Introduction to RAG

19/02/2025

Data Overflow
2025
INSAT ACM Student Chapter
IEEE Computer Society INSAT SB

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

🤗 https://huggingface.co/Goodnight7

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

mohammedarbinsibi@gmail.com

Content

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

By Mohammed Arbi Nsibi

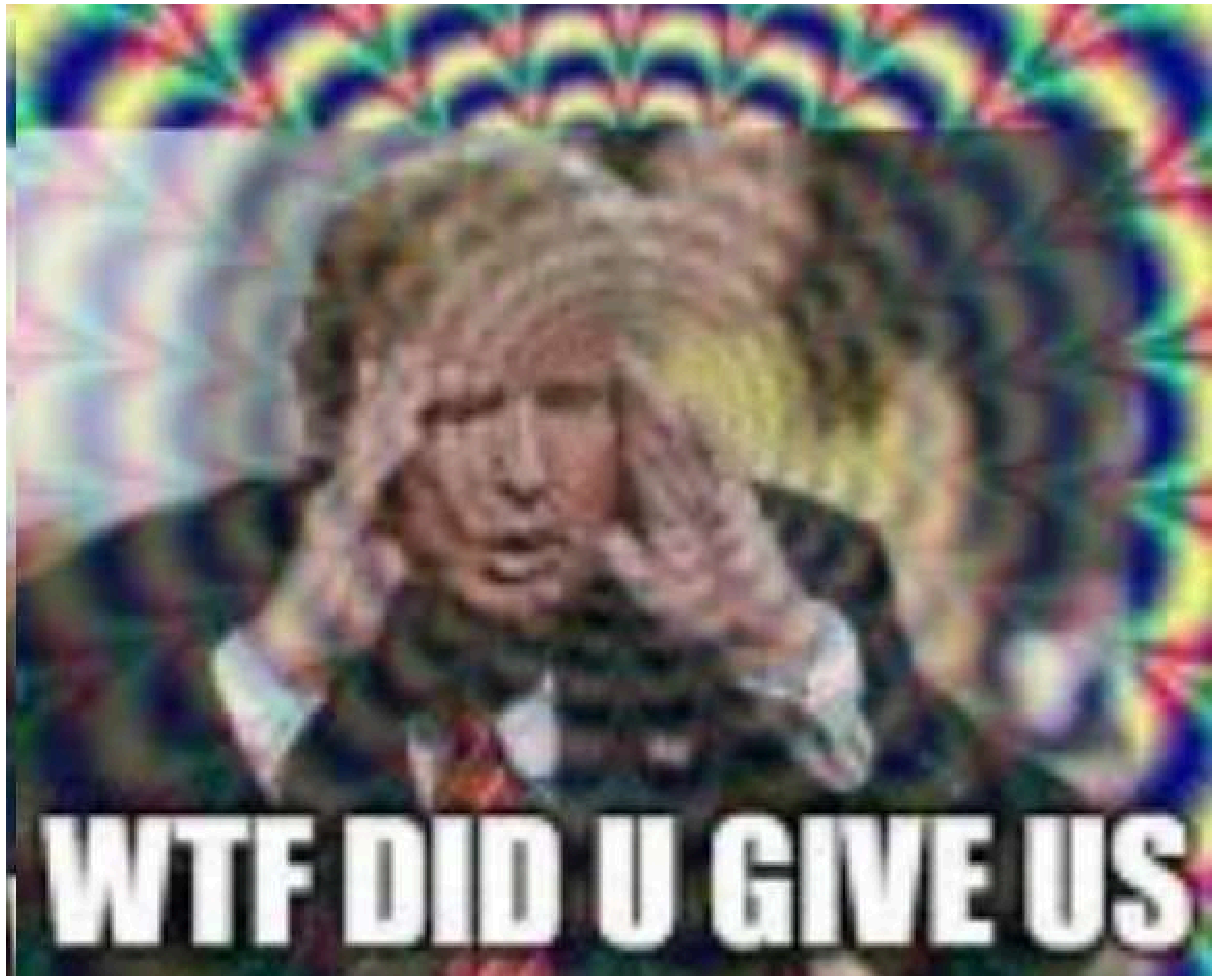
Motivation (Why do we need RAG?)

