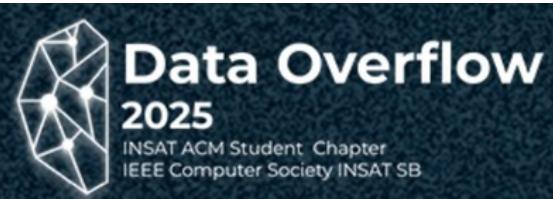


# Introduction to RAG

19 / 02 / 2025



By Mohammed Arbi Nsibi



Hello everynyan



# MOHAMED ARBI NSIBI

- Final year ICT engineering student@ SUP'COM
- GDG Carthage member
- Mentor of GDGoC SUP'COM & ISAMM
- Former GDSC Lead 23/24

mohammedarbinsibi@gmail.com



<https://huggingface.co/Goodnight7>



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

# Content

- Why do we need RAG?
- RAG components
- Frameworks
- Speaking on Your Behalf : Building a ChatBot
- QUIZ



# Motivation (Why do we need RAG?)

By Mohammed Arbi Nsibi

# **1- The need for an external Knowledge !**

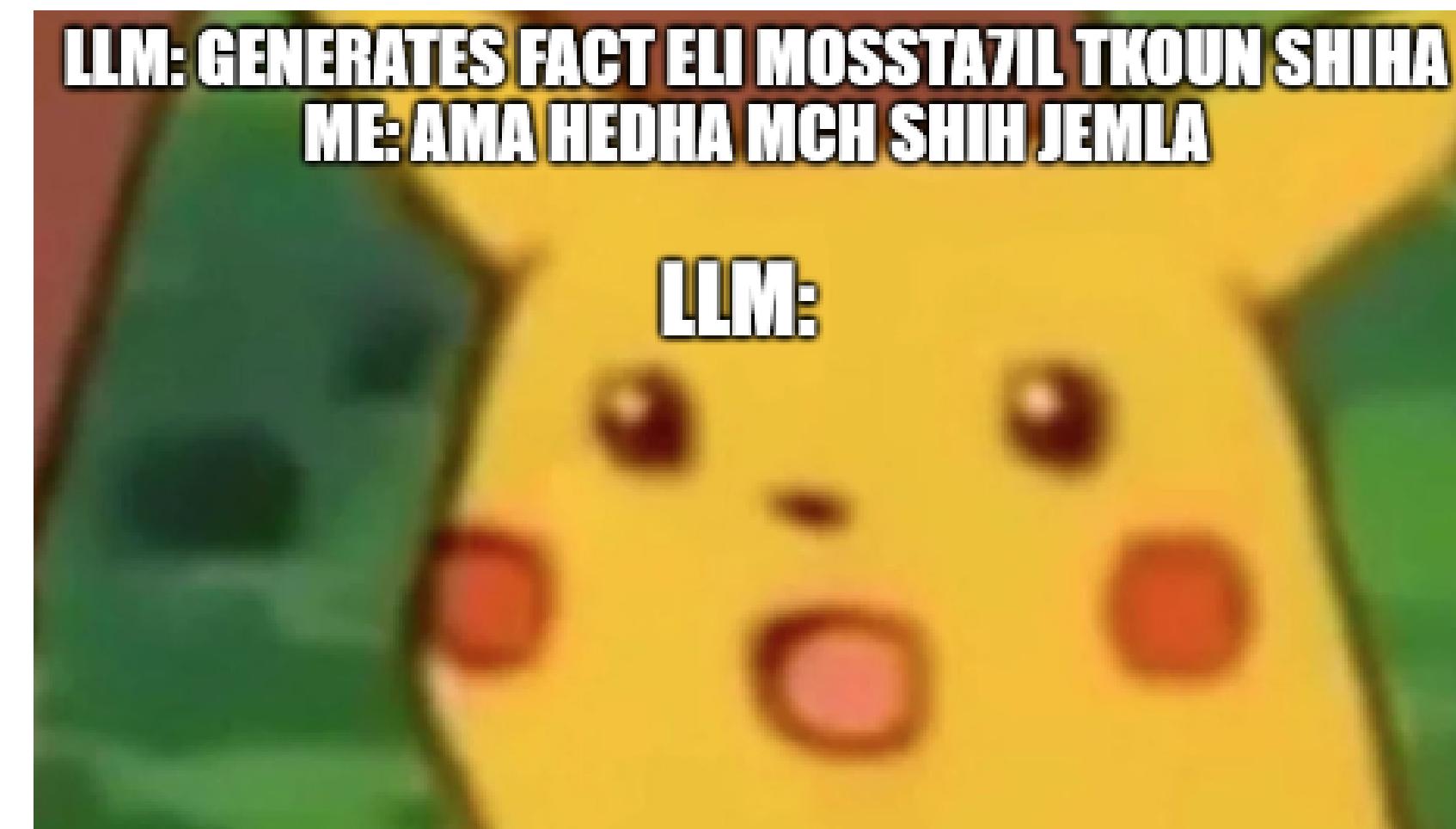
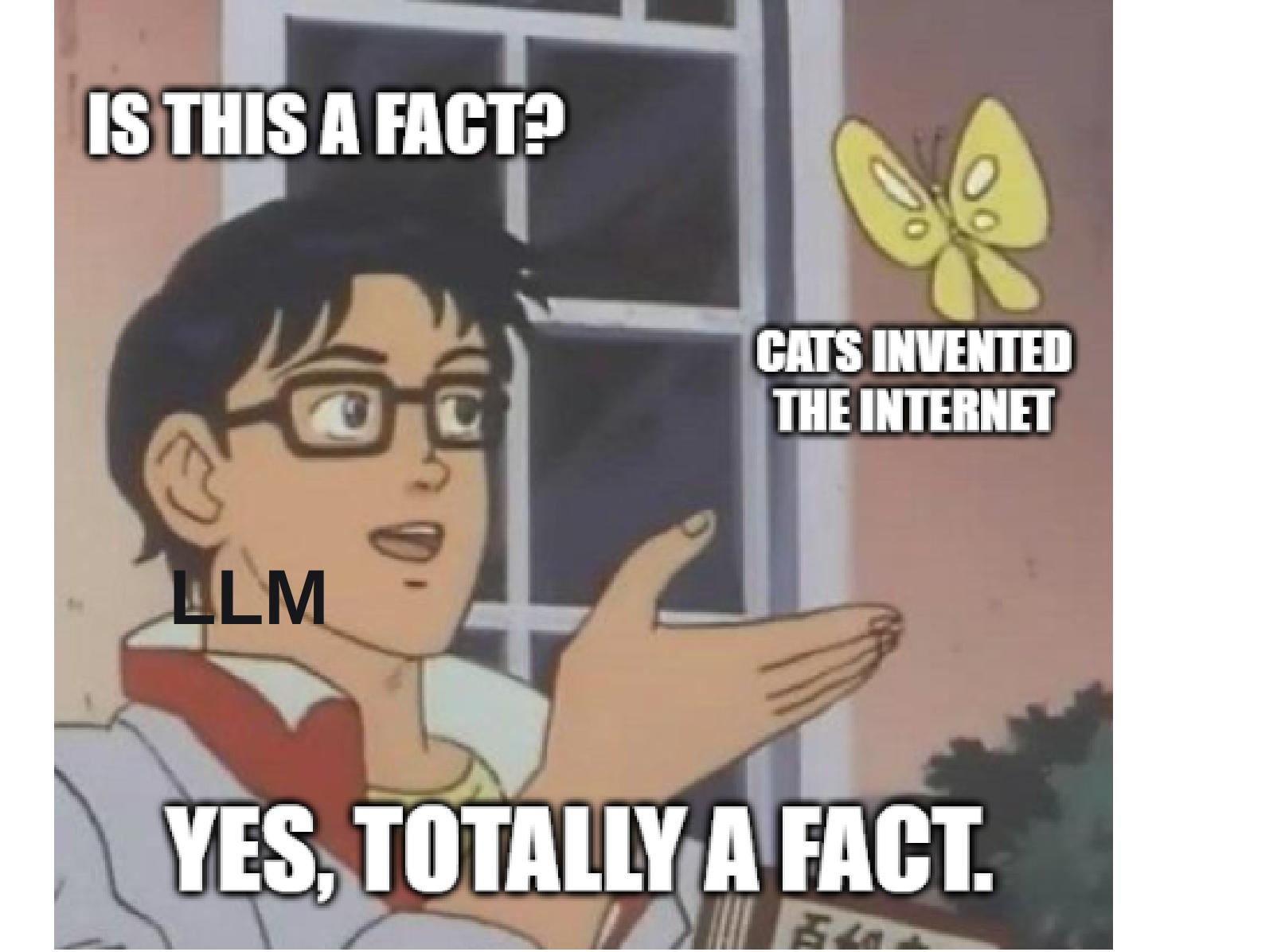
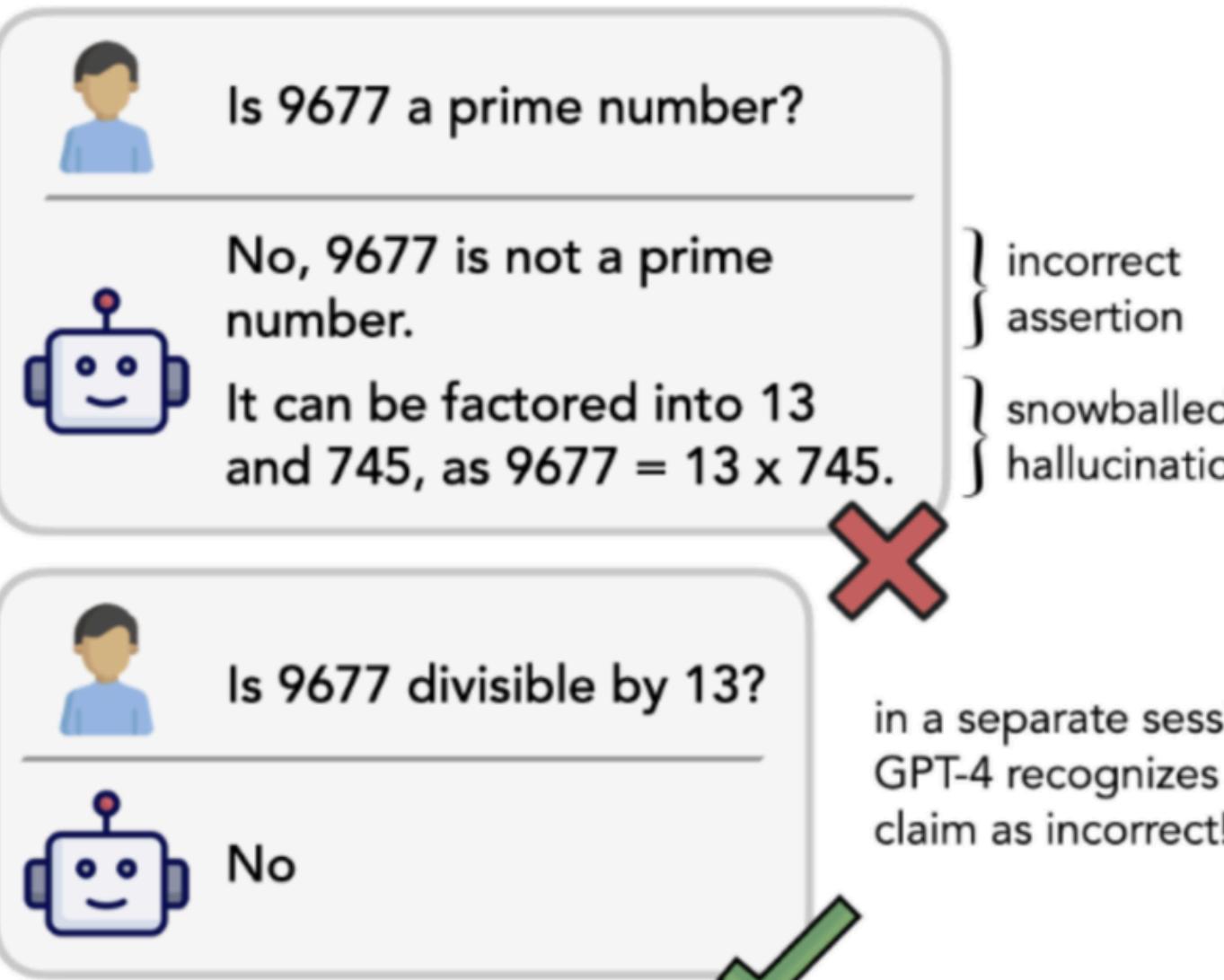
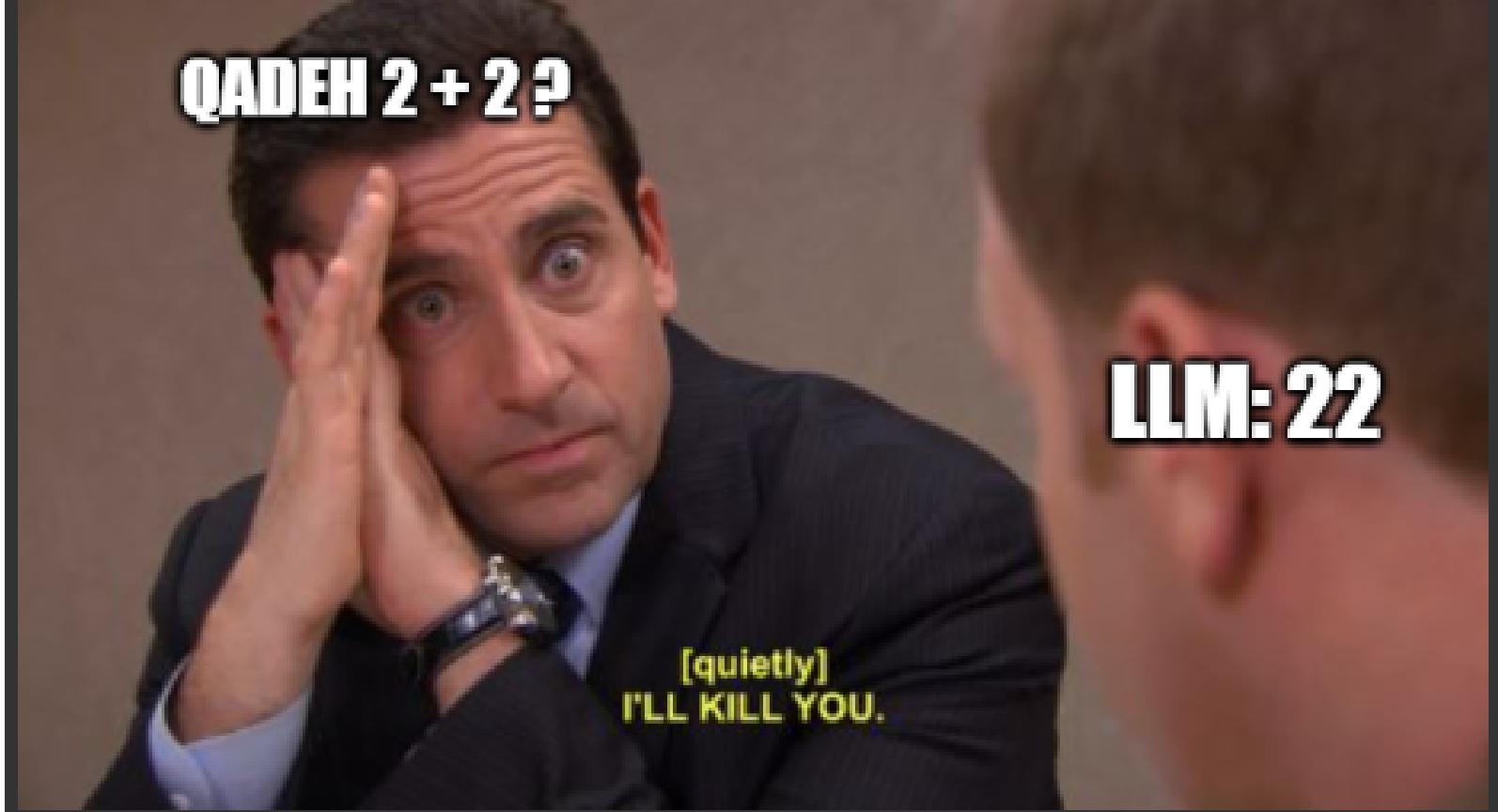
*By Mohammed Arbi Nsibi*

