From Vectors to Answers: Practical RAG with **Qdrant**

13/06/2025

By Mohammed Arbi Nsibi

MOHAMED ARBI NSIBI

- ML engineer @ Carbon Insights
- Final year ICT engineering student@ SUP’COM
- Former GDSC Lead 23/24
- GDGoC SUP’COM & ISAMM Mentor
- GDG Carthage Co-organizer

https://huggingface.co/Goodnight7

https://www.linkedin.com/in/mohammed-arbi-nsibi-584a43241/

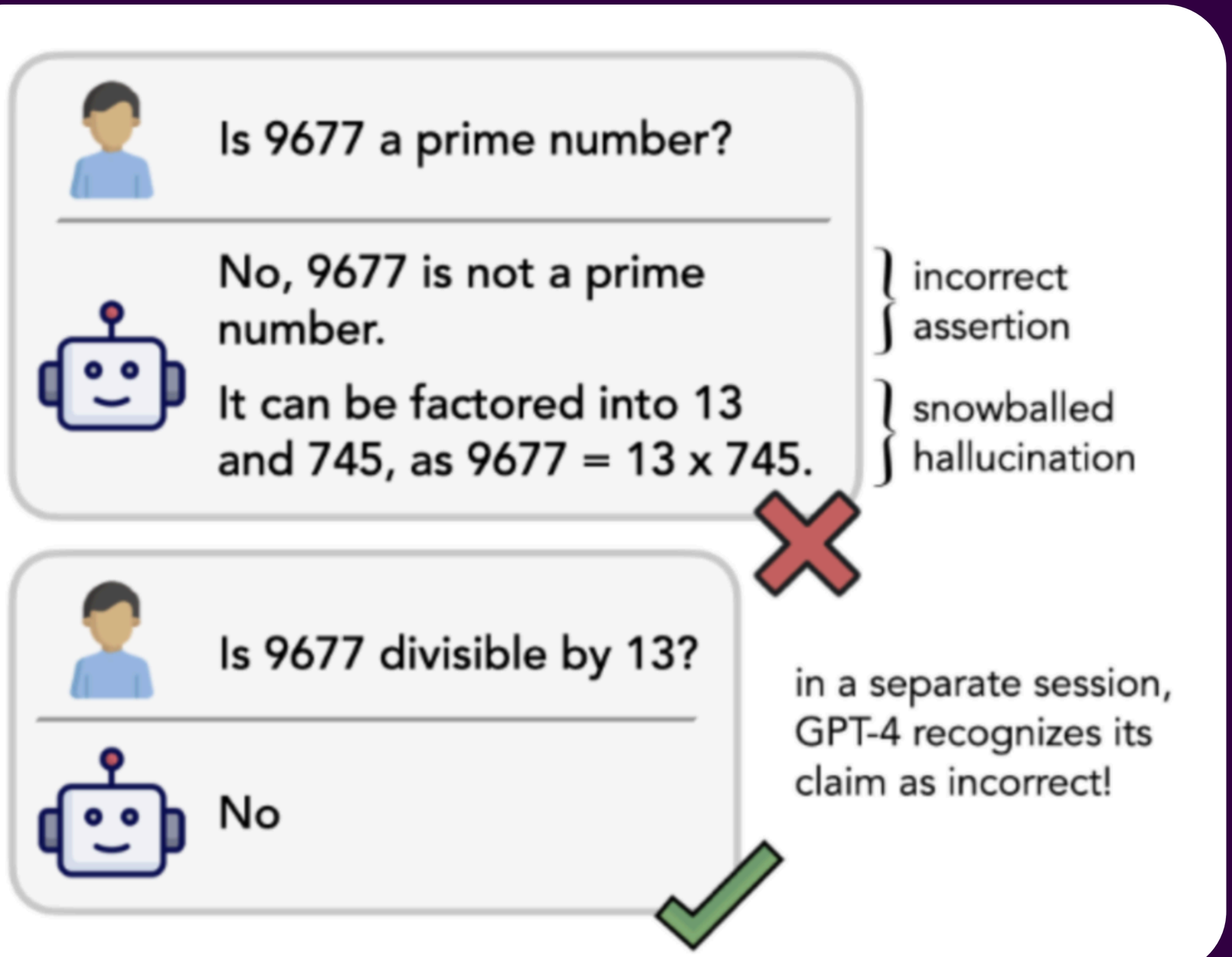
Content

- Motivation
- RAG components
- Vector stores deep dive
- Building basic RAG pipeline
- Some Advanced RAG Techniques

1- The need for an external Knowledge !

