SPRING IO

#springio24

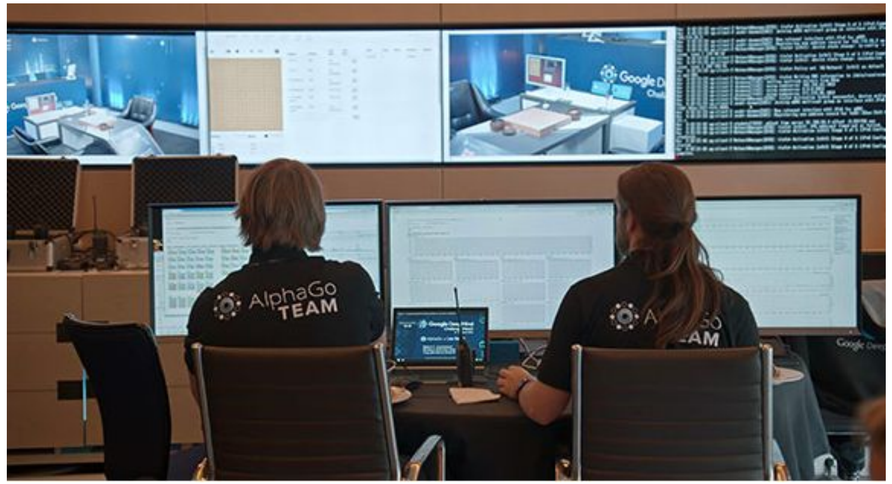
Introducing Spring AI

Christian Tzolov and Mark Pollack

Why are we at this talk?

ChatGPT

Pre ChatGPT AI Experience



Why are we at this talk?

ChatGPT

Explored code generation

```
> spring ai add "JPA functionality with an integration test. Include
all Java code in the same package."
```

Creates README.md

What makes this wave of Generative AI unique?

Accessible Web API

Human Language as the interface

No training required

You don't need to be a data scientist to solve your use cases.

