



Concerto for Java and AI

Building Production-Ready LLM Applications

Thomas Vitale
GOTO Copenhagen
Oct 2nd, 2024

@vitaletomas

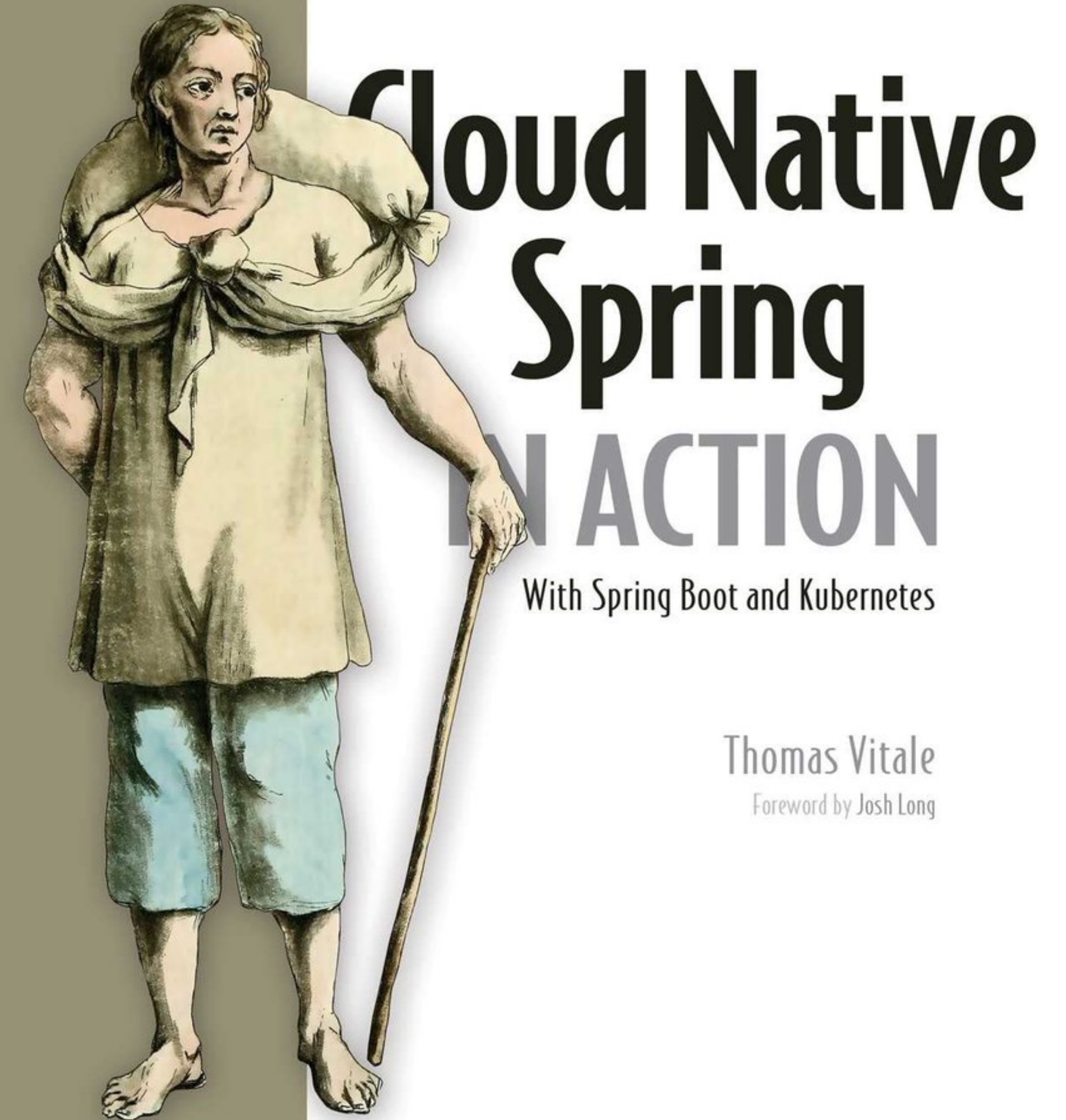
Thomas Vitale

Systematic

- **Software Engineer**
- CNCF Ambassador, Oracle ACE Pro, Testcontainers Community Champion
- Author of “**Cloud Native Spring in Action**” (Manning).
- **OSS contributor** (Java, Spring, Cloud Native Technologies)

thomasvitale.com

@vitaletomas



LLM

Agents

Vector Stores

Prompting

Generative AI

Hallucinations

Embeddings

RAG

A close-up photograph of a lit match. The matchstick is dark and textured, with a bright yellow flame at the tip. The background is dark and out of focus, emphasizing the flame and the match.