## 2- Hallucinations



By Mohammed Arbi Nsibi

## 2- Hallucinations



# Hallucinations

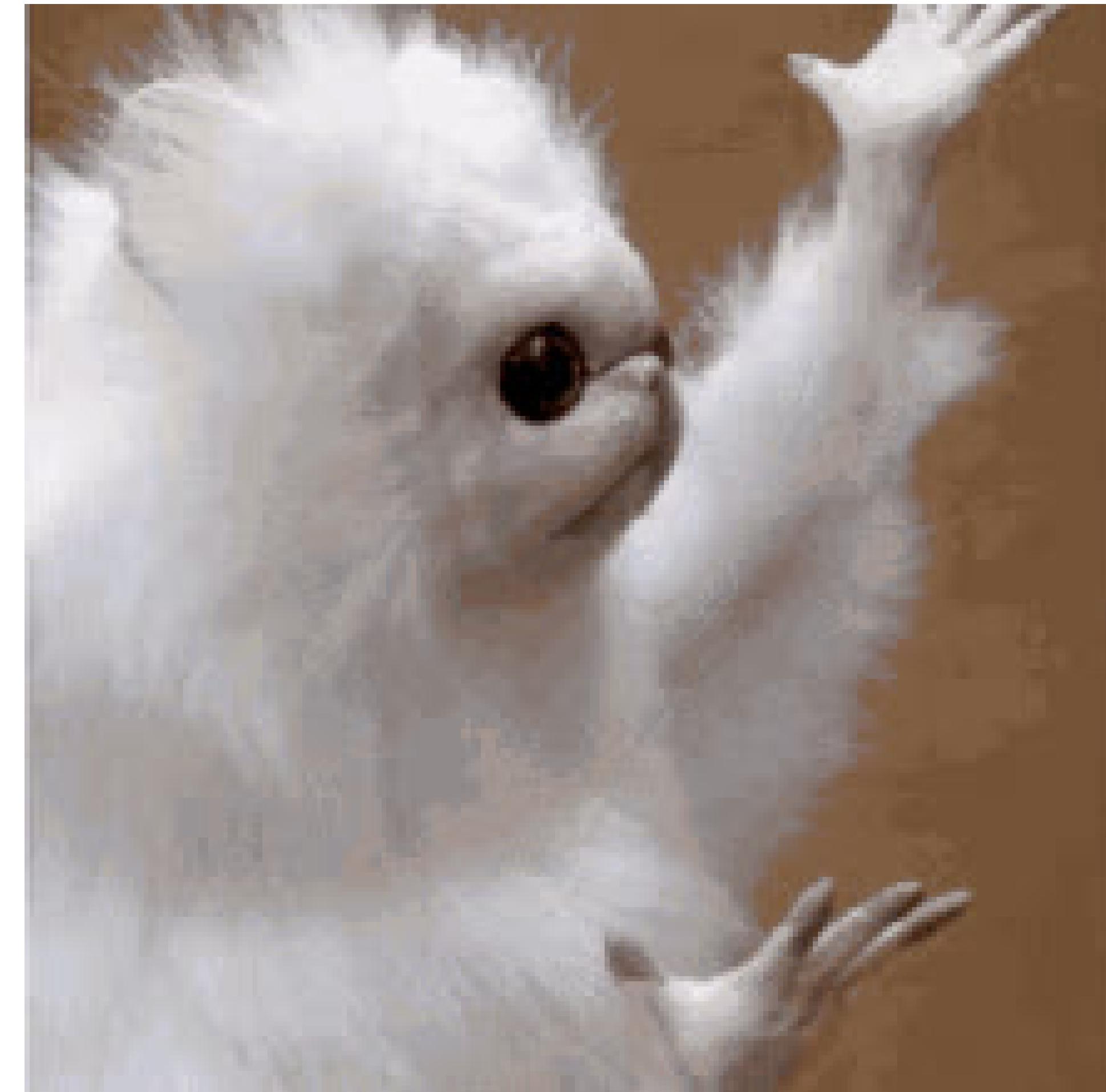
- The model is not trained on enough data.
- The model is trained on noisy or dirty data.
- The model is not given enough context .
- The model is not given enough constraints (rules, guidelines, or limitations)

LLM AFTER TRAINING  
ON 90% OF THE INTERNET...



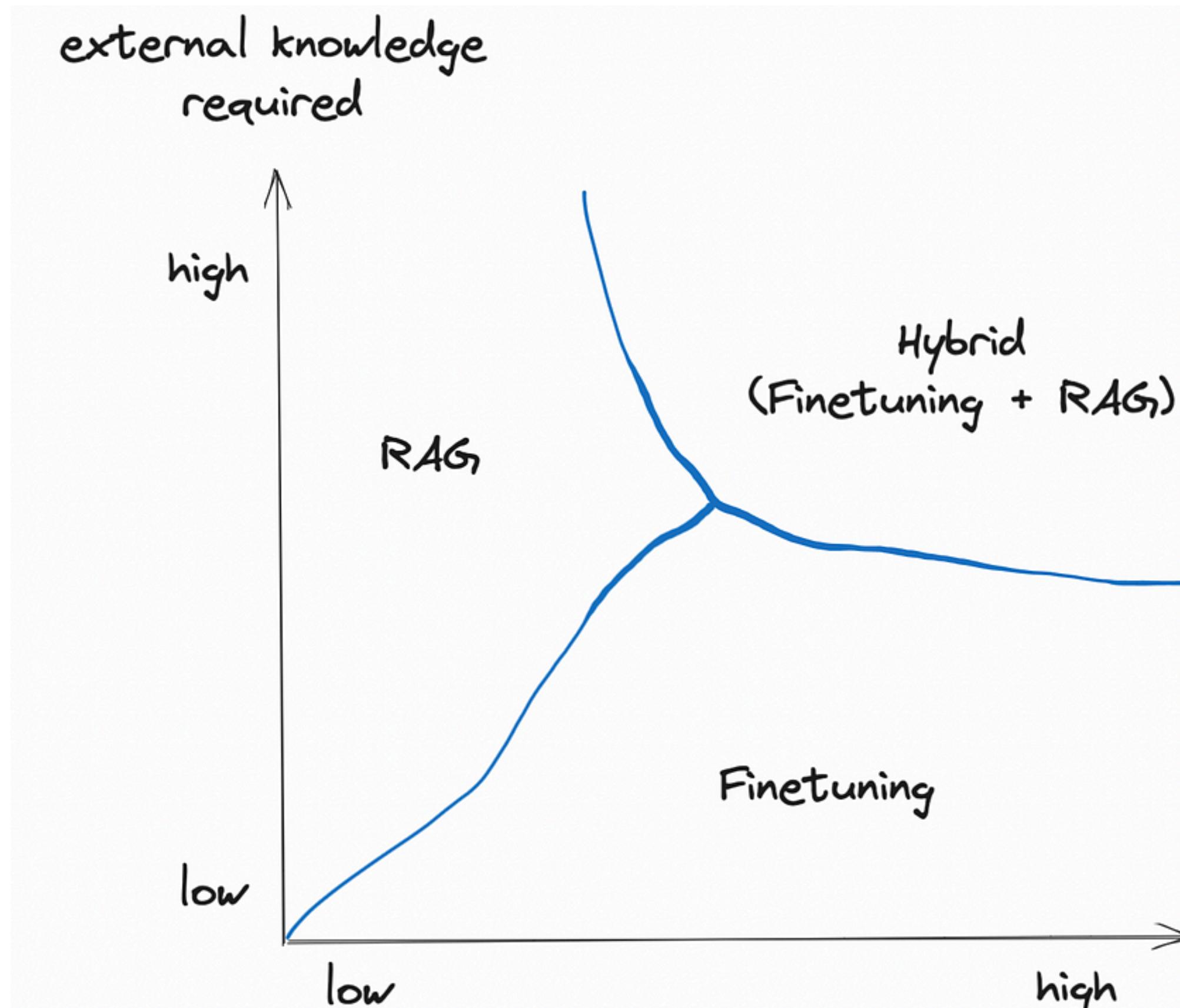
By Mohammed Arbi Nsibi

# Solutions ?



By Mohammed Arbi Nsibi

# RAG / Fine-tuning



model adaptation required  
(e.g. behaviour/  
writing style/  
vocabulary)

# But wait...

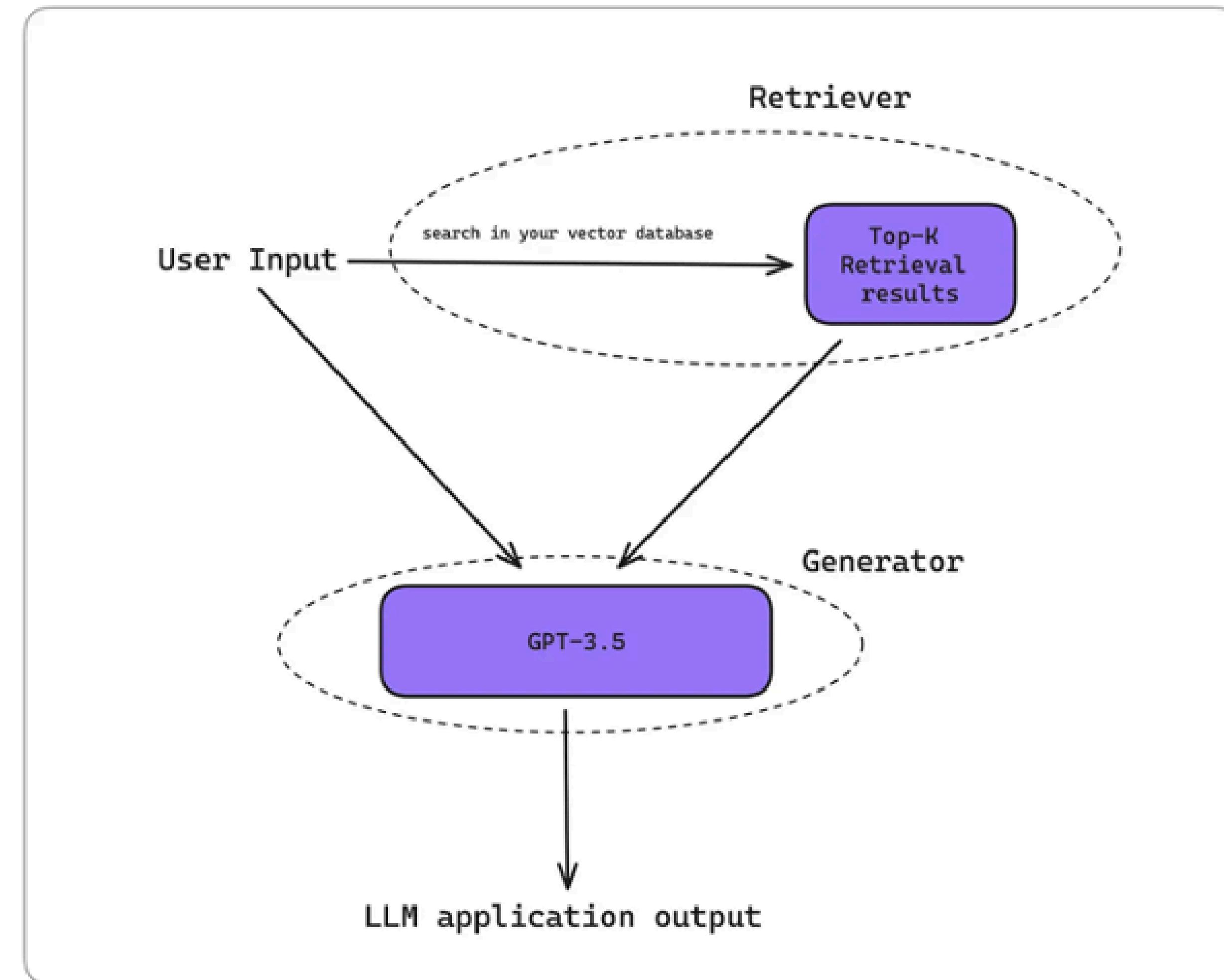
# WTF is RAG ?