By Mohammed Arbi Nsibi

1- The need for an external Knowledge !

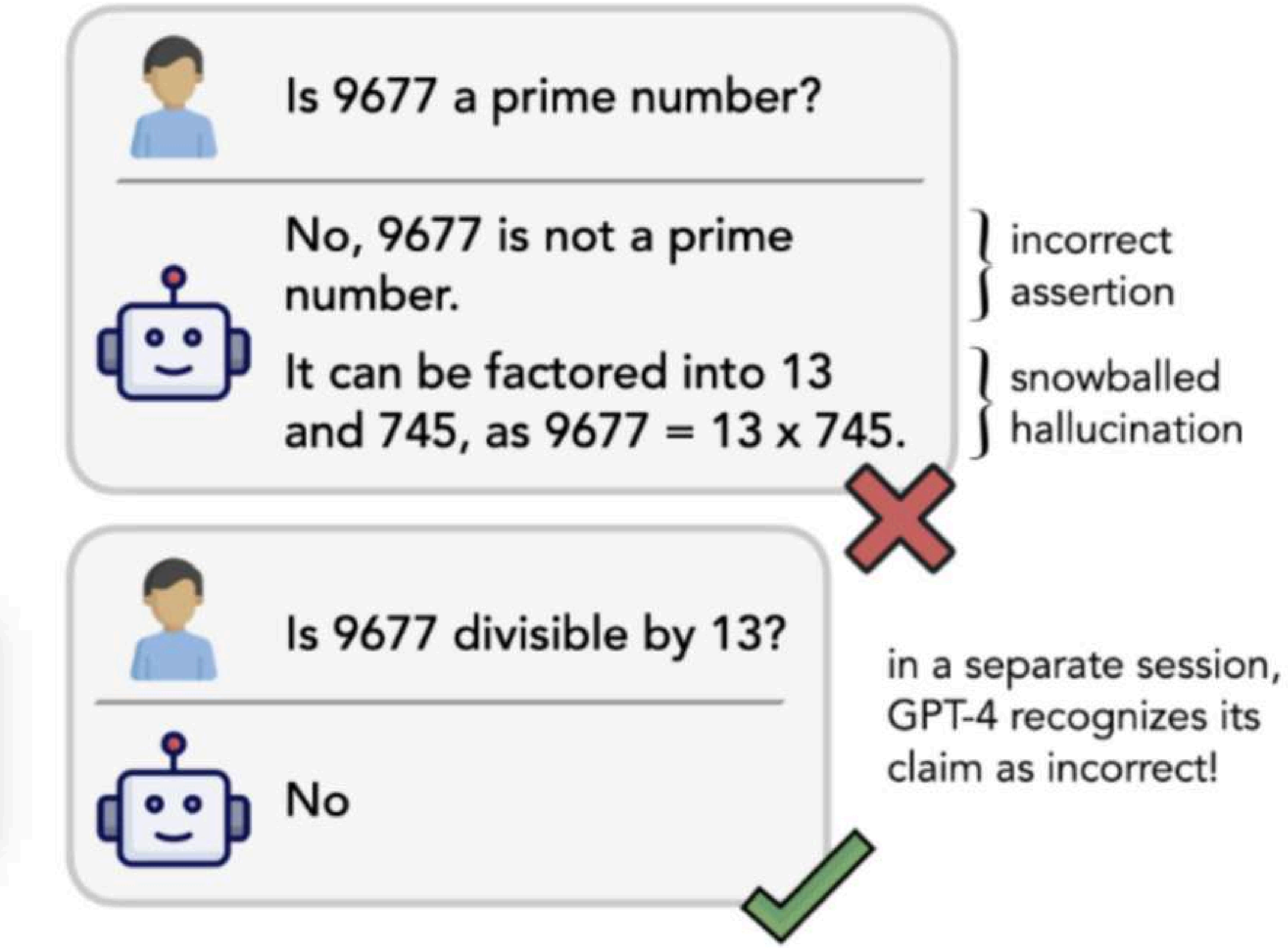
By Mohammed Arbi Nsibi

2- Hallucinations



By Mohammed Arbi Nsibi

2- Hallucinations



Is 9677 a prime number?

