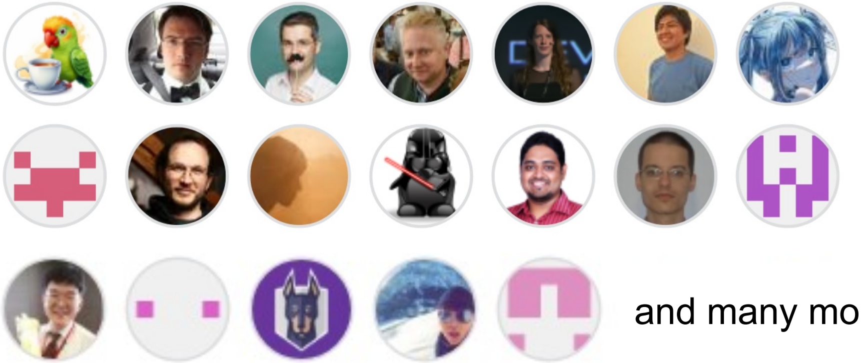
Java meets AI

How to Build LLM-Powered Applications with LangChain4j

The LangChain4j Team

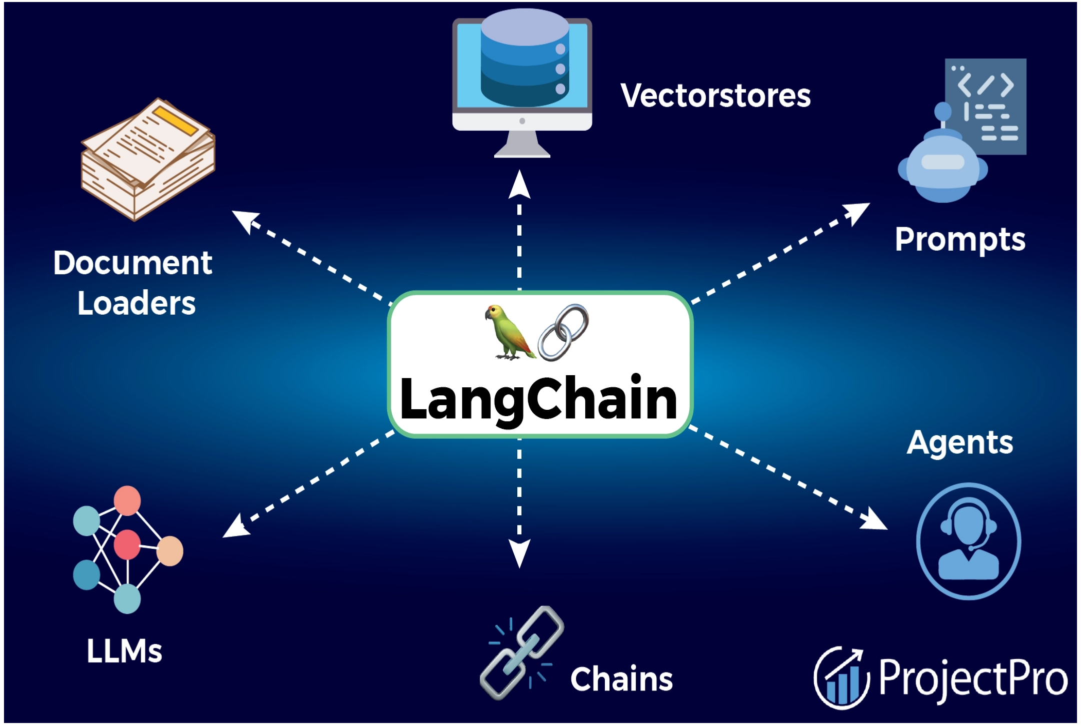
Creator and Lead Developer
Dmytro Liubarskyi