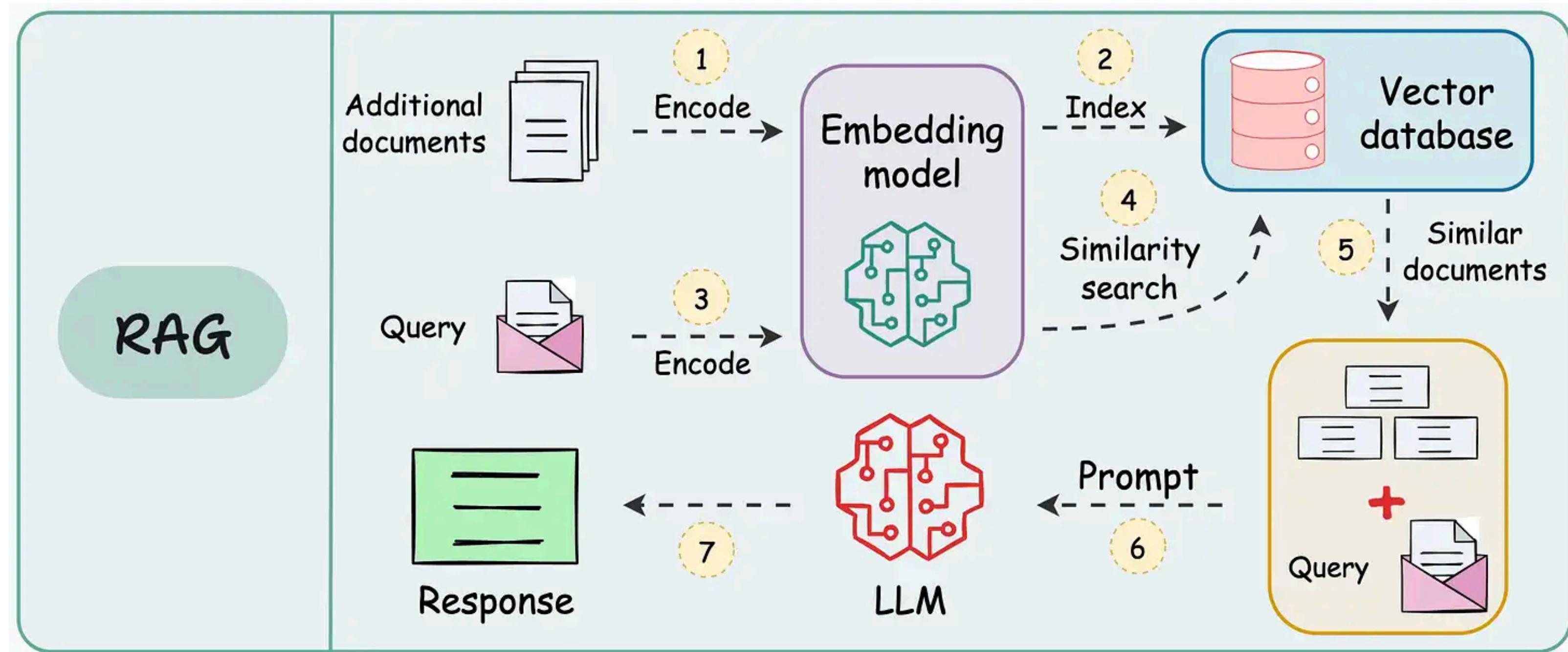
*Wait a minute, who are you?*

# RAG (Retrieval-augmented generation)

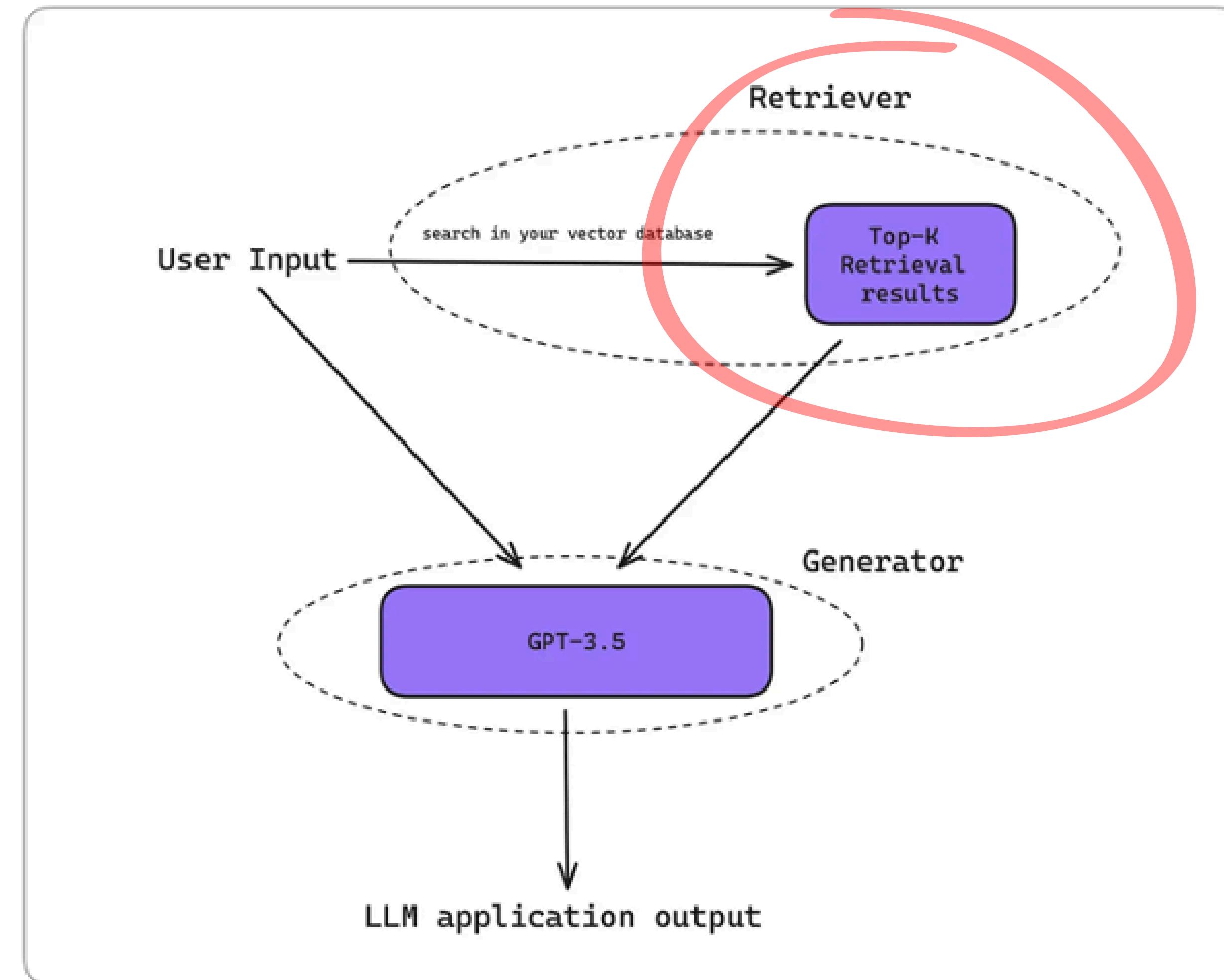




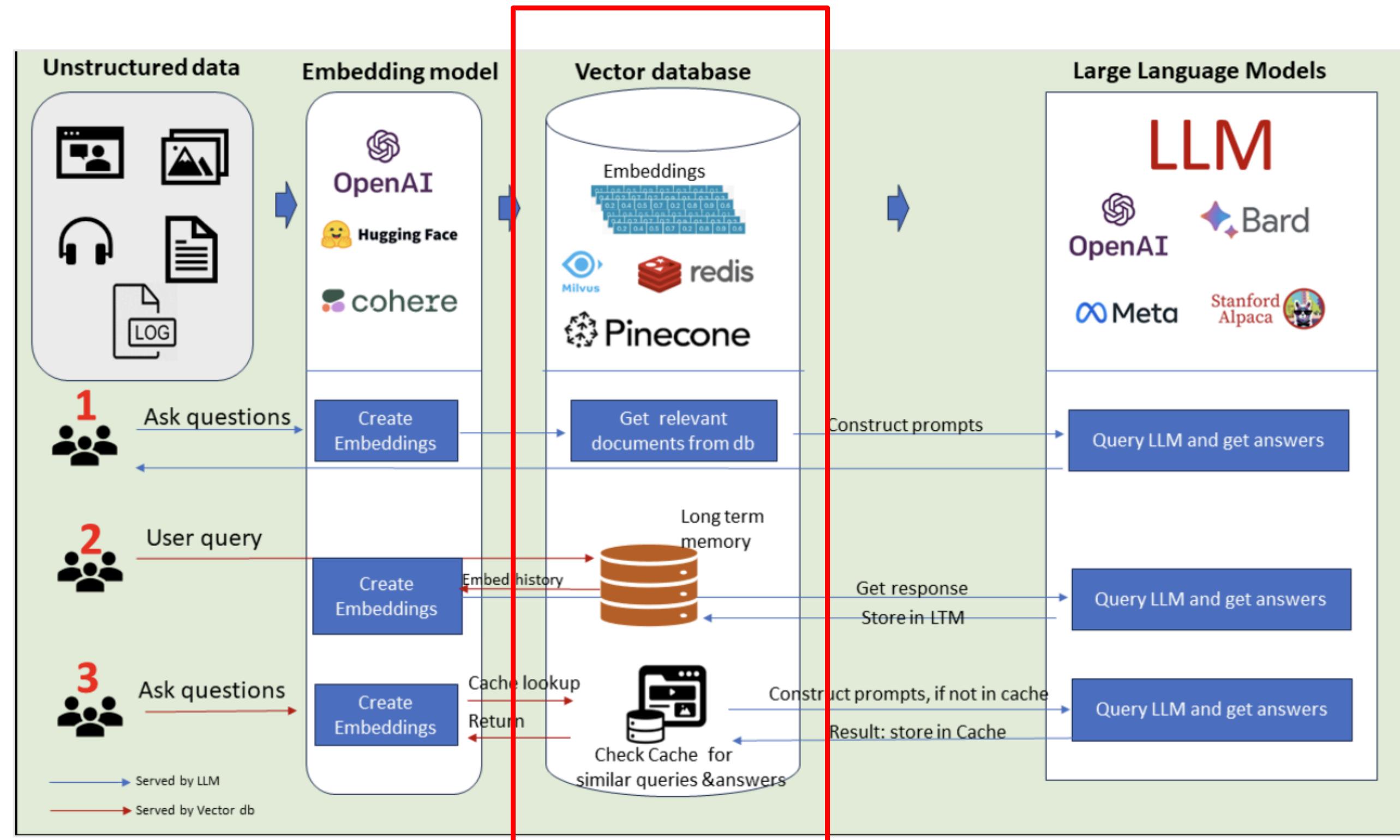
# RAG (Retrieval-augmented generation)



By Mohammed Arbi Nsibi



# Vector Database





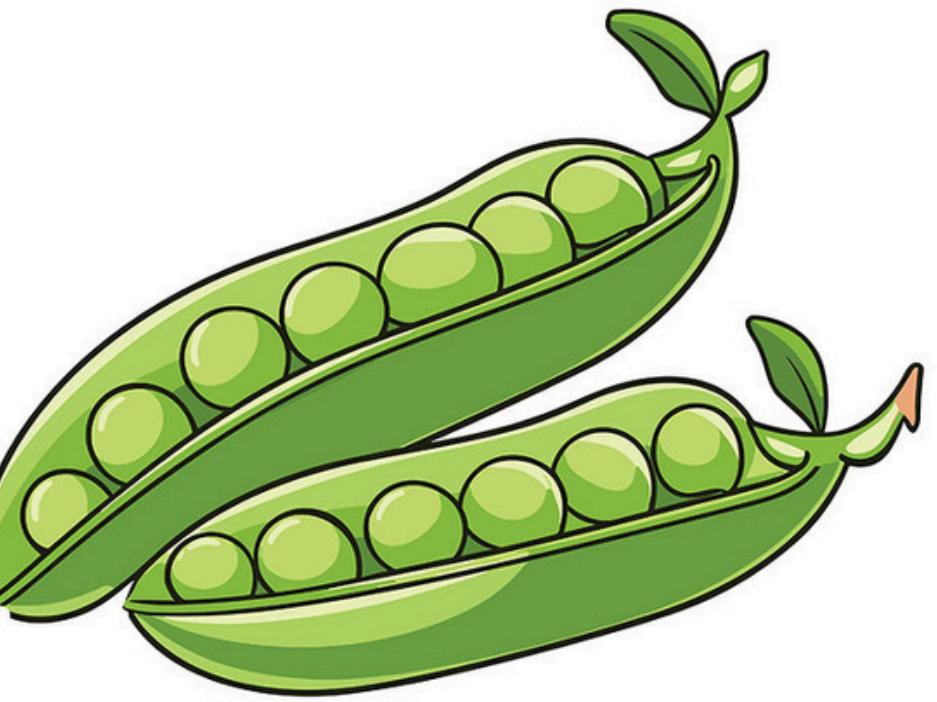
tea



coffee



tea

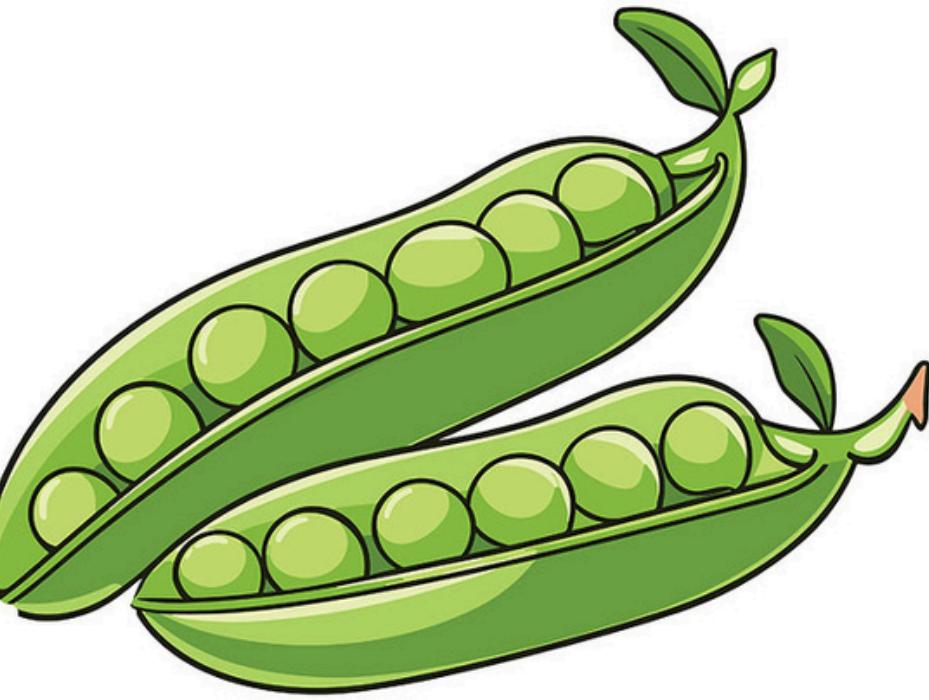


pea



tea

≠



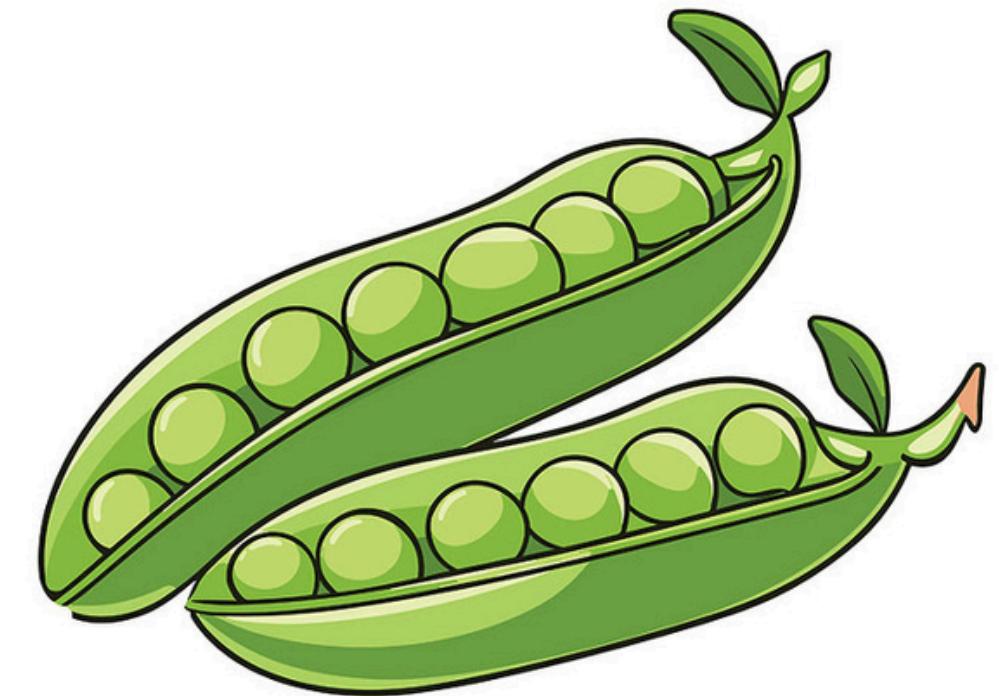
pea



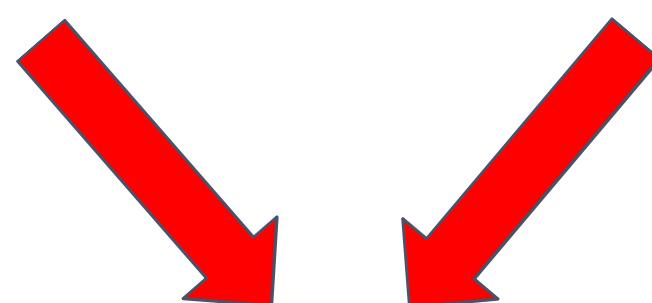
tea



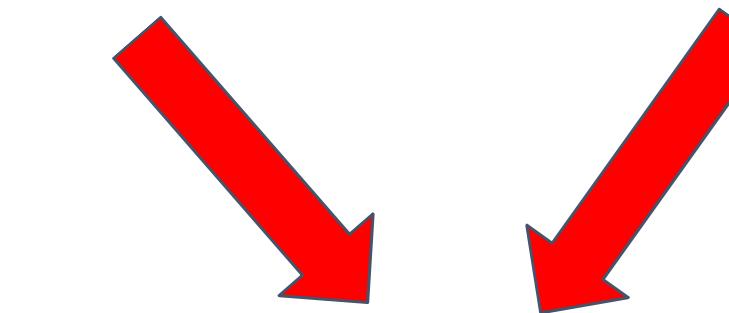
coffee



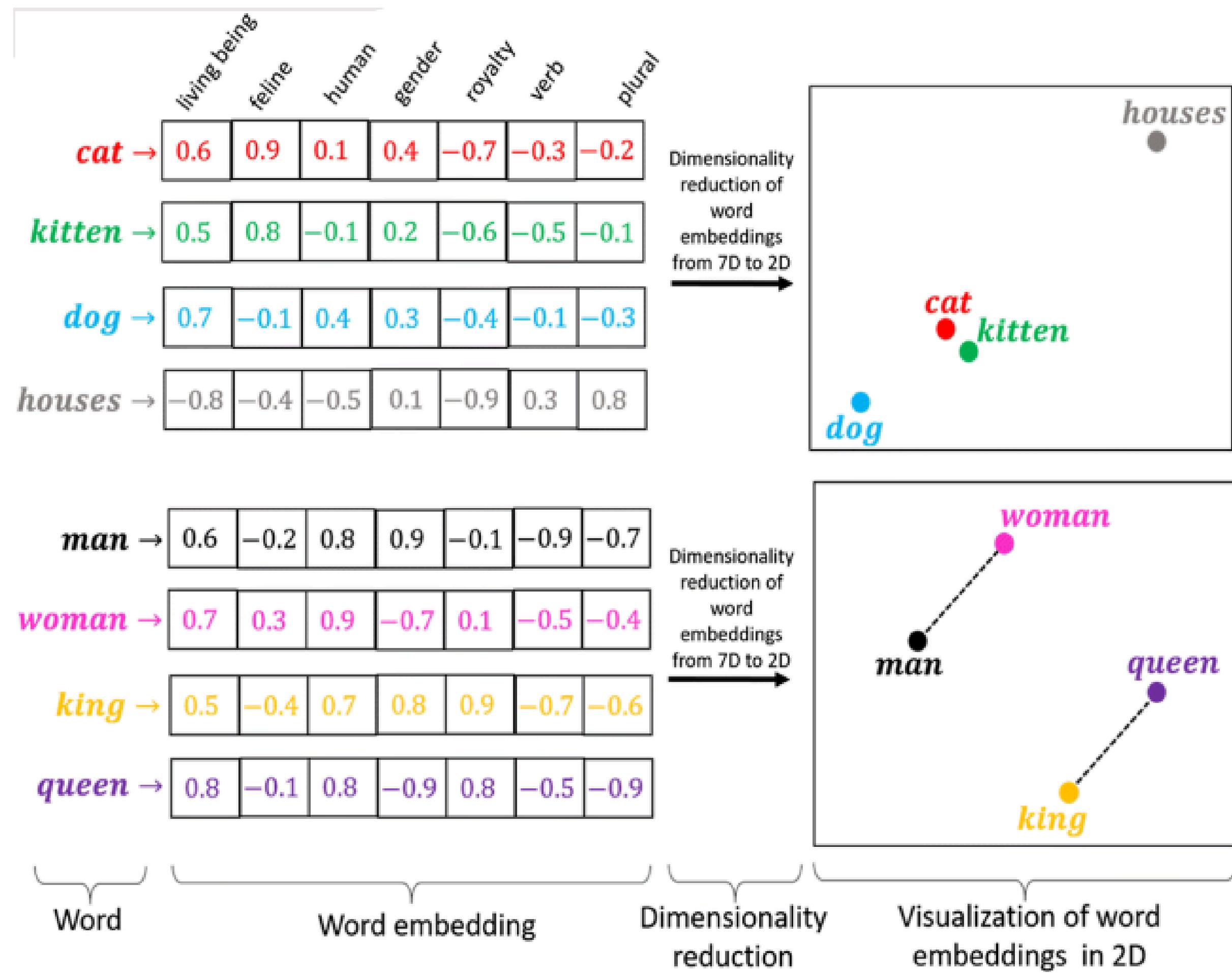
pea



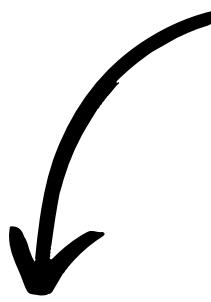
distance = 0.3



distance = 0.7



**How can we get these word embedding vectors?**

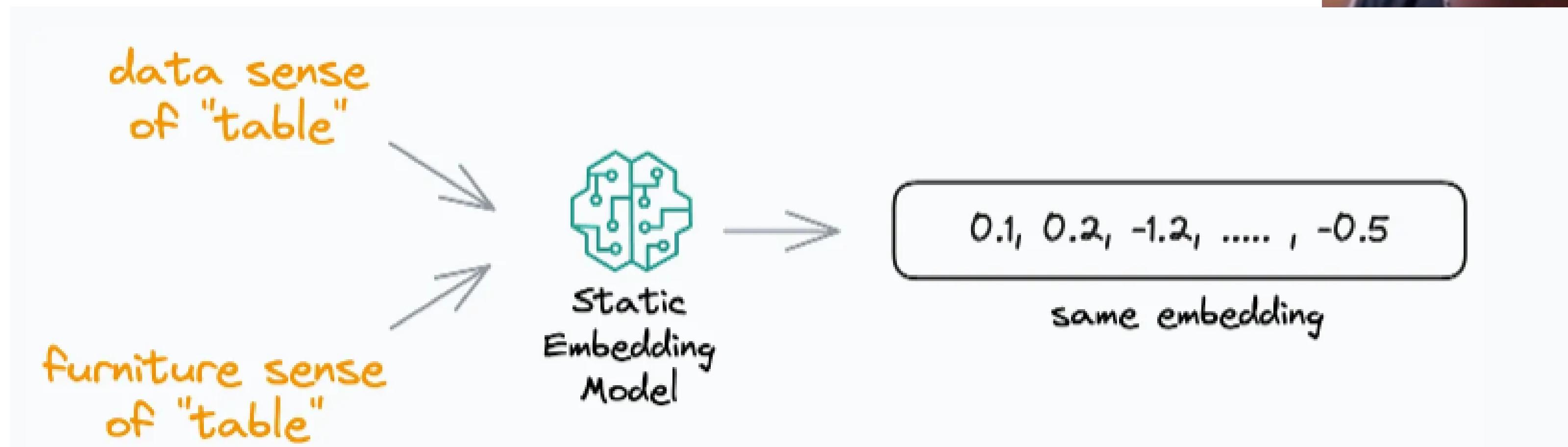


## **Pre-Trained Word Embeddings**

- **Word2vec by Google 2013**
- **GloVe by Stanford  
(Global Vectors for Word Representation)**
- **fasttext by Facebook**

## Limitations

- Convert this data into a **table** in Excel.
- Put this bottle on the **table**.