No, 9677 is not a prime number.

It can be factored into 13 and 745, as 9677 = 13 x 745.

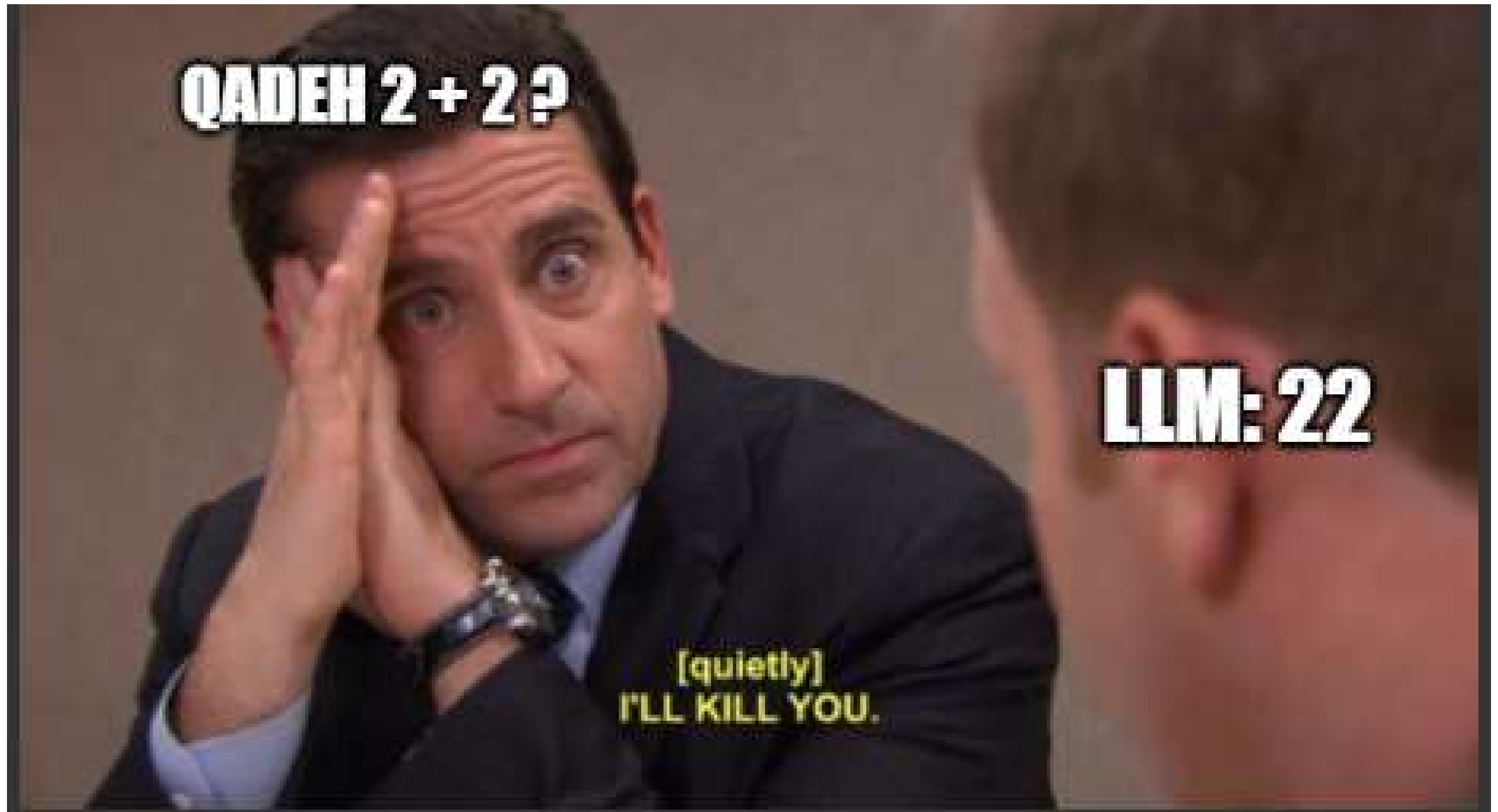
incorrect assertion