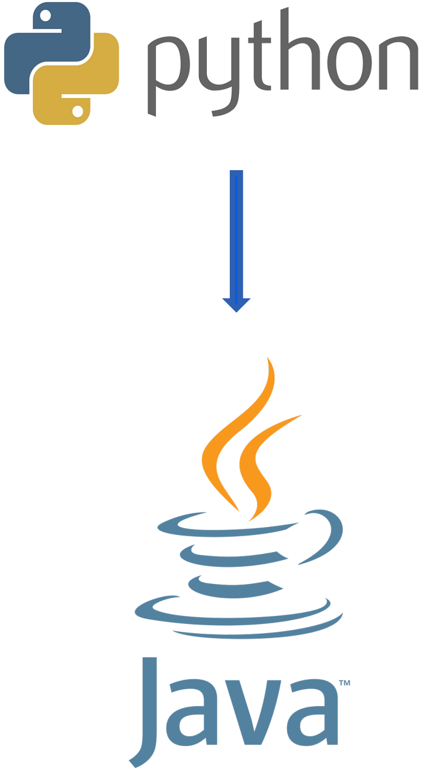


and many more

LangChain for Java: Supercharge your Java application with the power of LLMs

LangChain4j



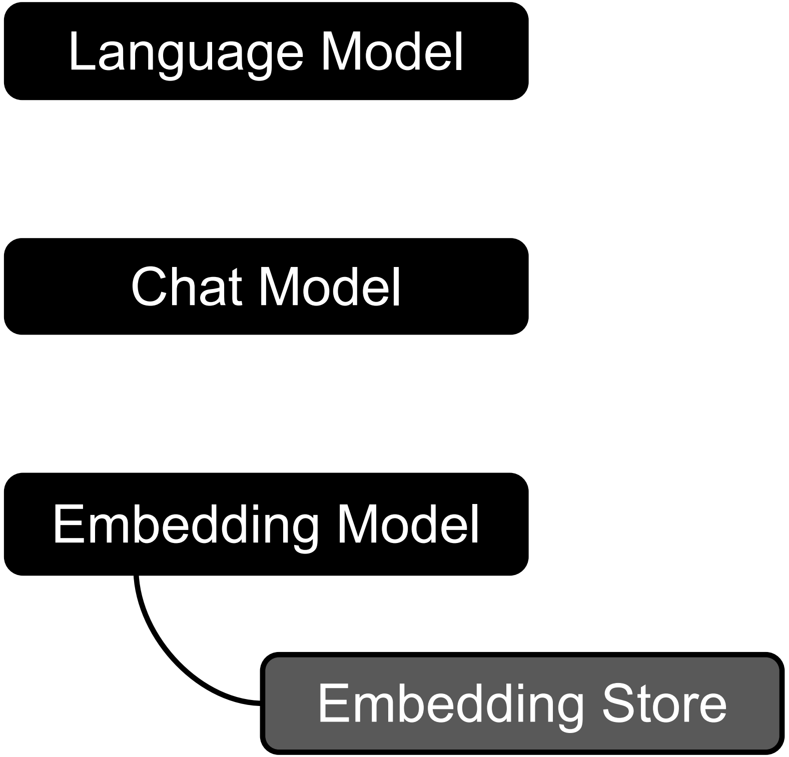


Quoting Dimitris Andreadis (Red Hat):

