Introduction to RAG

19/02/2025

By Mohammed Arbi Nsibi

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

By Mohammed Arbi Nsibi

Motivation
(Why do we need RAG?)

By Mohammed Arbi Nsibi

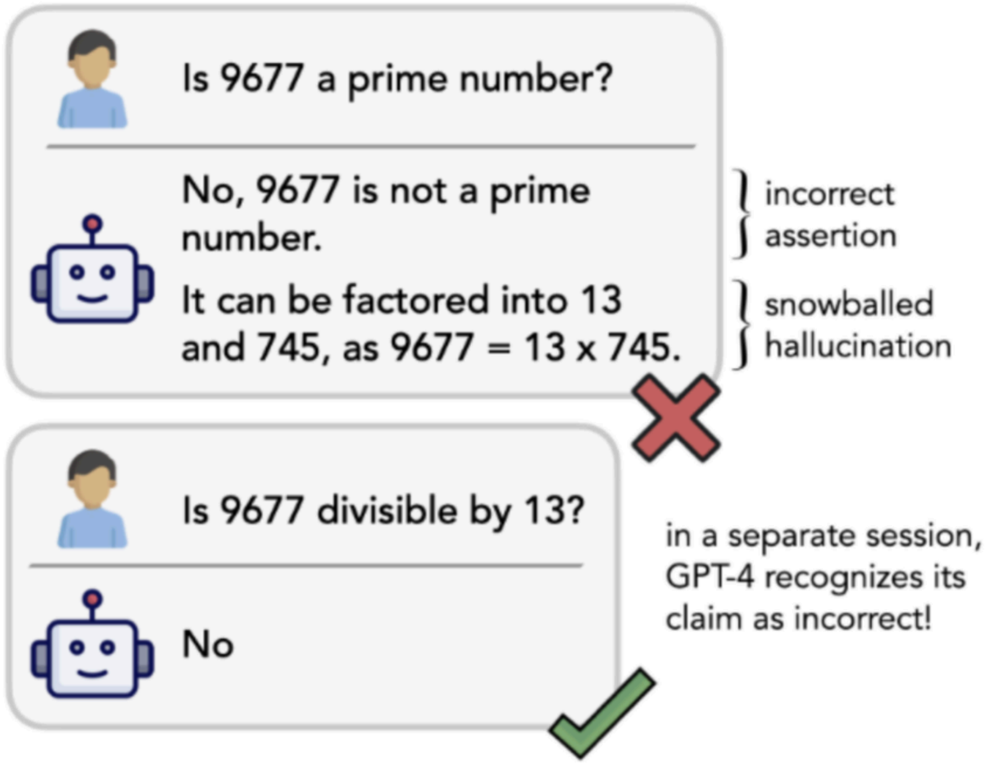
1- The need for an external Knowledge !

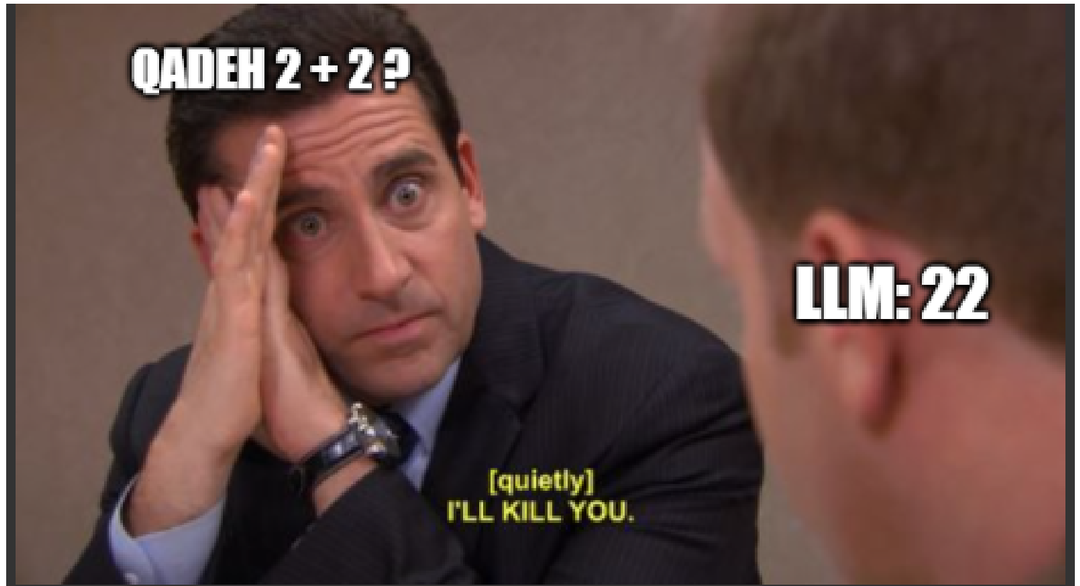
By Mohammed Arbi Nsibi

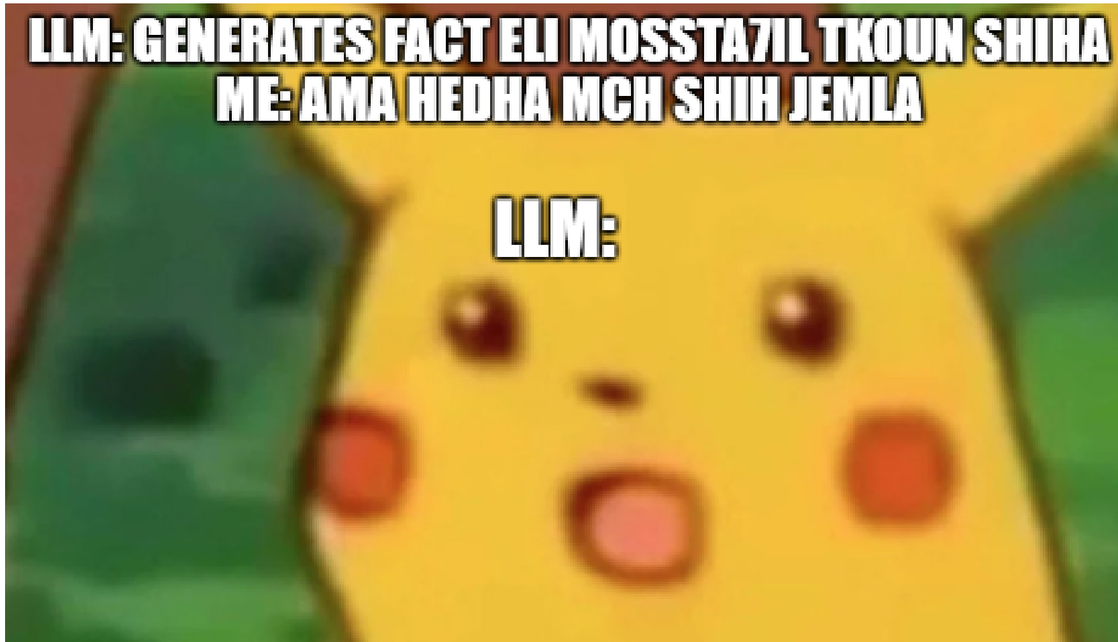
2- Hallucinations

By Mohammed Arbi Nsibi

2- Hallucinations







Hallucinations

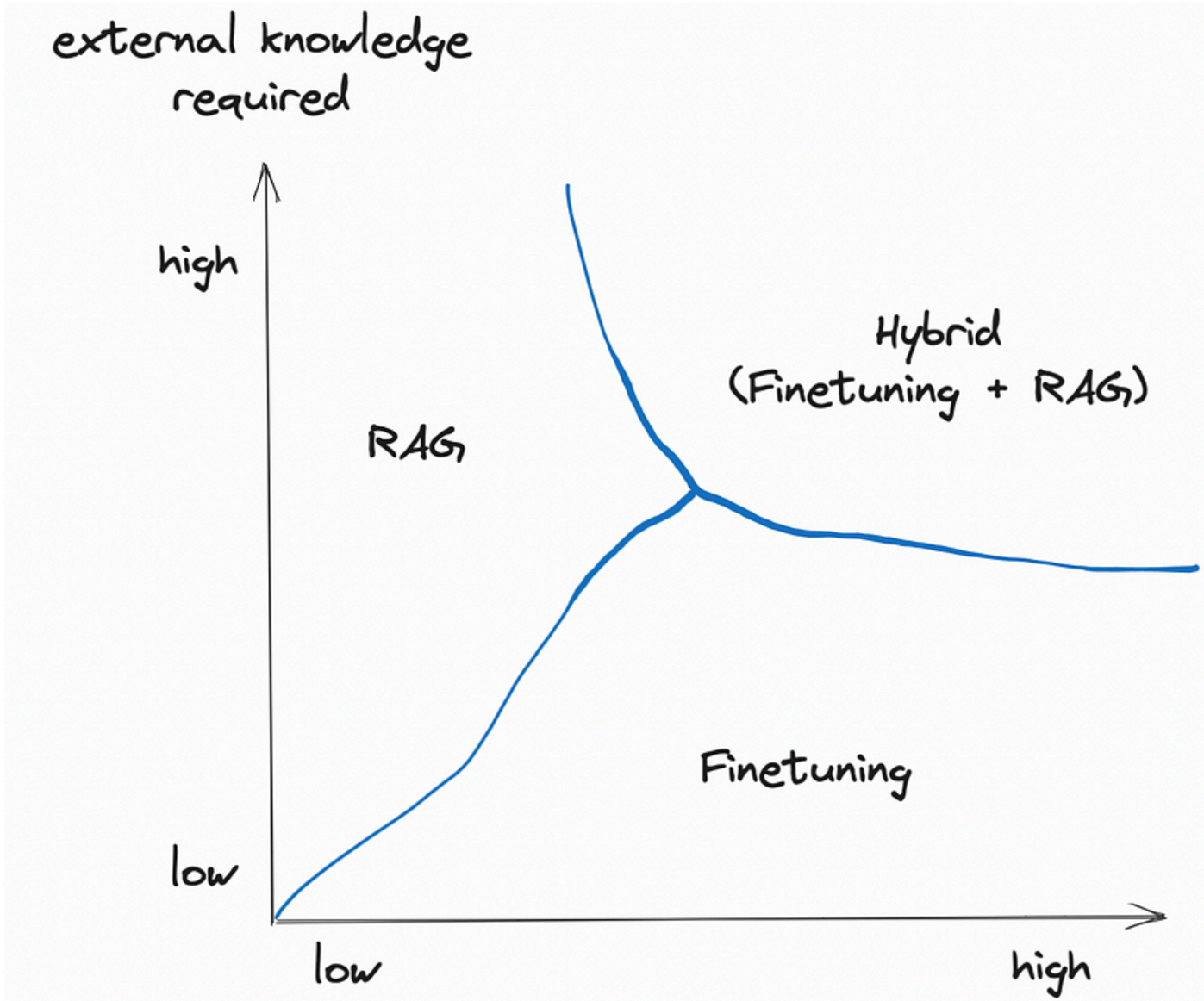
- The model is not trained on enough data.
- The model is trained on noisy or dirty data.
- The model is not given enough context .
- The model is not given enough constraints (rules, guidelines, or limitations)

By Mohammed Arbi Nsibi

Solutions ?

