BUILD YOUR FIRST LLM APP

CHANDRAHASAN SUBBAIYAN

AGENDA

POPULAR LLM’S

WHAT IS OLLAMA

CAUTION

WHAT IS LANGCHAIN

BENEFITS OF LANGCHAIN

DEMO

RAG

POPULAR LLM'S

LLM MODEL	ORGANIZATION	SOURCE TYPE	MULTI MODEL
GPT	OPEN AI	CLOSED	YES
GEMINI	GOOGLE	CLOSED	YES
LLAMA	META	OPEN	NO
CLAUDE	ANTHROPIC	CLOSED	YES
MIXTRAL	MISTRAL AI	OPEN / CLOSED	NO
GEMMA	GOOGLE	OPEN	NO
PHI-3	MICROSOFT	OPEN	NO

WHAT IS OLLAMA

Ollama https://ollama.com helps you get up and running with large language models, locally in very easy and simple steps

Ollama runs on your local machine at http://localhost:11434/

LLAMA3.1	SIZE	CPU CORE	RAM	GPU
8B	4.7 GB	6 – 8	16GB	12 GB VRAM
70B	40 GB	12-16	64GB/96GB	40 GB VRAM
405B	229 GB	24	512 GB	120 GB VRAM

* Approx

CAUTION

- **Do Not Use Company or Client Data for Testing**
- **Do not use/upload any sensitive data to the internet.**
- **If you have a valid use case to run company or client data, contact your manager. Obtain proper approvals before using any company or client data.**

WHAT IS LANGCHAIN

- LangChain is a framework for developing applications powered by large language models (LLMs).

https://python.langchain.com/v0.2/docs/integrations/platforms/

- LangChainjs is to simplify integrating LLMs into Node applications.

https://github.com/langchain-ai/langchainjs

- LangChain4j is to simplify integrating LLMs into Java applications.

https://github.com/langchain4j/langchain4j?tab=readme-ov-file

BENEFITS OF LANGCHAIN

- **Modular Design:** Flexibility and customization for integrating components.
- **Ease of Use**: User-friendly with comprehensive documentation.
- **Scalability:** Designed to handle large-scale applications and high volumes of data.
- **Integration Capabilities:** Seamless integration with existing systems and robust APIs.
- **Security:** Ensures secure handling of data, complying with industry standards.

TALK IS CHEAP
SHOW ME THE CODE

DEMO

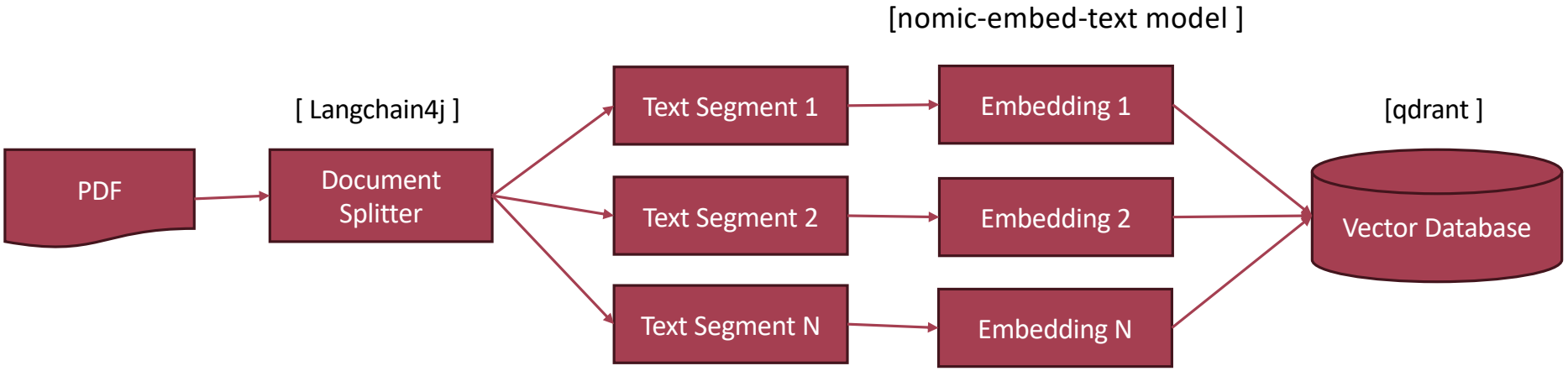
Prerequisite:

- Ollama
- LLM model
- Langchain4j
- Springboot, Java 17
- Docker
- IDE

RETRIEVAL AUGMENTED GENERATION (RAG)

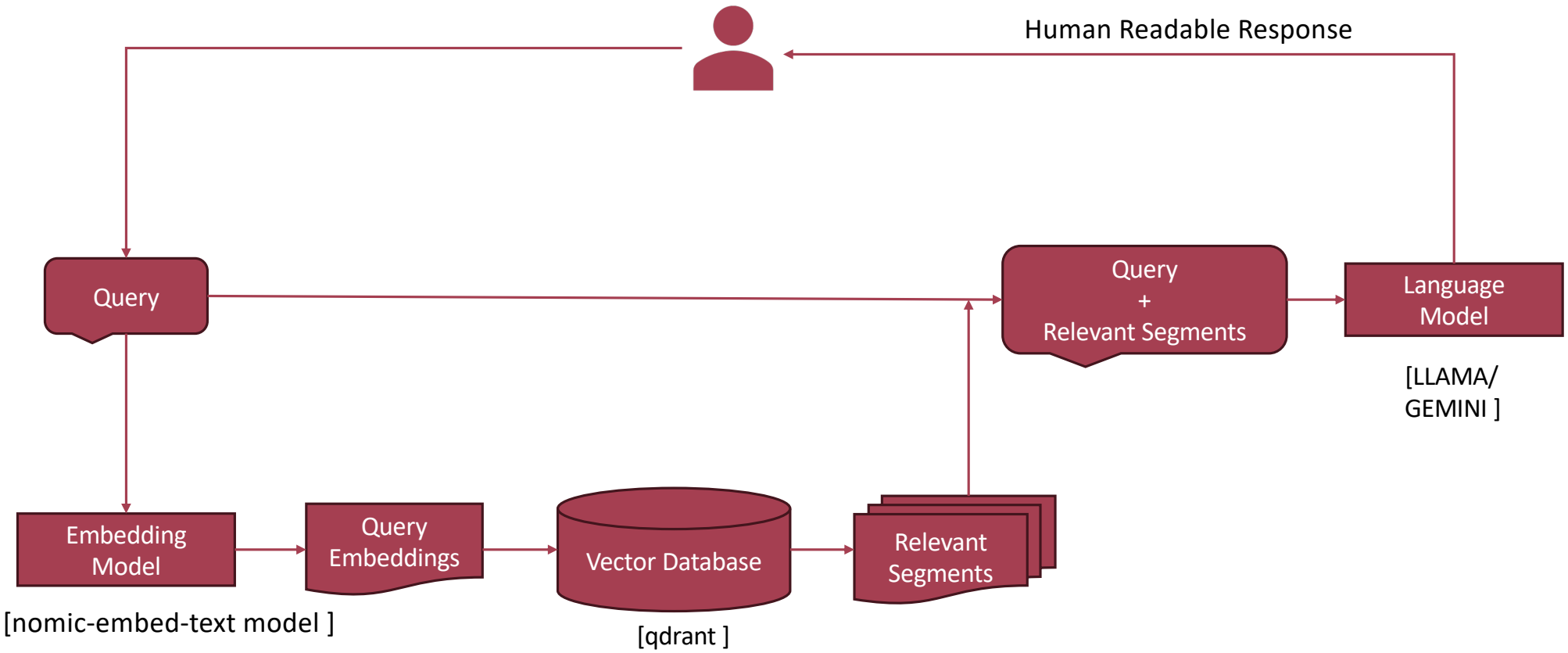
- **Two main limitations with LLM**
  - They have a cut-off date
  - They don't have your company internal data
- **How do we give more context to an LLM?**
  - Ingest new/private data using the same "embeddings" as the LLM
  - Store that data into a vector database
  - Request the vector database and send that data to the LLM