2- Hallucinations

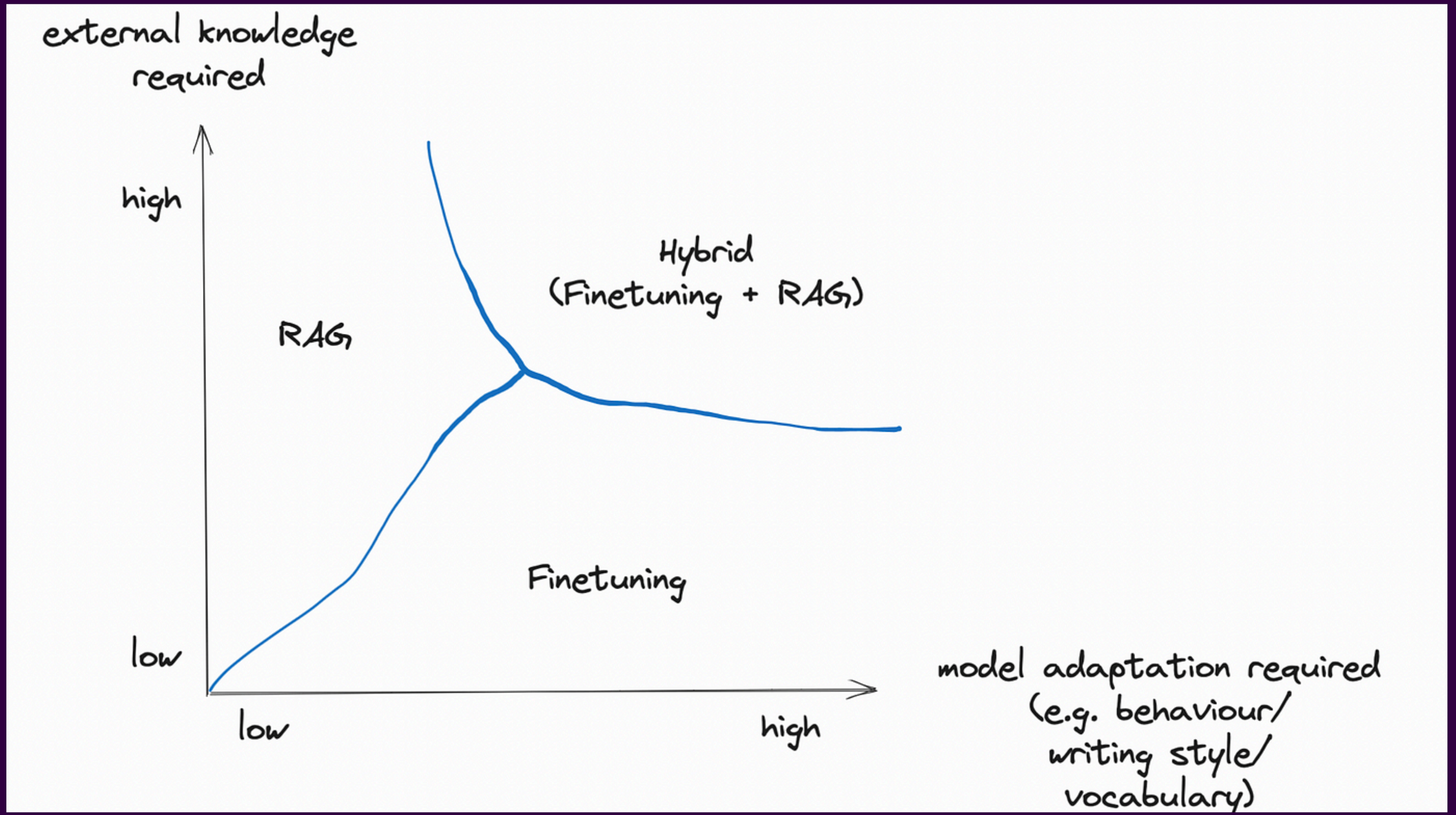
Hallucinations

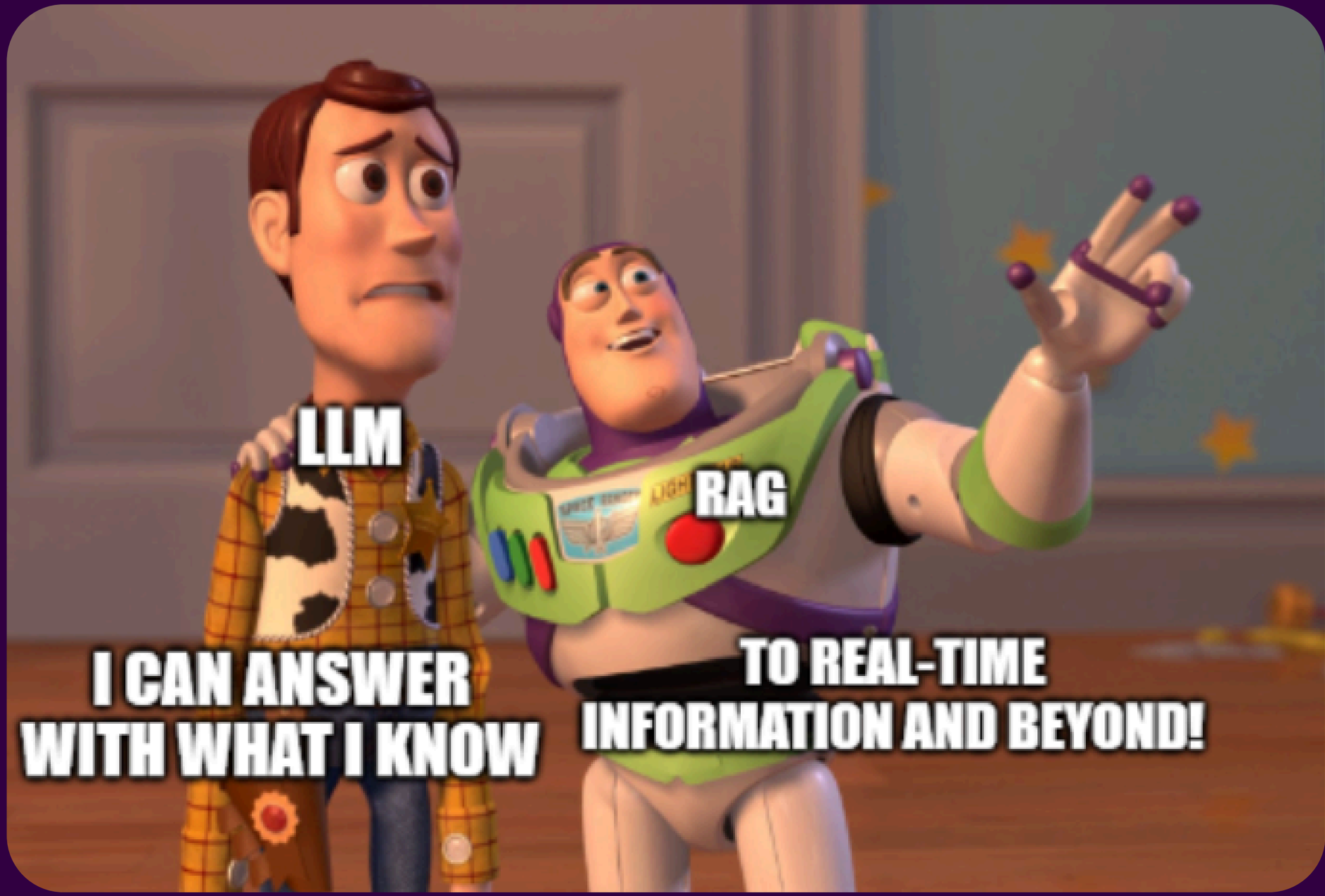


Hallucinations

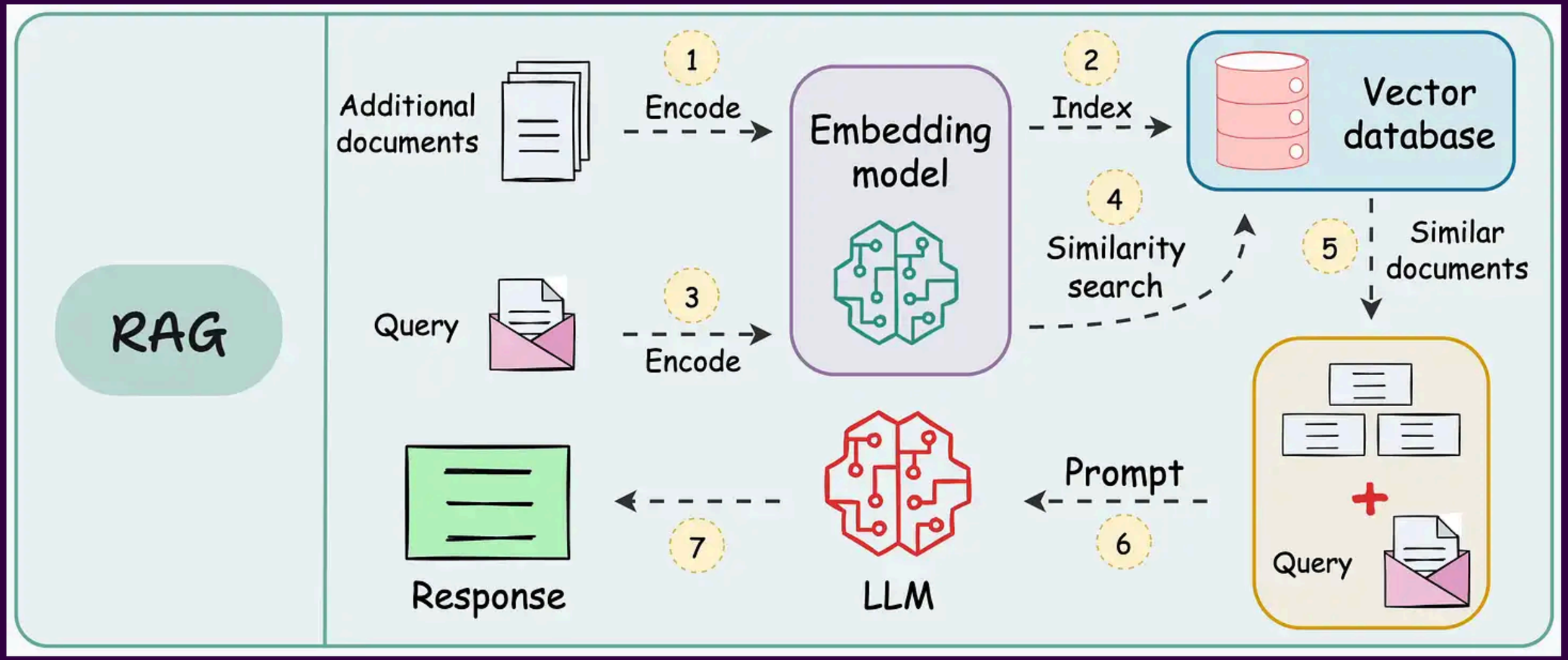
- The model is not trained on enough data.
- The model is trained on noisy or dirty data.
- The model is not given enough context .
- The model is not given enough constraints (rules, guidelines, or limitations)

RAG / Fine-tuning

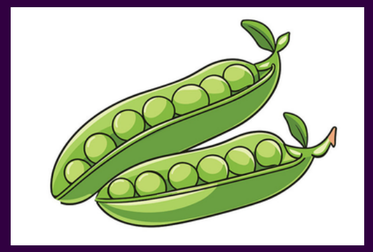



LLM
RAG
I CAN ANSWER WITH WHAT I KNOW
TO REAL-TIME INFORMATION AND BEYOND!