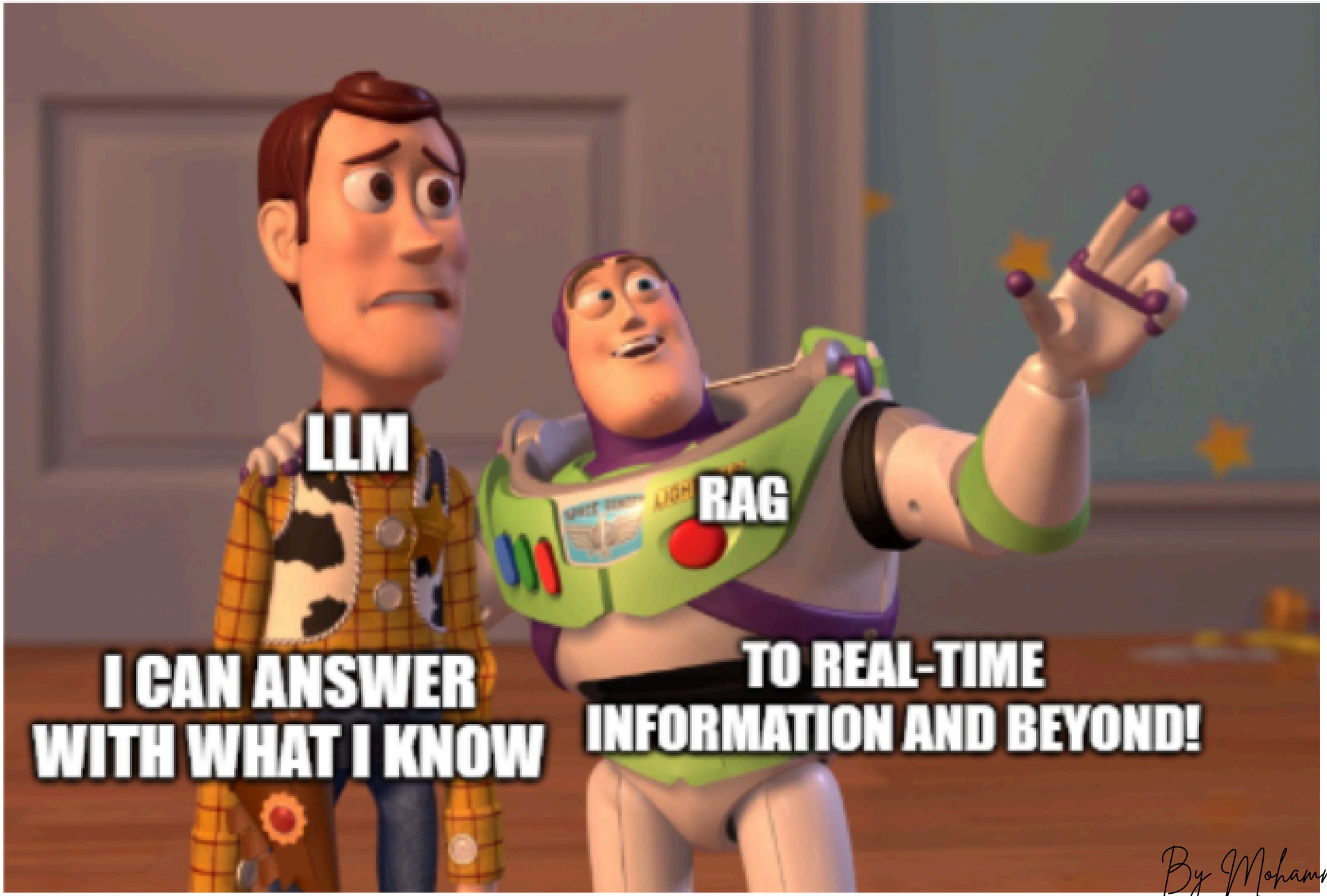
By Mohammed Arbi Nsibi

RAG / Fine-tuning

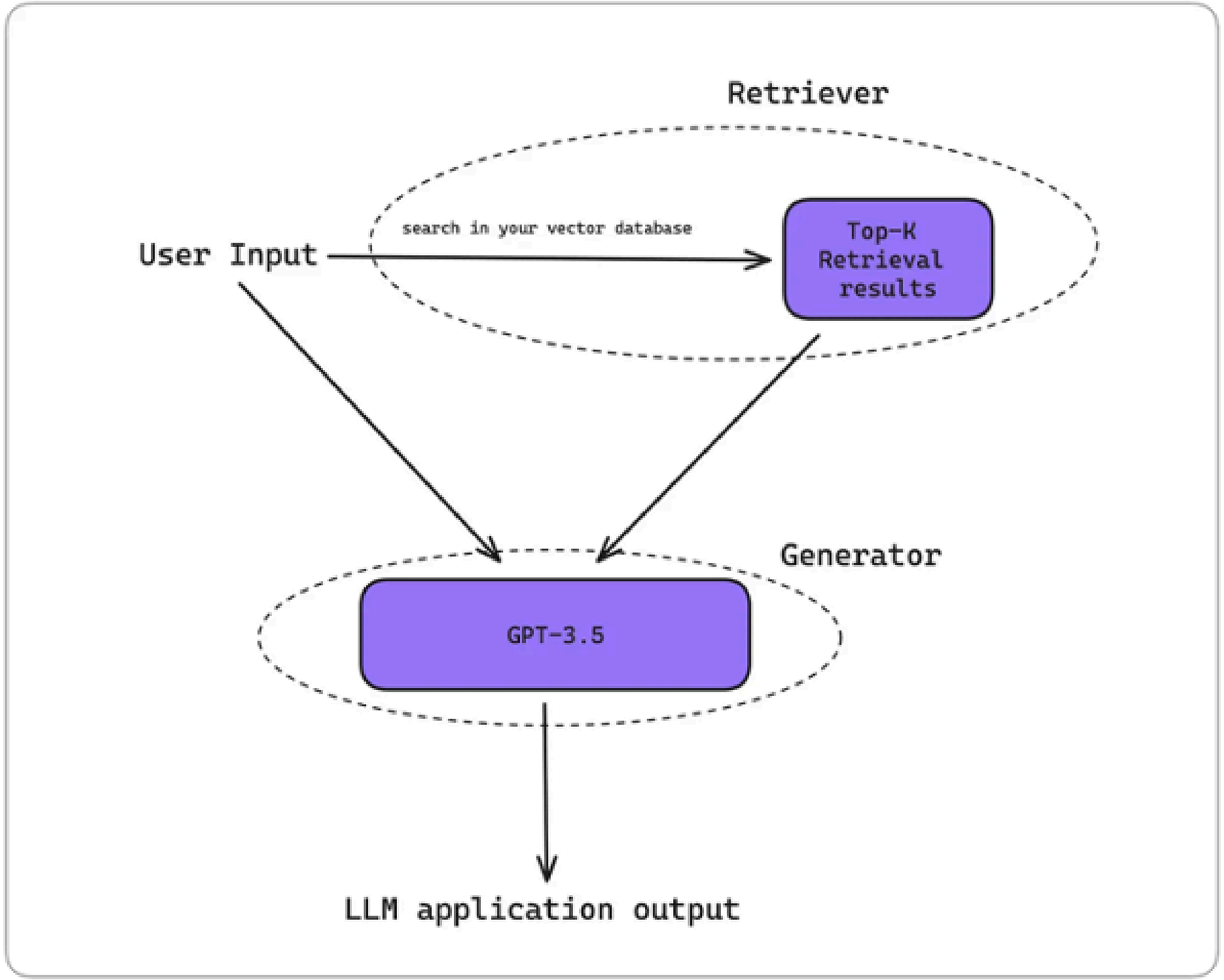


But wait..
WTF is RAG ?

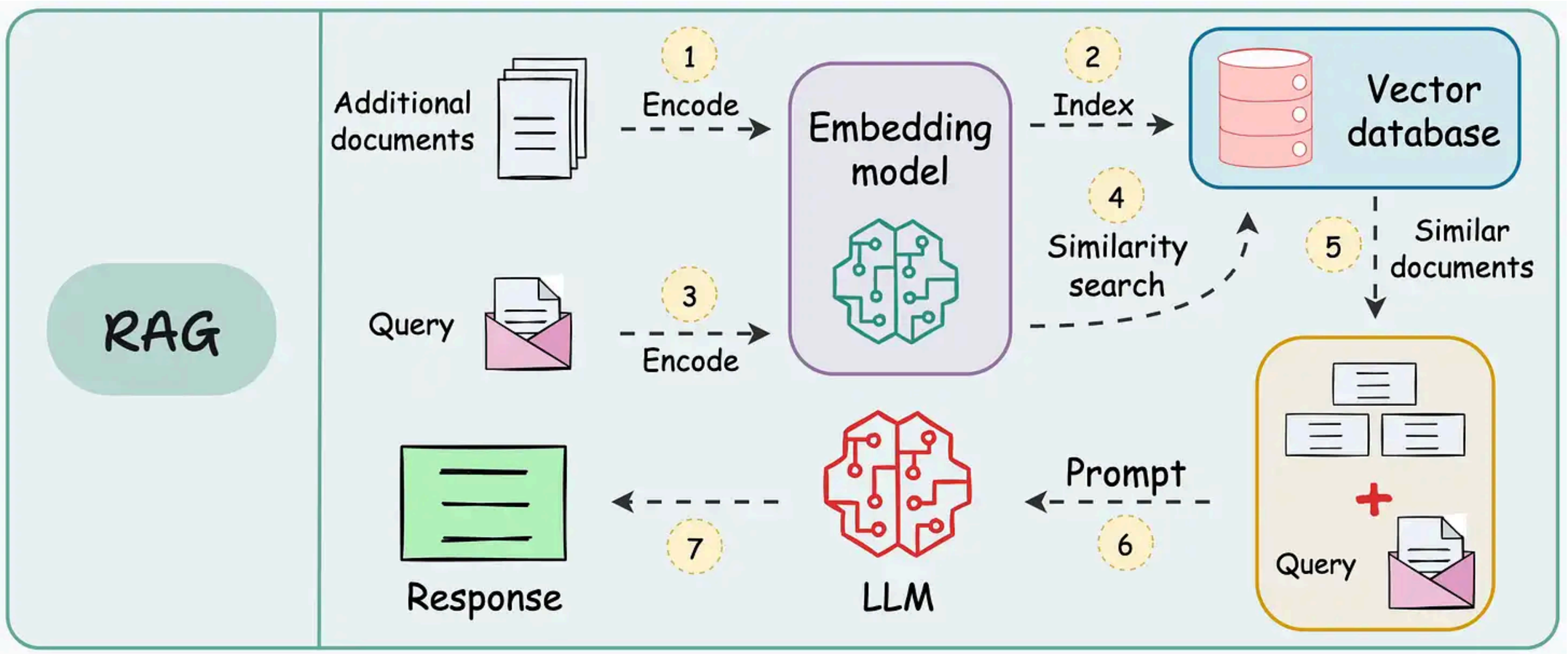
RAG (Retrieval-augmented generation)



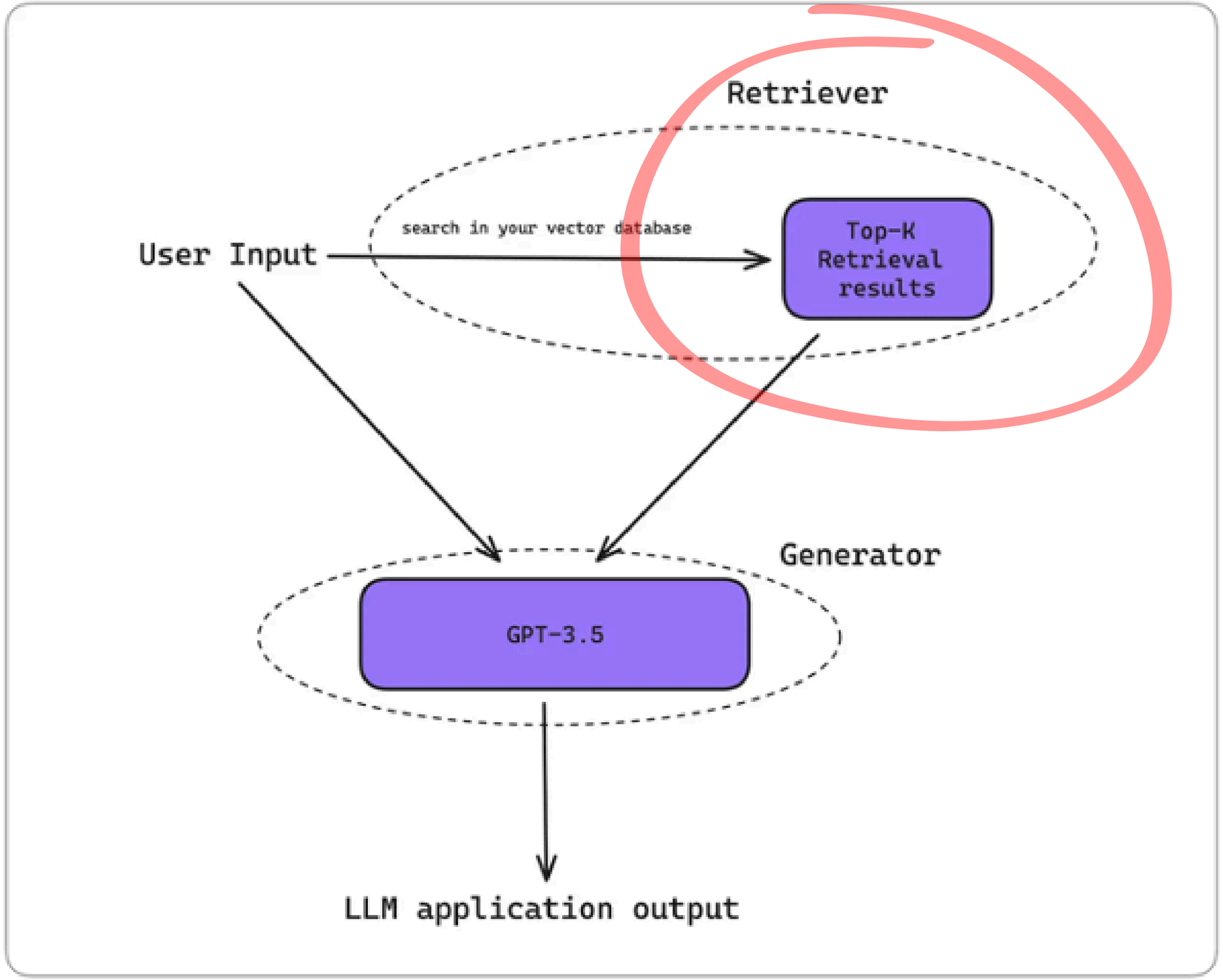
By Mohammed Arbi Nsibi


Retriever
search in your vector database
User Input
Top-K
Retrieval
results
Generator
GPT-3.5
LLM application output