One Buzzword To Rule Them All



The WHY Factor

The **WHY** Factor

What problem does it solve?

How ready is it for production?

Yeah, but how about the DevEx?

Large Language Models

Machine Learning

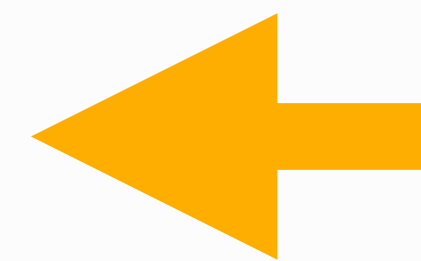
Subset of Artificial Intelligence



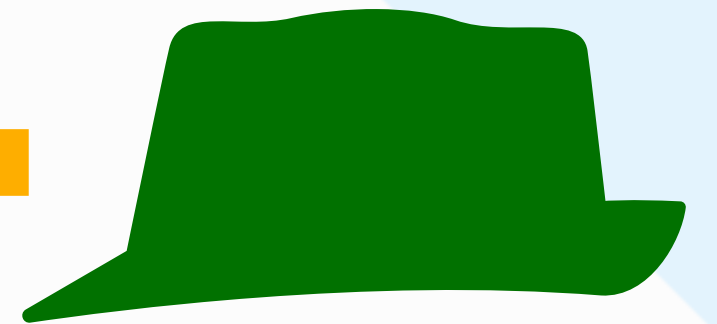
ML Engineers



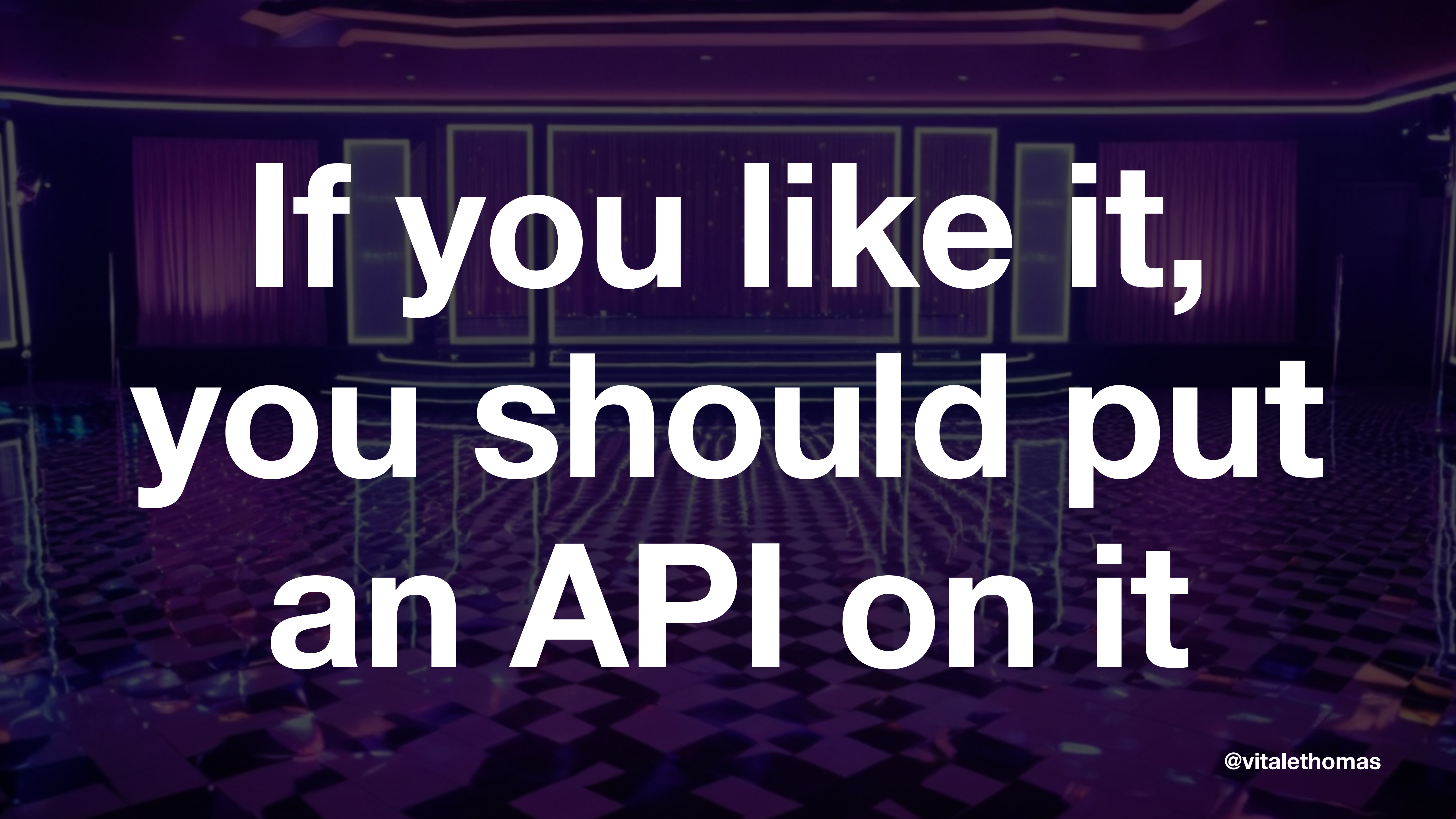
HTTP API



Application Developer



Platform Engineers

The background of the image is a dark, atmospheric room. The floor is covered in a checkered tile pattern, which is illuminated by a low, warm light source, creating a shimmering, reflective effect. In the background, there is a row of windows or glass doors, each framed by dark wood or metal. The overall color palette is dominated by deep purples, blues, and blacks, with some lighter, golden-brown highlights from the floor's reflection.

**If you like it,
you should put
an API on it**

Model Inference via HTTP APIs



Text Classification

Text Classification



LLM Security Risks (1)

OWASP Top 10 for LLM

Prompt Injection

Sensitive
Information
Disclosure

Model
Denial of Service

Resilience

LLM Applications

```
spring:  
  threads:  
    virtual:  
      enabled: true
```

```
@CircuitBreaker  
@RateLimiter  
@TimeLimiter  
public Type classify(String textToClassify) {  
    return chatClient.prompt()  
        .user(textToClassify)  
        .call()  
        .entity(Type.class);  
}
```

Resilience4J

```
spring:  
  ai:  
    retry:  
      max-attempts: 5  
      backoff:  
        multiplier: 5
```