The "Model" is only part of your Architecture

It's just another API

Large Language Models



Language Model

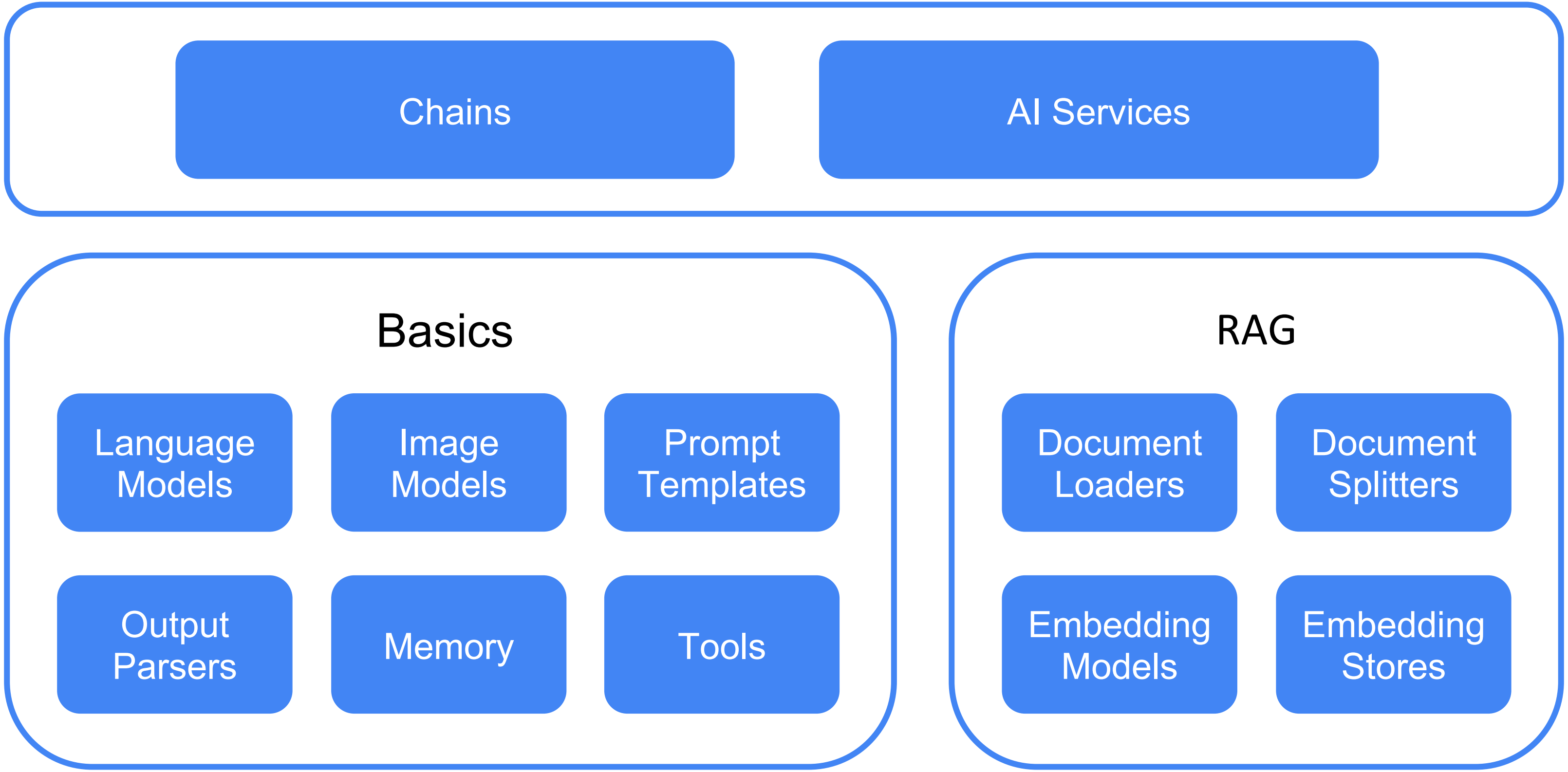
Chat Model

Embedding Model

Embedding Store

With LangChain4j

Components of LangChain4j



Chains

AI Services

Basics

Language Models

Image Models

Prompt Templates

Output Parsers

Memory

Tools

RAG

Document Loaders

Document Splitters

Embedding Models

Embedding Stores

Current LangChain4j Integrations

LLM Integrations	
Amazon Bedrock	Google Vertex AI PaLM 2
Azure OpenAI	HuggingFace
ChatGLM	LocalAI
DashScope	Ollama
Google Vertex AI Gemini	OpenAI
	Mistral

Image Model Integrations
Azure OpenAI Dall·E
OpenAI Dall·E
Vertex AI Gemini
Ollama
Qwen

Current LangChain4j Integrations

Embedding Stores	
Chroma	Astra DB
Elasticsearch	Cassandra
Milvus	Neo4j
Pinecone	OpenSearch
Vespa	PGVector
Weaviate	MongoDB
Redis	Qdrant

Document Loaders	
txt	ppt
html	url
doc	github loader
pdf	S3
xls	Azure blob loader
Tencent COS	

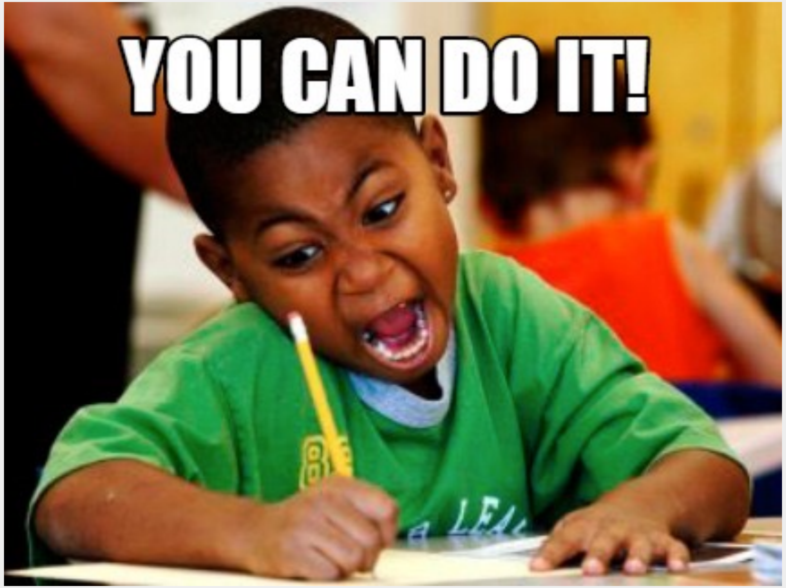
Code Execution Engines
GraalVM Polyglot/Truffle
Judge0