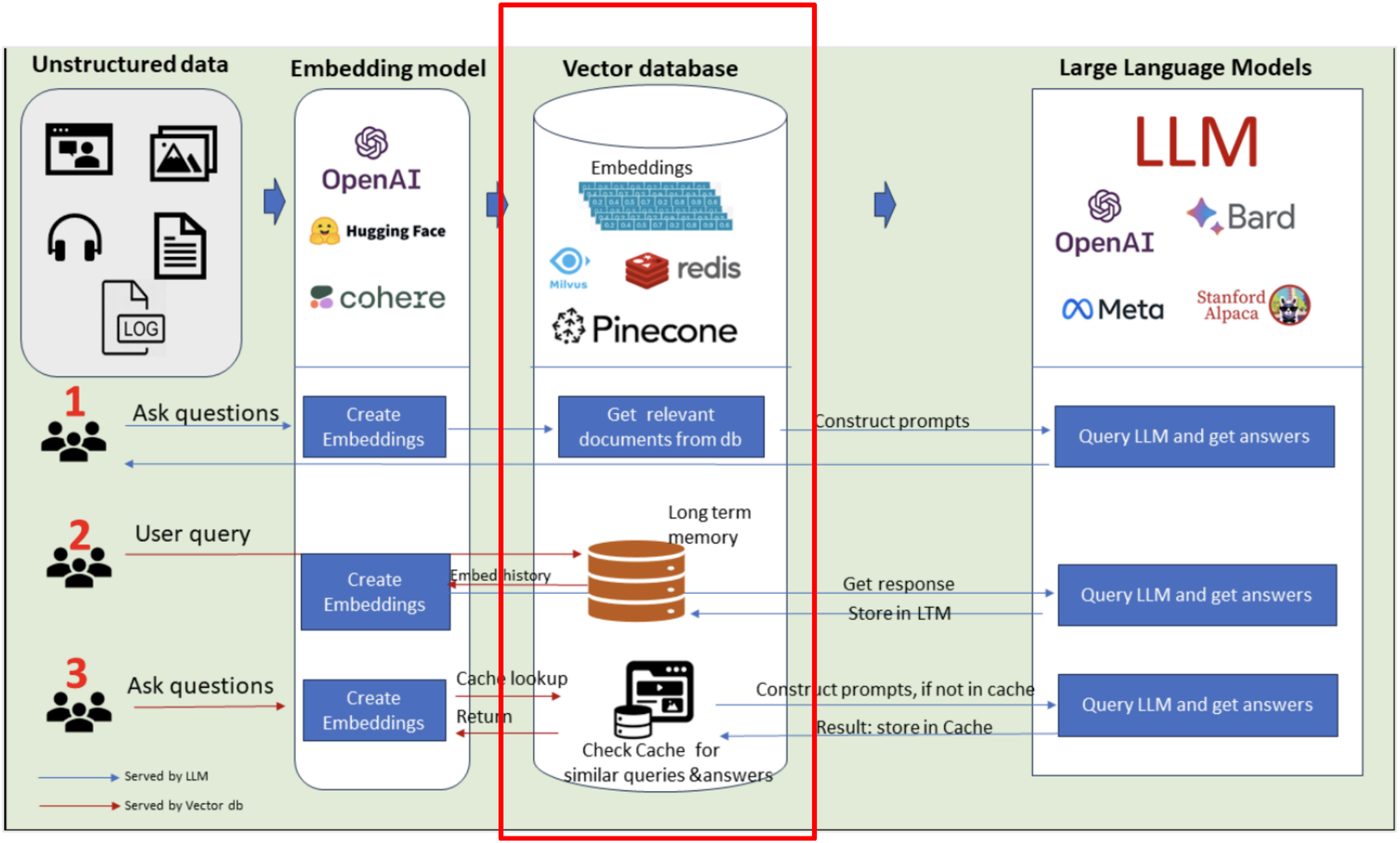
RAG (Retrieval-augmented generation)



By Mohammed Arbi Nsibi


Retriever
search in your vector database
User Input
Top-K
Retrieval
results
Generator
GPT-3.5
LLM application output

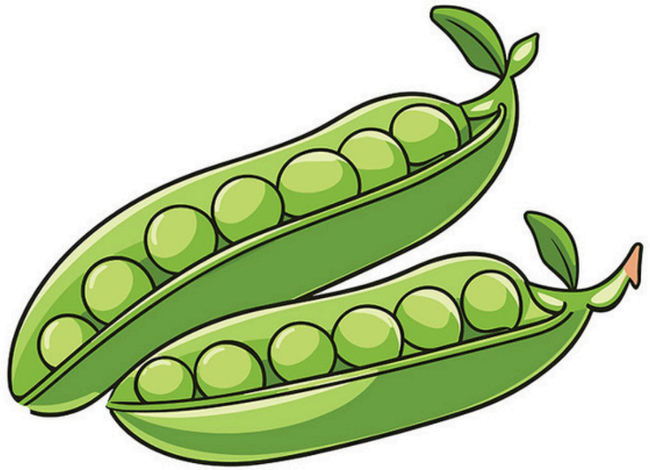
Vector Database



tea

coffee

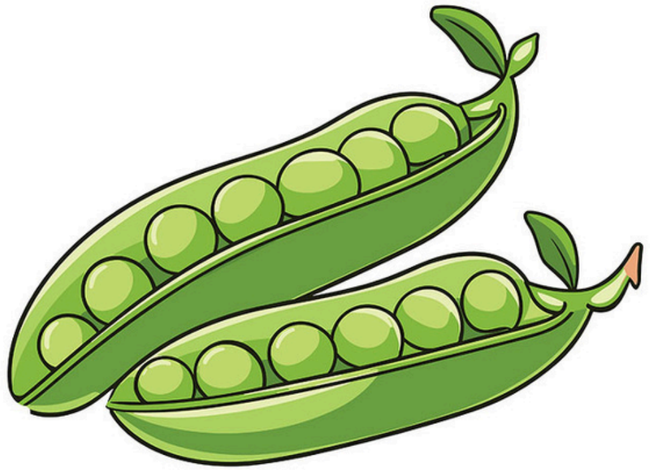
tea



pea

tea

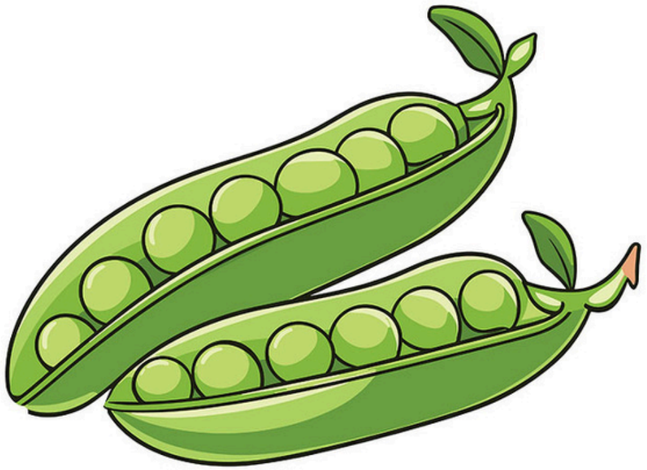
≠



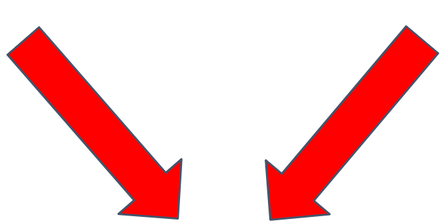
pea

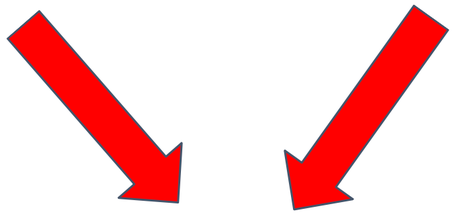
tea

coffee



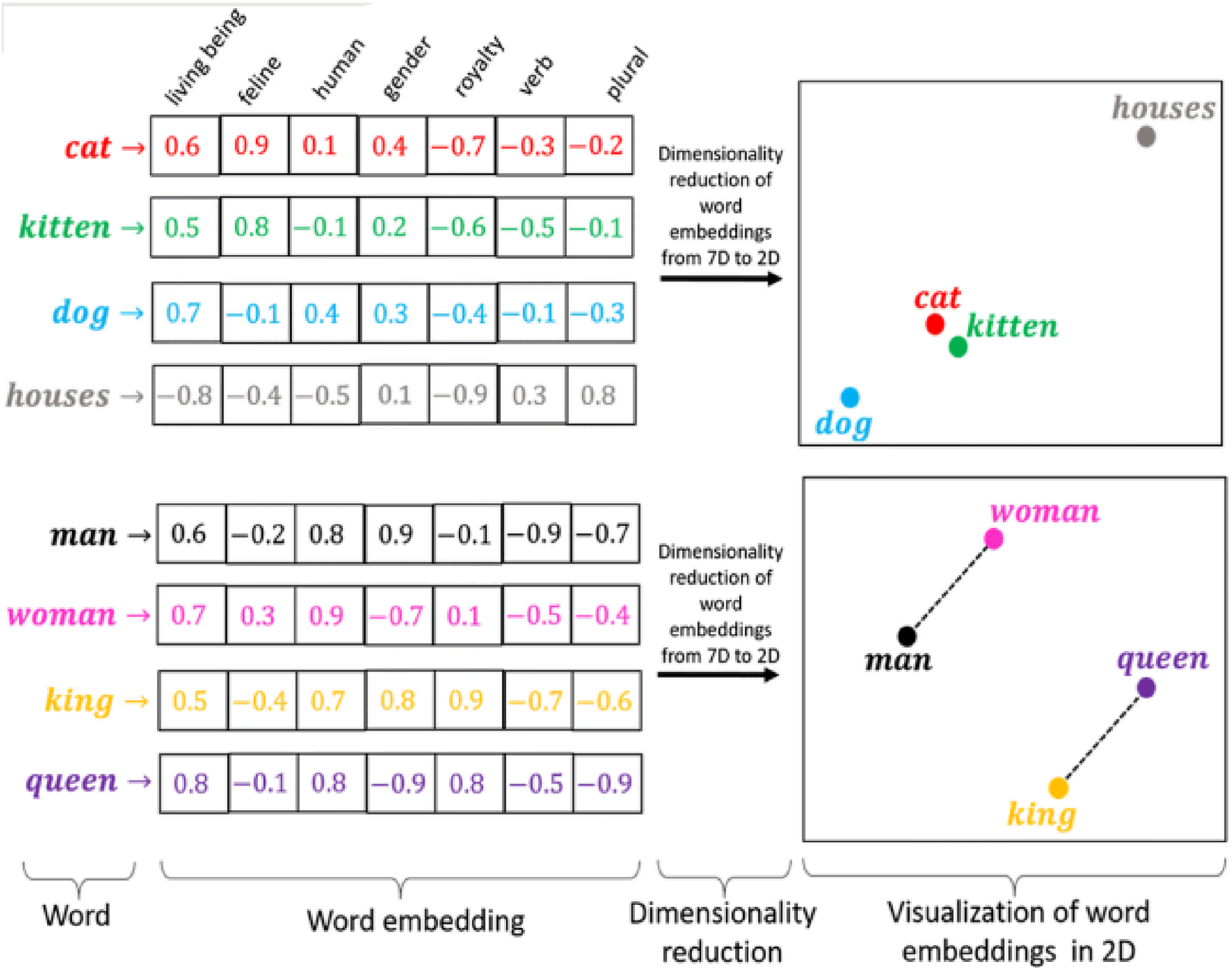
pea



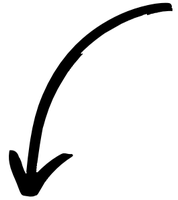


distance = 0.3

distance = 0.7


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

How can we get these word embedding vectors?



Pre-Trained Word Embeddings

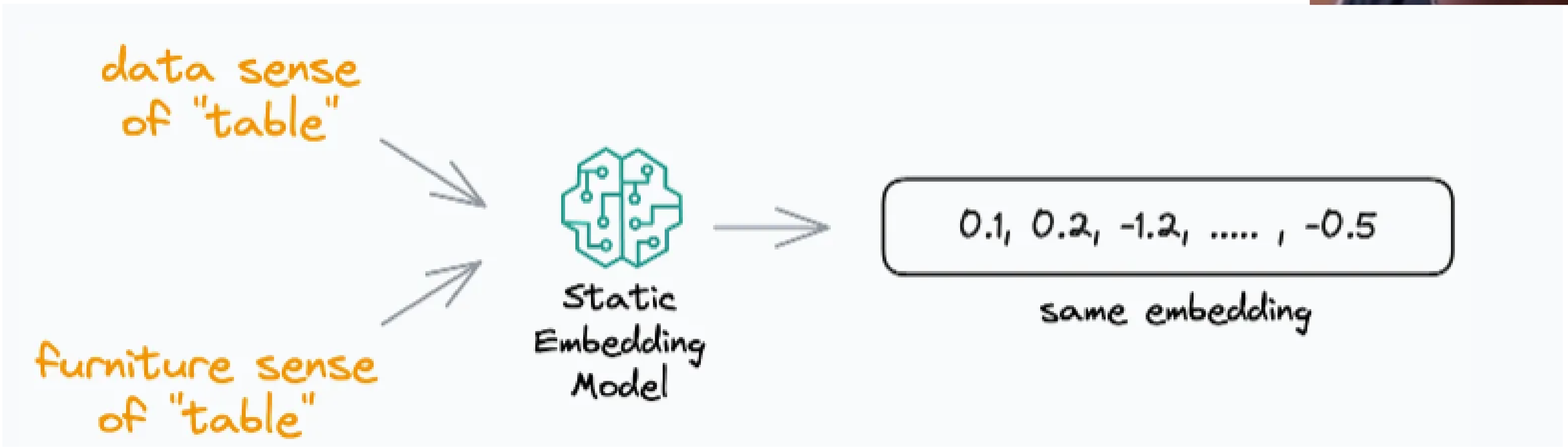
- **Word2vec by Google 2013**

- **GloVe by Stanford**
(Global Vectors for Word Representation)

- **fasttext by Facebook**

Limitations

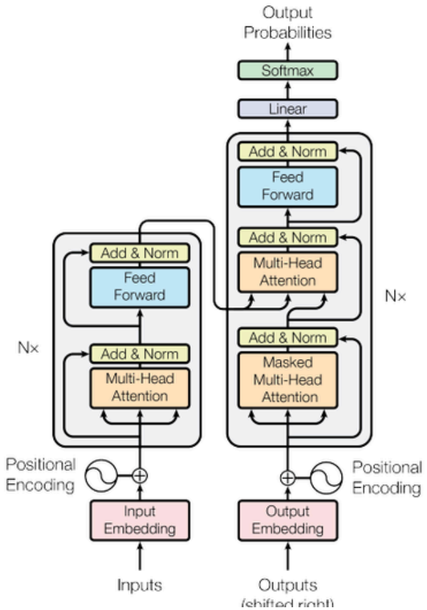
- Convert this data into a table in Excel.
- Put this bottle on the table.



Solution ?