RAG (Retrieval-augmented generation)



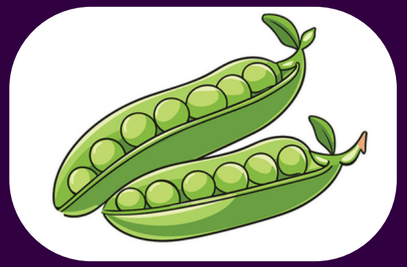
tea

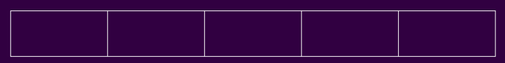


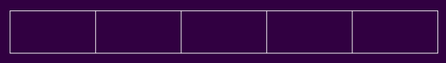
pea

coffee

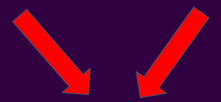
≠







distance = 0.3



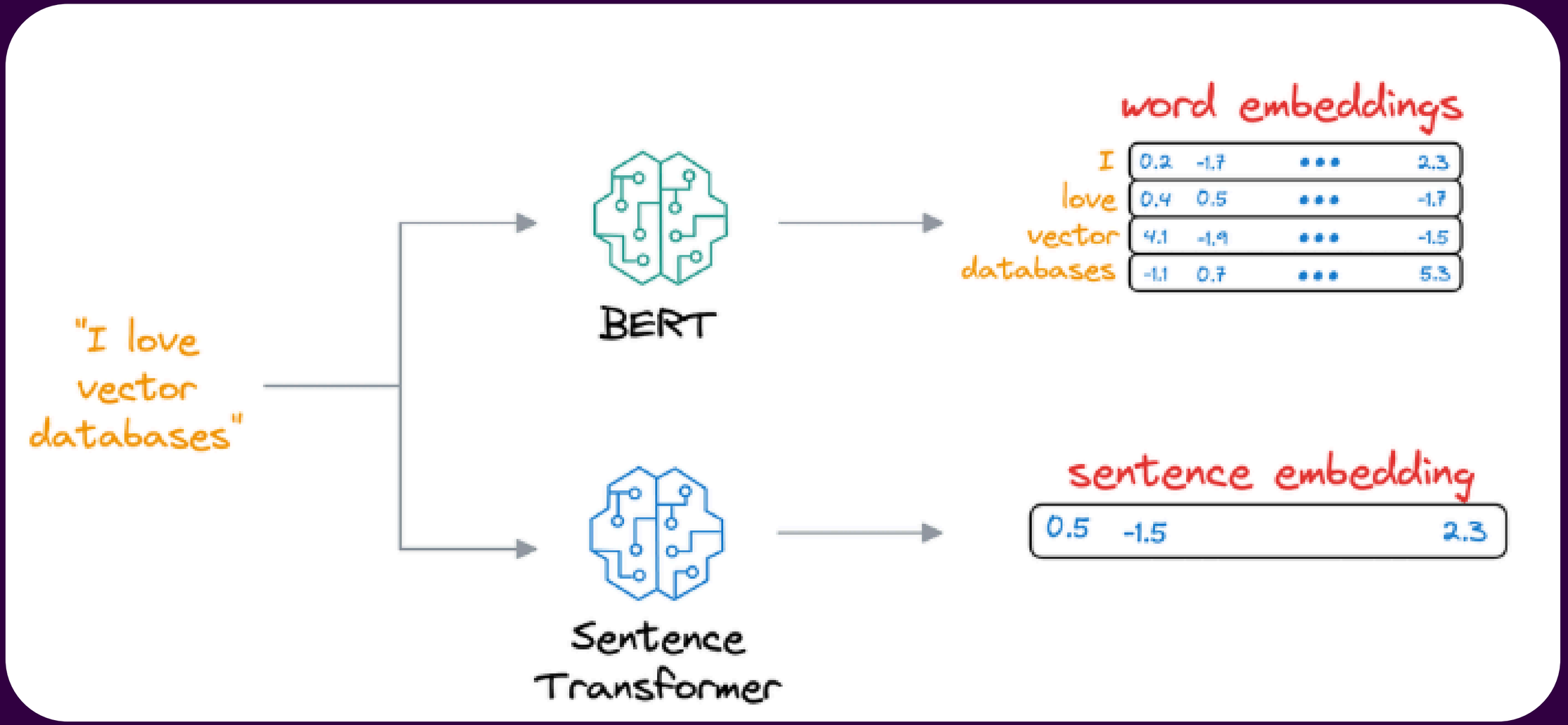
distance = 0.7

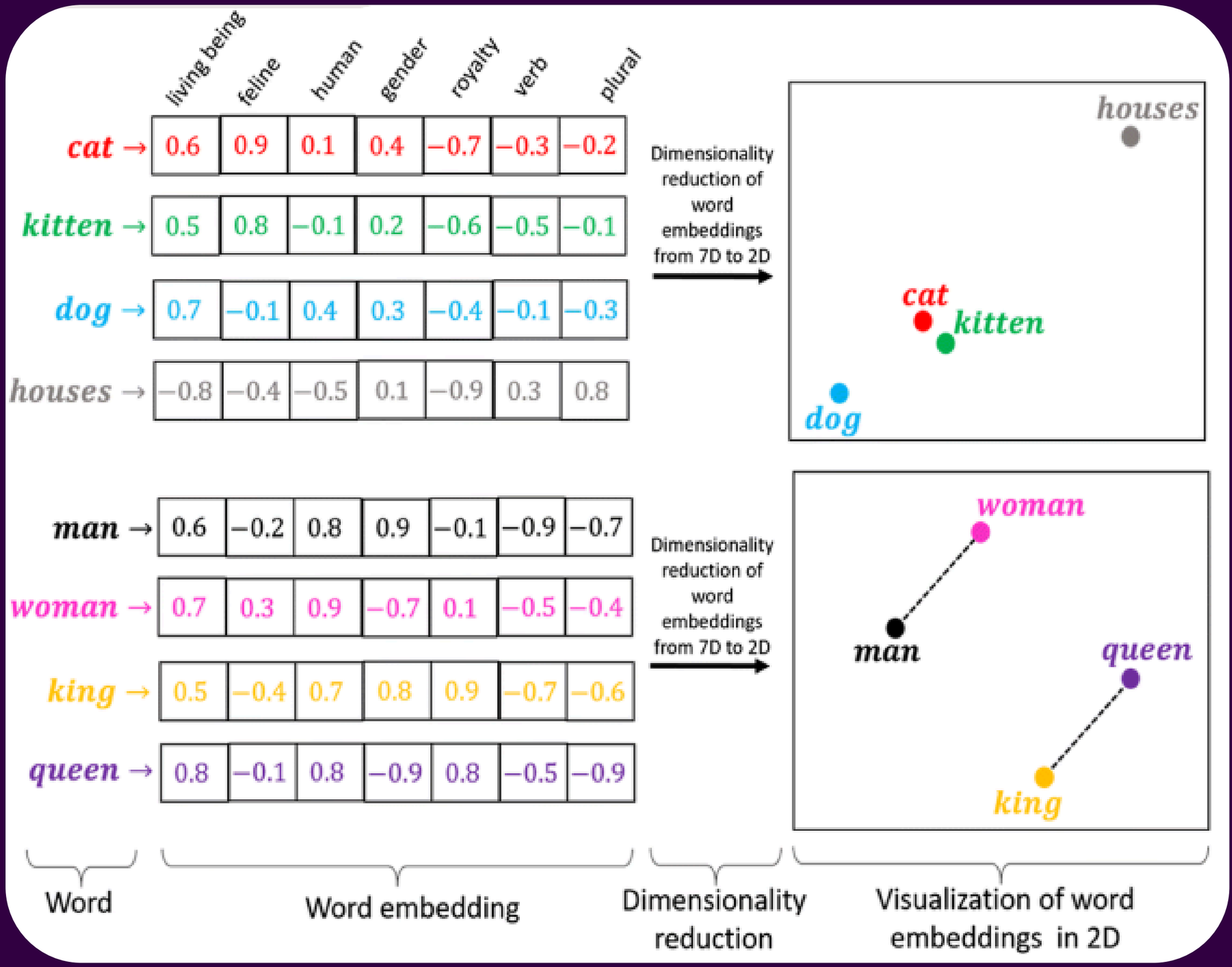
Embedding model

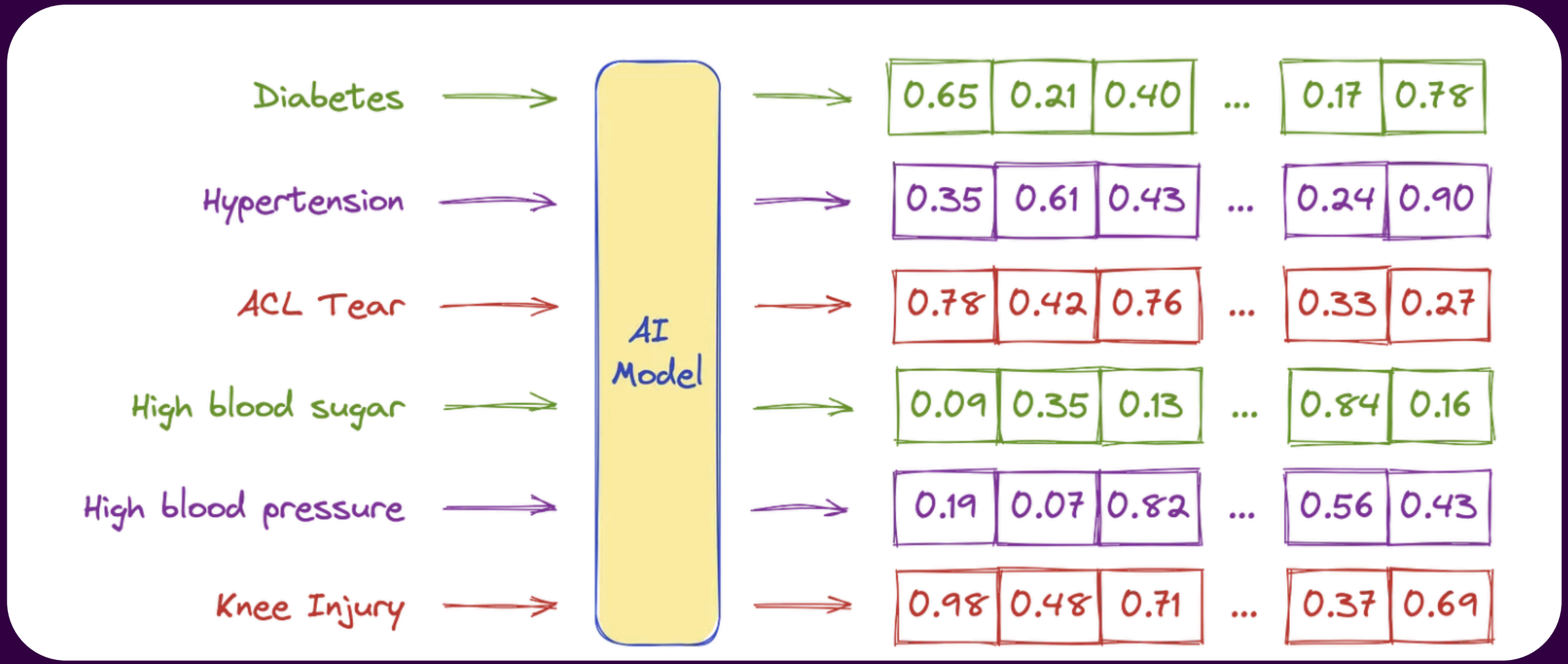
- These vectors live in a high-dimensional space where the proximity between vectors reflects the **relatedness** of the original items.

- Embedding model trained along LLM and learn to produce representation(vectors) based on context in word appear.

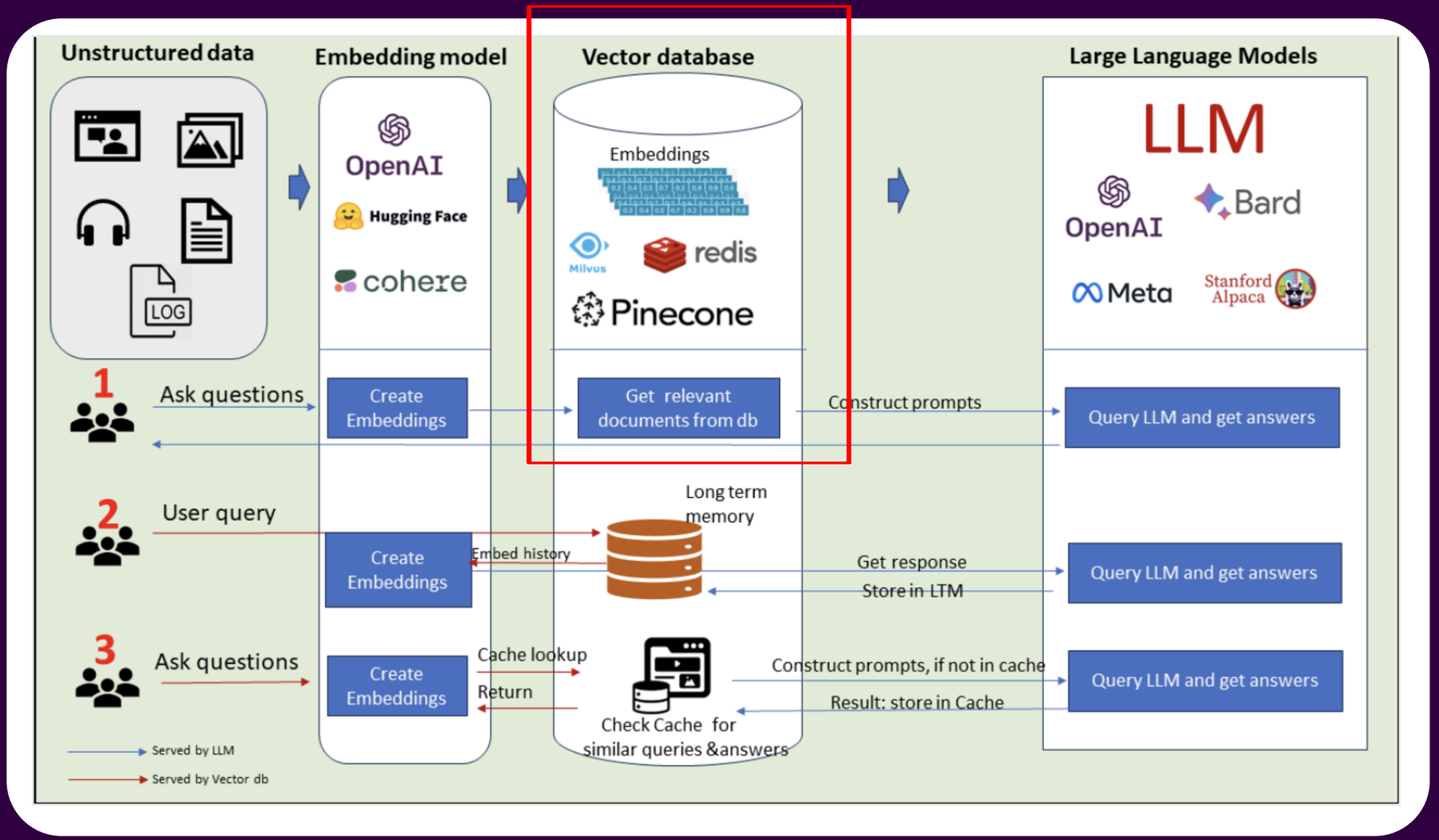
Embedding model




living being
feline
human
gender
royalty
verb
plural
cat → 0.6 0.9 0.1 0.4 −0.7 −0.3 −0.2