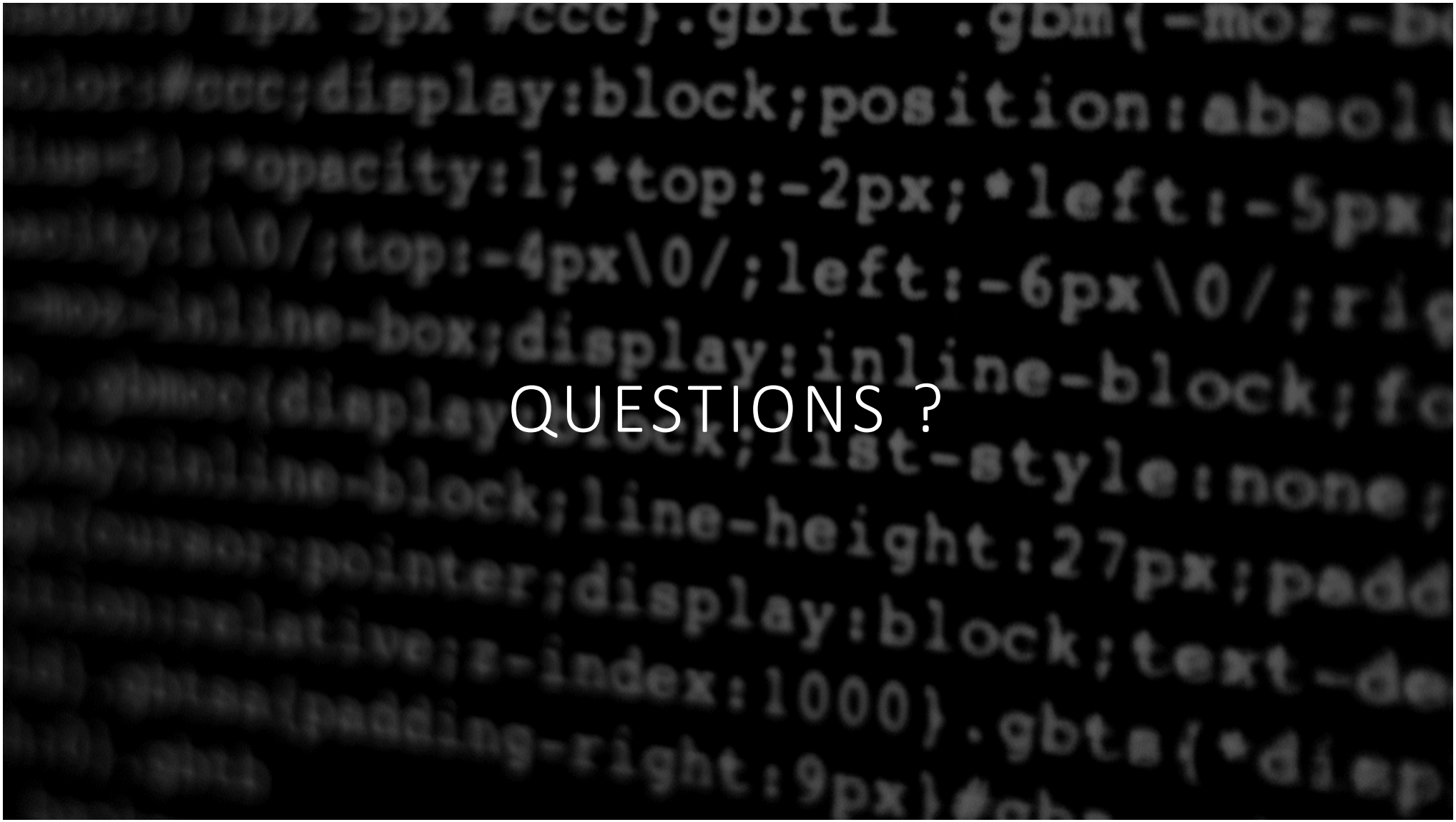
RAG - INGESTION PROCESS



[-0.058949846774339676,1.0301238298416138,-2.6069746017456055,-1.8815385103225708,
2.0599701404571533,-0.7361735701560974,...]

RAG – RETRIVAL PROCESS




QUESTIONS ?


THANK YOU
- CHANDRAHASAN SUBBAIYAN