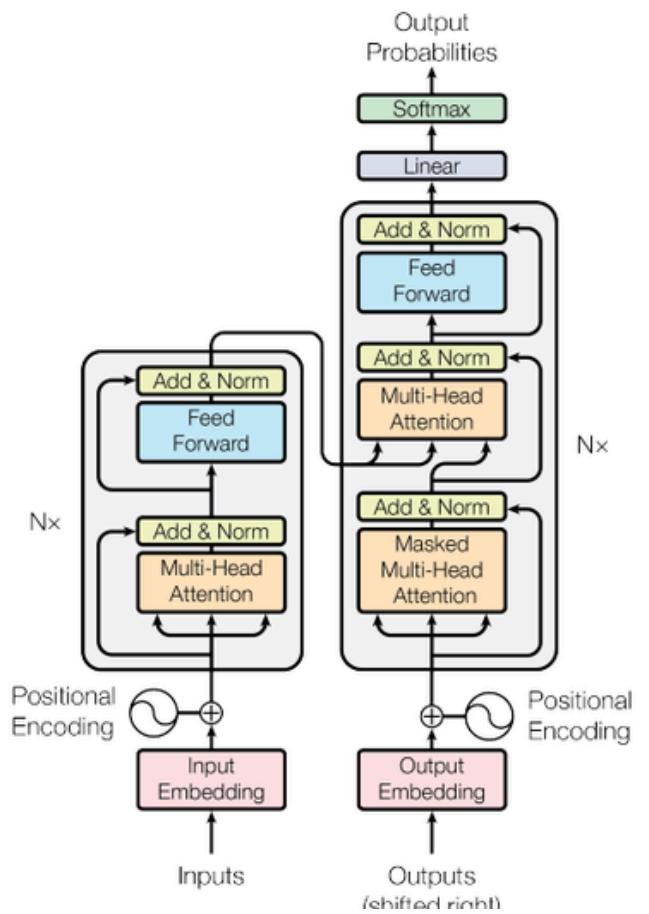
# Solution ?



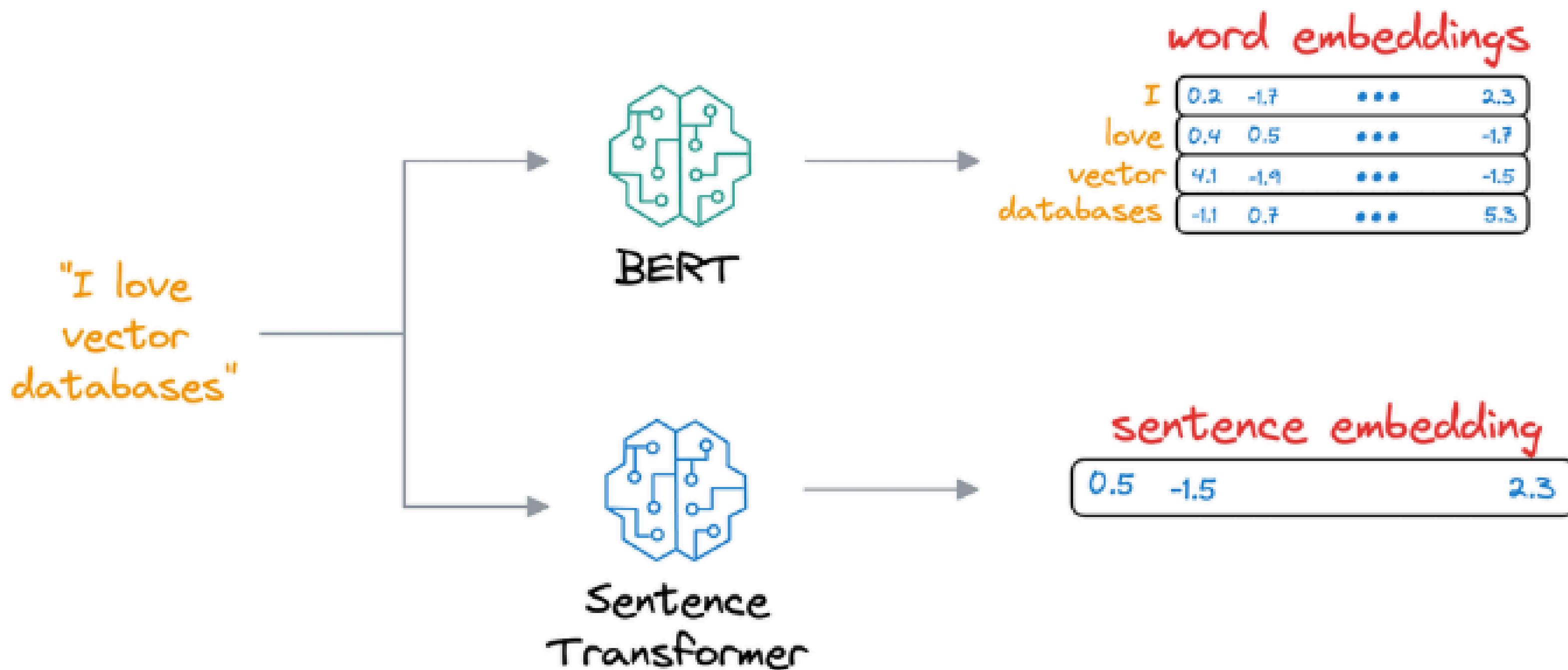
- **BERT (Bidirectional Encoder Representations from Transformers)**



- **DistilBERT: BERT which is around 40% smaller:**
- **ALBERT: A Lite BERT (ALBERT).**

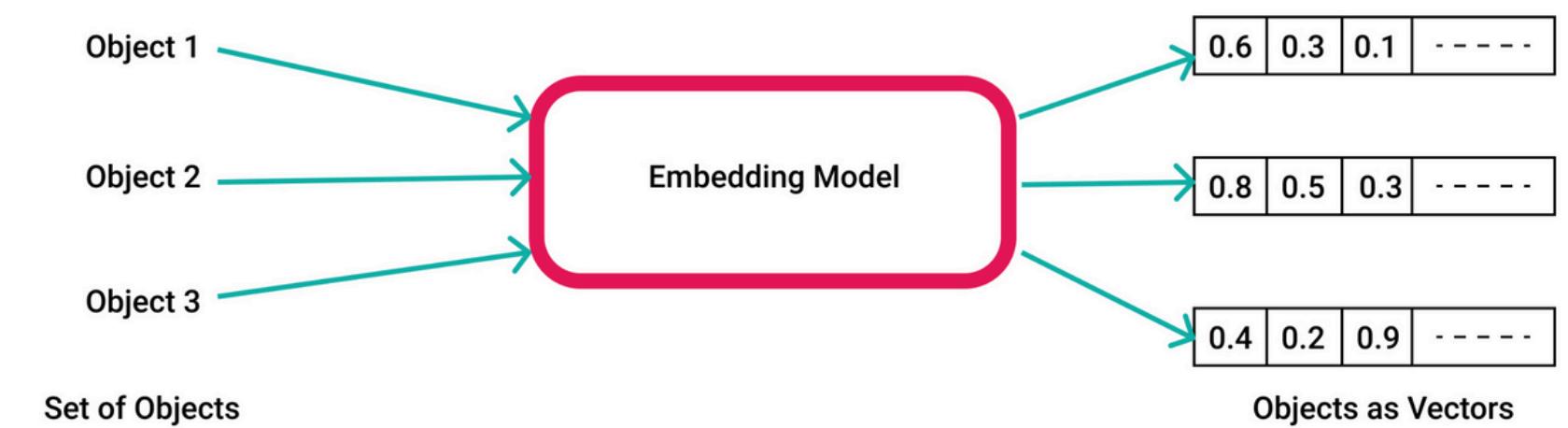


# Solution ?

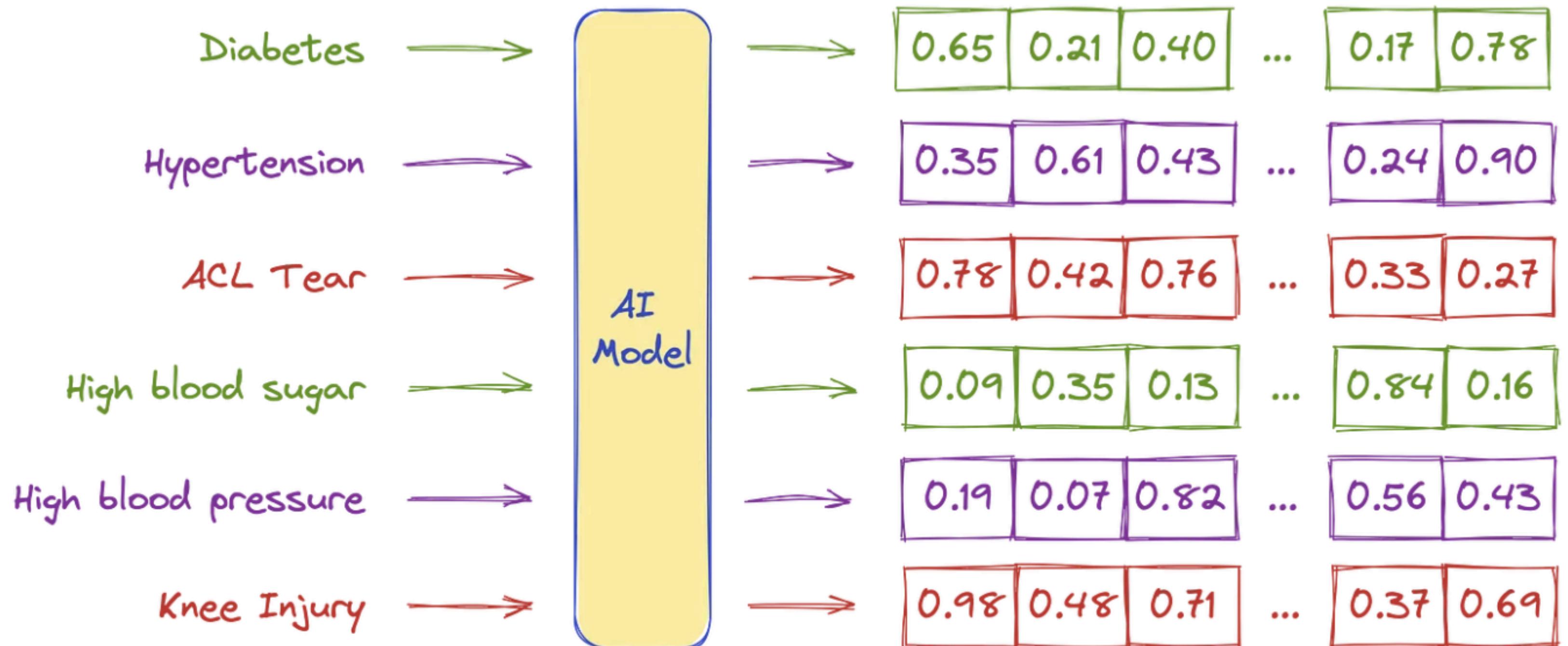


# Embedding Model

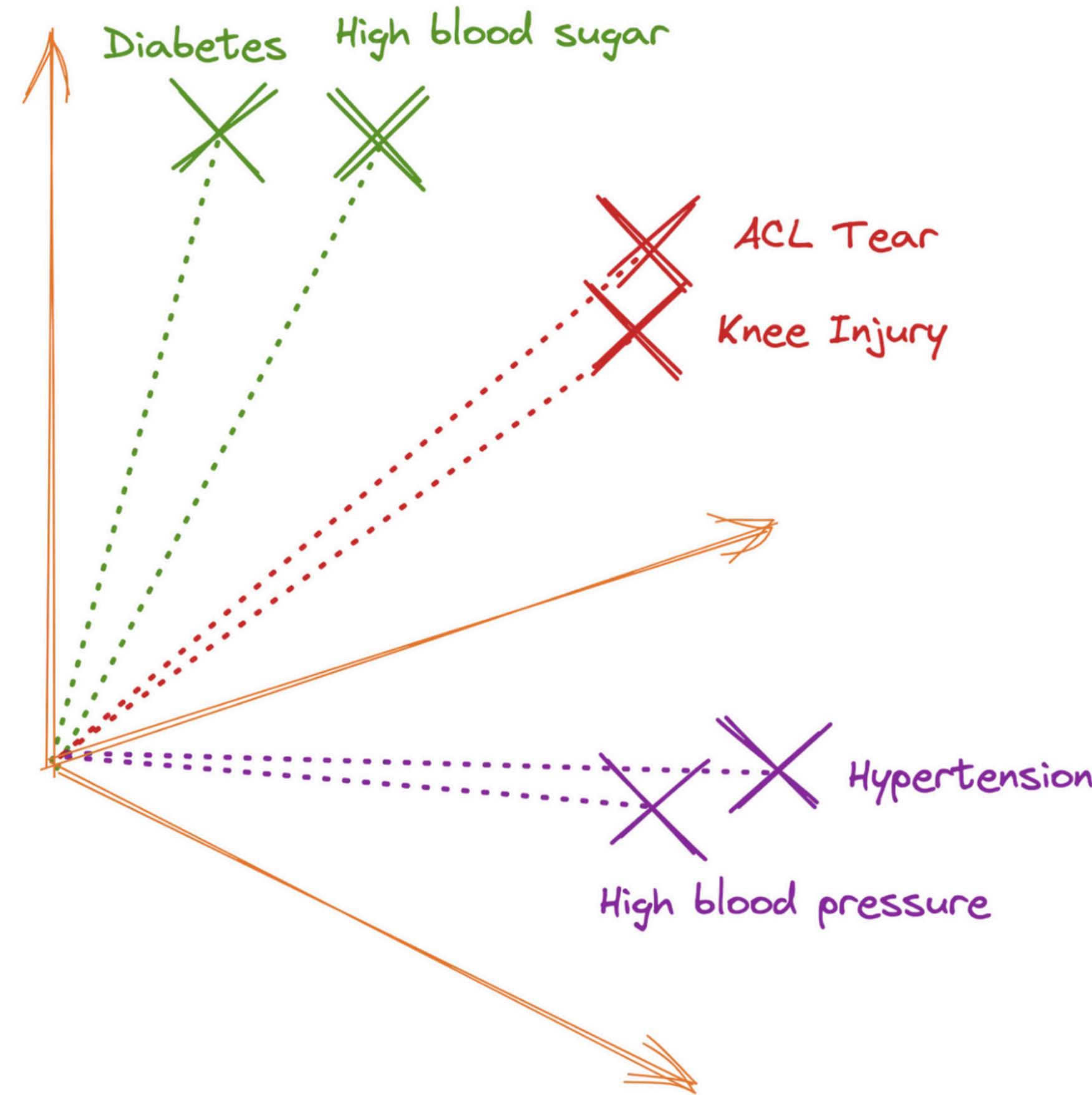
- These vectors live in a high-dimensional space where the proximity between vectors reflects the relatedness of the original items.
- **Embedding model** trained along LLM and learn to **produce representation(vectors)** based on **context** in word appear.