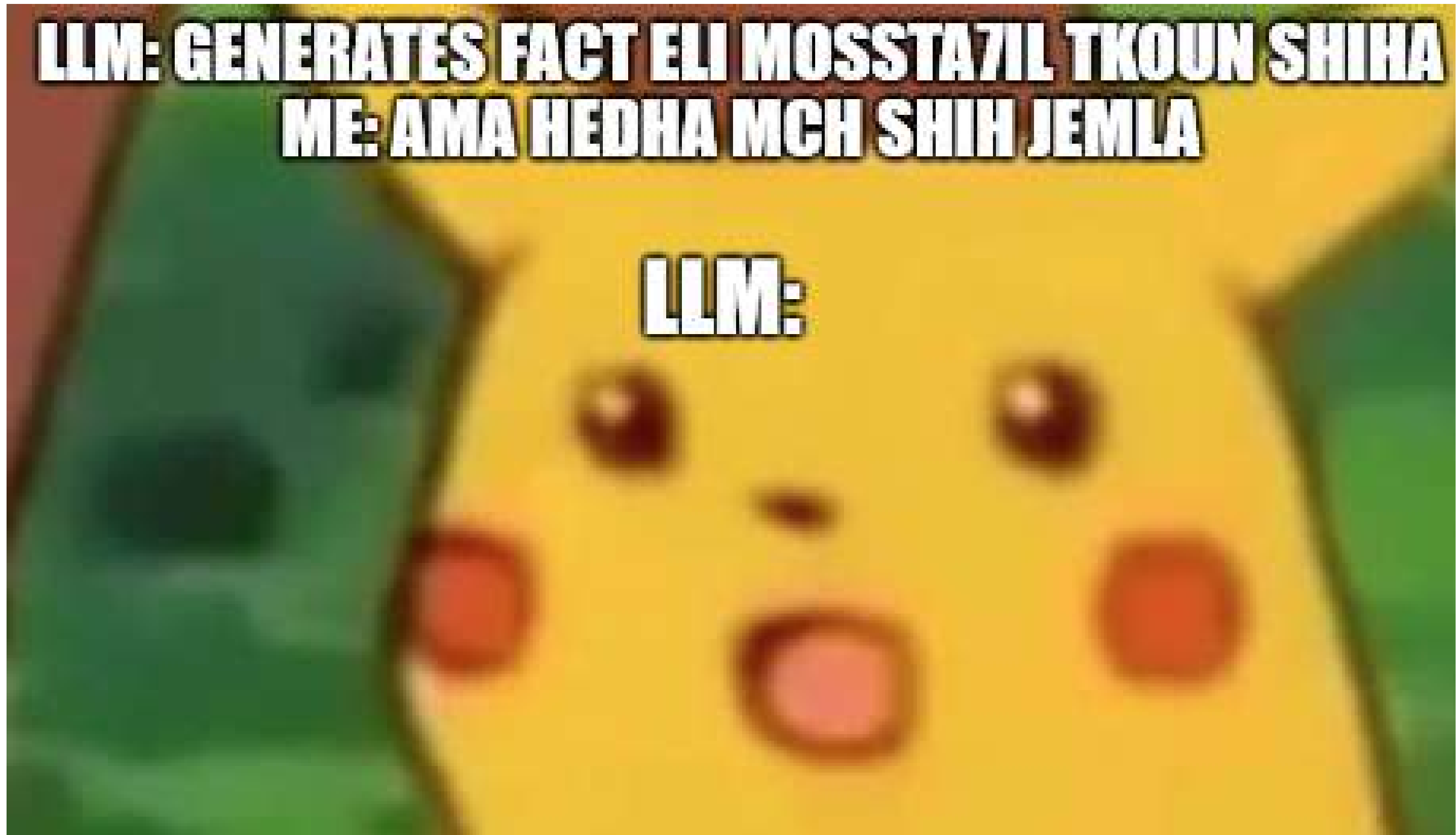
snowballed hallucination

Is 9677 divisible by 13?

No

in a separate session, GPT-4 recognizes its claim as incorrect!