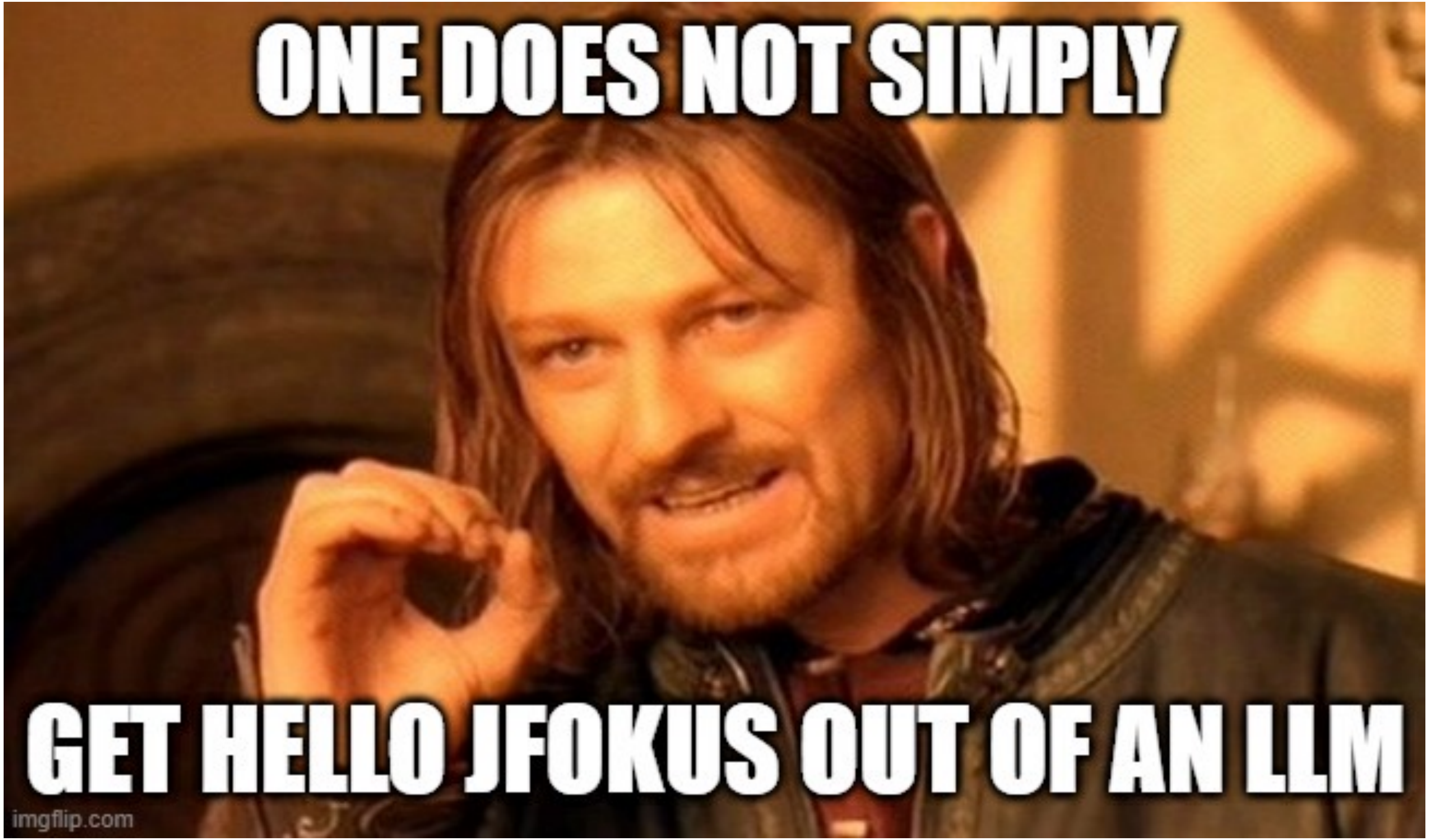
Frameworks
Quarkus
Spring Boot

Demo Time

Hello Jfokus

The Basics




ONE DOES NOT SIMPLY
GET HELLO JFOKUS OUT OF AN LLM
imgflip.com

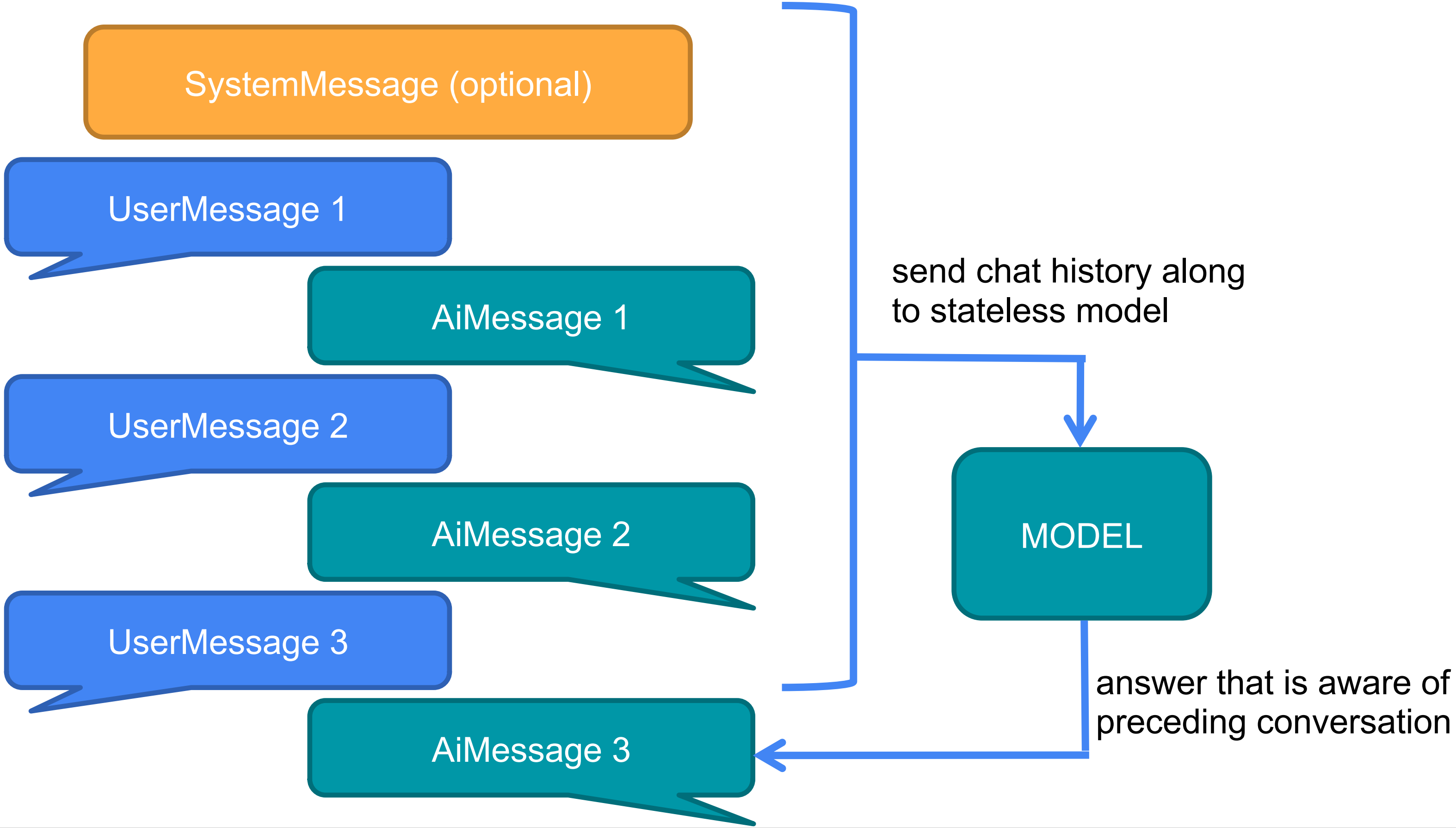
Tips from the team

- 🪄 Choose the best model for your use case using leaderboards
- 🪄 Use the best available model if you have the choice
- 🪄 You can use an openAI demo key for trying out LangChain4j for prototyping
- 🪄 Choose good prompt templates (eg. ‘answer stricly in the following format’)