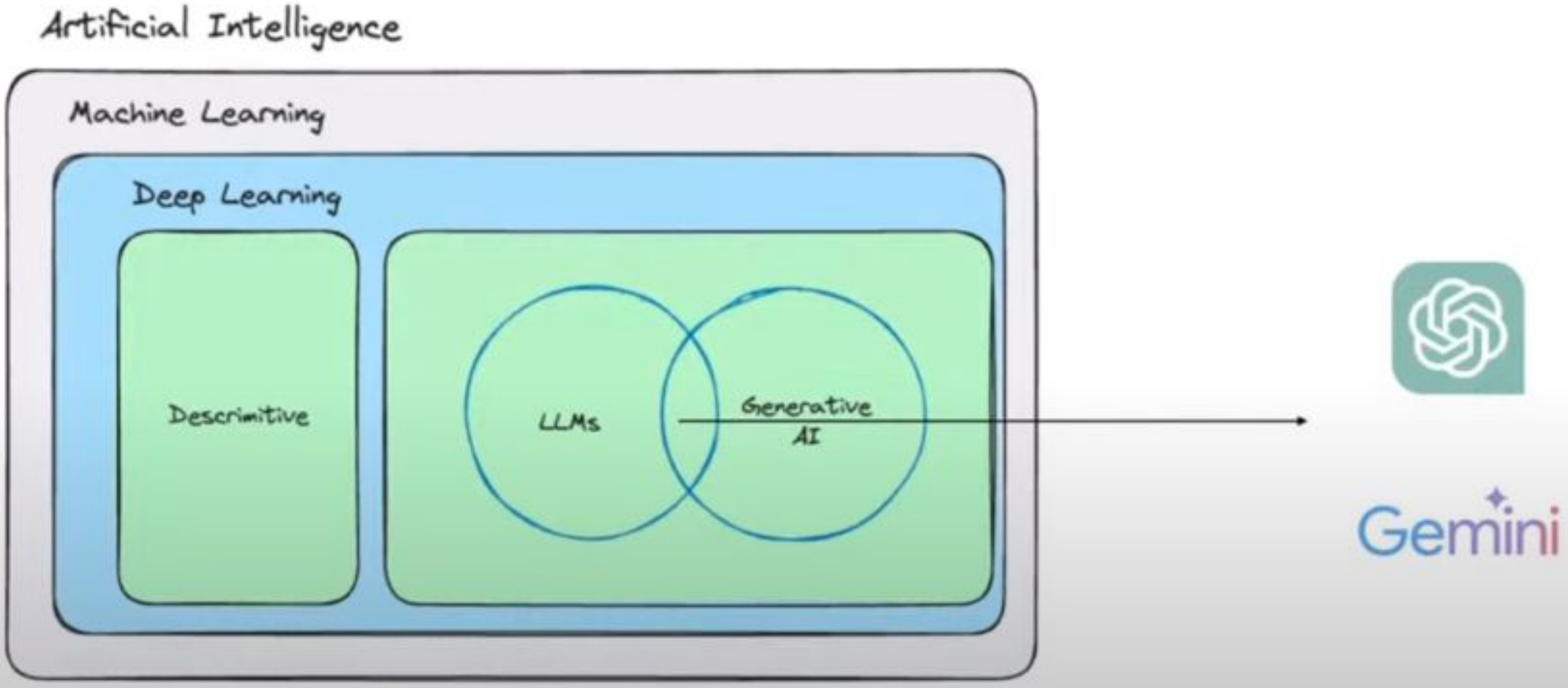
A seasoned Data Scientist perspective

Daniel Whitenack: Practical AI Podcast

> “I can solve a lot of the problems that I need to solve without doing any training at all”

> “Now I’m doing this sort of software engineering around the information that goes into a generative model”