LLM: GENERATES FACT ELI MOSSTA7IL TKOUN SHIHA
ME: AMA HEDHA MCH SHIH JEMLA
LLM:

Hallucinations

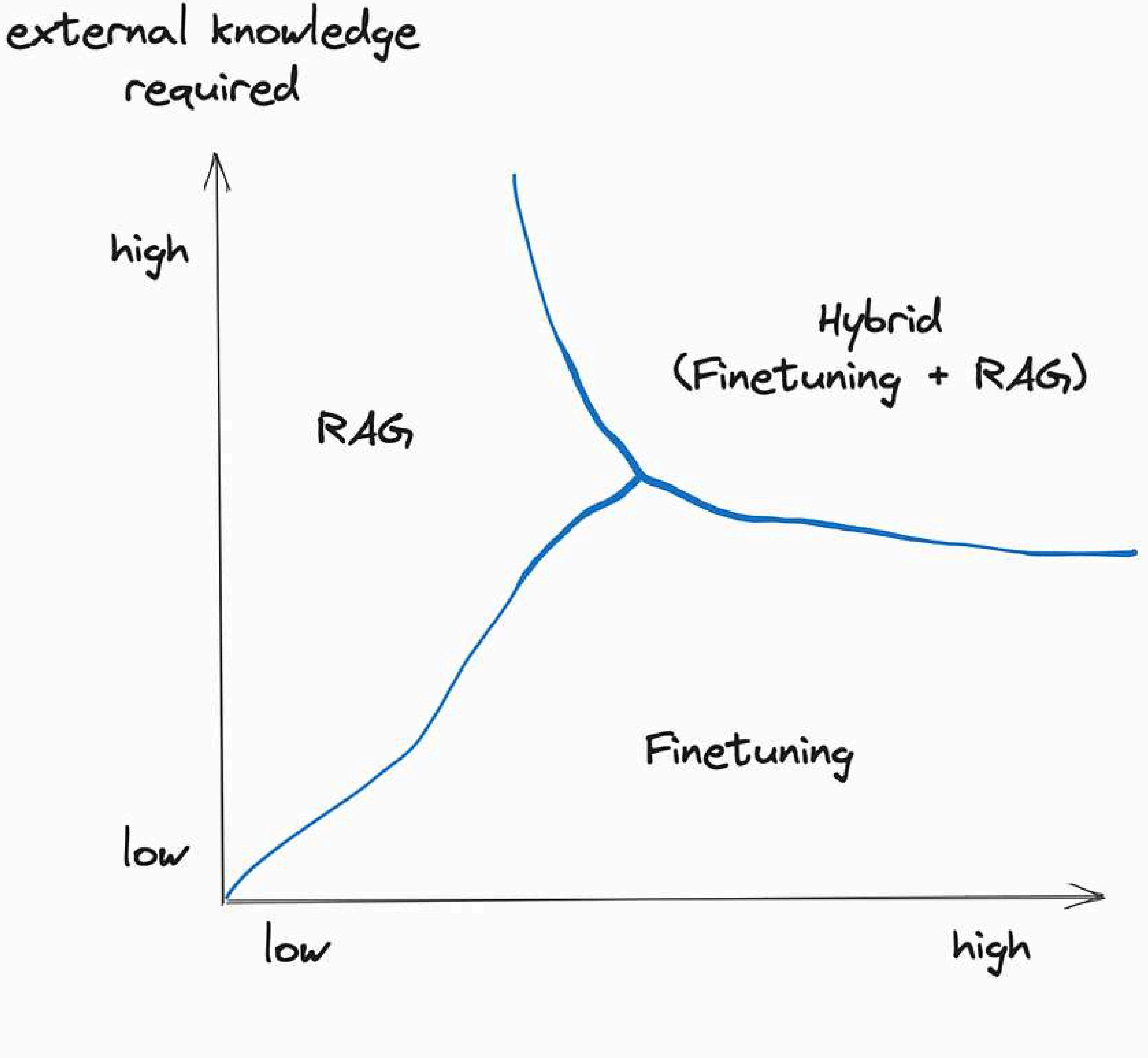
- The model is not trained on enough data.
- The model is trained on noisy or dirty data.
- The model is not given enough context .
- The model is not given enough constraints (rules, guidelines, or limitations)

By Mohammed Arbi Nsibi

Solutions ?

