

Introduction

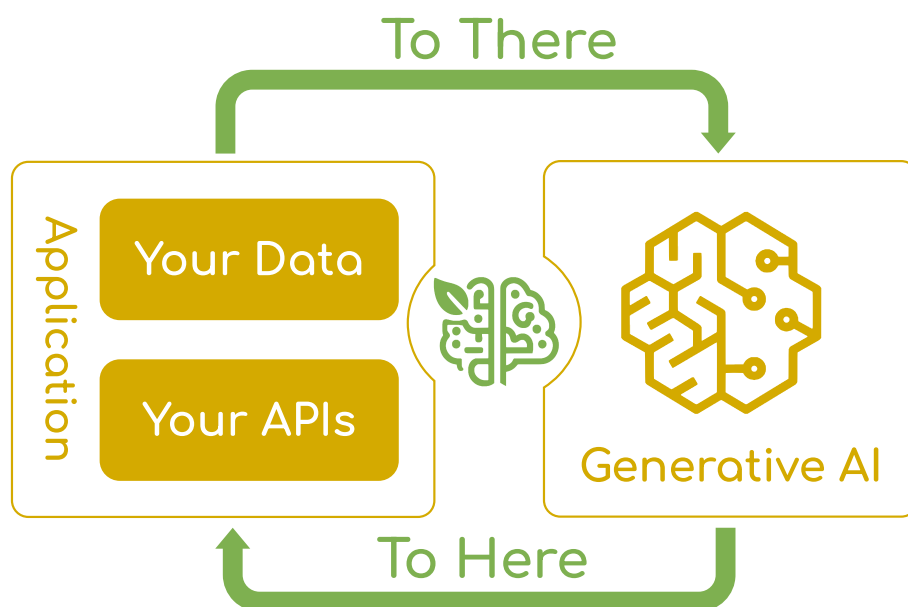


The Spring AI project aims to streamline the development of applications that incorporate artificial intelligence functionality without unnecessary complexity.

The project draws inspiration from notable Python projects, such as LangChain and LlamaIndex, but Spring AI is not a direct port of those projects. The project was founded with the belief that the next wave of Generative AI applications will not be only for Python developers but will be ubiquitous across many programming languages.

NOTE

Spring AI addresses the fundamental challenge of AI integration: Connecting your enterprise Data and APIs with the AI Models.



Spring AI provides abstractions that serve as the foundation for developing AI applications. These abstractions have multiple implementations, enabling easy component swapping with minimal code changes.

Spring AI provides the following features:

- Portable API support across AI providers for Chat, text-to-image, and Embedding models. Both synchronous and streaming API options are supported. Access to model-specific features is also available.
- Support for all major AI Model providers such as Anthropic, OpenAI, Microsoft, Amazon, Google, and Ollama. Supported model types include:
 - Chat Completion
 - Embedding
 - Text to Image
 - Audio Transcription
 - Text to Speech
 - Moderation
- Structured Outputs - Mapping of AI Model output to POJOs.
- Support for all major Vector Database providers such as Apache Cassandra, Azure Cosmos DB, Azure Vector Search, Chroma, Elasticsearch, GemFire, Milvus, MongoDB Atlas, Neo4j, OpenSearch, Oracle, PostgreSQL/PGVector, PineCone, Qdrant, Redis, SAP Hana, Typesense and Weaviate.
- Portable API across Vector Store providers, including a novel SQL-like metadata filter API.
- Tools/Function Calling - permits the model to request the execution of client-side tools and functions, thereby accessing necessary real-time information as required.
- Observability - Provides insights into AI-related operations.
- Document injection ETL framework for Data Engineering.
- AI Model Evaluation - Utilities to help evaluate generated content and protect against hallucinated response.
- Spring Boot Auto Configuration and Starters for AI Models and Vector Stores.
- ChatClient API - Fluent API for communicating with AI Chat Models, idiomatically similar to the WebClient and RestClient APIs.
- Advisors API - Encapsulates recurring Generative AI patterns, transforms data sent to and from Language Models (LLMs), and provides portability across various models and use cases.
- Support for Chat Conversation Memory and Retrieval Augmented Generation (RAG).

This feature set lets you implement common use cases such as "Q&A over your documentation" or "Chat with your documentation."

The concepts section provides a high-level overview of AI concepts and their representation in Spring AI.

The [Getting Started](#) section shows you how to create your first AI application. Subsequent sections delve into each component and common use cases with a code-focused approach.



Copyright © 2005 - 2024 Broadcom. All Rights Reserved. The term "Broadcom" refers to Broadcom Inc. and/or its subsidiaries.

[Terms of Use](#) • [Privacy](#) • [Trademark Guidelines](#) • [Thank you](#) • [Your California Privacy Rights](#) • [Cookie Settings](#)

Apache®, Apache Tomcat®, Apache Kafka®, Apache Cassandra™, and Apache Geode™ are trademarks or registered trademarks of the Apache Software Foundation in the United States and/or other countries. Java™, Java™ SE, Java™ EE, and OpenJDK™ are trademarks of Oracle and/or its affiliates. Kubernetes® is a registered trademark of the Linux Foundation in the United States and other countries. Linux® is the registered trademark of Linus Torvalds in the United States and other countries. Windows® and Microsoft® Azure are registered trademarks of Microsoft Corporation. "AWS" and "Amazon Web Services" are trademarks or registered trademarks of Amazon.com Inc. or its affiliates. All other trademarks and copyrights are property of their respective owners and are only mentioned for informative purposes. Other names may be trademarks of their respective owners.