Transformer

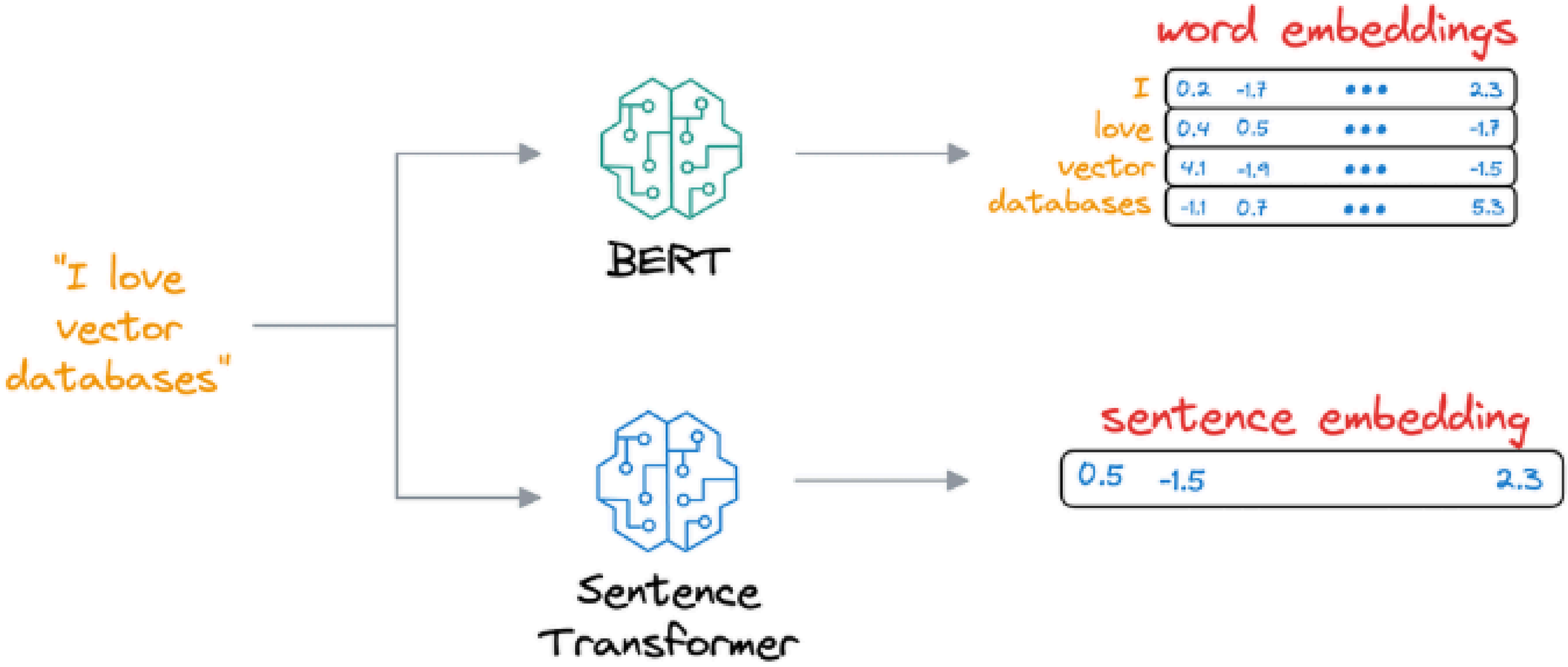
- BERT (Bidirectional Encoder Representations from Transformers)



- DistilBERT: BERT which is around 40% smaller:

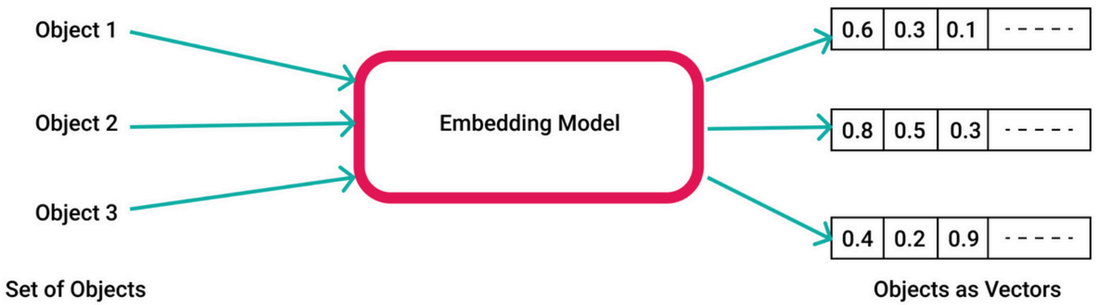
- ALBERT: A Lite BERT (ALBERT).

Solution ?

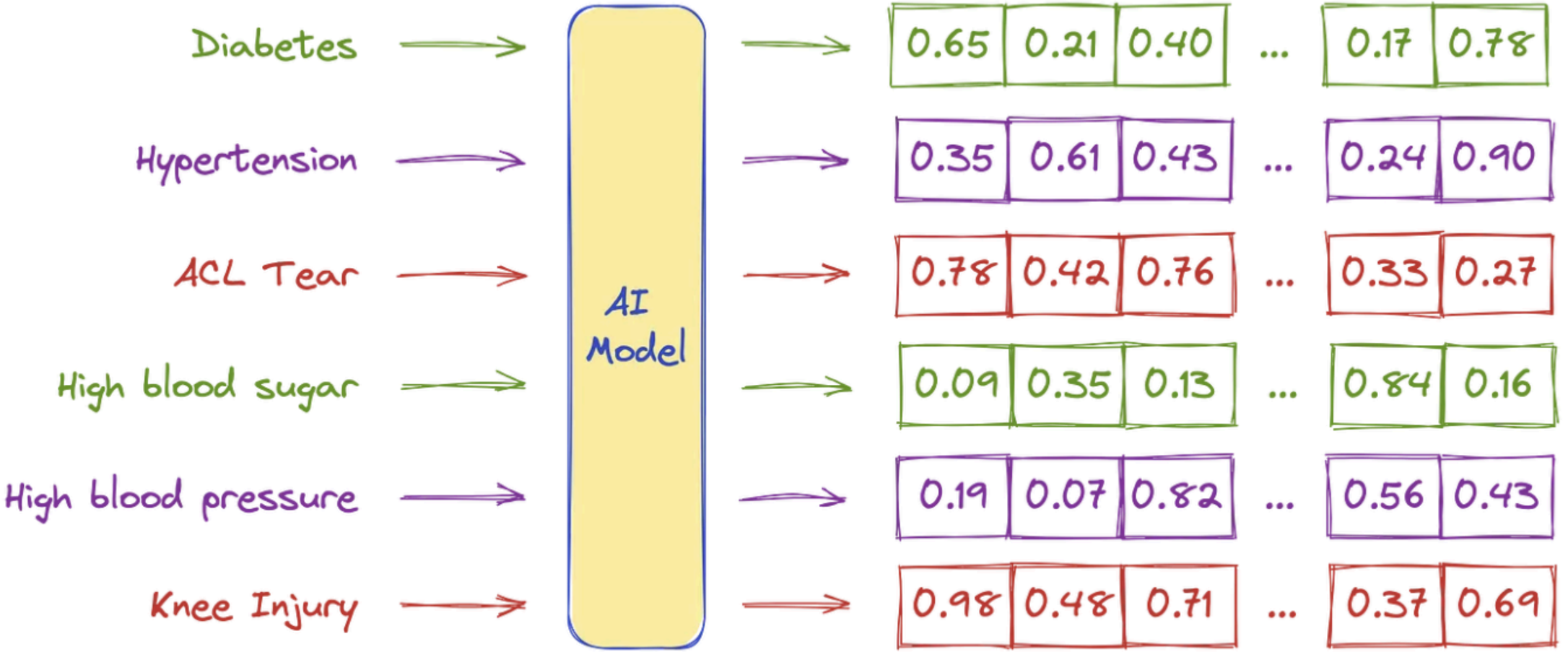

word embeddings
I 0.2 -1.7 ••• 2.3
love 0.4 0.5 ••• -1.7
vector 4.1 -1.9 ••• -1.5
databases -1.1 0.7 ••• 5.3
"I love vector databases"
BERT
Sentence Transformer
sentence embedding
0.5 -1.5 2.3

Embedding Model

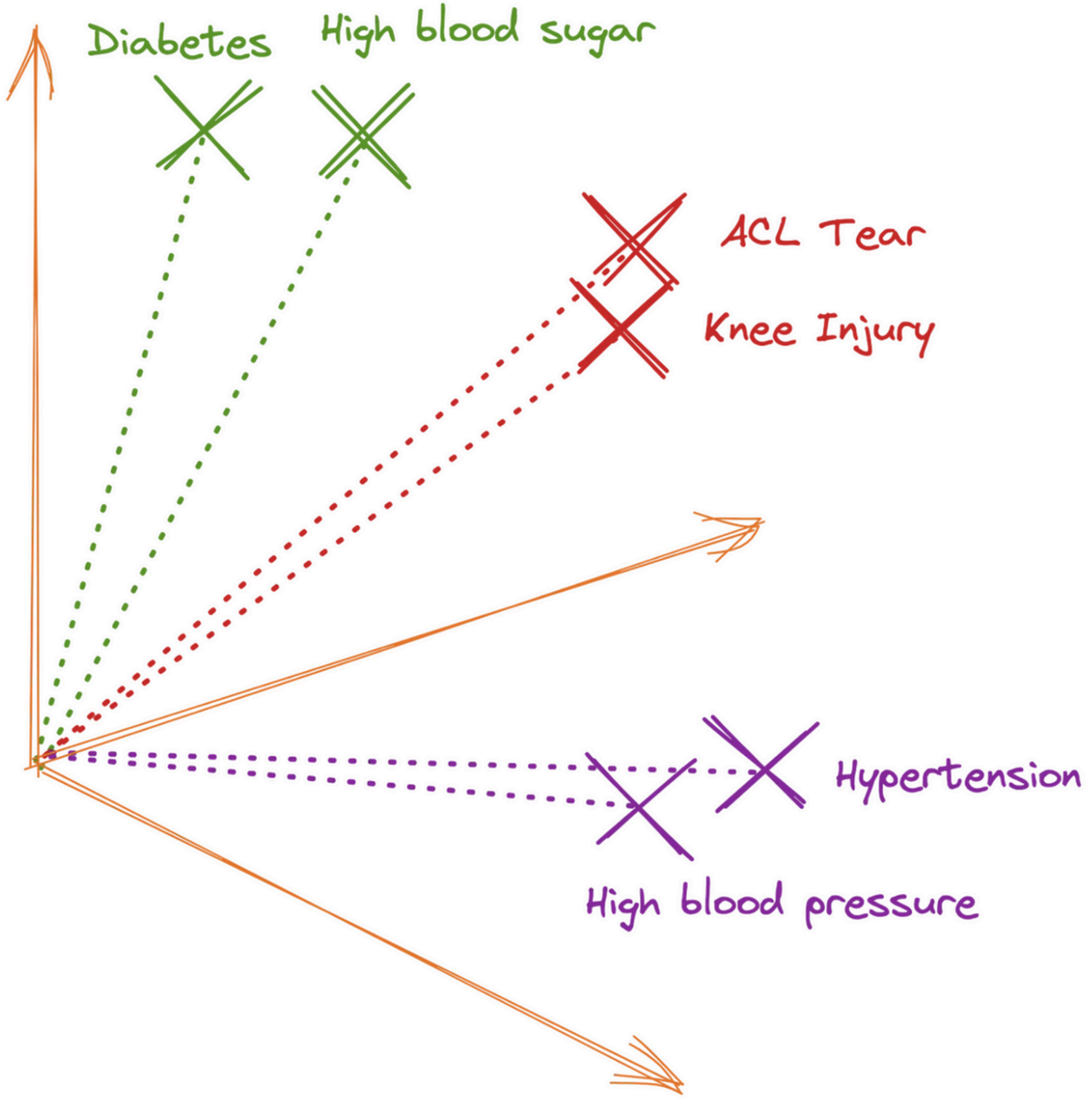
- These vectors live in a high-dimensional space where the proximity between vectors reflects the relatedness of the original items.
- **Embedding model** trained along LLM and learn to **produce representation(vectors) based on context** in word appear.



Embedding Space



Embedding Space

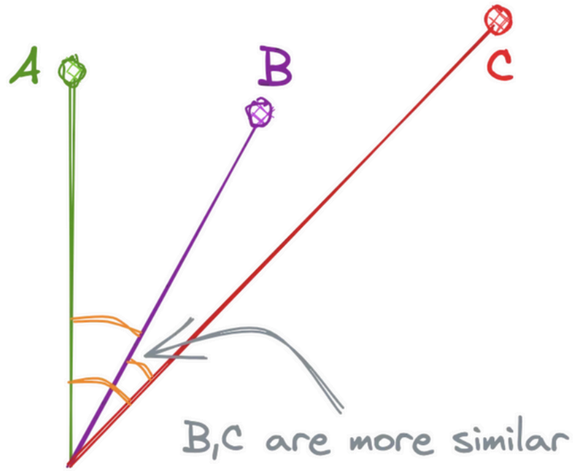


Querying a vector database

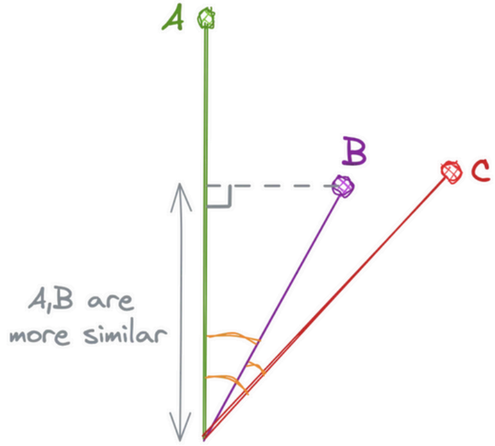
Similarity Calculation = > objective is to return the nearest neighbors

