relevance.
- Use the RelevancyEvaluator to assess the relevance of AI-generated content.
- Implement JUnit tests to perform Retrieval Augmented Generation (RAG) and validate responses.
- Set up Docker environments for Spring applications to enhance scalability and deployment.
- Integrate vector databases within Docker containers for efficient data retrieval.
- Build and deploy native images using GraalVM for improved performance and reduced memory footprints.

Audience

This is course is for experienced Java developers with a solid understanding of Spring Boot. The class is taught using Java 21, ensuring participants leverage the latest features and best practices in their AI-driven projects.

Pre-Requisites

This is course is for experienced Java developers with a solid understanding of Spring Boot. The class is taught using Java 21, ensuring participants leverage the latest features and best practices in their AI-driven projects.

Agenda

1. Exploring Artificial Intelligence

- Explore the foundational principles of AI.
- Gain an in-depth understanding of how AI thinks and processes information.
- Learn how vector databases enhance AI capabilities by efficiently managing high-dimensional data.
- Understanding Large Language Models (LLMs), discovering how they generate human-like text and their applications.
- Understand the concept of tokens in AI models
- Explore the techniques for effectively interacting with large language models to achieve desired outcomes.

2. AI Fundamentals: Exploring Core Concepts

- Introduction to AI Models: The basics of building and fine-tuning AI models that drive smart applications.

- Learn the fundamentals of creating impactful prompts and templates to enhance AI performance.
- Discover the foundational concepts of embeddings and how they transform data into actionable insights.
- Gain a basic understanding of the roles messages play in improving AI system functionality.
- Get introduced to the techniques that boost AI-generated content relevance and efficiency.
- Understand the basics of function calling to create dynamic and responsive AI applications.

3. Prompt engineering for developers

- Introduce the foundational concepts of prompt engineering and its importance in AI development.
- Learn about various types of prompts and their applications in different AI scenarios.
- Gain insights into crafting prompts that optimize AI model performance and accuracy.
- Discover how to use prompt templates to streamline the prompt creation process and maintain consistency.
- Learn techniques for testing and refining prompts to achieve the desired AI responses.
- Lab: AI Prompts for Developers

4. Getting Started with Spring AI

- Introduction to Spring AI: Understand its role in simplifying AI development for Java developers.
- Learn the basics of the AI Model API and how it facilitates interaction with various AI models.
- Discover how Spring Boot's auto-configuration and starters streamline the integration of AI models and vector stores.
- Understand the concept of vector stores and their application within the Spring framework.
- Explore the Function Calling API and its use in enabling dynamic and responsive AI applications.
- Get introduced to the fundamentals of ETL (Extract, Transform, Load) data engineering and its importance in preparing data for AI models.

5. Exploring Spring's Chat Client API

- Learn how to create and configure a ChatClient to build interactive chat applications.
- Use an autoconfigured ChatClient.Builder to streamline the setup and customization of your ChatClient
- Explore the different types of responses supported by ChatClient
- Work with streaming responses to enable real-time, dynamic interactions in your chat applications.
- Utilize default system text with parameters to enhance the flexibility and personalization of chat responses
- Learn about advisors and how they can be used to manage and influence the behavior of your ChatClient.
- Explore the concept of chat memory and its application in maintaining context and continuity in chat conversations.
- Lab: Building a Chat Client

6. Crafting Effective Prompts

- Explore the structure and significance of the Prompt class in crafting effective interactions.
- Discover how the Message interface facilitates seamless communication within your AI applications.
- Roles in Prompt Engineering: System, Assistant, and Function: Learn about the distinct roles in prompt engineering
- Creating Dynamic Prompts with PromptTemplate
- Understand the benefits of using resource files instead of raw strings for more maintainable and scalable prompt management.
- Creating Prompts:Reducing extensive text into concise summaries that capture essential points and main ideas.
- Creating Prompts: Extract precise answers from provided text based on user questions.
- Creating Prompts: Ctegorizing text into predefined groups, enhancing the organization and analysis of content.
- Creating Prompts: Develop interactive dialogues where AI engages in natural, back-and-forth communication with users.
- Creating Prompts: Discover how to translate user requirements or descriptions into functional code snippets.
- Implementing Zero-shot and Few-shot Learning: Enable AI to understand and respond to new tasks with minimal to no prior examples.

- Maintaining Context with Chain-of-Thought: Link multiple AI responses to create coherent and contextually aware conversations, maintaining the thread of discussion.
- Combining Reasoning and Action with ReAct: Explore the ReAct method where AI first analyzes the input and then determines the best course of action.
- Lab: Crafting Effective Prompts

7. Transforming AI Responses

- Discover why the ability of Language Learning Models (LLMs) to produce structured outputs is crucial for applications that depend on accurate data parsing.
- Learn how Spring AI Structured Output Converters transform LLM outputs into structured formats, ensuring reliability and consistency.
- Explore the Structured Output API
- Gain insights into using Bean Output Converters and generic bean types to streamline the conversion process of LLM outputs.
- Explore the functionalities of Map and List Output Converters in organizing and structuring AI-generated data.
- Produce structured outputs, for example in JSON format, to enhance the usability and integration of AI-generated data.
- Lab: Transforming AI Responses