GENERATIVE AI



Capabilities

Large general knowledge base

Ability to create new original content

Beats the majority of human level scores in exams

Versed in multiple modalities

Instant summarization

Can align responses based on your goals

Understands 20+ programming languages

Constraints

Stateless APIs

Limited size of data you can send

No structured output

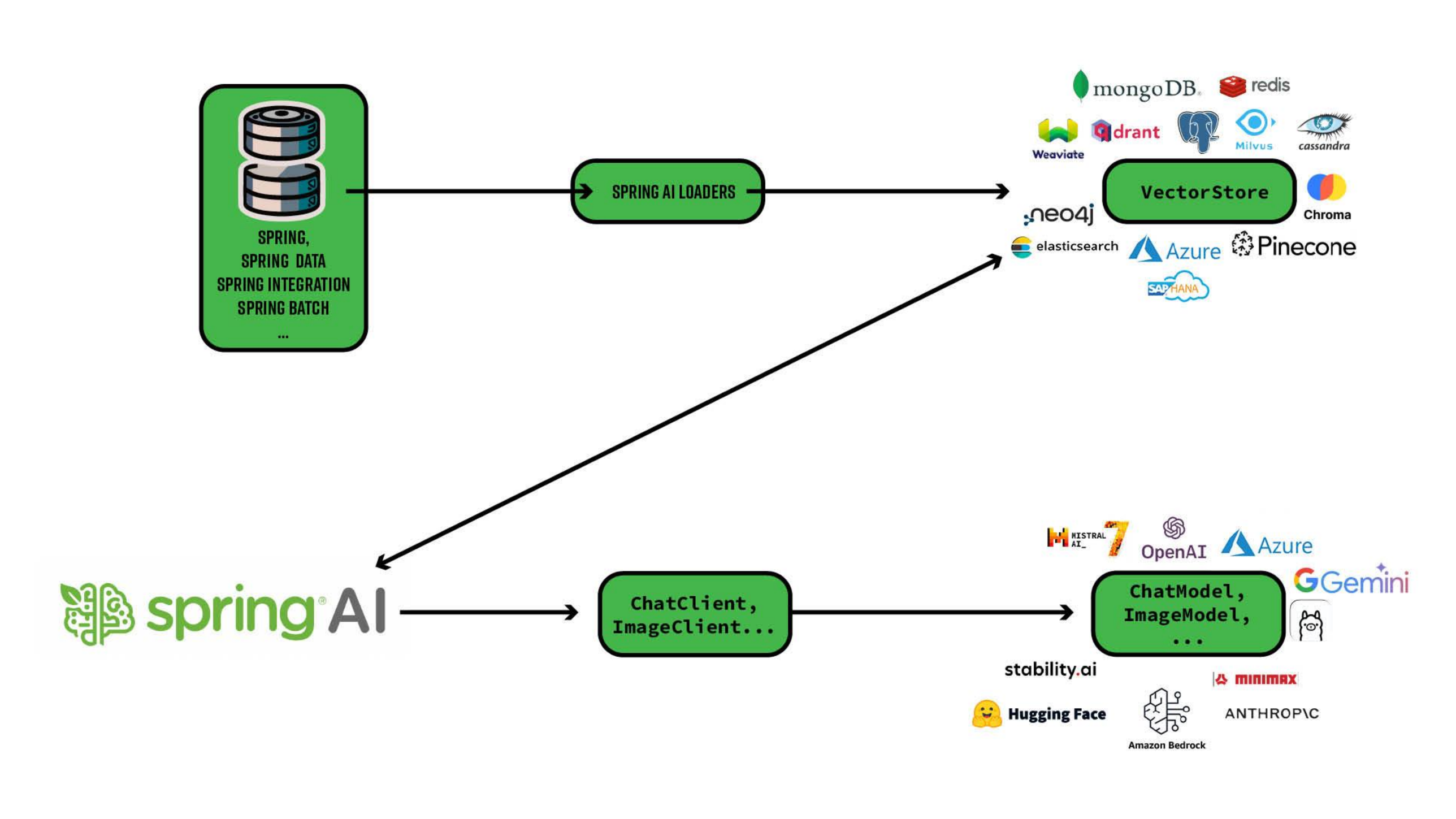
Not trained on your data

Not aware of your APIs

Prone to hallucinations

How to use this technology?

My Data & APIs


SPRING,
SPRING DATA
SPRING INTEGRATION
SPRING BATCH
...
SPRING AI LOADERS
mongoDB.
redis
Weaviate
drant
Milvus
cassandra
neo4j
VectorStore
Chroma
elasticsearch
Azure
Pinecone
SAP HANA
spring AI
ChatClient,
ImageClient...
MISTRAL AI_
OpenAI
Azure
Gemini
ChatModel,
ImageModel,
...
stability.ai
MINIMAX
Hugging Face
Amazon Bedrock
ANTHROP\C

How to get started?

Spring AI: **https://spring.io/projects/spring-ai**

https://start.spring.io

Hello World

SpringAI.java

```java
public class SpringAI {

  private ChatClient chatClient;

  public SpringAI(ChatClient.Builder builder) {
    this.chatClient = builder.build();
  }

  public String tellMeAJoke() {
    return chatClient.prompt()
        .user("Tell me a joke")
        .call()
        .content();

  }
}
```

Demo

Can align responses based on your goals

System Prompt

No structured output

Output Converters

Not trained on your data

Prompt Stuffing

RAG

Prone to hallucinations

Evaluation

Customize Behaviour

SpringAI.java

```java
public class SpringAI {

  private ChatClient chatClient;

  public SpringAI(ChatClient.Builder builder) {
    this.chatClient = builder.build();
  }

  public String tellMeAJoke() {
    return chatClient.prompt()
        .system("""
          You are a friendly assistant that answers
          questions in the voice of a Pirate.
          """)
        .user("Tell me a joke")
        .call()
        .content();

  }
}
```

Design vs Runtime

SpringAI.java

```java
public class SpringAI {

  private ChatClient chatClient;

  public SpringAI(ChatClient.Builder builder) {
    this.chatClient = builder
        .defaultSystem("""
            You are a friendly assistant that answers
            questions in the voice of a {voice}.
            """)
        .build();
  }

  public String tellMeAJoke(String voice) {
    return chatClient.prompt()
        .system(sp → sp.param("voice", voice))
        .user("Tell me a joke")
        .call()
        .content();

  }
}
```