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

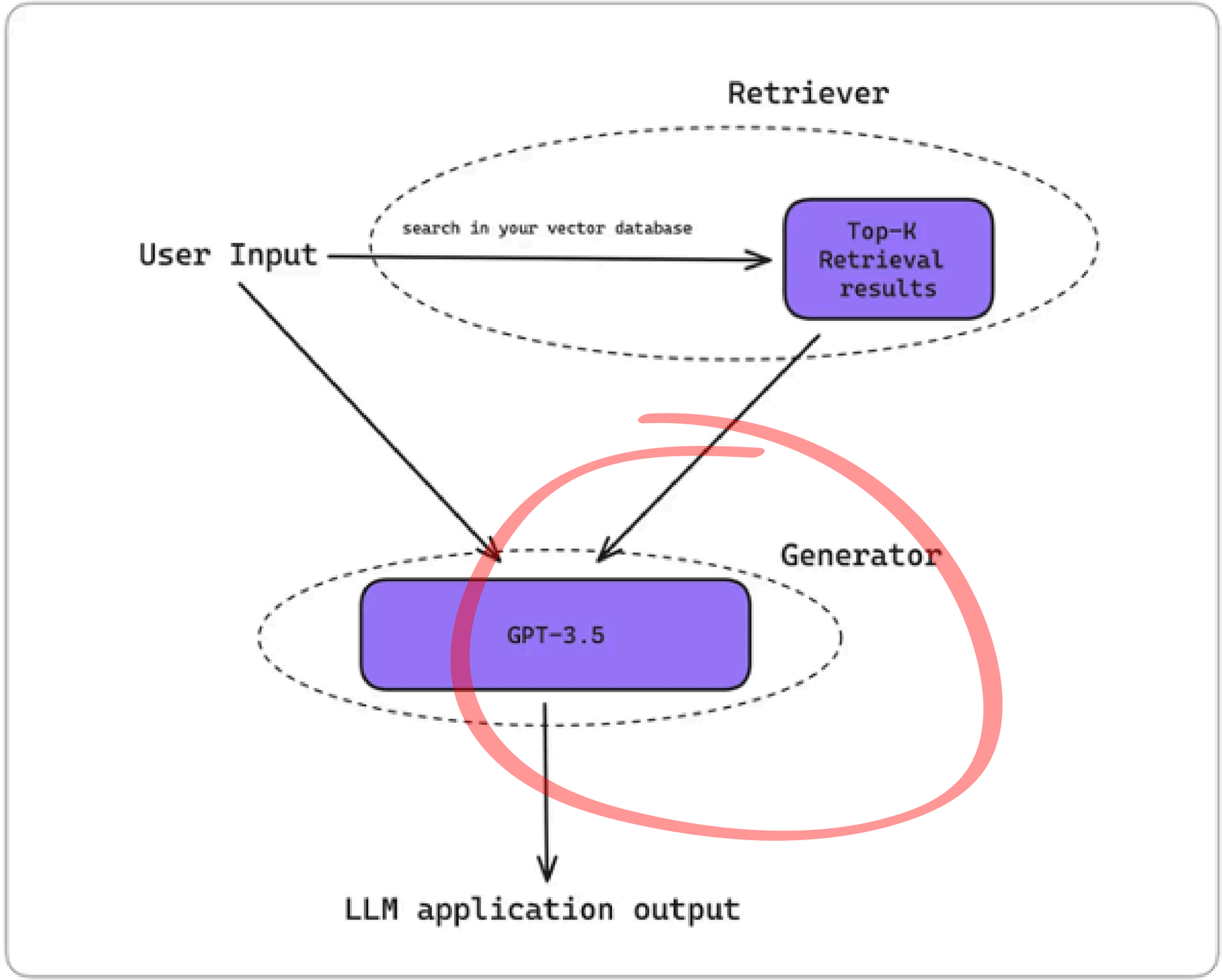
Euclidean distance

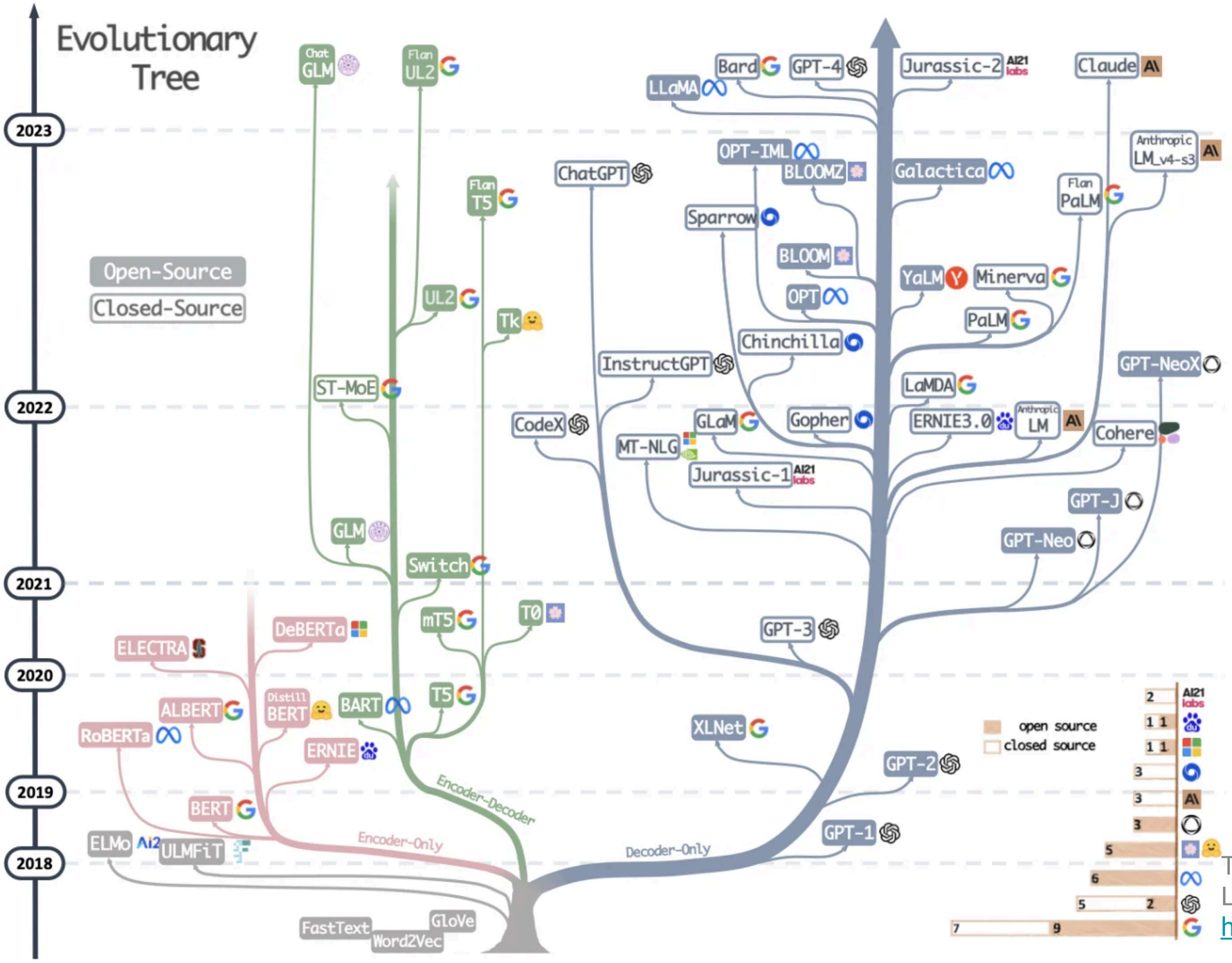


Cosine similarity



Dot Product


Retriever
search in your vector database
User Input
Top-K
Retrieval
results
Generator
GPT-3.5
LLM application output



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

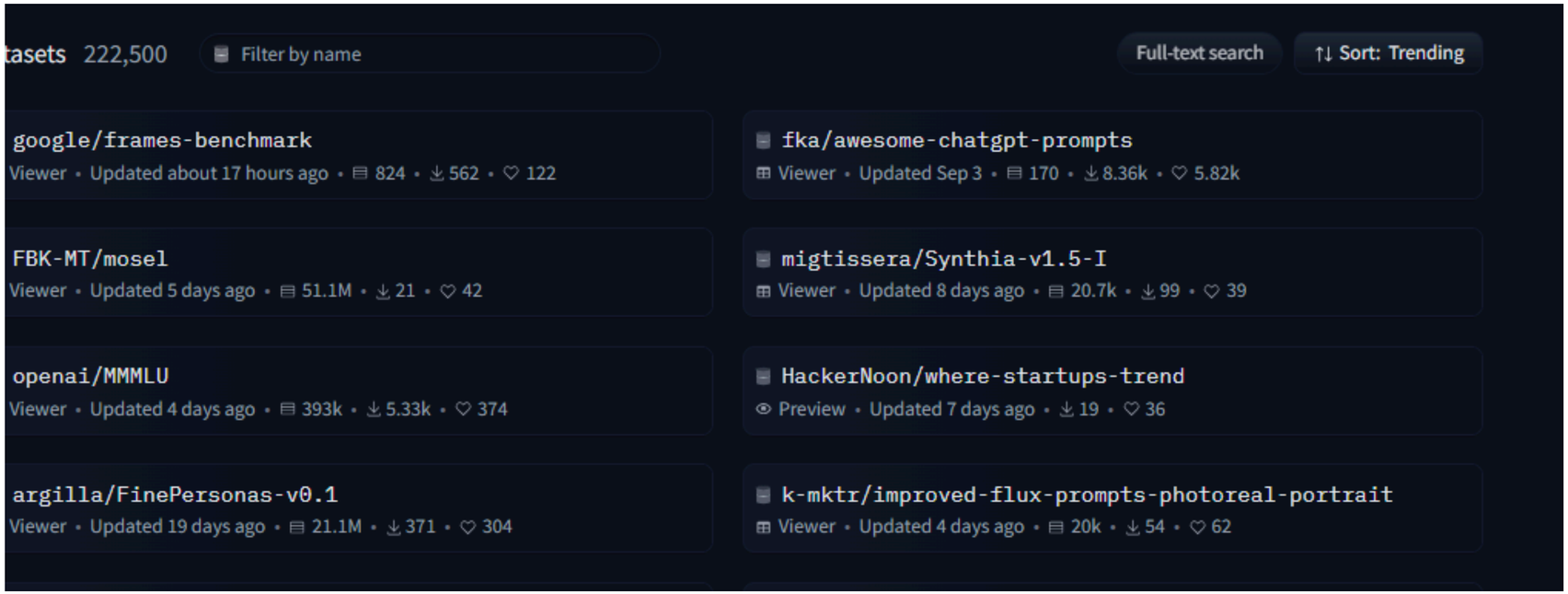
Where to find pretrained LLMs ?