kitten → 0.5 0.8 −0.1 0.2 −0.6 −0.5 −0.1
dog → 0.7 −0.1 0.4 0.3 −0.4 −0.1 −0.3
houses → −0.8 −0.4 −0.5 0.1 −0.9 0.3 0.8
Dimensionality reduction of word embeddings from 7D to 2D
houses
cat
kitten
dog
man → 0.6 −0.2 0.8 0.9 −0.1 −0.9 −0.7
woman → 0.7 0.3 0.9 −0.7 0.1 −0.5 −0.4
king → 0.5 −0.4 0.7 0.8 0.9 −0.7 −0.6
queen → 0.8 −0.1 0.8 −0.9 0.8 −0.5 −0.9
Dimensionality reduction of word embeddings from 7D to 2D
woman
man
queen
king
Word
Word embedding
Dimensionality reduction
Visualization of word embeddings in 2D

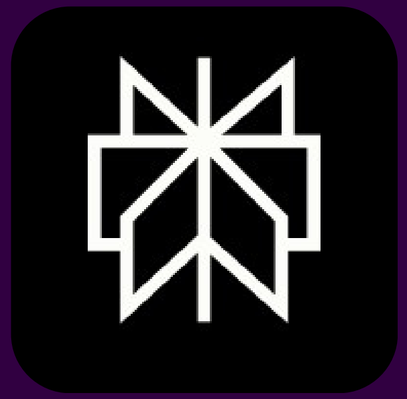

Diabetes
Hypertension
ACL Tear
High blood sugar
High blood pressure
Knee Injury
AI Model
0.65 0.21 0.40 ... 0.17 0.78
0.35 0.61 0.43 ... 0.24 0.90
0.78 0.42 0.76 ... 0.33 0.27
0.09 0.35 0.13 ... 0.84 0.16
0.19 0.07 0.82 ... 0.56 0.43
0.98 0.48 0.71 ... 0.37 0.69