Chat API

CUSTOMER TRYING TO MAKE CATBOT
UNDERSTAND THEY WANT TO SPEAK TO A REAL AGENT

ChatLanguageModel



AI Services

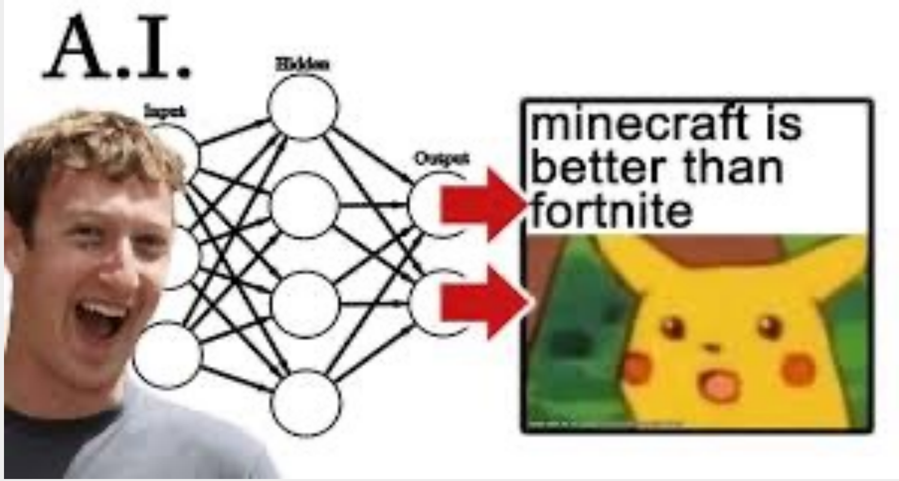
LLM's returning POJOs

Tools

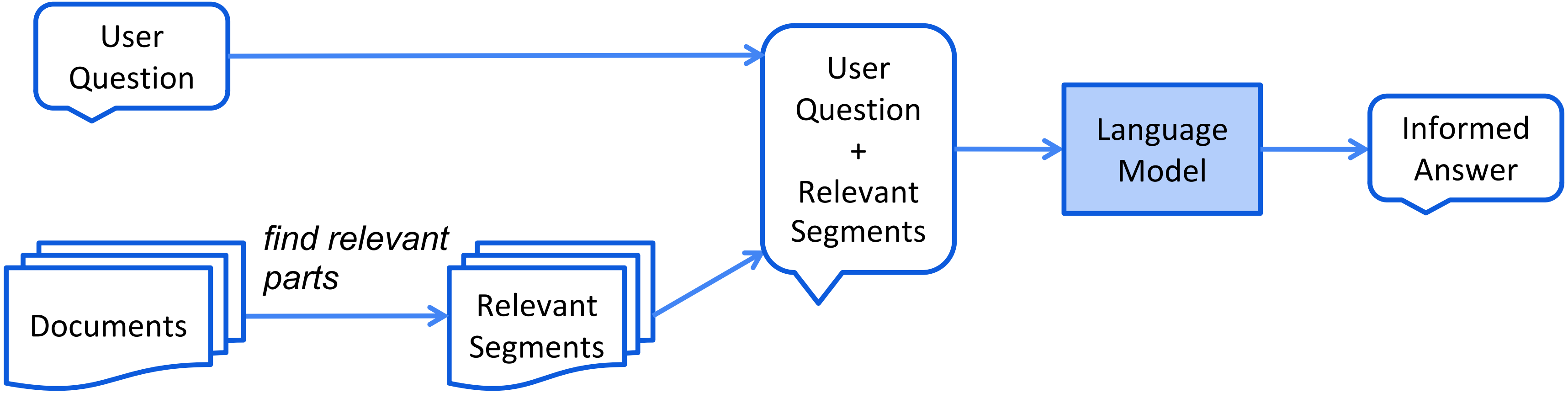
Let’s take it to another level

Retrieval-Augmented Generation

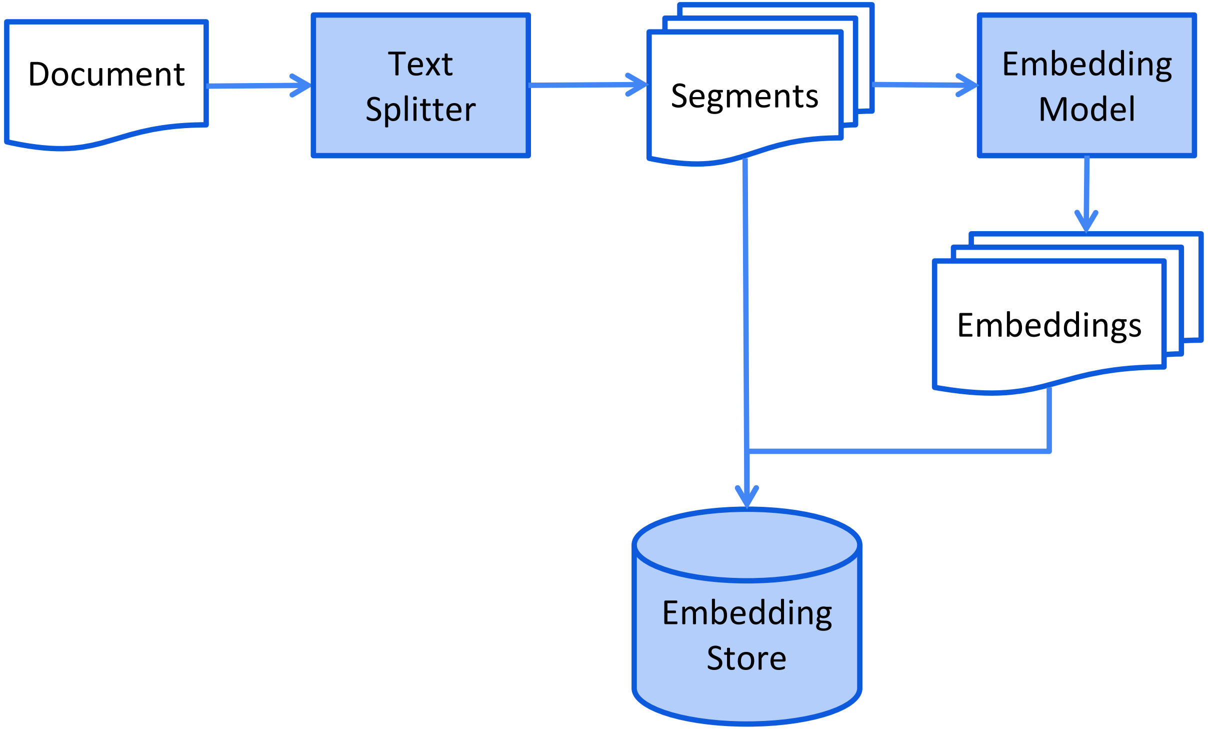
Stop hallucinating!