Observability

Distributed Tracing

Request Rate
Errors
Duration

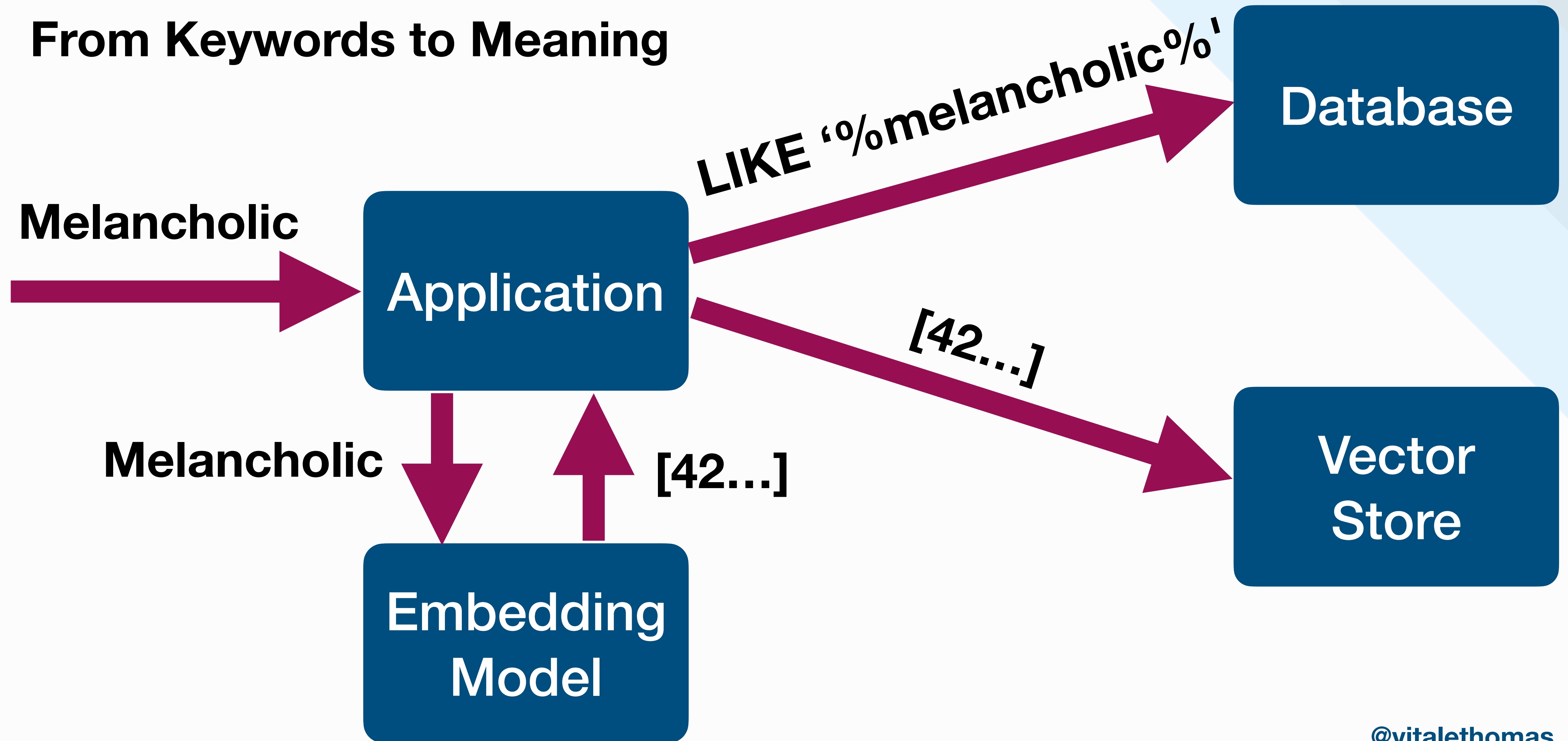
Token Usage
Context Window

Prompt Content

Semantic Search

Semantic Search

From Keywords to Meaning



Data Ingestion

LLM Applications

Document
Reader

Document
Transformer

Document
Writer

```
@Job(name = "Data Ingestion Pipeline")
@Recurring(id = "data-ingestion", cron = "*/5 * * * *")
public void runDataIngestion() {
    ...
}
```