Vector Database



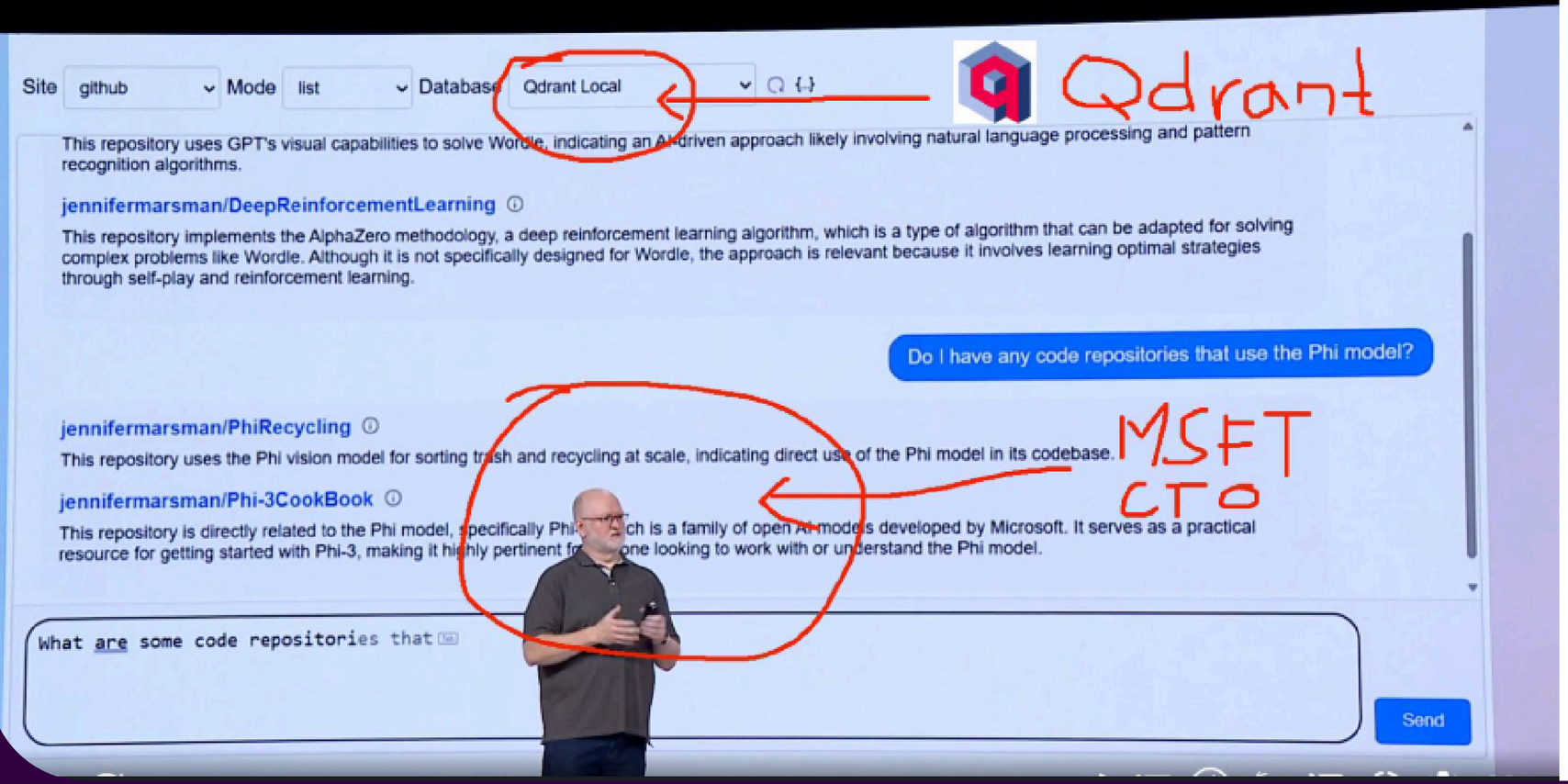
Fully Open-source
Self-hosting : run on ur own infra
Super fast: Latency ~0.024s (~24ms)
Hybrid Search
UI support
Free Tier ~1M(vectors) 768-dim vectors

Used by



Used by

drant



Andre Zayarni • 1st
Co-founder & CEO, Qdrant | Open-Source Vector Search
3w • Edited •

Kevin Scott, Microsoft's CTO, used Qdrant as the vector engine for the #NLWeb demo at the opening keynote of the annual Build Conference.

#opensource rules!

You and 907 others 53 comments · 23 reposts

Love Comment Repost Send

Tell them what you loved...

Most relevant

Andre Zayarni Author 3w
Co-founder & CEO, Qdrant | Open-Source Vector Search

Recording https://youtu.be/ceV3RsG946s?si=Xh6NkXCJH1JhKHeV&t=4879

Like · 4 Reply

Philippe Bourcier • 2nd 3w
CTO #GenAI / Business Angel / Maker

Soon a Windows build for Qdrant ?

Like Reply · 2 replies

NLWeb_link

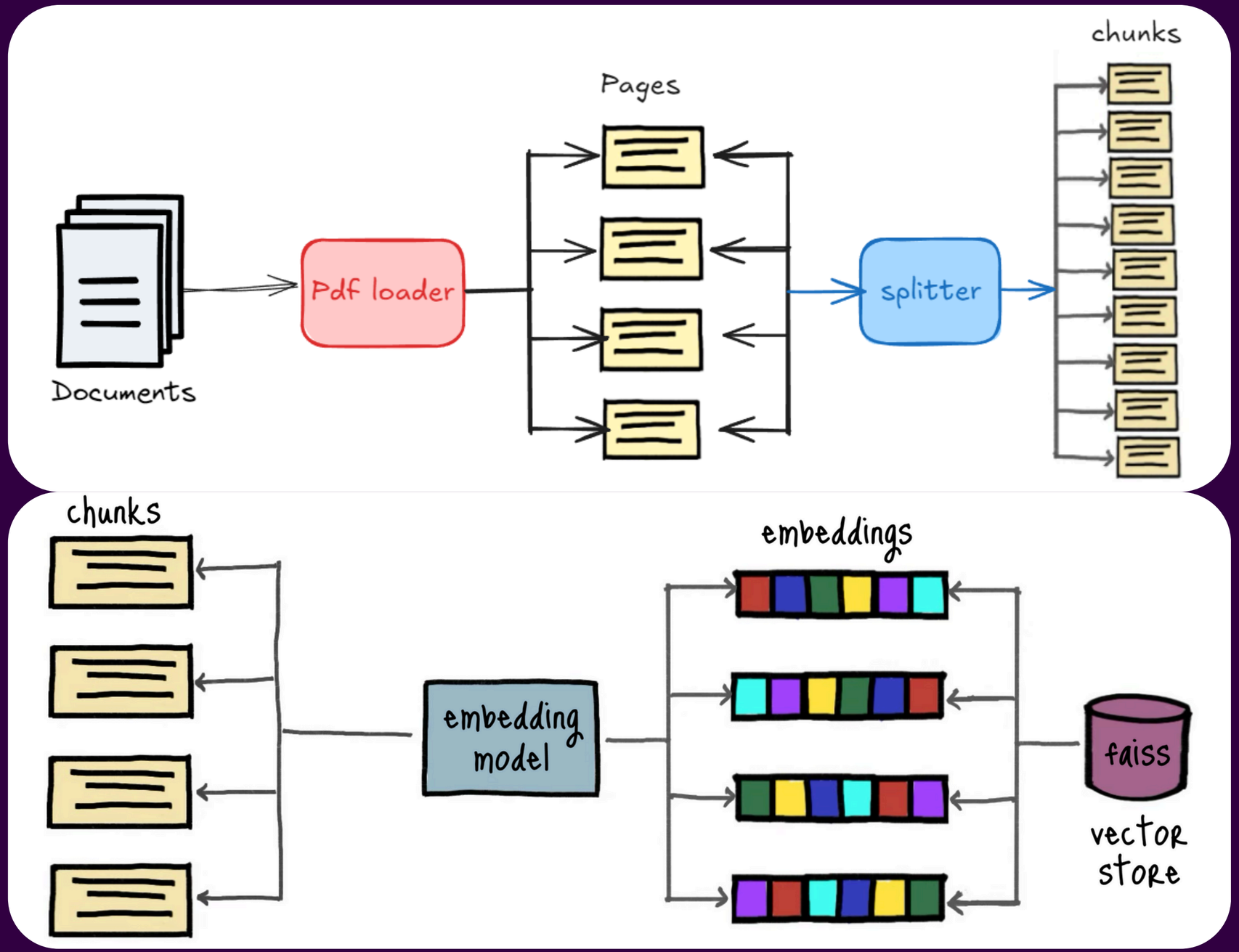
Why Qdrant is super fast ?