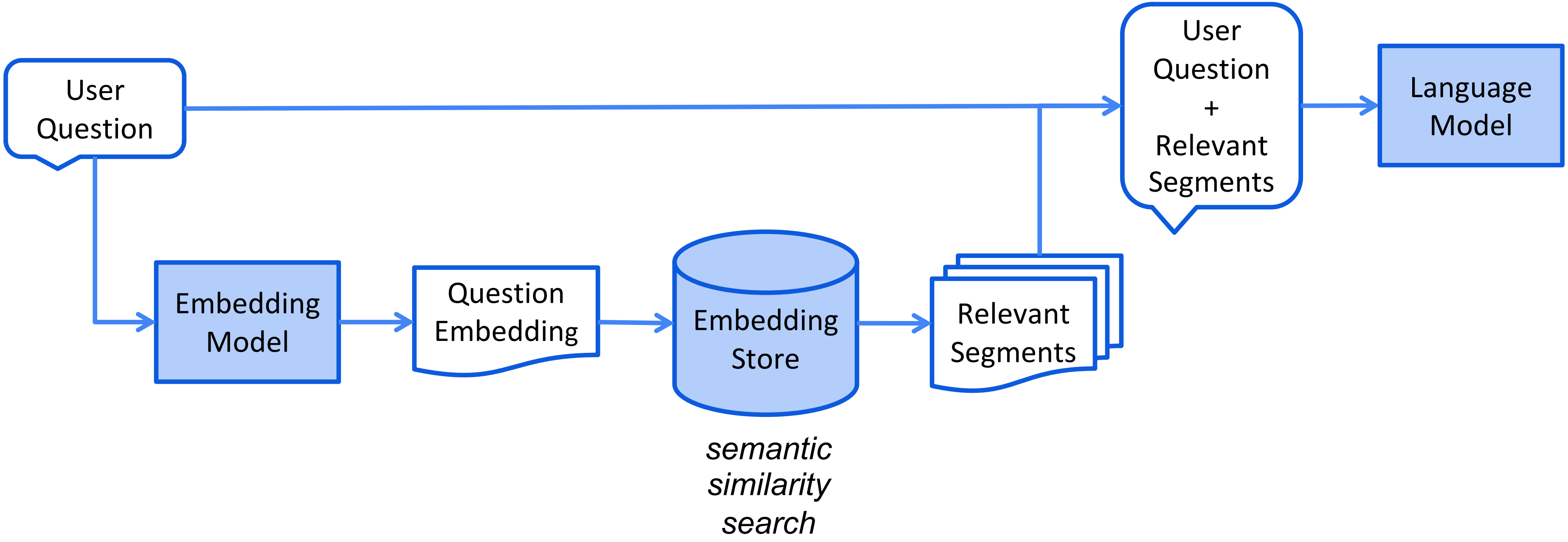
RAG - Overview



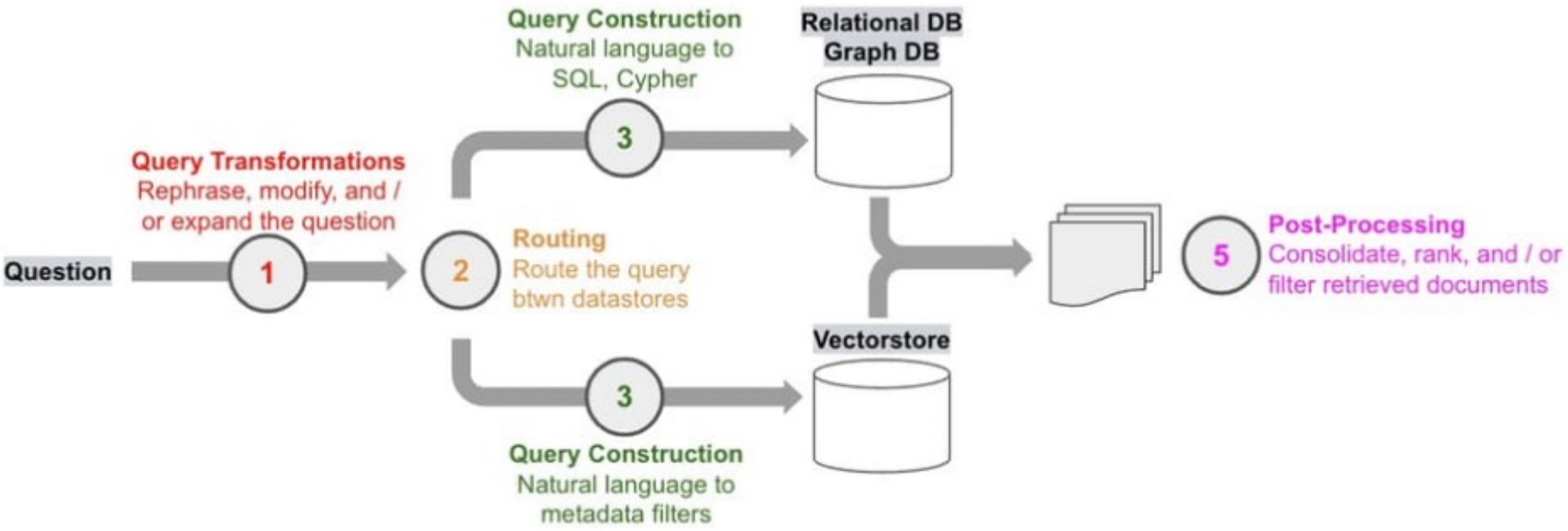
RAG - Ingestion



RAG - Retrieval



Advanced RAG - *new in 0.26* 🎉


Question
1
Query Transformations
Rephrase, modify, and / or expand the question
2
Routing
Route the query btwn datastores
3
Query Construction
Natural language to SQL, Cypher
Relational DB
Graph DB
3
Query Construction
Natural language to metadata filters
Vectorstore
5
Post-Processing
Consolidate, rank, and / or filter retrieved documents

Tips from the team

- 🪄 For RAG prototyping: use in-process embedding models and in-memory embedding stores
- 🪄 Optimize segmentation and embedding parameters for your use case and document corpus
- 🪄 Clean up documents before segmentation, and add context to segments
- 🪄 Advanced RAG techniques allow for better quality results, but higher latency. Pick what your app needs.

Framework Integrations

Quarkus and Spring Boot

Useful Links

LangChain4j **Follow @langchain4j**

⭐ Github repo: github.com/langchain4j/langchain4j

Documentation linked in Readme

⭐ Github examples: github.com/langchain4j/langchain4j-examples/

Useful Links

Language Model Leaderboard

⭐ huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard

Hallucination Leaderboard

⭐ Leaderboardhttps://huggingface.co/spaces/hallucinations-leaderboard/leaderboard

Local Models Community

⭐ www.reddit.com/r/LocalLLaMA/

Useful Links

Info on Embedding Models

- huggingface.co/spaces/mteb/leaderboard/
- blog.vespa.ai
- www.pinecone.io/learn

Info on RAG

- https://blog.langchain.dev/deconstructing-rag/

Short Courses on LLMs

- www.deeplearning.ai/short-courses/

We look forward to your pull requests ❤️

Enjoy, and happy with your feedback and contributions!

Let’s be in touch!

lize.raes@open-tide.com

EPIC ENGINEERING

Personal blog

Slides under ‘Presentations’

https://epic.engineering

@LizeRaes

Lize Raes

Follow @langchain4j

github.com/langchain4j/langchain4j