# Embedding Space



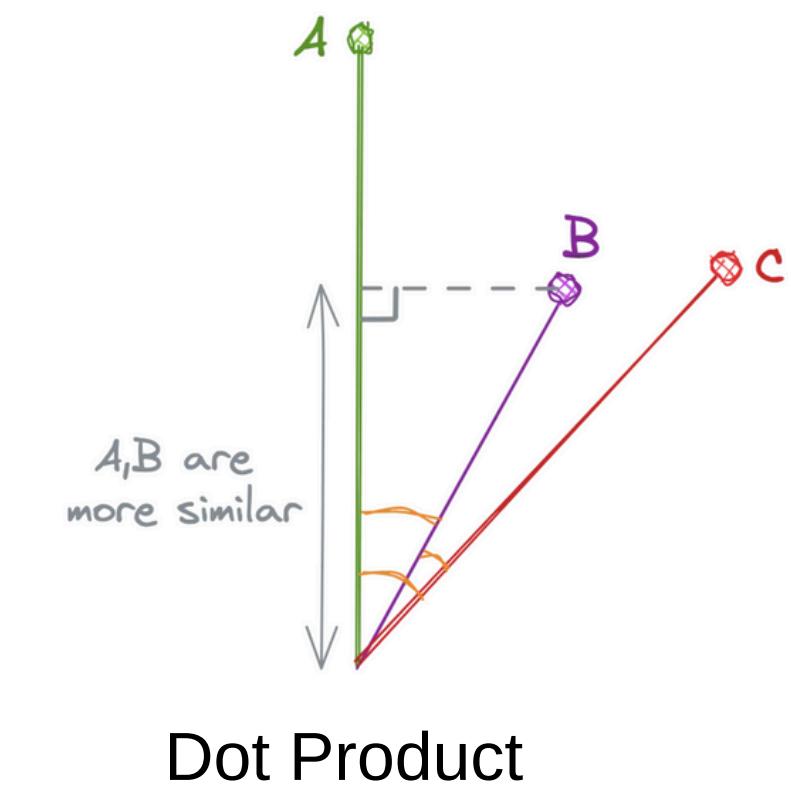
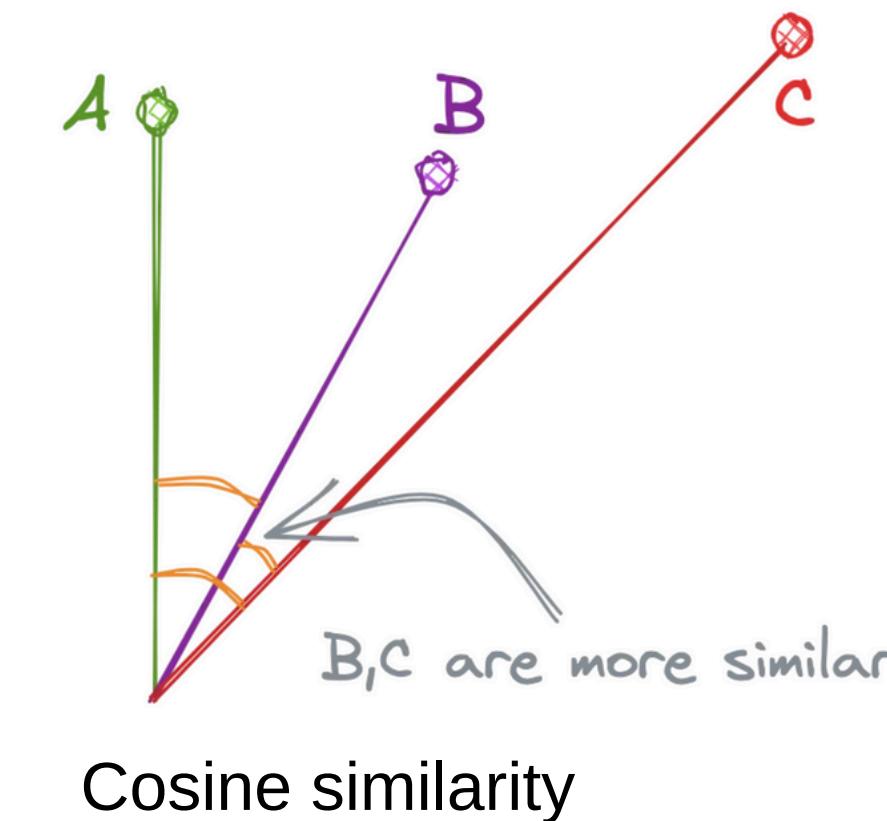
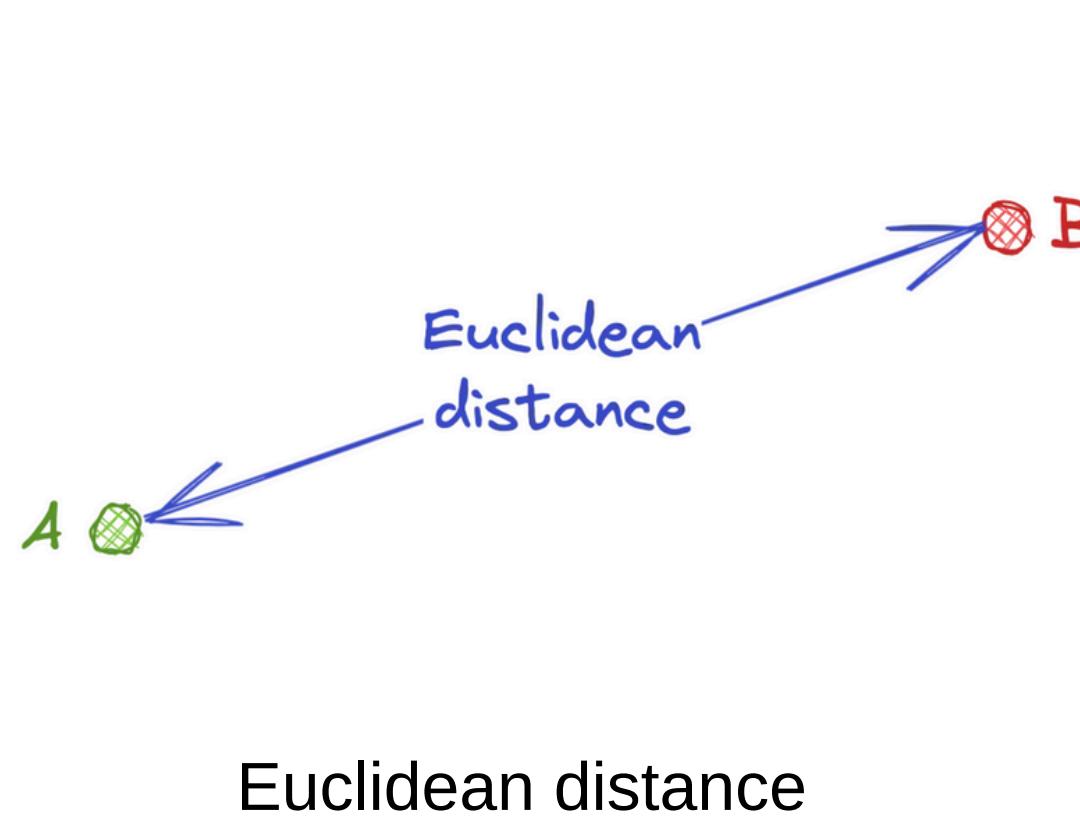
# Embedding Space

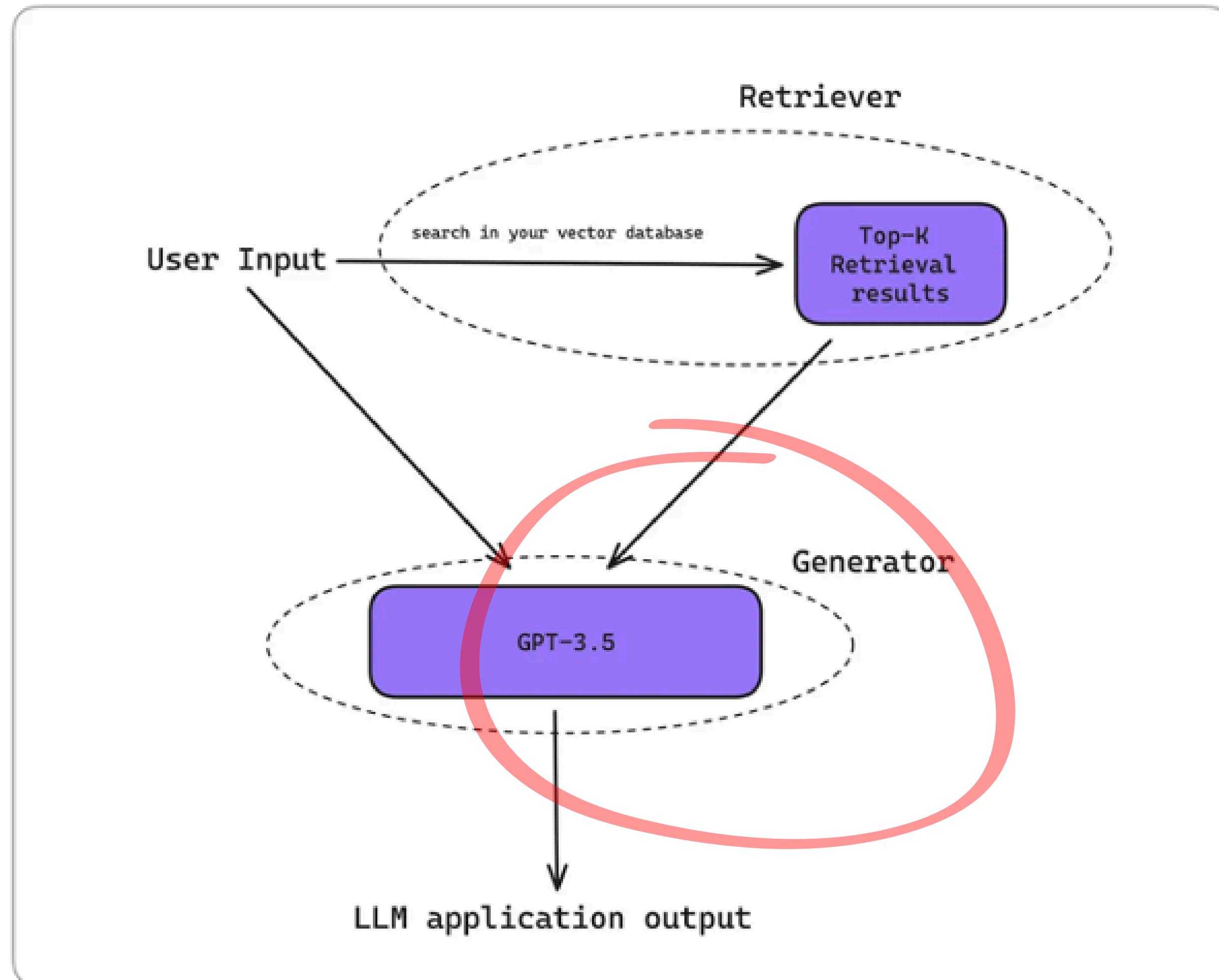


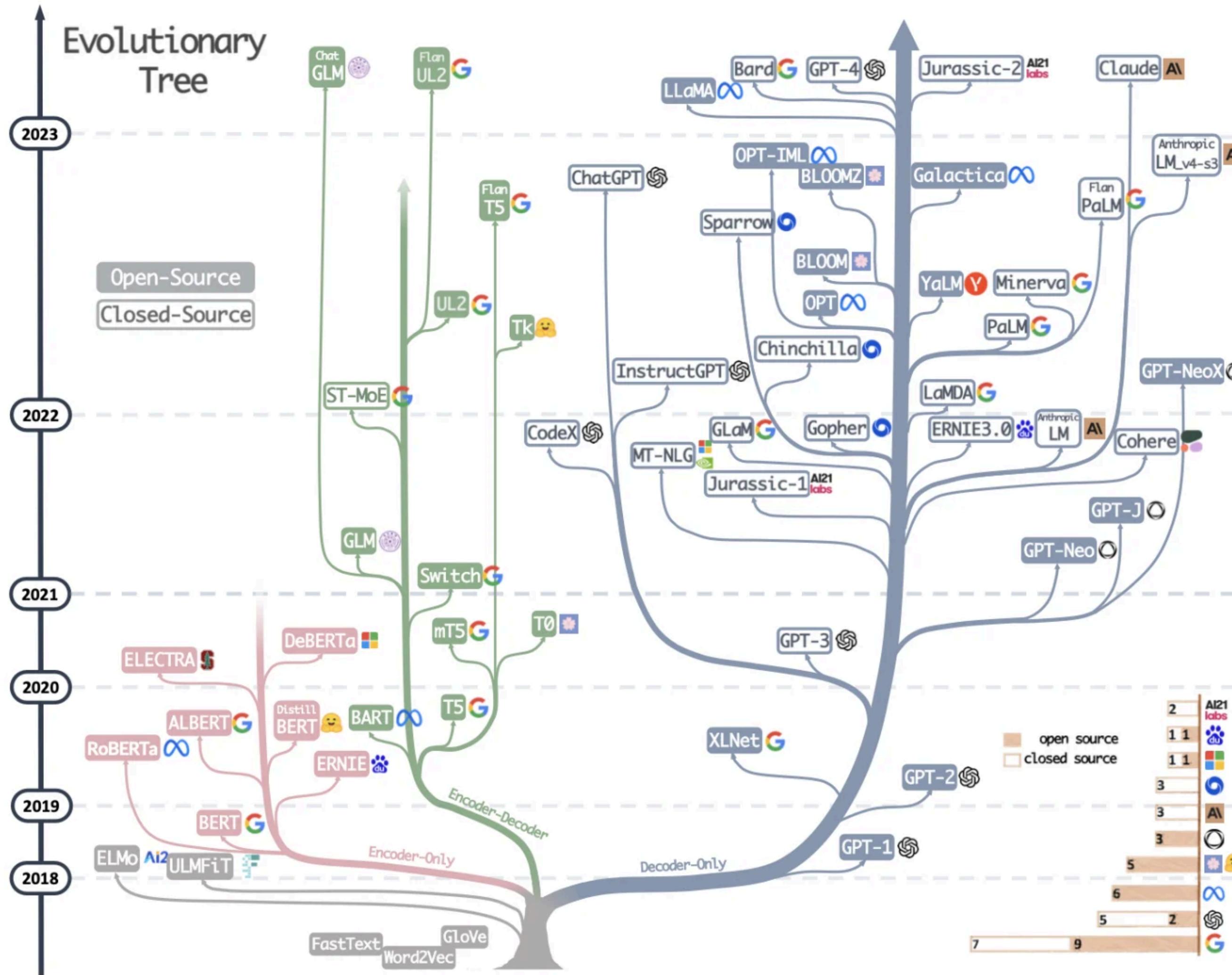
# Querying a vector database

**Similarity Calculation => objective is to return the nearest neighbors**

- For calculating similarity, there are several methods:
  - measuring distance - euclidean distance
  - cosine similarity or inner product







The evolutionary tree of modern  
LLMs via  
<https://arxiv.org/abs/2304.13712>.

# Where to find pretrained LLMs ?



# Hugging Face

# groq®

By Mohammed Arbi Nsibi

# Where to find pretrained LLMs ?

Models 1,028,261 [Filter by name](#)

Full-text search [↑↓ Sort: Trending](#)

 openai/whisper-large-v3-turbo  
Automatic Speech Recognition • Updated 1 day ago • 10k • 324

 black-forest-labs/FLUX.1-dev  
Text-to-Image • Updated Aug 16 • 1.14M • 5.03k

 jasperai/Flux.1-dev-Controlnet-Upscaler  
Image-to-Image • Updated 3 days ago • 9.86k • 244

 allenai/Molmo-7B-D-0924  
Image-Text-to-Text • Updated 1 day ago • 14.5k • 273

 meta-llama/Llama-3.2-11B-Vision-Instruct  
Image-Text-to-Text • Updated 4 days ago • 139k • 479

 nvidia/NVLM-D-72B  
Image-Text-to-Text • Updated about 18 hours ago • 860 • 242

 meta-llama/Llama-3.2-1B  
Text Generation • Updated 3 days ago • 61.2k • 299

 openbmb/MiniCPM-Embedding  
Feature Extraction • Updated 2 days ago • 130k • 204

datasets 222,500 [Filter by name](#)

Full-text search [↑↓ Sort: Trending](#)

 google/frames-benchmark  
Viewer • Updated about 17 hours ago • 824 • 562 • 122

 FBK-MT/mosel  
Viewer • Updated 5 days ago • 51.1M • 21 • 42

 openai/MMMLU  
Viewer • Updated 4 days ago • 393k • 5.33k • 374

 argilla/FinePersonas-v0.1  
Viewer • Updated 19 days ago • 21.1M • 371 • 304

 fka/awesome-chatgpt-prompts  
Viewer • Updated Sep 3 • 170 • 8.36k • 5.82k

 migtissera/Synthia-v1.5-I  
Viewer • Updated 8 days ago • 20.7k • 99 • 39

 Hacker Noon/where-startups-trend  
Preview • Updated 7 days ago • 19 • 36

 k-mktr/improved-flux-prompts-photoreal-portrait  
Viewer • Updated 4 days ago • 20k • 54 • 62

 [Image-Text-to-Text](#)  [Visual Question Answering](#)

 [Document Question Answering](#)  [Video-Text-to-Text](#)

 [Any-to-Any](#)

Computer Vision

 [Depth Estimation](#)  [Image Classification](#)

 [Object Detection](#)  [Image Segmentation](#)

 [Text-to-Image](#)  [Image-to-Text](#)  [Image-to-Image](#)

 [Image-to-Video](#)  [Unconditional Image Generation](#)

 [Video Classification](#)  [Text-to-Video](#)

 [Zero-Shot Image Classification](#)  [Mask Generation](#)

 [Zero-Shot Object Detection](#)  [Text-to-3D](#)

 [Image-to-3D](#)  [Image Feature Extraction](#)

 [Keypoint Detection](#)

Natural Language Processing

 [Text Classification](#)  [Token Classification](#)

 [Table Question Answering](#)  [Question Answering](#)

 [Zero-Shot Classification](#)  [Translation](#)

 [Summarization](#)  [Feature Extraction](#)

 [Text Generation](#)  [Text2Text Generation](#)

 [Fill-Mask](#)  [Sentence Similarity](#)