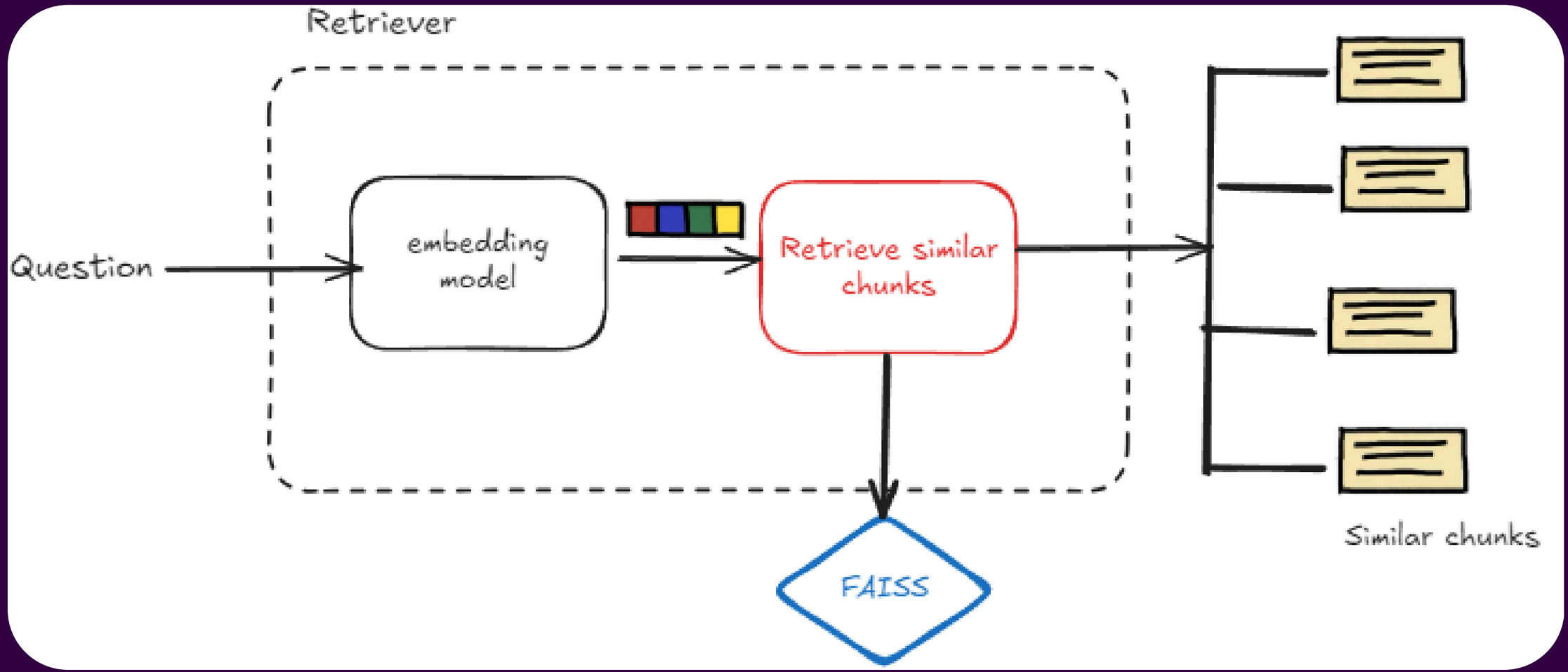
- Rust-Based Engine
- HNSW (Hierarchical Navigable Small World) Indexing : fast ANN search (on the best matches without scanning everything.)
- Vector Quantization (useful for large-scale datasets) : Saves RAM (up to 16x)
- Batch & Parallel Processing