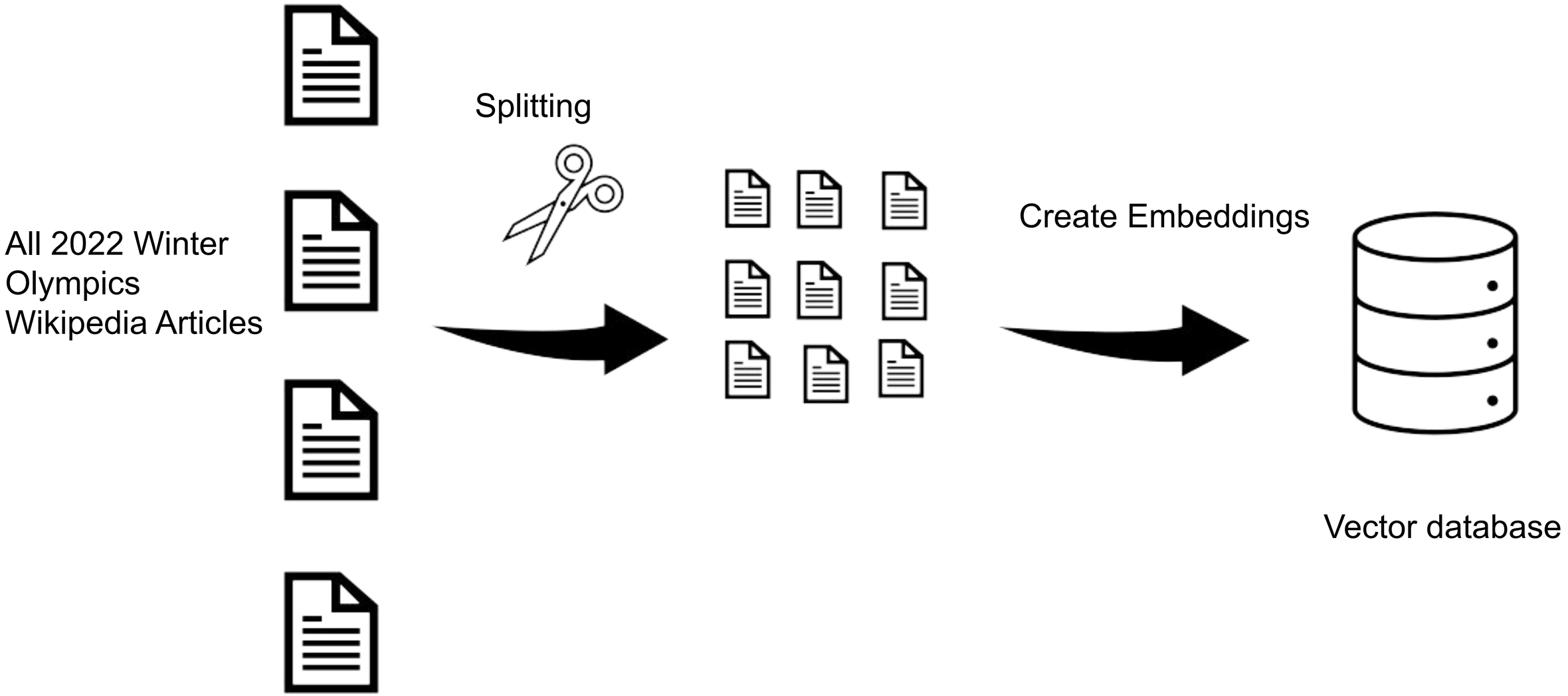
Chat Model Context Window Sizes

GPT-3.5 Turbo

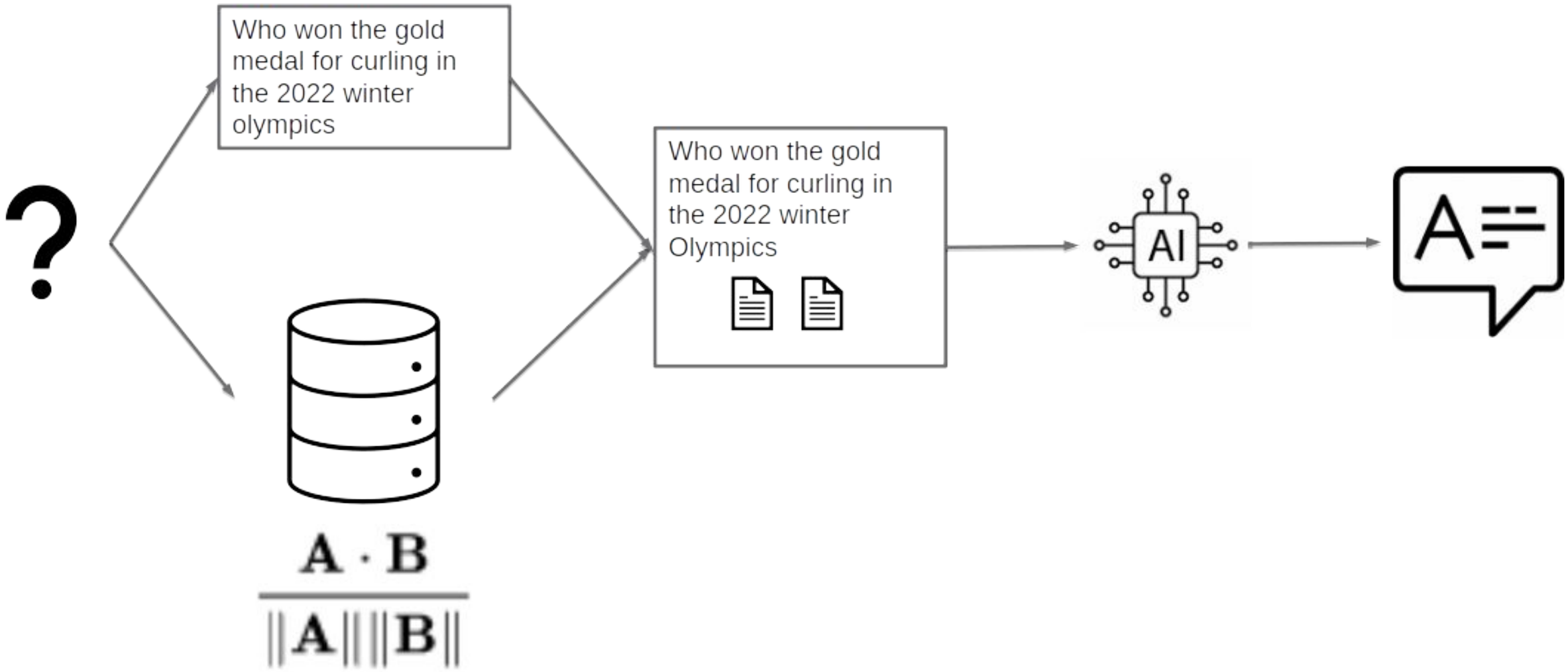
GPT-3.5 Turbo models can understand and generate natural language or code and have been optimized for chat using the Chat Completions API but work well for non-chat tasks as well.

MODEL	DESCRIPTION	CONTEXT WINDOW	TRAINING DATA
`gpt-3.5-turbo-0125`	New **Updated GPT 3.5 Turbo** The latest GPT-3.5 Turbo model with higher accuracy at responding in requested formats and a fix for a bug which caused a text encoding issue for non-English language function calls. Returns a maximum of 4,096 output tokens. Learn more.	16,385 tokens	Up to Sep 2021
`gpt-3.5-turbo`	Currently points to `gpt-3.5-turbo-0125`.	16,385 tokens	Up to Sep 2021
`gpt-3.5-turbo-1106`	GPT-3.5 Turbo model with improved instruction following, JSON mode, reproducible outputs, parallel function calling, and more. Returns a maximum of 4,096 output tokens. Learn more.	16,385 tokens	Up to Sep 2021
`gpt-3.5-turbo-instruct`	Similar capabilities as GPT-3 era models. Compatible with legacy Completions endpoint and not Chat Completions.	4,096 tokens	Up to Sep 2021

Retrieval Augmented Generation - Loading



Retrieval Augmented Generation - RAG



Conversational Memory

Stateless APIs

Function Calling

Not aware of your APIs

THANKS!!!

https://spring.io/projects/spring-ai