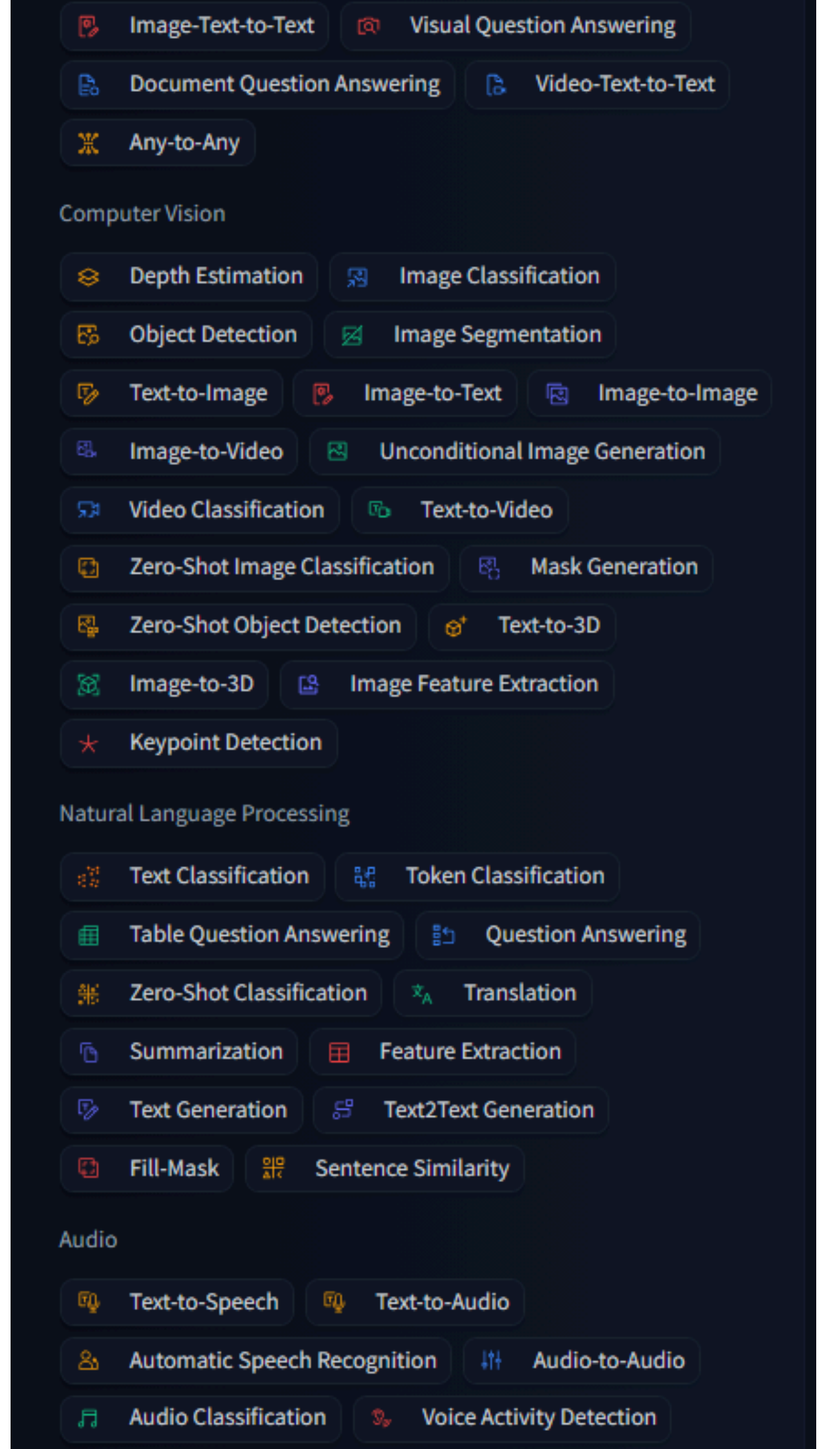
Hugging Face

groq®

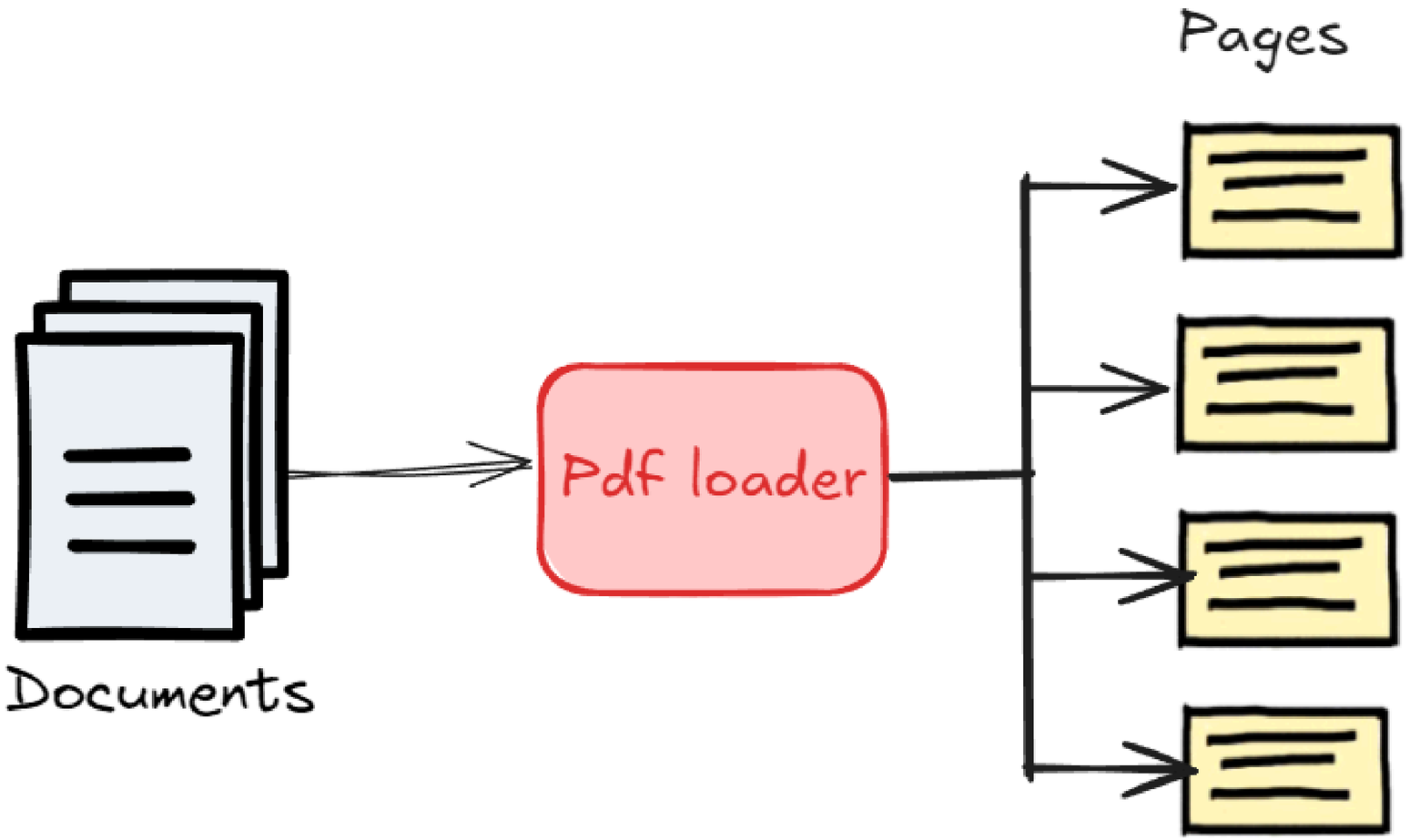
By Mohammed Arbi Nsibi

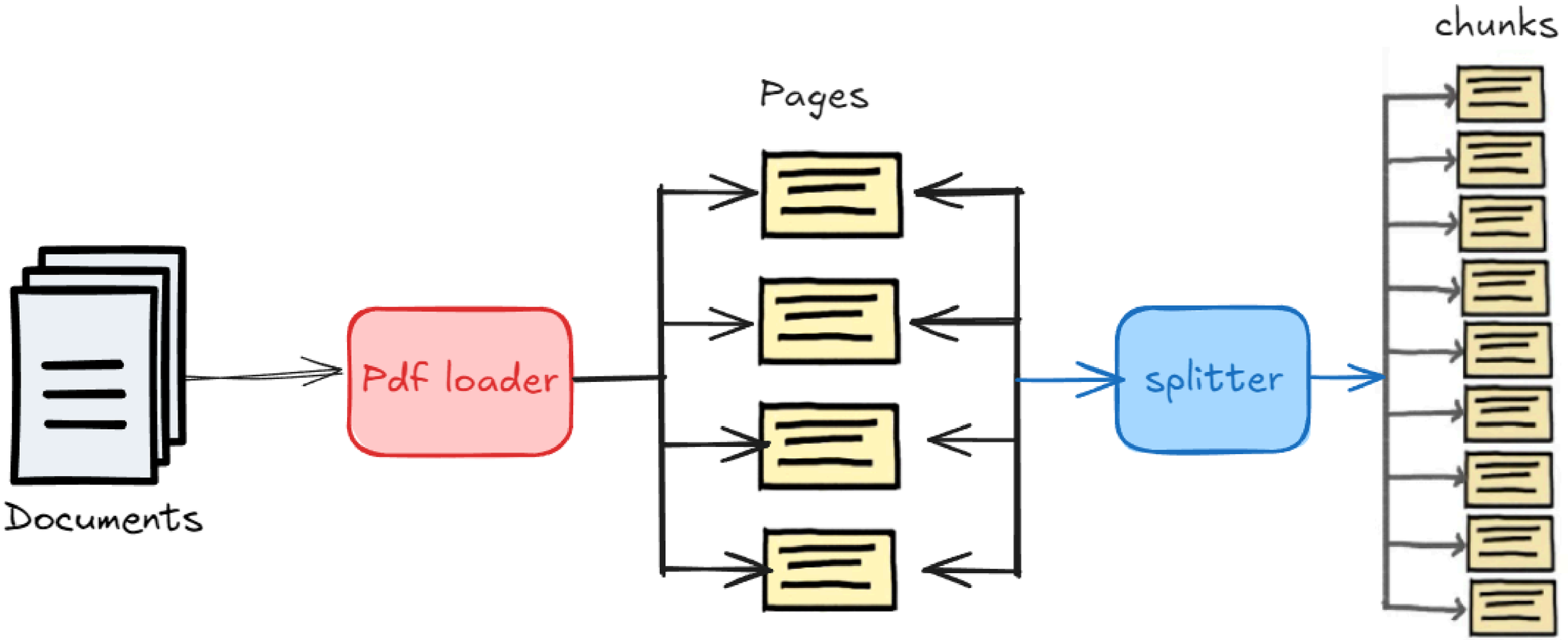
Where to find pretrained LLMs ?