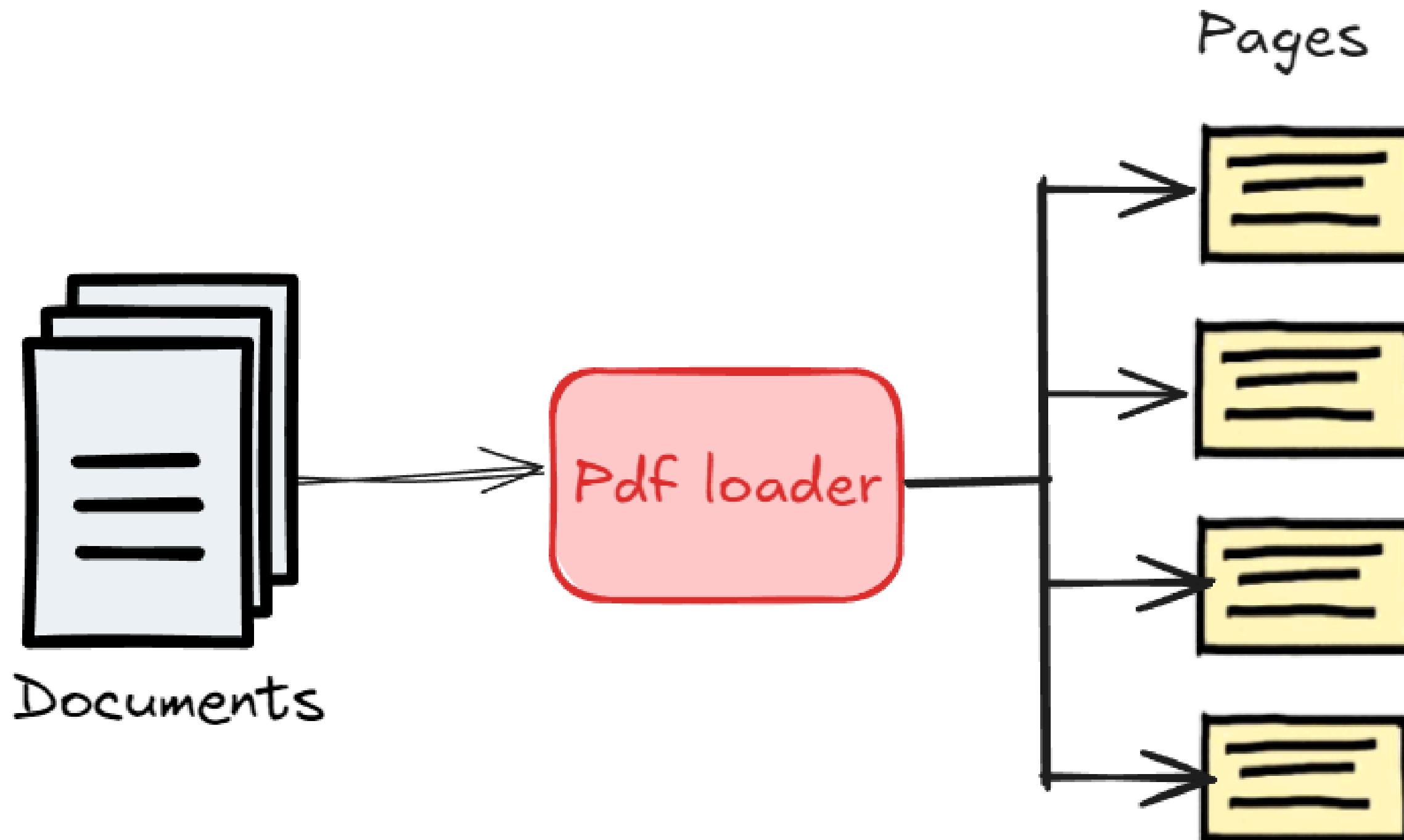
Audio

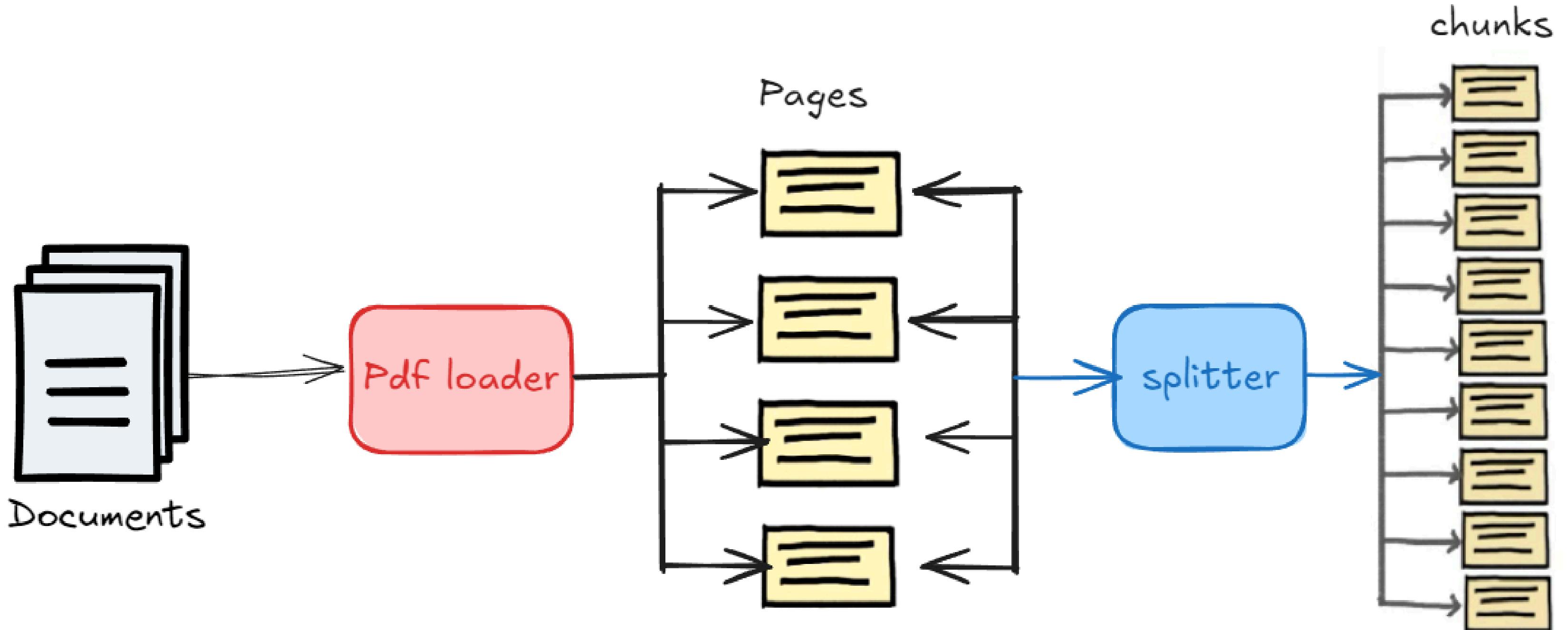
 [Text-to-Speech](#)  [Text-to-Audio](#)

 [Automatic Speech Recognition](#)  [Audio-to-Audio](#)

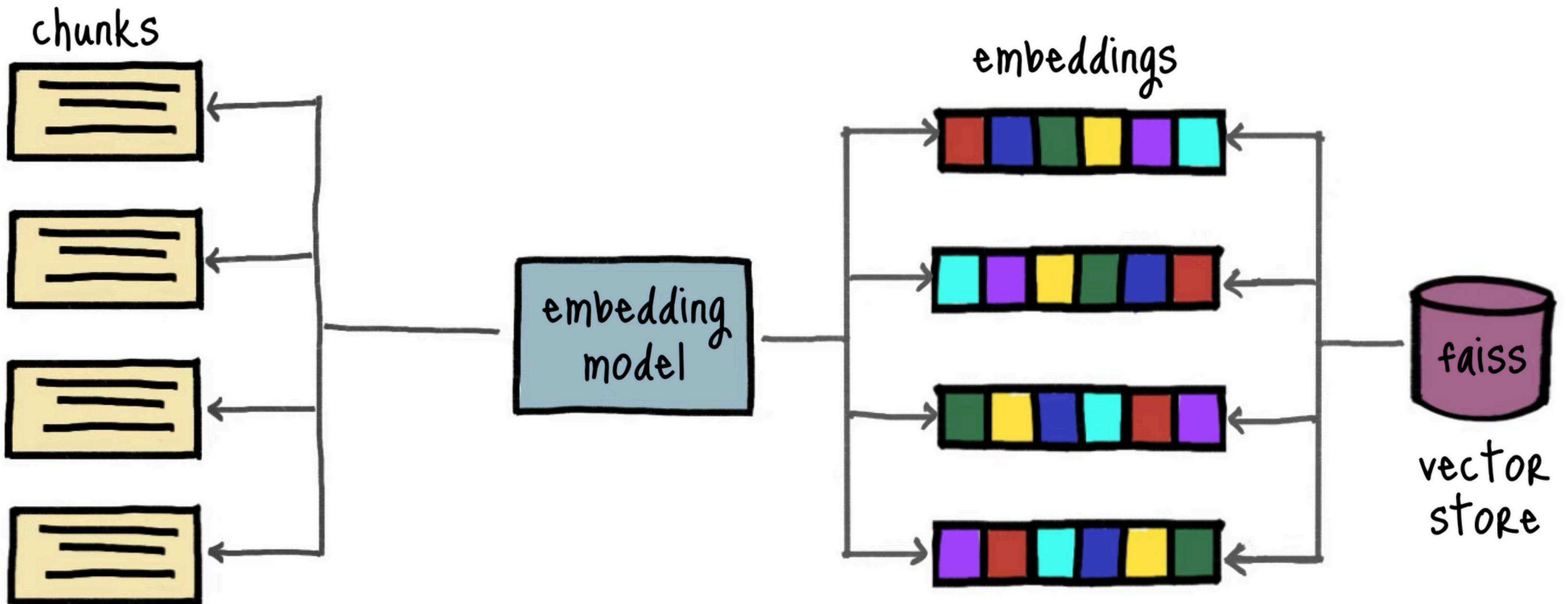
 [Audio Classification](#)  [Voice Activity Detection](#)