8. Exploring the Chat Model API

- ChatModel: The AI model behind the scenes, understanding the message structure and generating responses.
- Delve into the AI model that powers chat interactions, learning how it interprets message structures and generates responses.
- Compare and contrast the functionalities of ChatModel and StreamingChatModel, understanding their use cases and benefits.
- Explore the ChatResponse Class which encapsulates the AI model's output
- Gain insights into the Generation class, an extension of ModelResult, which represents the AI-generated response and includes relevant metadata.
- Discover how to customize AI model interactions by passing various options (e.g. Temperature and TopP) to the ChatModel.
- Lab: Chat Model API

9. Adding your own data (Stuffing the prompt)

- Learn the basics of prompt stuffing and its importance in tailoring AI responses to your specific data needs.
- Understand the inherent limitations of AI models, including their reliance on publicly available knowledge up to a certain date and lack of access to private corporate data.
- Discover how you can fine-tune AI models by incorporating your own data to enhance their accuracy and relevance to your specific use cases.
- Explore practical techniques for adding your own data into prompts, ensuring that the AI model generates responses that are aligned with your requirements.
- Learn strategies for optimizing the length and content of your prompts to balance detail with performance.
- Understand how to use function calling within your prompts to achieve more dynamic and responsive AI interactions.
- Lab: Stuffing the Prompt

10. Retrieval Augmented Generation

- Understand the concept of RAG and its significance in enhancing AI model responses by integrating retrieval mechanisms.
- Learn the basics of the Extract, Transform, and Load (ETL) framework and its role in preparing data for retrieval processes.
- Store and manage data in a way that supports efficient retrieval for AI applications.
- Learn the basics of vector databases and their role in efficiently storing and retrieving high-dimensional data for AI applications.
- Discover the VectorStore interface and how it facilitates interaction with vector databases to support RAG.
- Gain insights into initializing and configuring the schema for vector databases to optimize data storage and retrieval.
- Learn how to use metadata filters to refine and enhance data retrieval processes, ensuring more relevant and accurate AI responses.
- Practical steps for integrating vector databases into your RAG system
- Explore the concept of embeddings and how they transform raw data into a format that can be effectively used in retrieval and generation tasks.
- Retrieving relevant data based on user input and incorporating it into prompts to improve AI responses.
- Connect and utilize data sources to provide the AI with comprehensive and up-to-date information
- Lab: RAG

11. Command-Line Interfaces with Spring Shell

- Discover the basics of Spring Shell and its role in creating interactive command-line applications with Spring Boot
- Configure a Spring Shell environment to start building your own CLI applications
- Explore how to create and implement custom commands that allow users to interact with your application through the command line.
- Understand how to use Spring Shell's annotations to define commands
- Use Spring Shell to load your own data into a Vector database
- Lab: Spring Shell

12. Introduction to Spring Docker

- Understand the basics of Docker and how it integrates with Spring
- Set up a Docker environment for your Spring application
- Understand how to incorporate vector databases within Docker containers to support efficient data retrieval for RAG.
- Learn how to implement Extract, Transform, and Load (ETL) processes in Dockerized Spring applications to prepare data for retrieval.
- Lab: Running the Vector database in a container

13. Transforming Images with Spring AI

- Introduction to the Spring Image Model API
- Create, configure, and manage image models using the Spring Image Model API.
- Explore techniques for processing images, including transformations, enhancements, and analysis using the Spring Image Model API.
- Lab: Generating images with Spring AI
- Lab: Analyzing images using Spring AI

14. Spring's Audio Model API

- Introduction to Spring's Audio Model API
- Discover how to create, configure, and manage audio models using the Spring Audio Model API

- Utilize OpenAI's Transcription Model for converting speech to text, enhancing audio data usability.
- Leverage OpenAI's Speech API to add advanced speech processing capabilities to your applications
- Lab: Speech Transcriptions
- Lab: Text-To-Speech

15. Spring AI Multimodality

- Learn the foundational principles of multimodality, its importance in human learning, and its application in modern AI models.
- Understand the transition from specialized single-modality models to contemporary multimodal large language models (LLMs).
- Discover the capabilities of multimodal LLMs, such as OpenAI's GPT-4 Vision and GPT-4o
- Understand how Spring AI supports multimodality through its Message API, providing necessary abstractions for integrating various data formats.
- Gain practical skills in implementing and working with multimodal models using Spring AI's tools and APIs.
- Lab: Multimodality

Additional Topic: Time Permitting

These topics will be included in your course materials but may or may not be presented during the live class depending on the pace of the course and attendee skill level and participation.

16. Evaluation Techniques for Spring AI

- Evaluate the content generated by AI models, ensuring it is accurate and free from hallucinations.
- Use AI models for self-evaluation, selecting the best model for the task

All applicable course software, digital courseware, labs, data sets, solutions, and extended post-training resources are delivered through our “easy access, single source, no install required” online Learning Experience Platform (LXP). Courses also incorporate use of widely available AI tools to enhance the learning process and provide practical, real-world experience. Access periods vary by course. We’ll ensure participants are set up and ready well before class begins, with clear guidance on how to get started.

For More Information

Please contact us or call 844-475-4559 toll free for more information about our training services (instructor-led, self-paced or blended), coaching and mentoring services, public course enrollment or questions, partner programs, courseware licensing options and more.