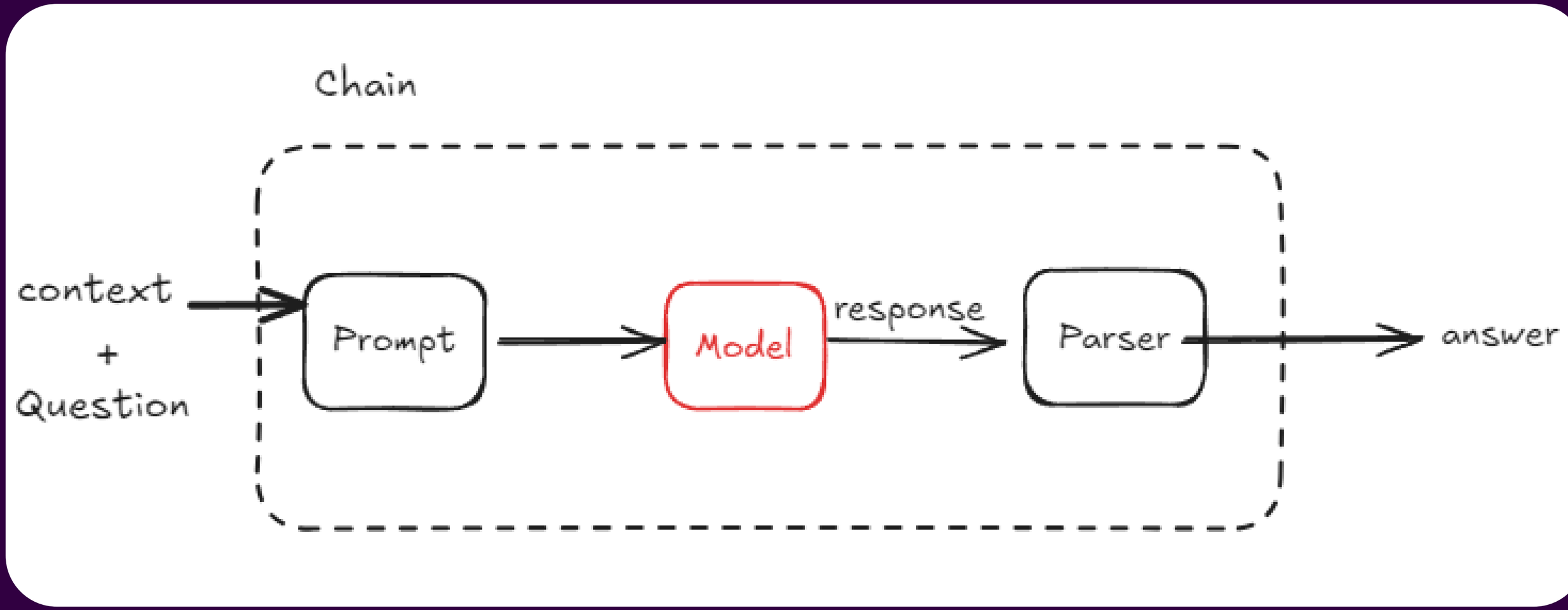
QUANTIZATION EXAMPLE

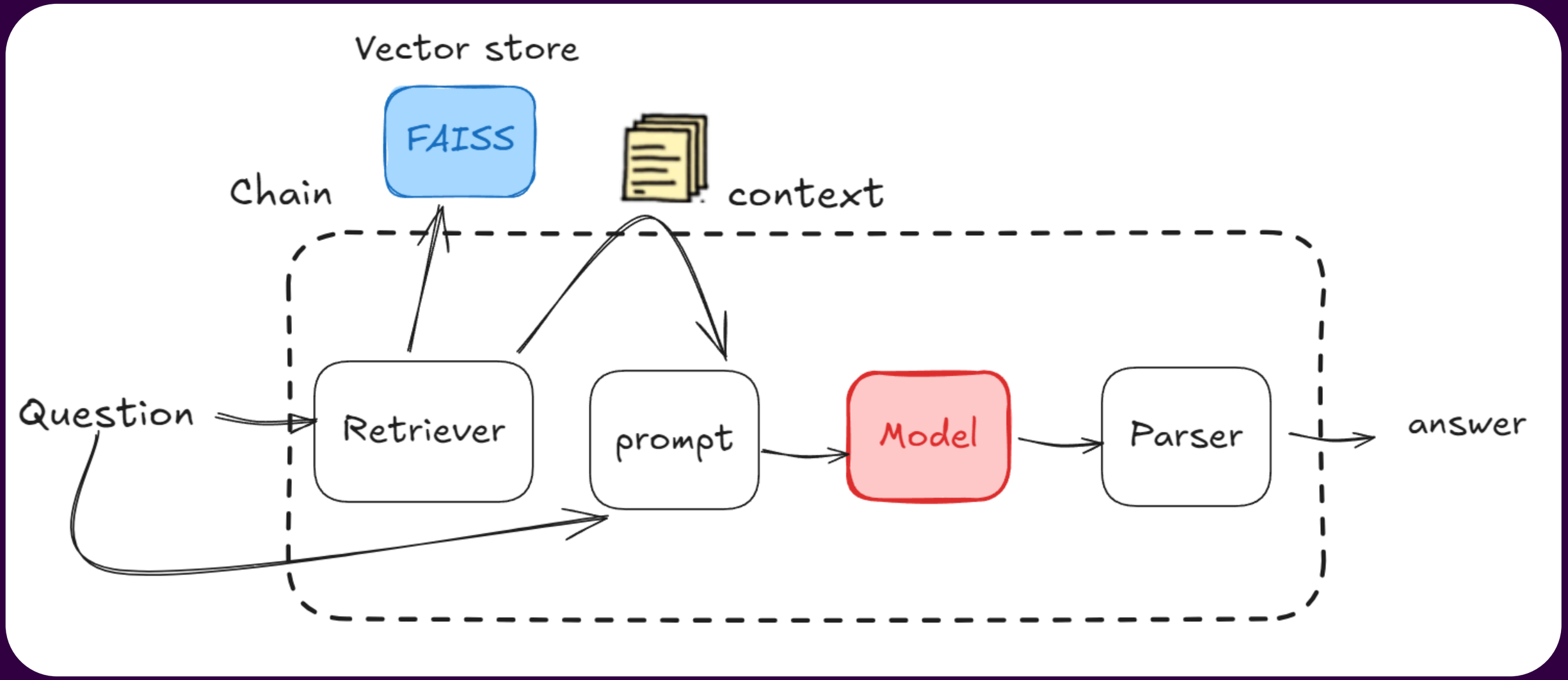
```python
client.recreate_collection(
    collection_name=collection_name,
    vectors_config=models.VectorParams(
        size=512,  # dimension of your vectors #PQ works best with float32 vectors and dimensions divisible by 4 or 8.
        distance=models.Distance.COSINE,
        quantization_config=models.QuantizationConfig(
            quantization=models.Quantization(
                quantization=models.ProductQuantization( # using PQ method
                    enabled=True,
                    compression="x8",  # or "x4", "x16" # x8 means 8x compression
                    always_ram=False,  # keep quantized vectors on disk
                )
            )
        )
    )
)
```

RAG architecture




Retriever
Question
embedding model
Retrieve similar chunks
FAISS
Similar chunks


Chain
context
+
Question
Prompt
Model
response
Parser
answer



https://arxiv.org/abs/2005.11401

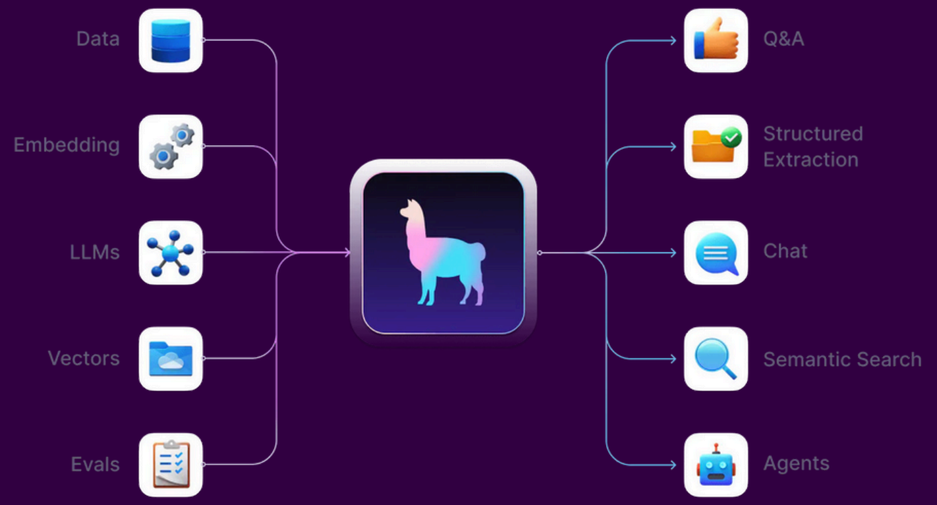
How to get started ?

LangChain

- LangChain is a framework designed to simplify the creation of applications using large language models.

LlamaIndex

- LlamaIndex is a handy tool that acts as a bridge between your custom data and large language models (LLMs) which are powerful models capable of understanding human-like text.



DEMO

```
Question: mention all the managers of AINS this year
Answer is : The managers of AINS mentioned are:

1. Fayez Zouari - Project Manager
2. Makki Aloulou - Program Manager
3. Kacem Mathlouthi - Technical Manager
4. Aziz Amari - manages the Poster Session (not explicitly titled as a manager, but has a management role)
5. Ahmed Amin Chabbah - oversees community engagement (not explicitly titled as a manager, but has a management role)
6. Rayen Khammar - handles event content (not explicitly titled as a manager, but has a management role)
```

```
[ ]  query= "who is the contact manager of AINS this year "
     response = qa.run(query)
```

```
Question: who is the contact manager of AINS this year
Answer is : I don't know who the contact manager of AINS is this year. The provided context mentions various roles and team members, including Project
```

```
     query = "who is the ambassador from IEEE ESSTHS SB in this edition?"
     response = qa.run(query)
```

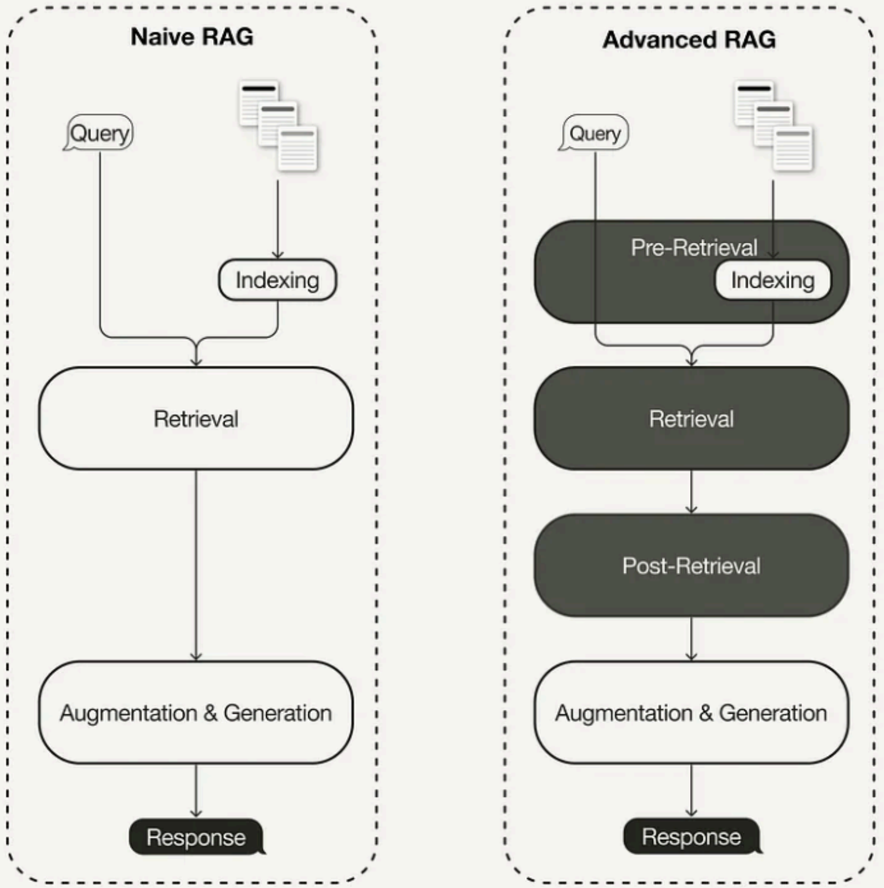
```
Question: who is the ambassador from IEEE ESSTHS SB in this edition?
Answer is : The ambassador from IEEE ESSTHS SB is Molka Attia.
```

```
[ ]  query= "What talks will Hafedh Hichri and Mohamed Arbi Nsibi discuss in their talks in AINS 3.0 ?"
     response = qa.run(query)
     print(f"Question: {query}")
     print(f"Answer is : {response}")
```

```
Question: What talks will Hafedh Hichri and Mohamed Arbi Nsibi discuss in their talks in AINS 3.0 ?
Answer is : Hafedh Hichri will discuss "Revolutionizing RAG with Chonkie", which is about his ultra-light Python library for efficient text chunking i
```