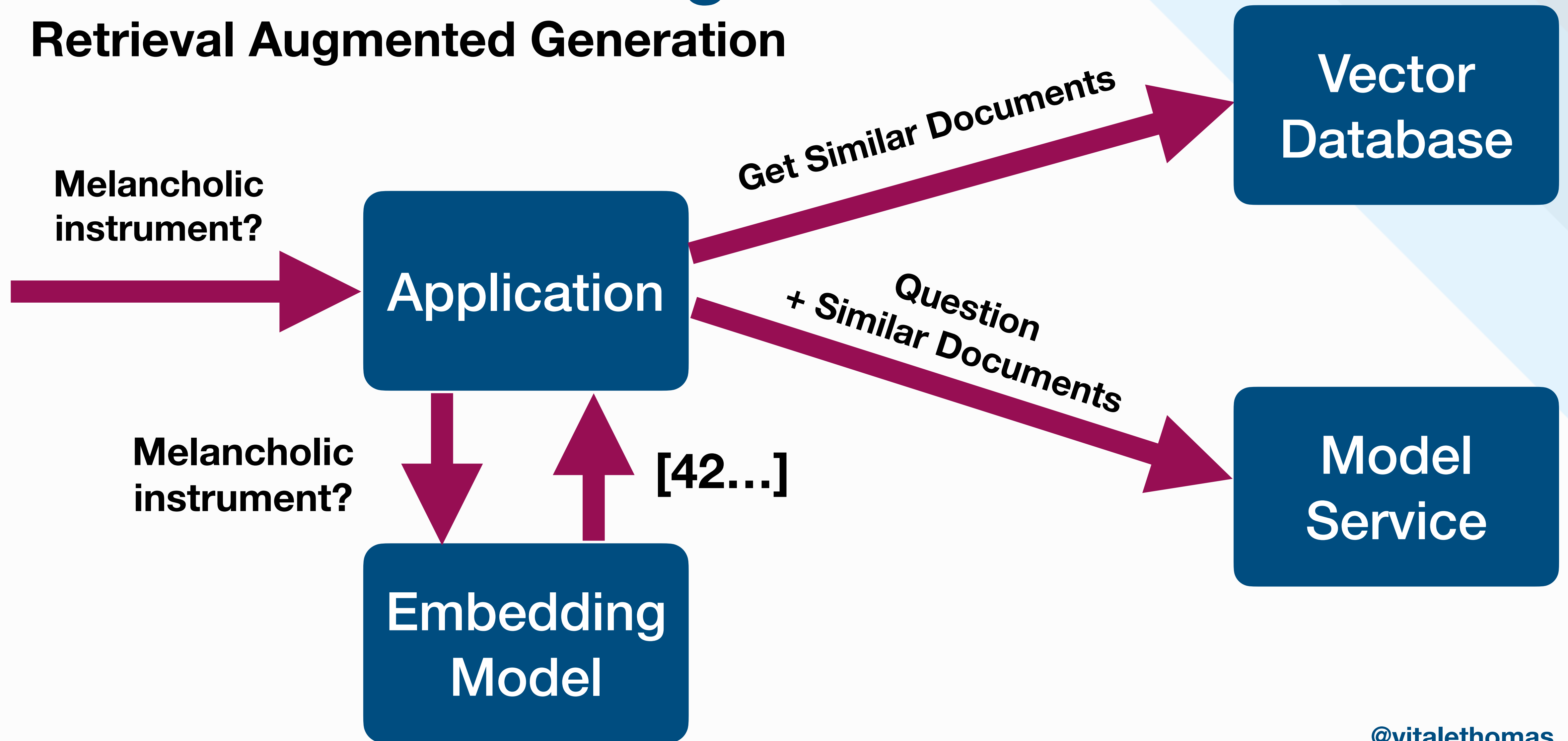
JobRunr



Question Answering

Question Answering with Docs

Retrieval Augmented Generation



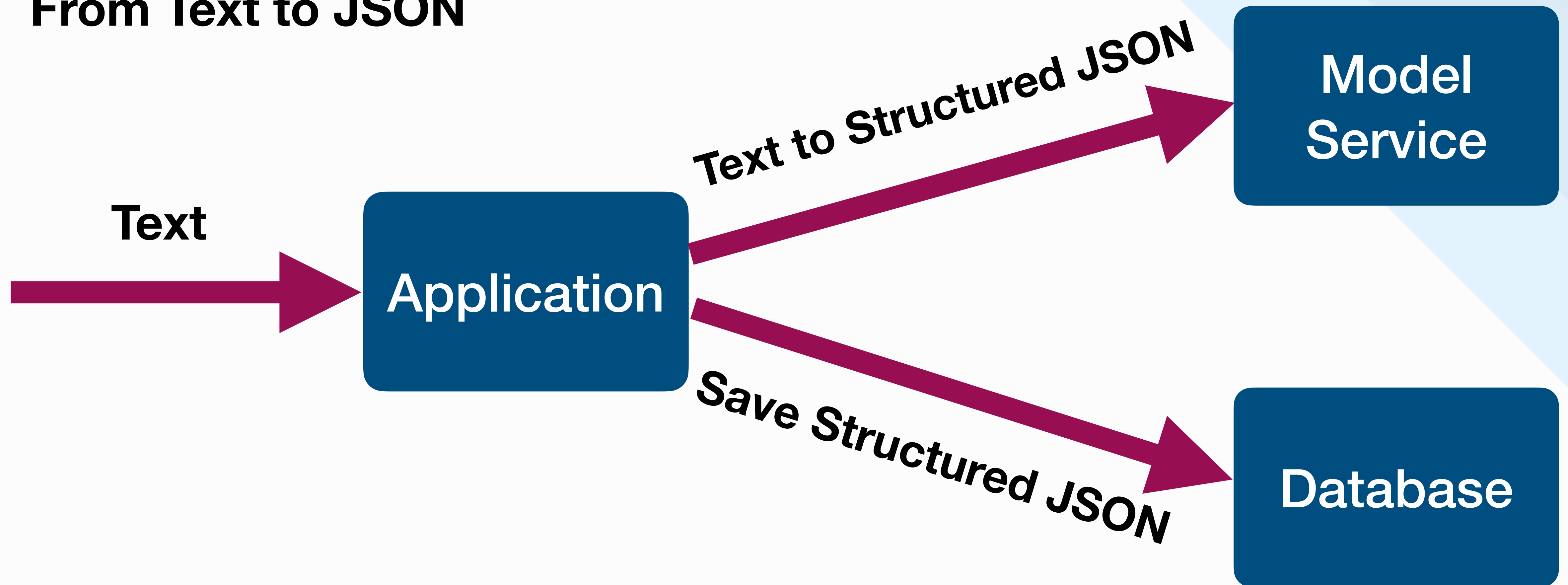
A dramatic background image of a volcano erupting at night. A massive plume of white smoke and ash rises from the crater, illuminated from below by the intense orange and yellow glow of the lava fire. Lava flows are visible on the dark slopes of the volcano.