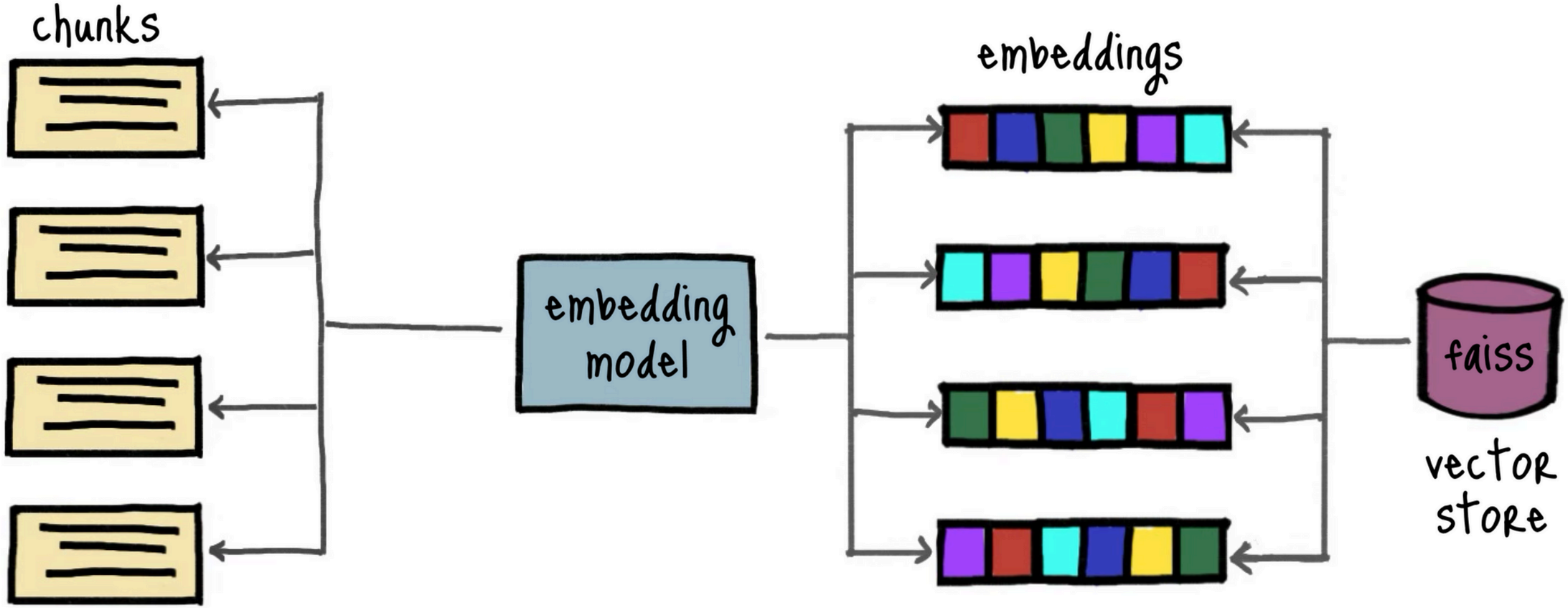


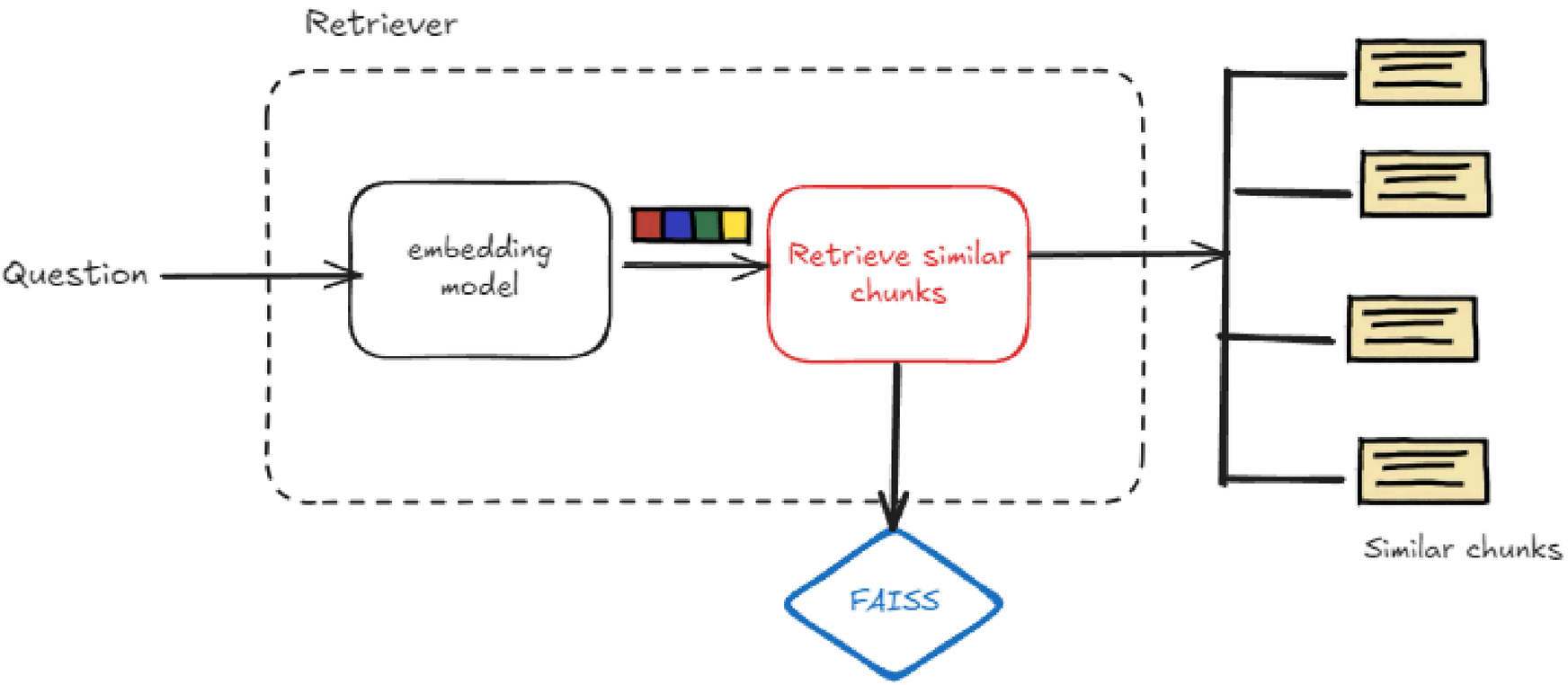
RAG architecture



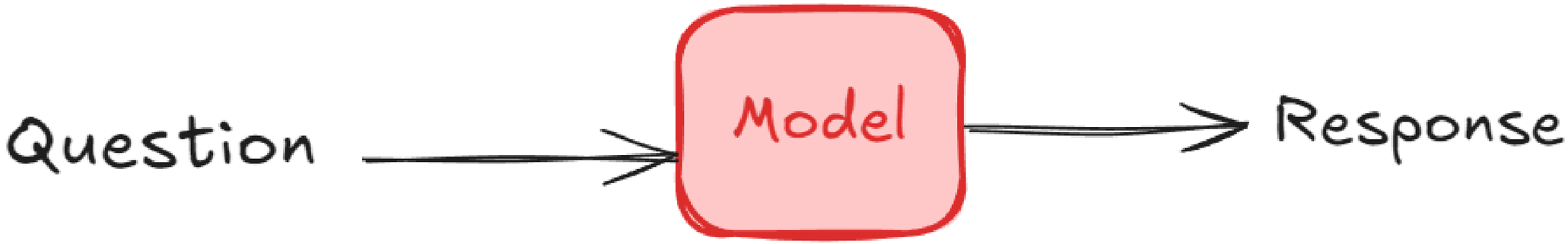


By Mohammed Arbi Nsibi

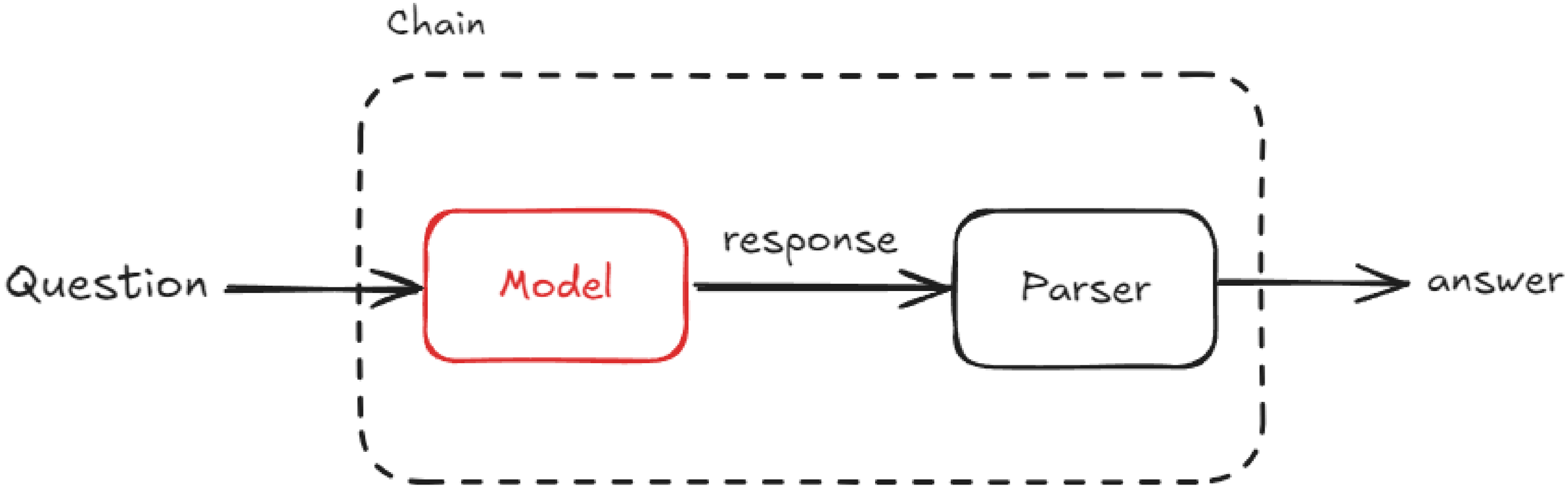

chunks
embedding model
embeddings
faiss
vector store



By Mohammed Arbi Nsibi

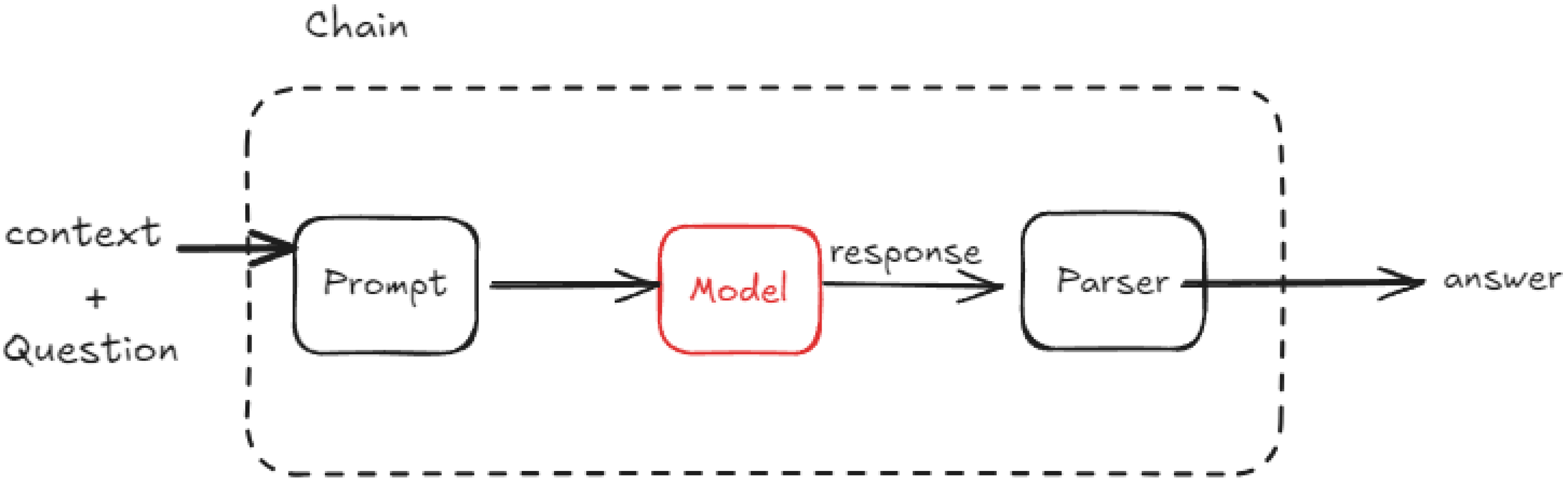



```
AIMessage(content='As of my last update in April 2023, Joe Biden is the President of the United States. He took office
```



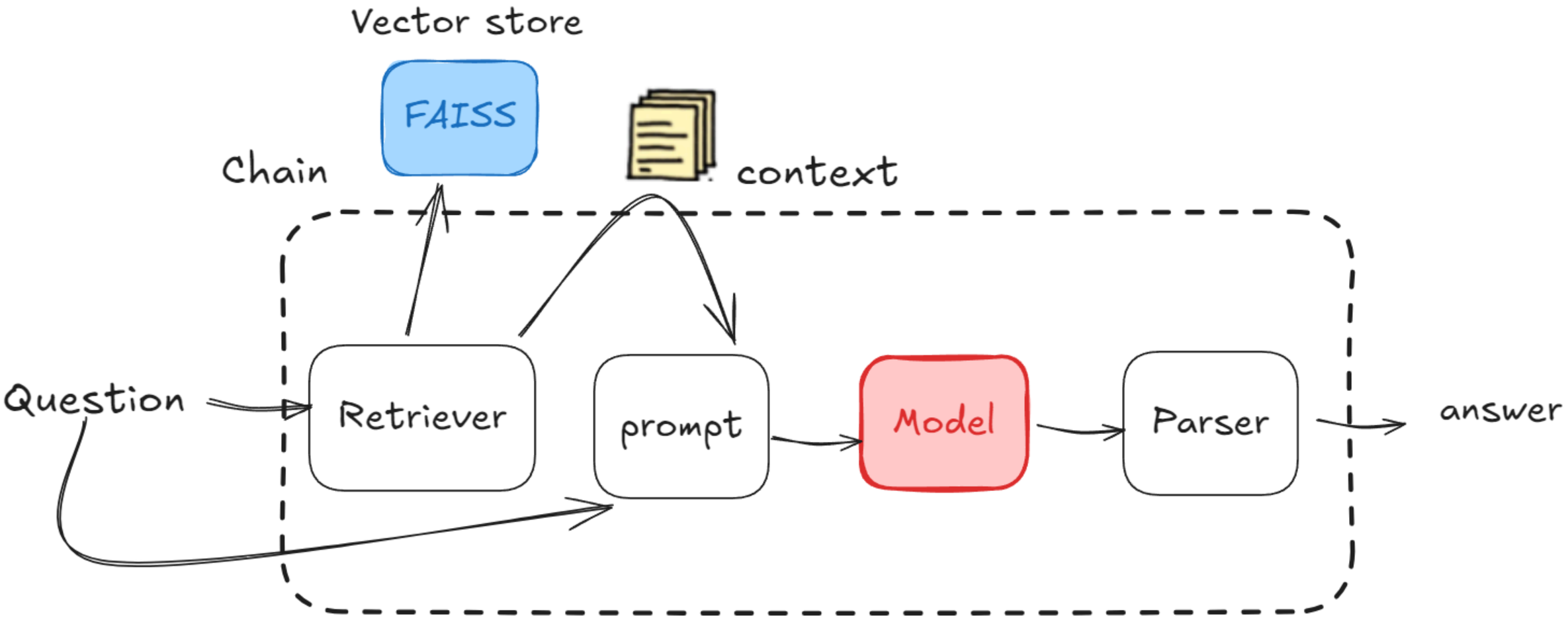

```
'As of my last update in April 2023, Joe Biden is the President of the United States. He took office on January 20, 2021,
```


context
+
Question
Prompt template
Prompt



By Mohammed Arbi Nsibi

https://arxiv.org/abs/2005.11401



By Mohammed Arbi Nsibi

How to get started ?

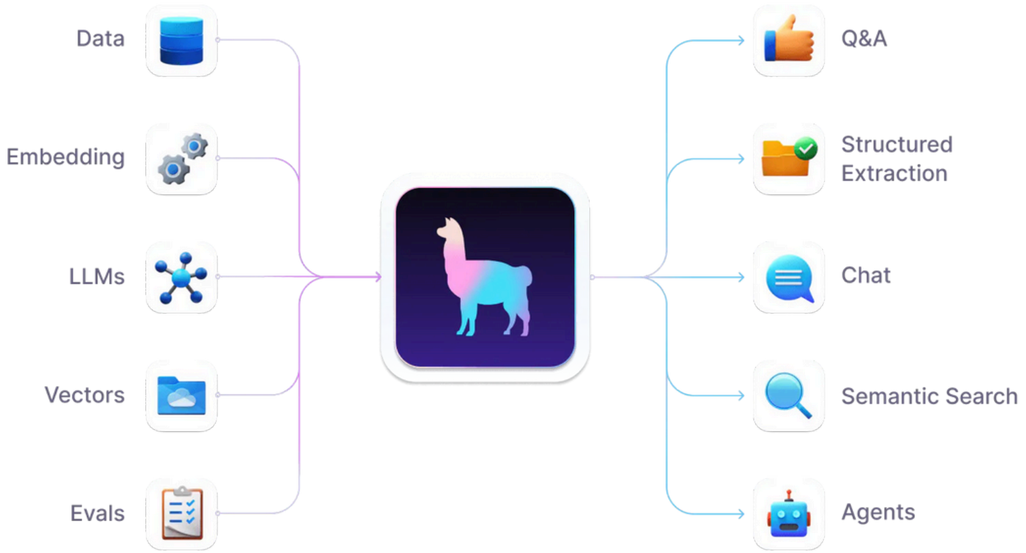
By Mohammed Arbi Nsibi

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.

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

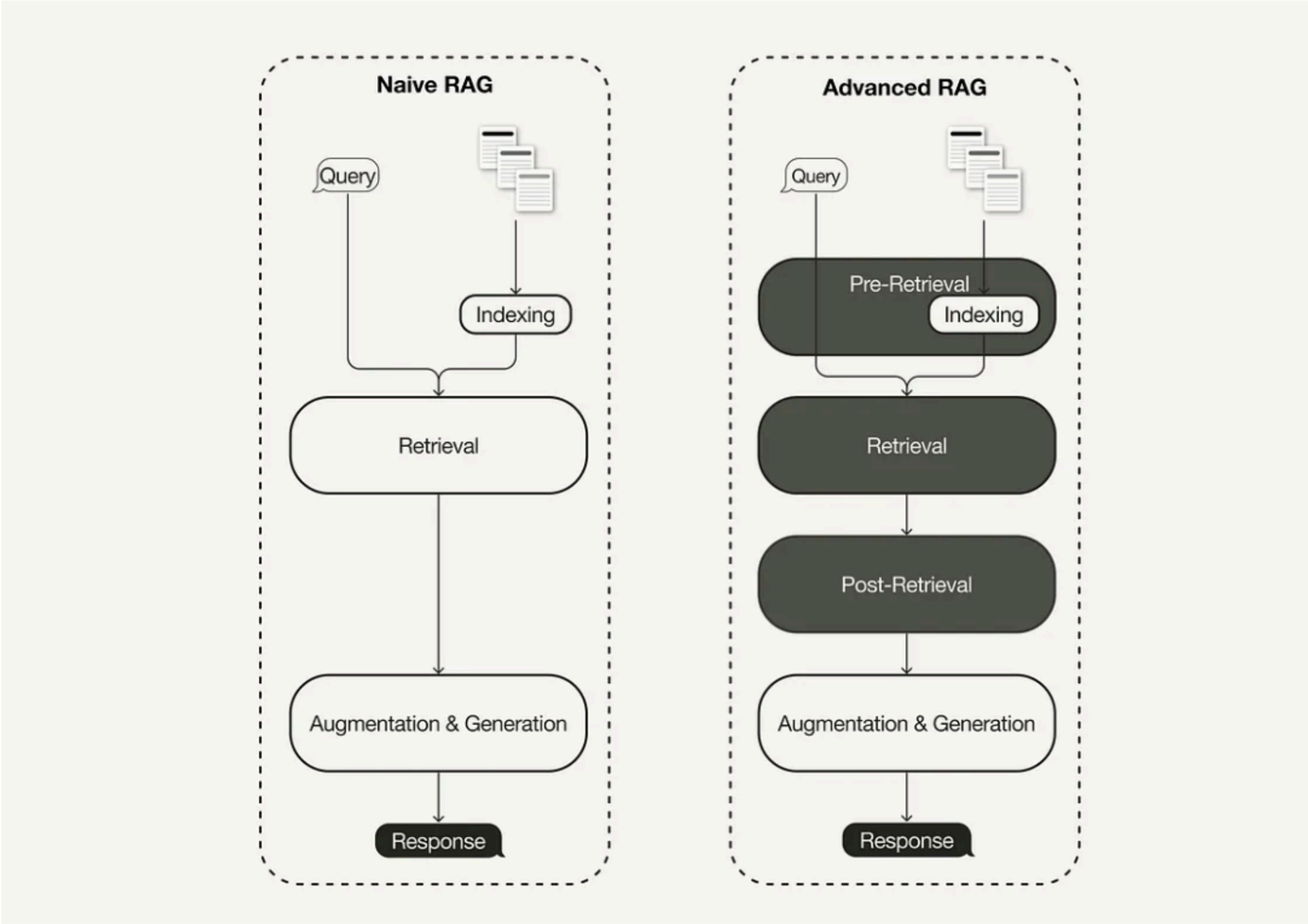

always has been
wait so creating chatbots is that easy ?