Naive RAG vs Advanced RAG

- There are many implementation to further improve performance of Naive RAG.
- Advanced RAG has evolved as a new paradigm with targeted enhancements to address some of the limitations of the naive RAG paradigm.
  - Advanced RAG techniques can be categorized into
    - pre-retrieval optimization,
    - retrieval optimization, and
    - post-retrieval optimization
- some examples :
  - Feedback loops (re-ranking, similarity score thresholds)
  - Hybrid Search (dense + keyword)
  - Contextual compression (summarize before feeding to LLM)
  - Multi-vector per chunk (dense embeddings per aspect)

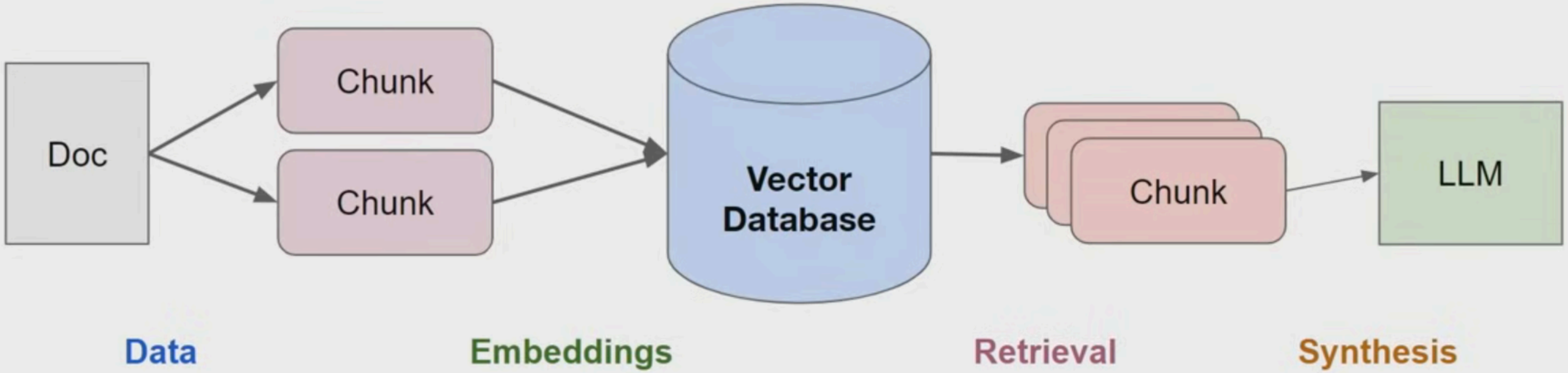


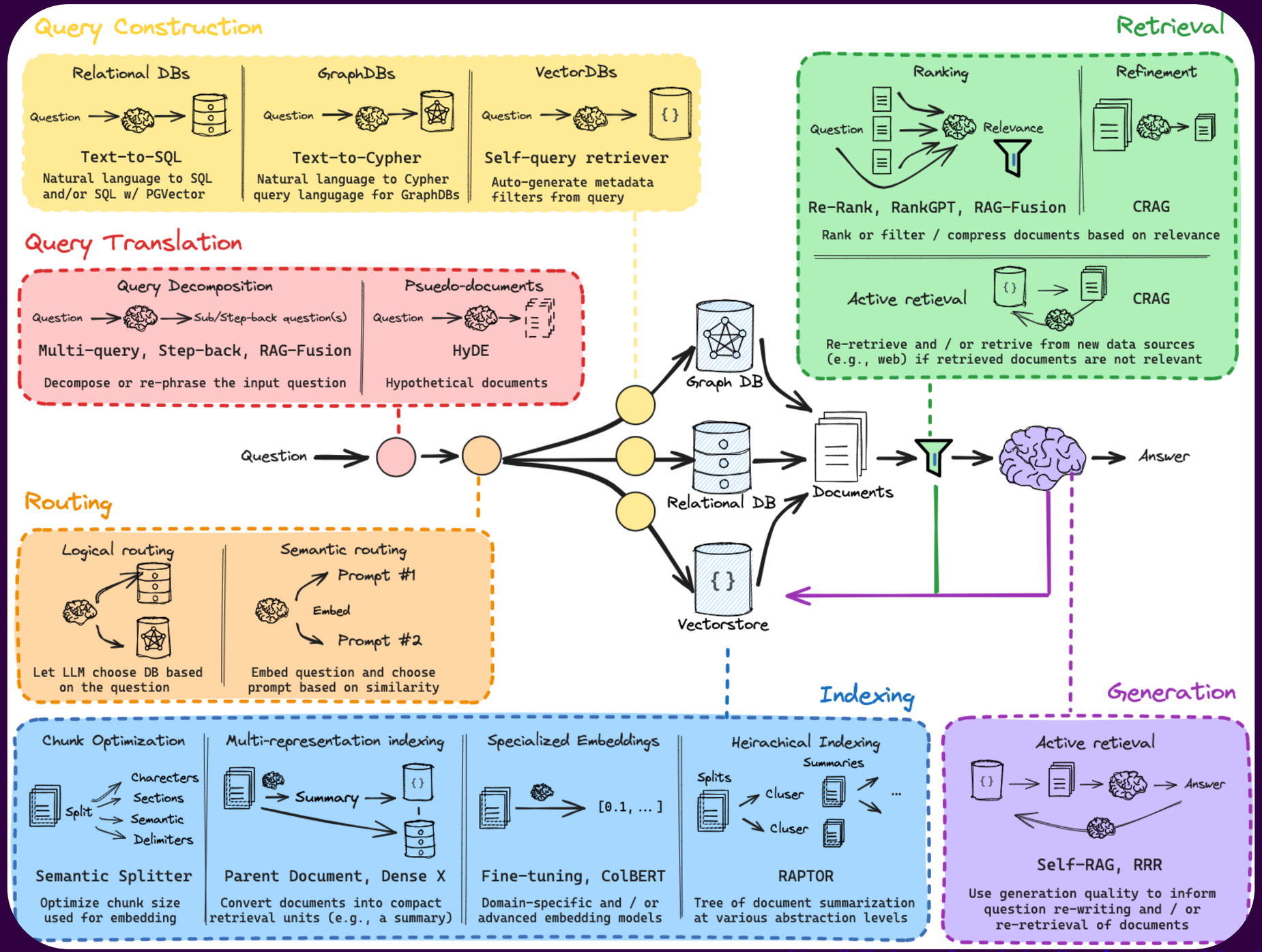
Difference between Naive and Advanced RAG (Image by the author, inspired by [1])

Naive RAG vs Advanced RAG

What do we do?

- **Data:** Can we store additional information beyond raw text chunks?
- **Embeddings:** Can we optimize our embedding representations?
- **Retrieval:** Can we do better than top-k embedding lookup?
- **Synthesis:** Can we use LLMs for more than generation? v





https://github.com/langchain-ai/rag-from-scratch/tree/main?tab=readme-ov-file

Resources

- All rag techniques
- Qdrant + DataTalks.club free course
- Similarity search HNSW
- Building nerual search service with ST and Qdrant
- Your RAG powered by Google Search Technology
- https://arxiv.org/abs/2005.11401
- Embedding models leaderboard
- Let’s talk about LlamaIndex and LangChain
- Retrieval-Augmented Generation (RAG) framework in Generative AI
- Retrieval-Augmented Generation (RAG): From Theory to LangChain Implementation
- Advanced Retrieval-Augmented Generation: From Theory to LlamaIndex
- Retrieval-augmented generation for large language models: A survey [arXiv]

THANK YOU FOR YOUR ATTENTION!!

Q&A

https://www.linkedin.com/in/mohammed-arbi-nsibi-584a43241/