One Does Not Simply Test LLM Applications

Structured Data Extraction

Structured Data Extraction

From Text to JSON



LLM Security Risks (2)

OWASP Top 10 for LLM

Insecure
Output
Handling

Insecure
Plugin
Design

Excessive
Agency



Hallucinations

Data Validation

Mitigating hallucination risks

JSON
Schema

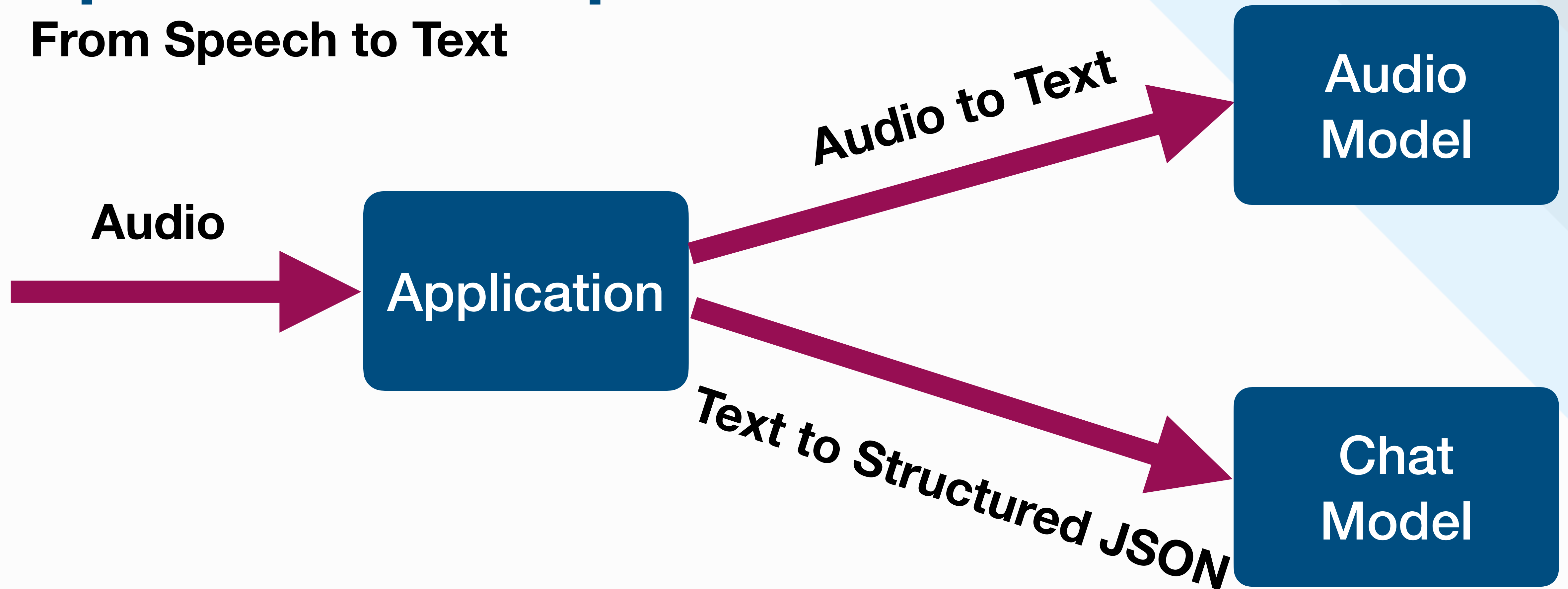
Optional
Values

Humans
in the Loop

Speech Transcription

Speech Transcription

From Speech to Text





Privacy

Build & Deploy

Going to Production

Cloud Native
Buildpacks

Native
Executables
with GraalVM

Kubernetes
Service
Binding

Service Bindings for Spring AI

```
dependencies {  
    implementation 'org.springframework.ai:spring-ai-spring-cloud-bindings'  
}
```

```
apiVersion: servicebinding.io/v1beta1  
kind: ServiceBinding  
spec:  
  service:  
    apiVersion: v1  
    kind: Secret  
    name: ollama  
  workload:  
    apiVersion: serving.knative.dev/v1  
    kind: Service  
    name: spring-ai-application
```

Composer Assistant

<https://github.com/ThomasVitale/concerto-for-java-and-ai>

Thomas Vitale

@vitalethomas

thomasvitale.com



Concerto for Java and AI

Building Production-Ready LLM Applications