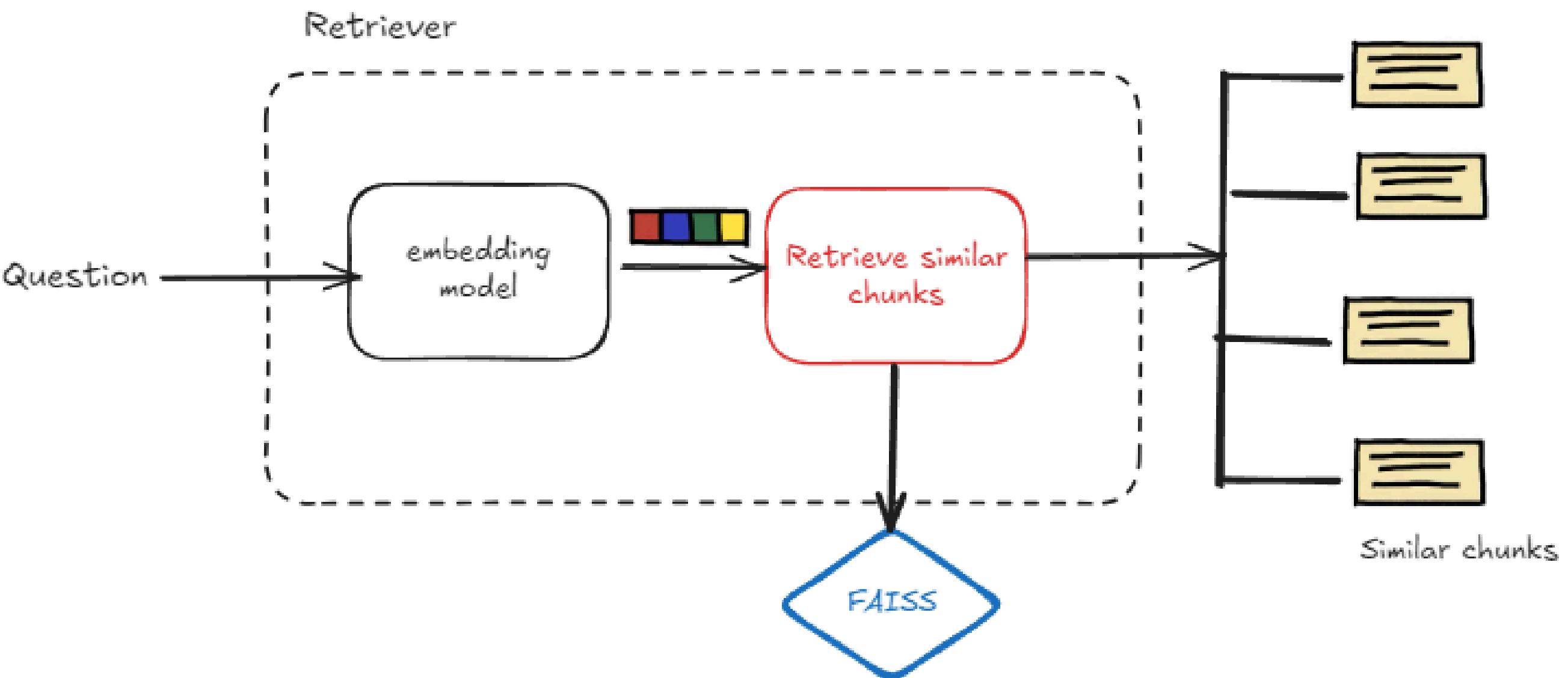
# RAG architecture



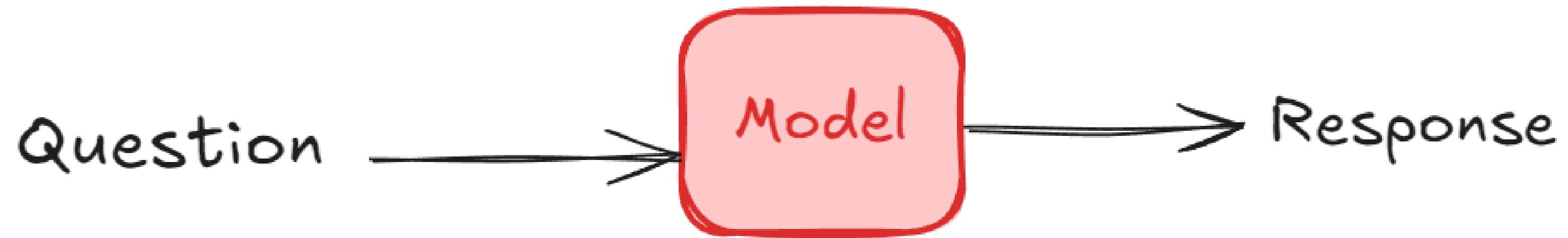


By Mohammed Arbi Nsibi

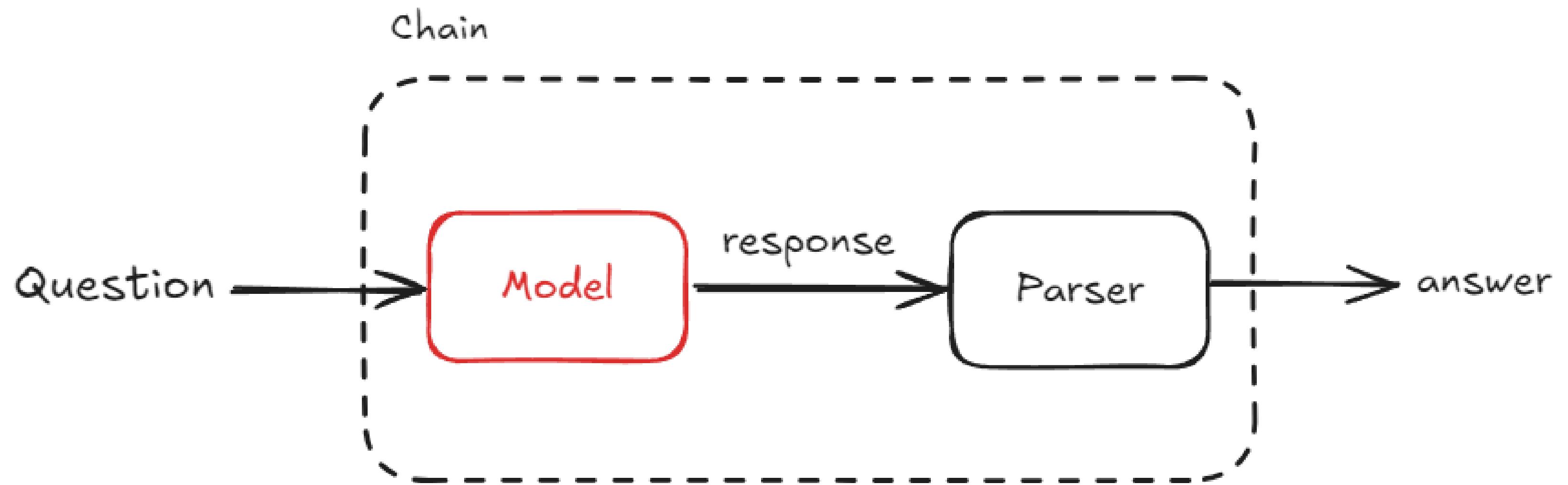




By Mohammed Arbi Nsibi

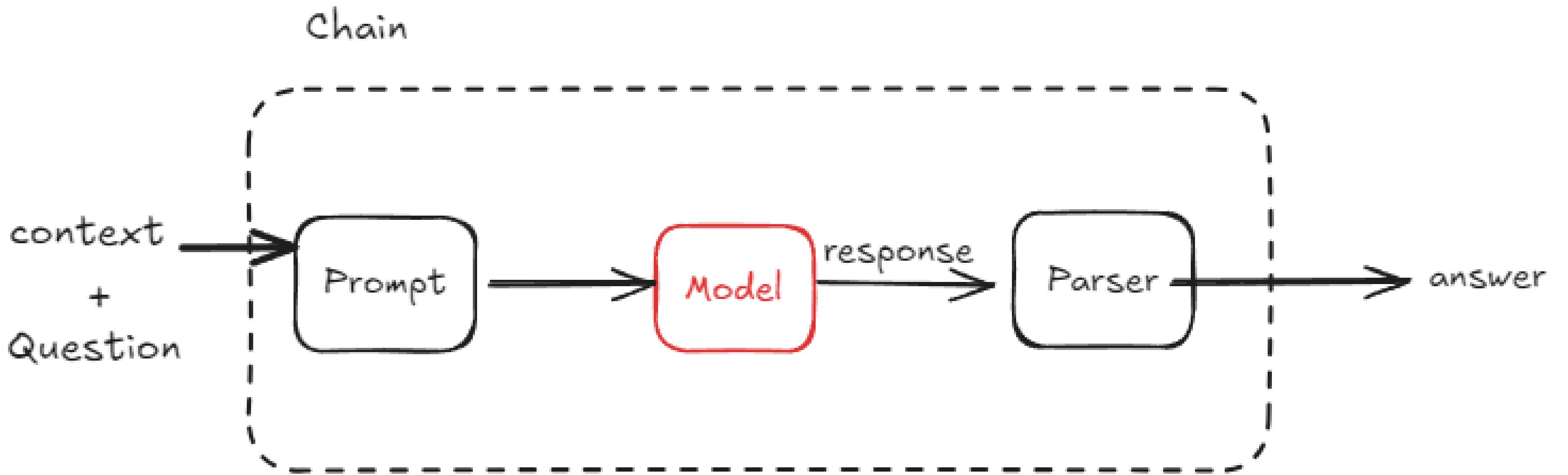


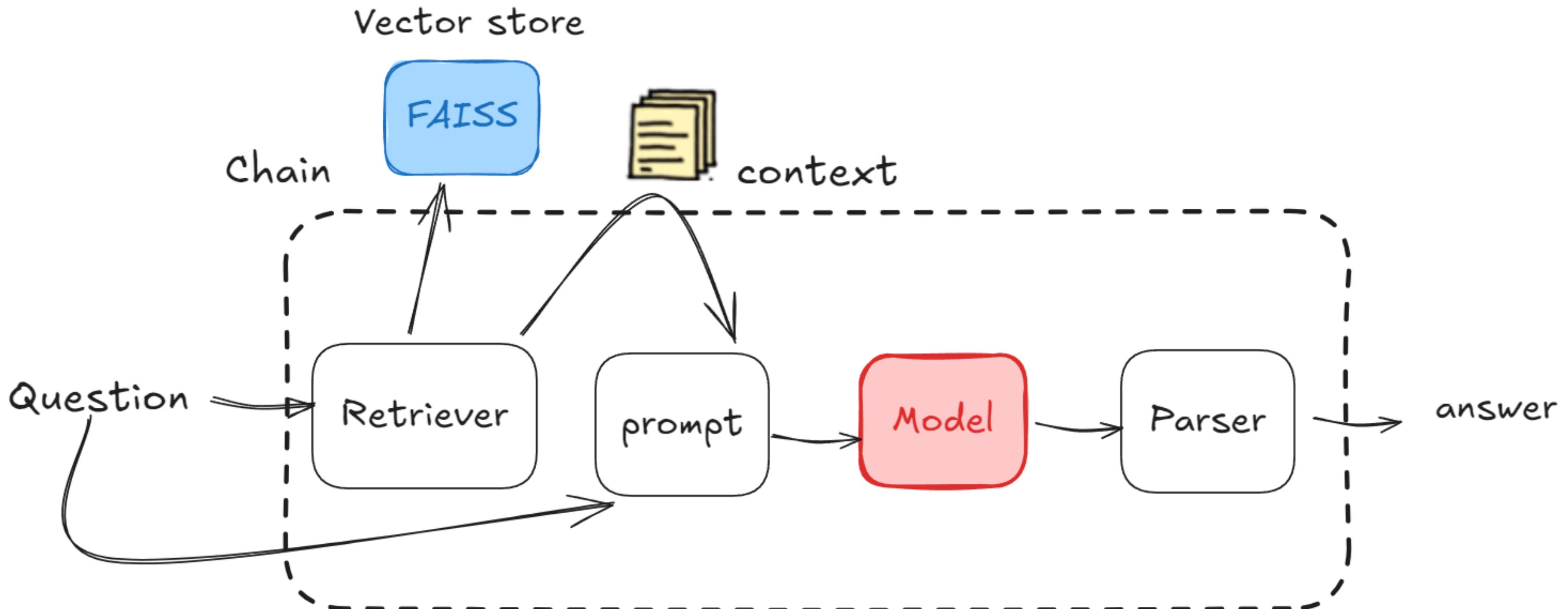
```
AIMessage(content='As of my last update in April 2023, Joe Biden is the President of the United States. He took office in 2021 after Donald Trump left office. Biden is the 46th President of the United States.')
```



'As of my last update in April 2023, Joe Biden is the President of the United States. He took office on January 20, 2021,







# How to get started ?

*By Mohammed Arbi Nsibi*

# LangChain

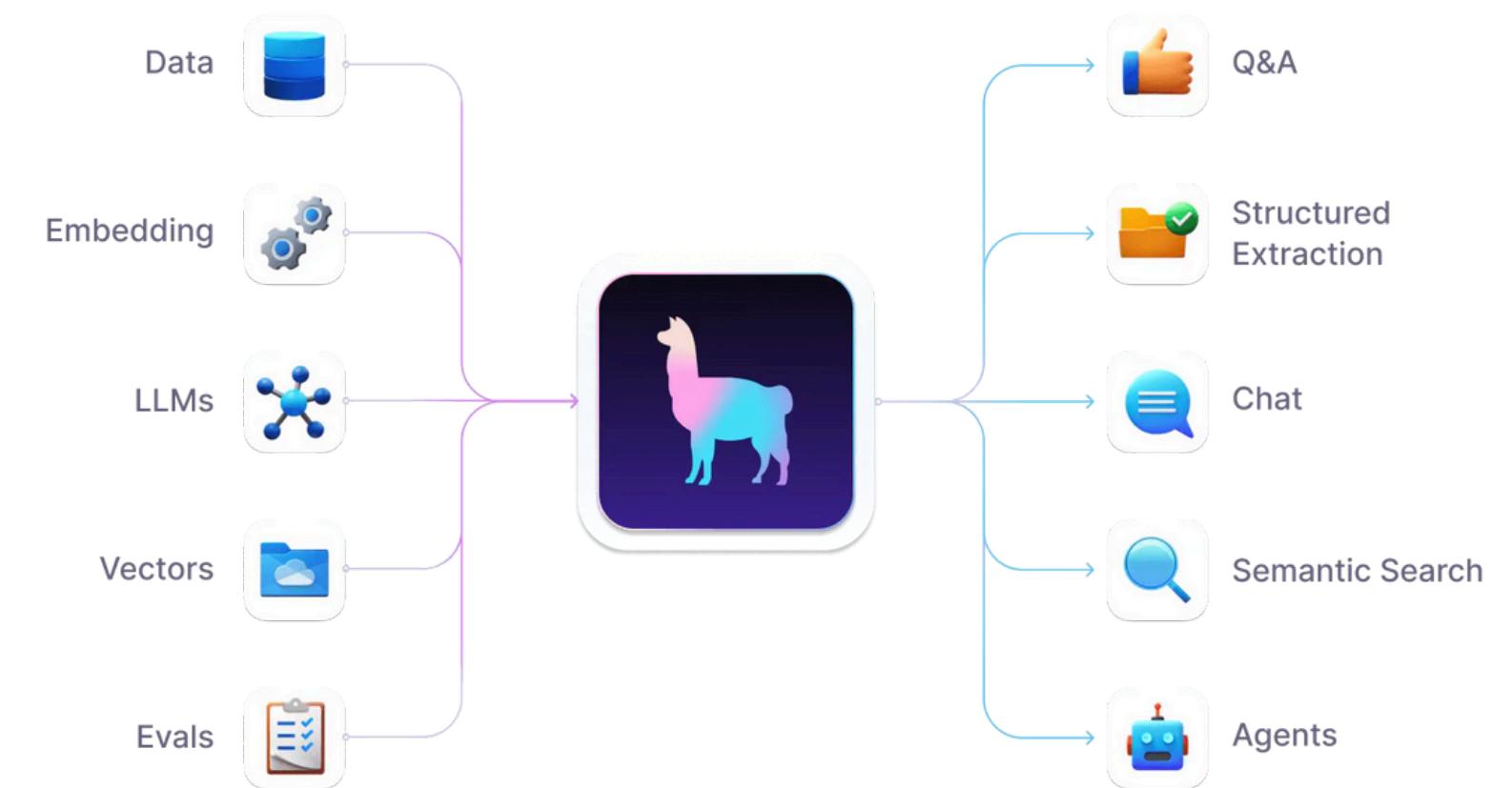
- LangChain is a framework designed to simplify the creation of applications using large language models.
- It is based on LCEL (LangChain Expression Language)(build, compose, or manage sequences of operations)
- Use-cases including chatbots, RAG, document summarization and synthetic data generation.



**LangChain**

# 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.
- Since majority applications are RAG, Llamaindex provides the right tools to build RAG



# LET'S CODE 😺



By Mohammed Arbi Nsibi

always has been

wait so creating chatbots is that easy ?





THEY SAID U CREATED A CHATBOT?

I DID

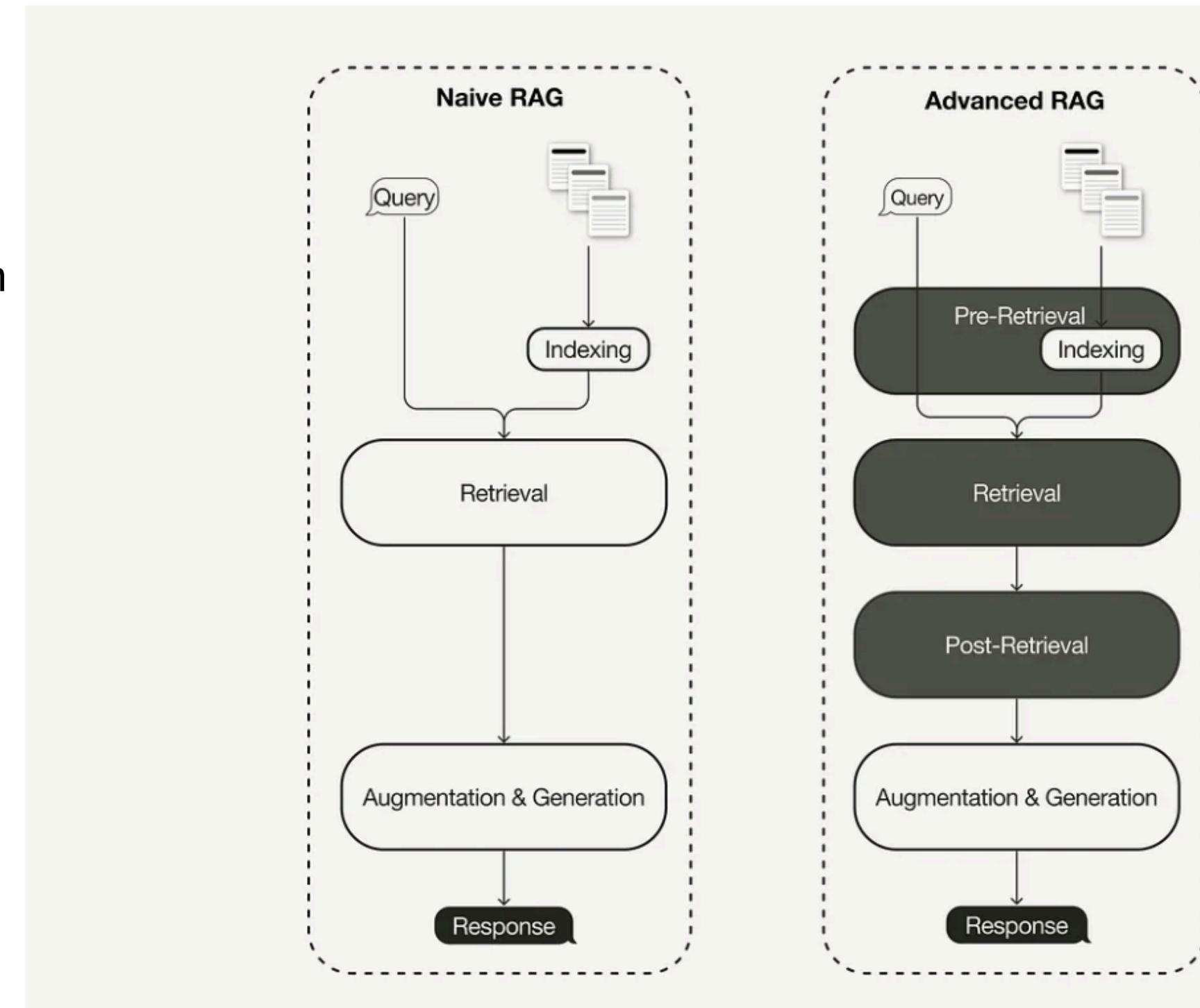


WHAT DID IT COST ?

NOTHING, IT'S FREE

## 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

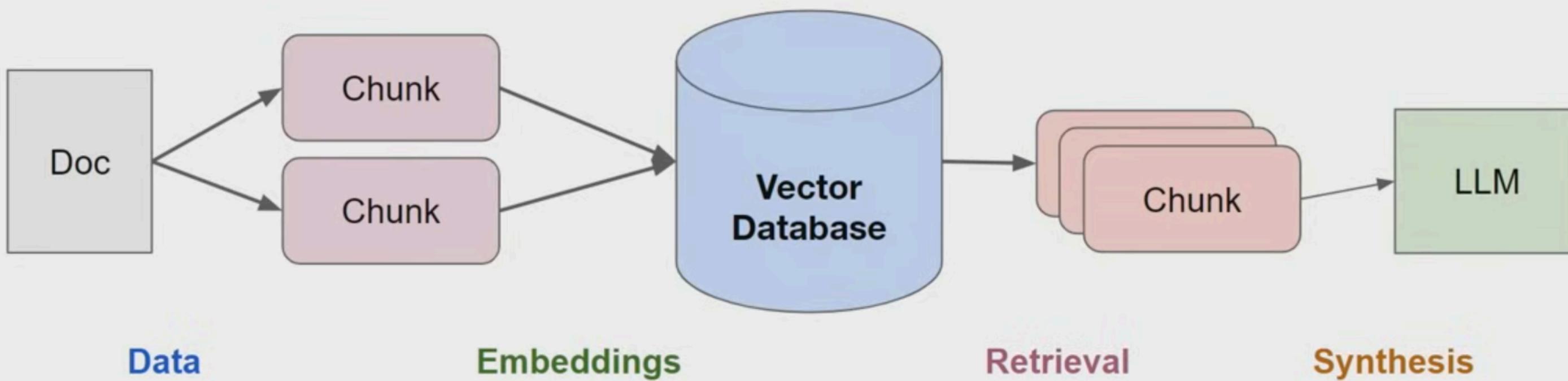


Difference between Naive and Advanced RAG (Image by the author, inspired by [1])

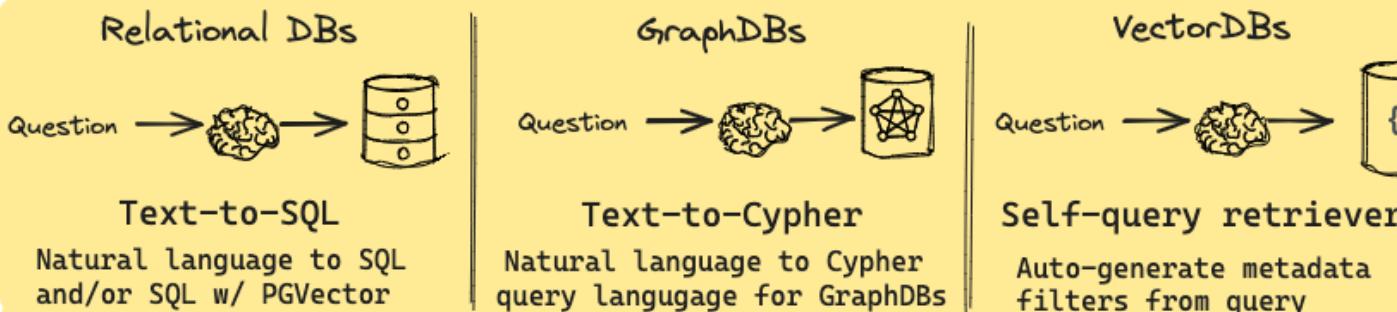
## Naive RAG vs Advanced RAG

### What do we do?

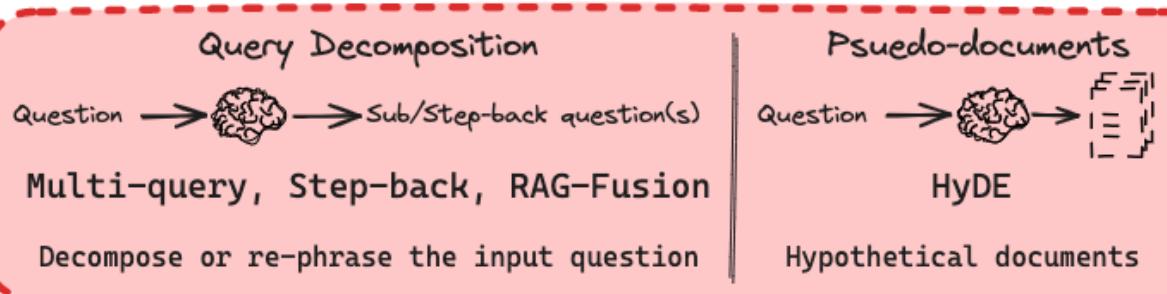
- **Data:** Can we store additional information beyond raw text chunks?
- **Embeddings:** Can we optimize our embedding representations?
- **Retrieval:** Can we do better than top-k embedding lookup?
- **Synthesis:** Can we use LLMs for more than generation? ✓