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

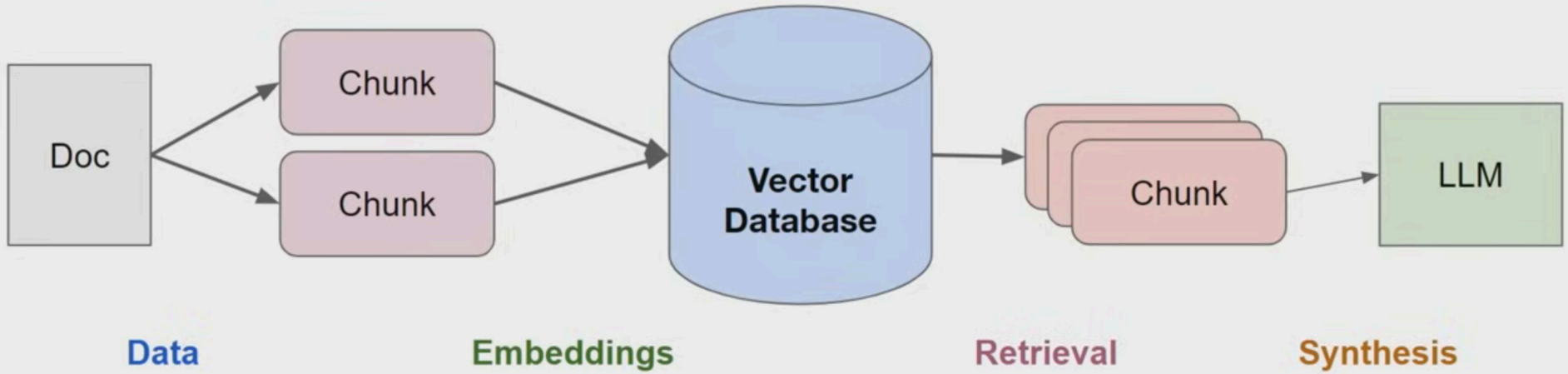


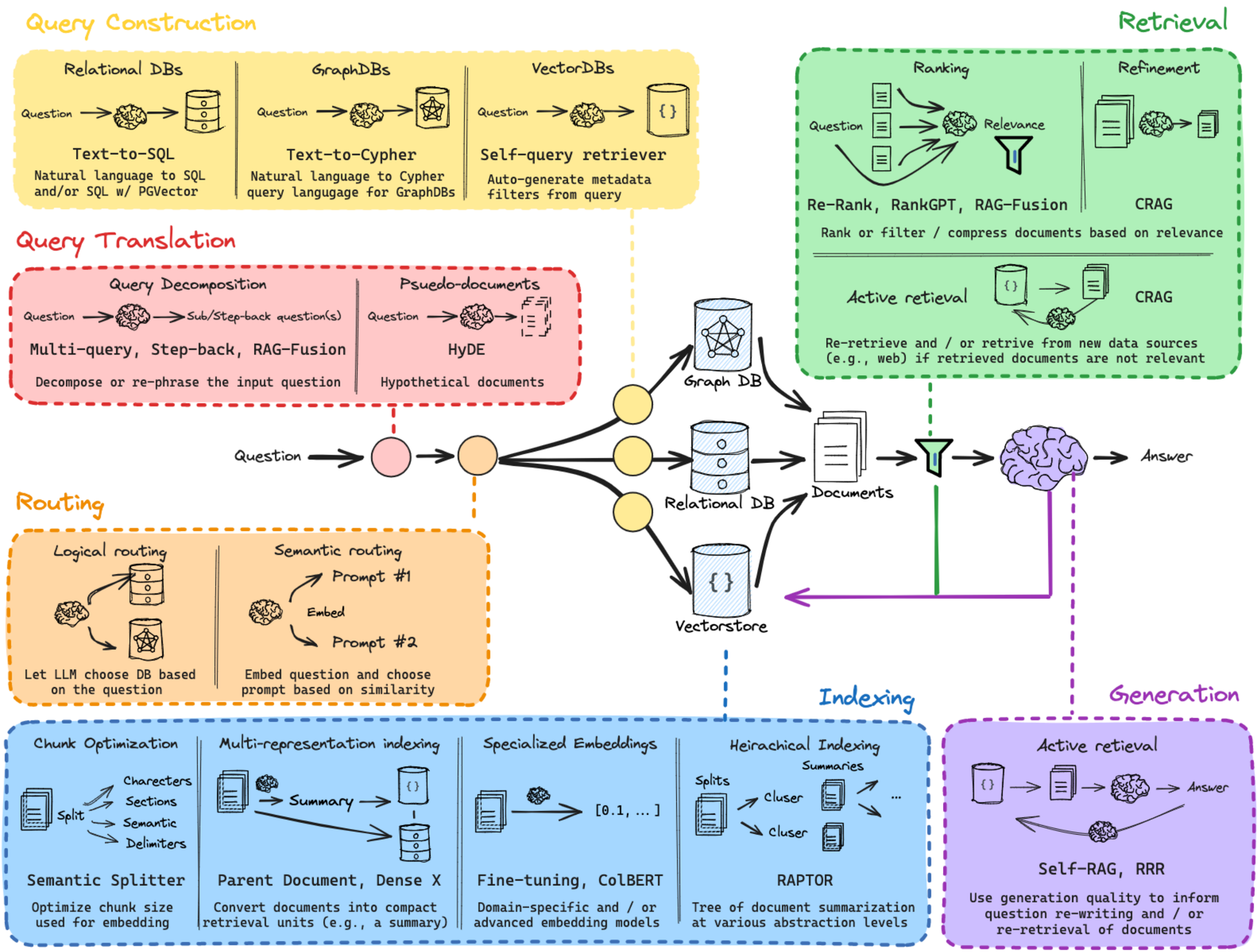
Difference between Naive and Advanced RAG (Image by the author, inspired by [1])

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? v





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


LLM
hallucinations
are now illegal

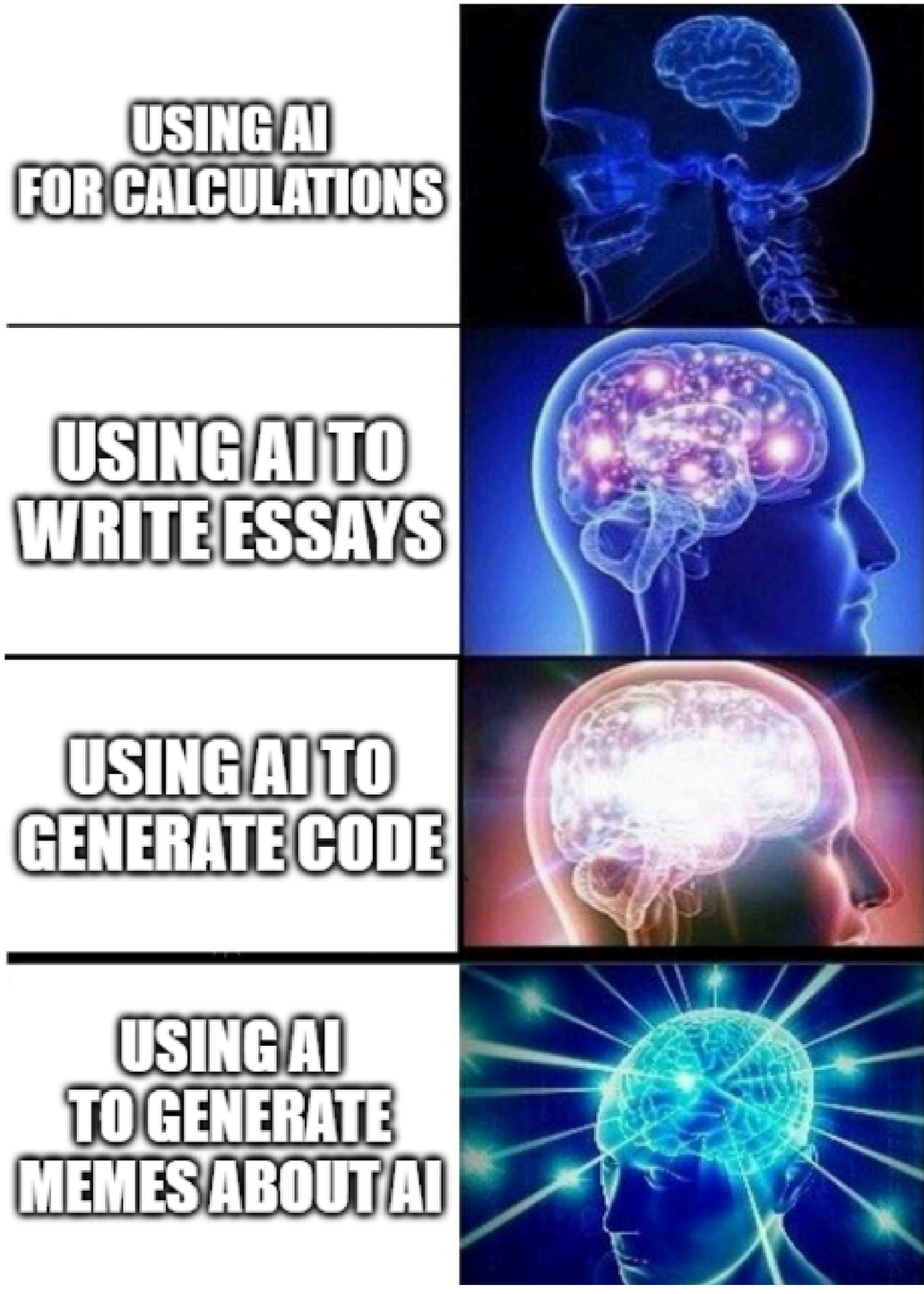
Resources

- 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

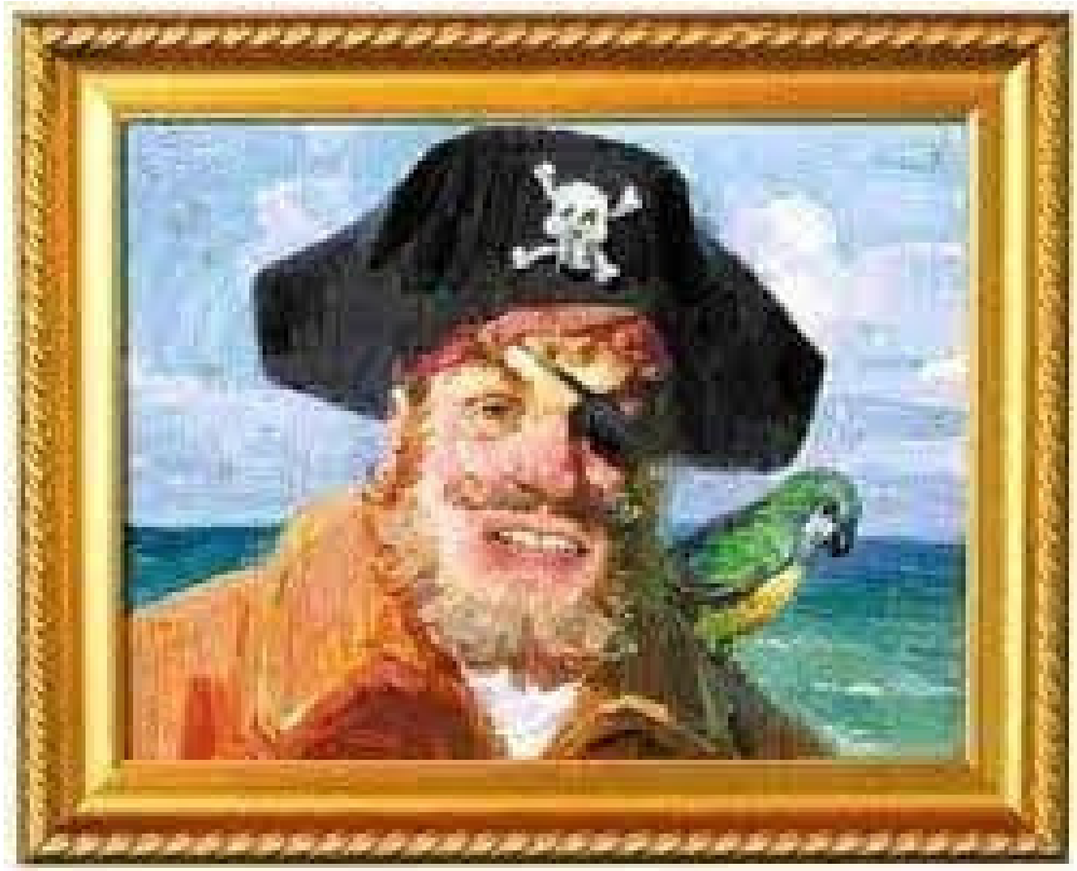
By Mohammed Arbi Nsibi

THANK YOU for your attention!!



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

QUIZ TIME



By Mohammed Arbi Nsibi