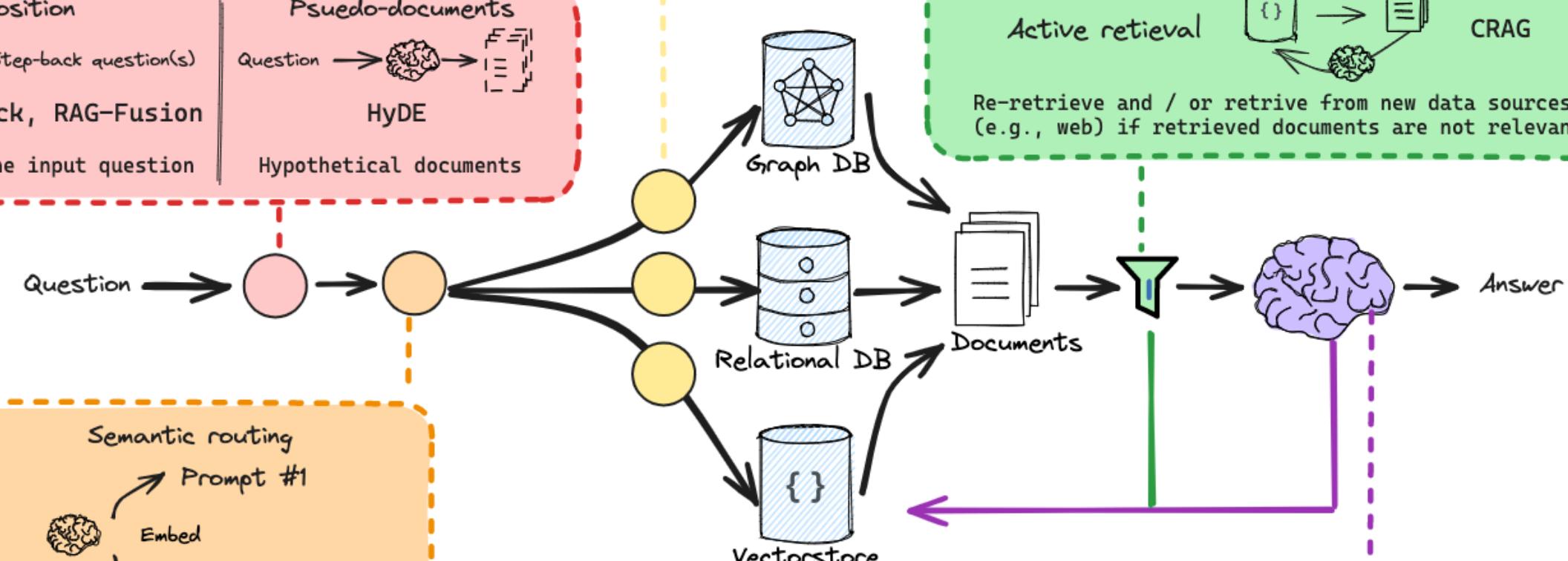
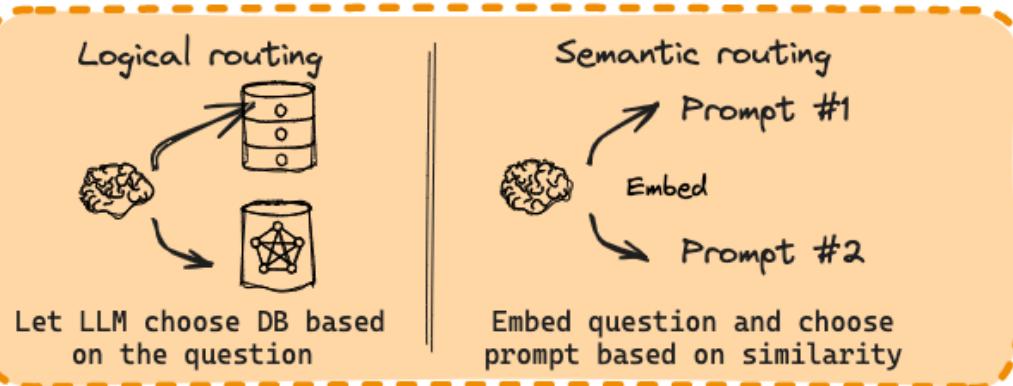
## Query Construction



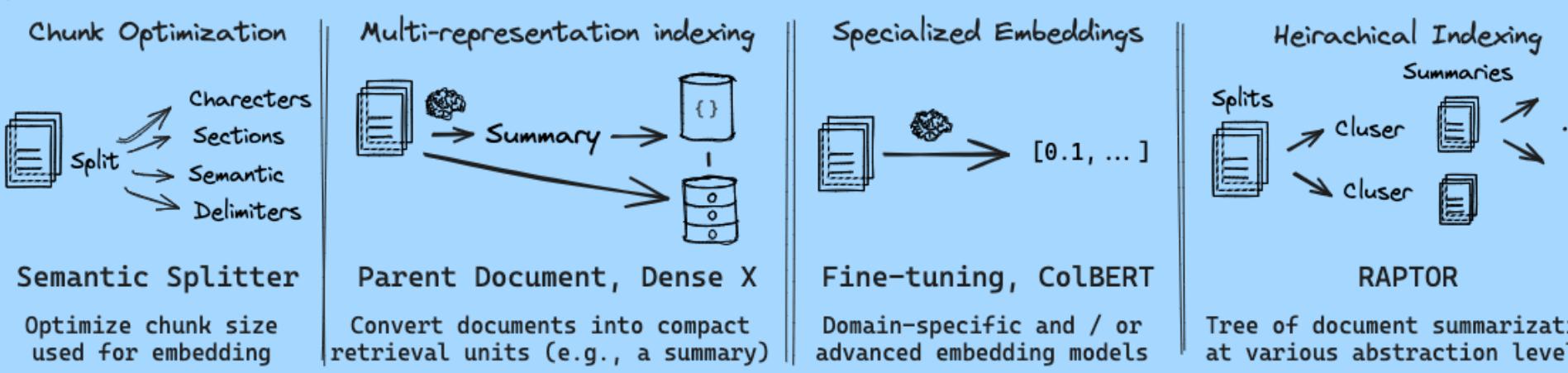
## Query Translation



## Routing



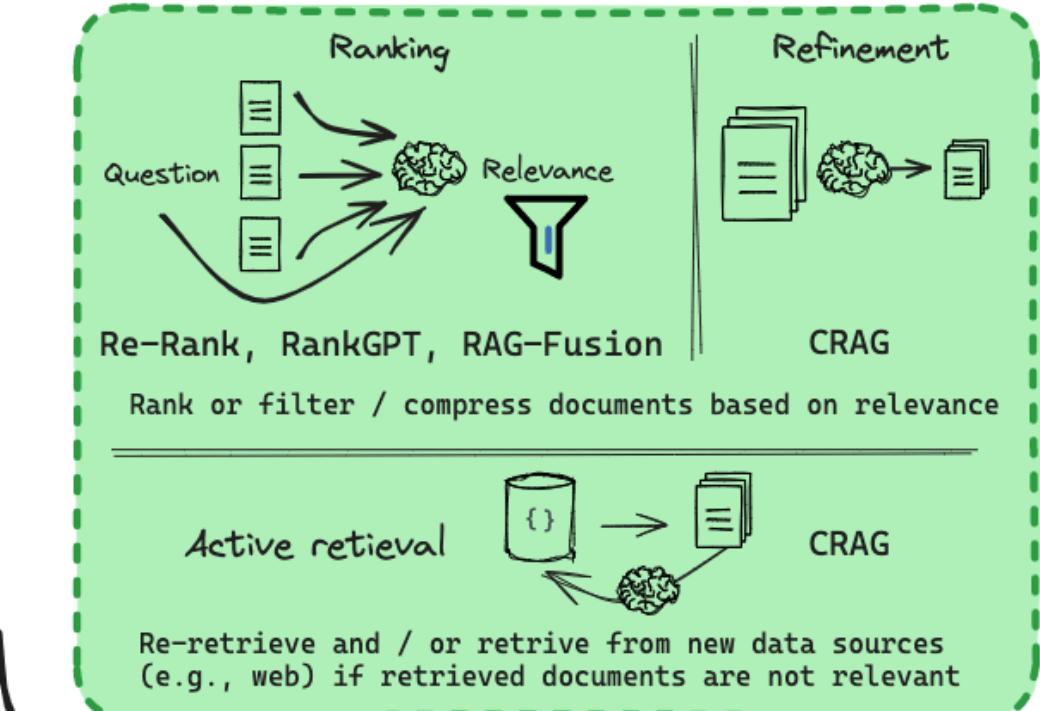
## Indexing



## Generation



## Retrieval





# Resources

- [Your RAG powered by Google Search Technology.](#)
- <https://arxiv.org/abs/2005.11401>
- [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 Implementation](#)
- [Retrieval-augmented generation for large language models: A survey \[arXiv\]](#)

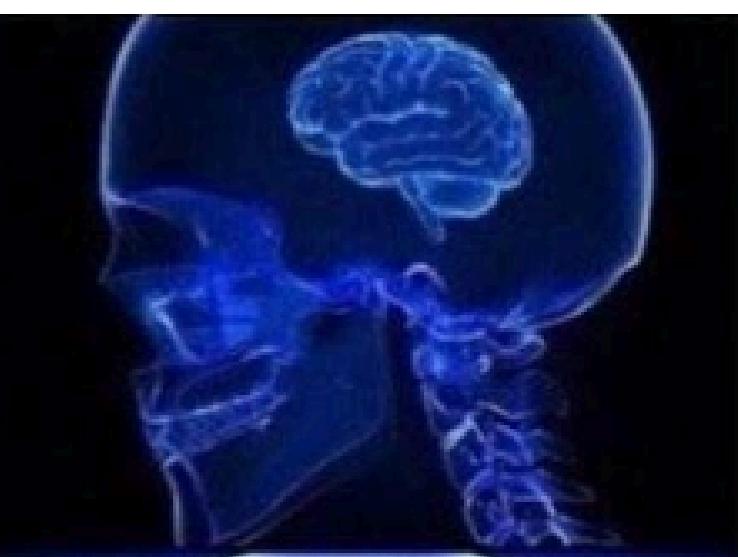
# Q&A

*By Mohammed Arbi Nsibi*

**THANK YOU** for your  
attention!!

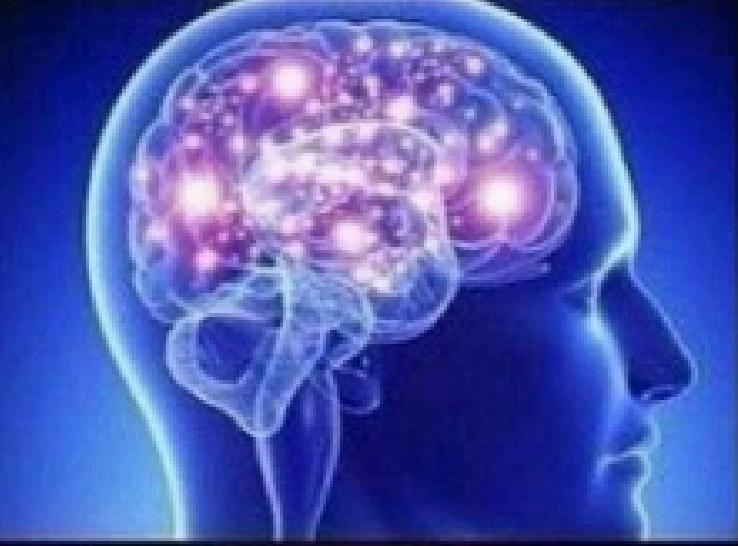
**USING AI  
FOR CALCULATIONS**

---



**USING AI TO  
WRITE ESSAYS**

---



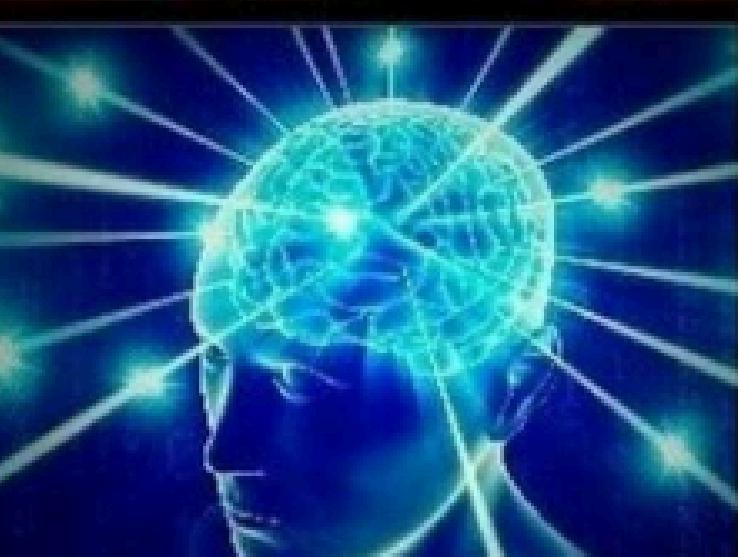
**USING AI TO  
GENERATE CODE**

---

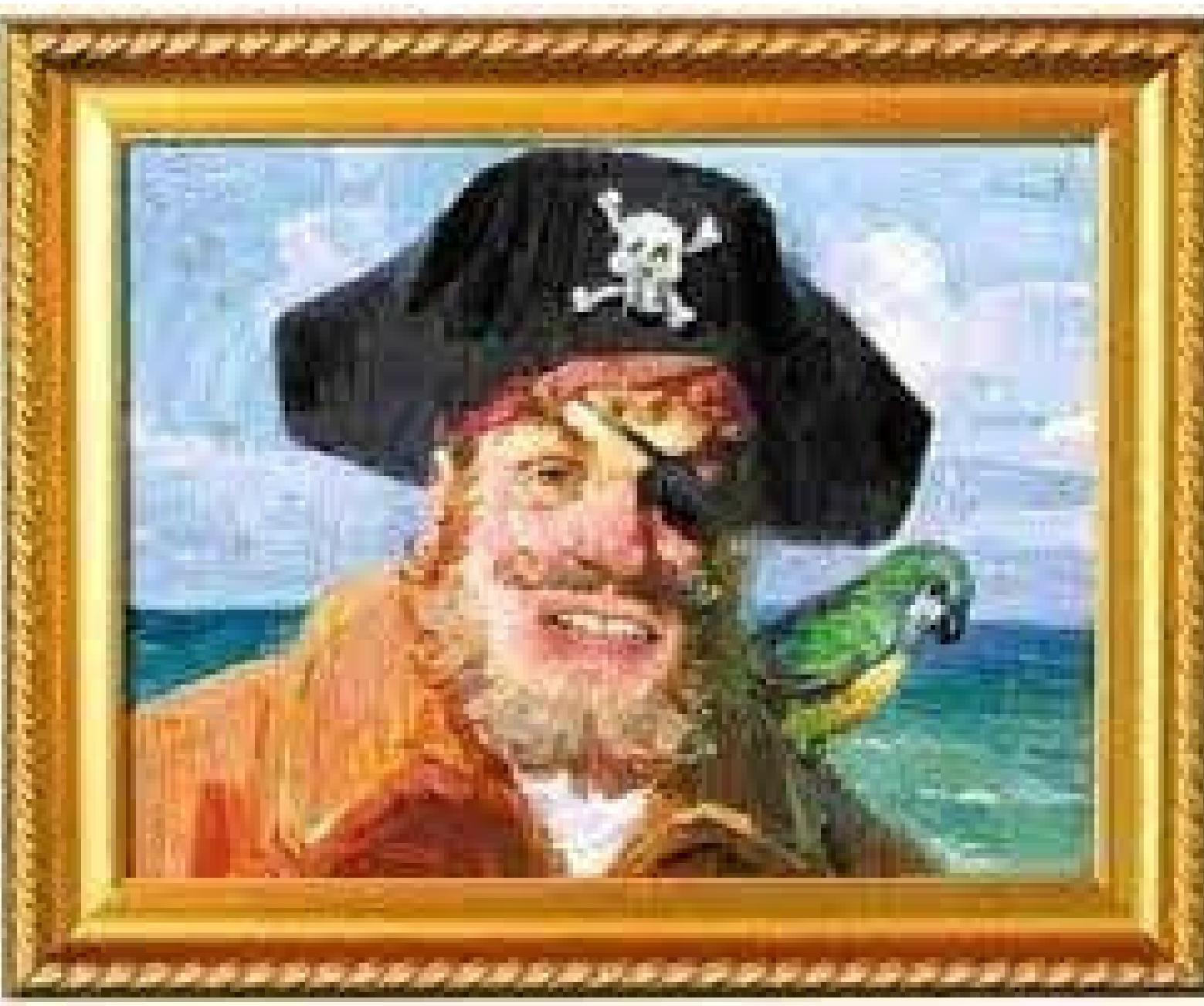


**USING AI  
TO GENERATE  
MEMES ABOUT AI**

---



# QUIZ TIME



By Mohammed Arbi Nsibi