By Mohammed Arbi Nsibi

RAG / Fine-tuning

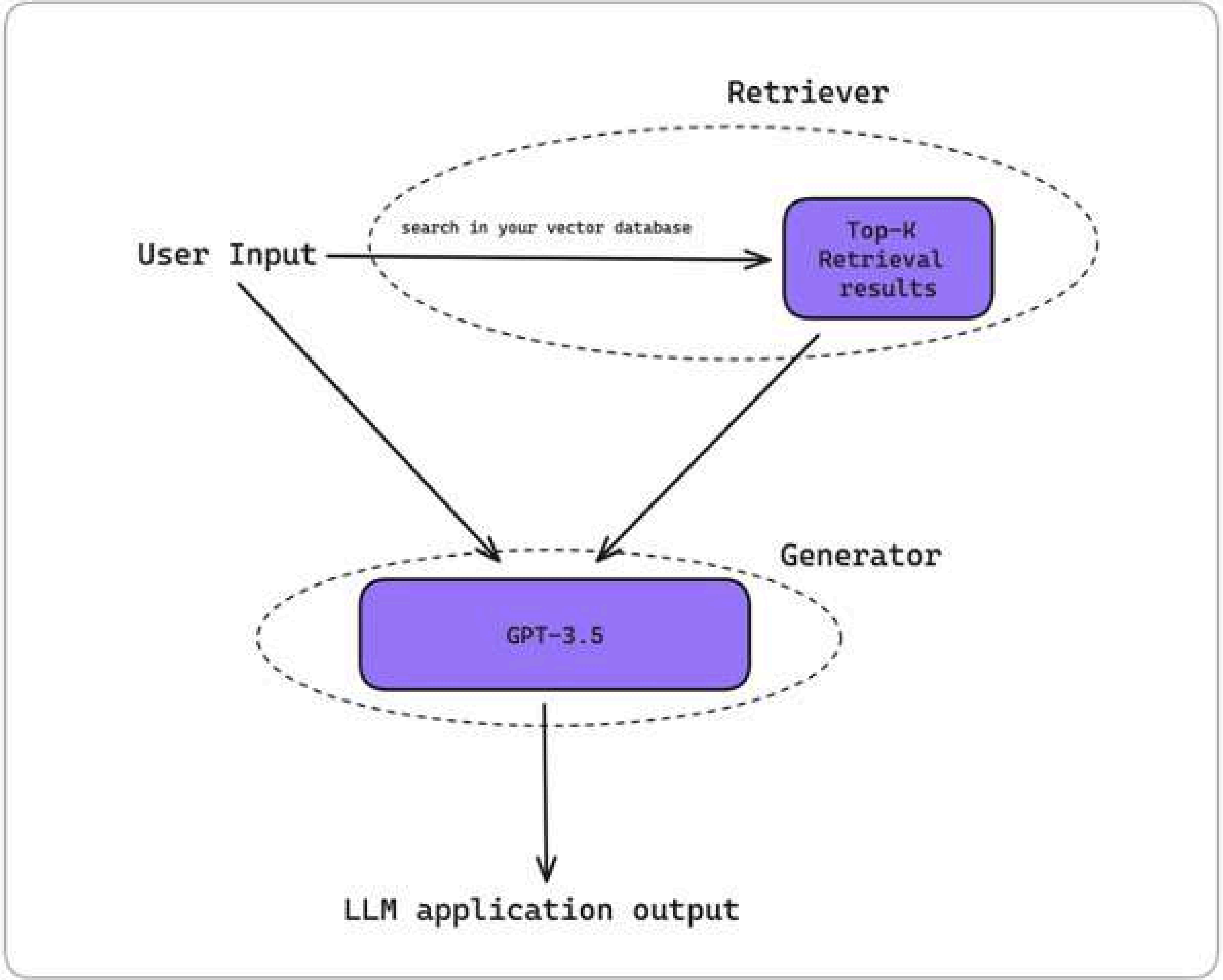

external knowledge required
high
low
RAG
Hybrid (Finetuning + RAG)
Finetuning
low
high
model adaptation required (e.g. behaviour/ writing style/ vocabulary)

But wait..
WTF is RAG ?

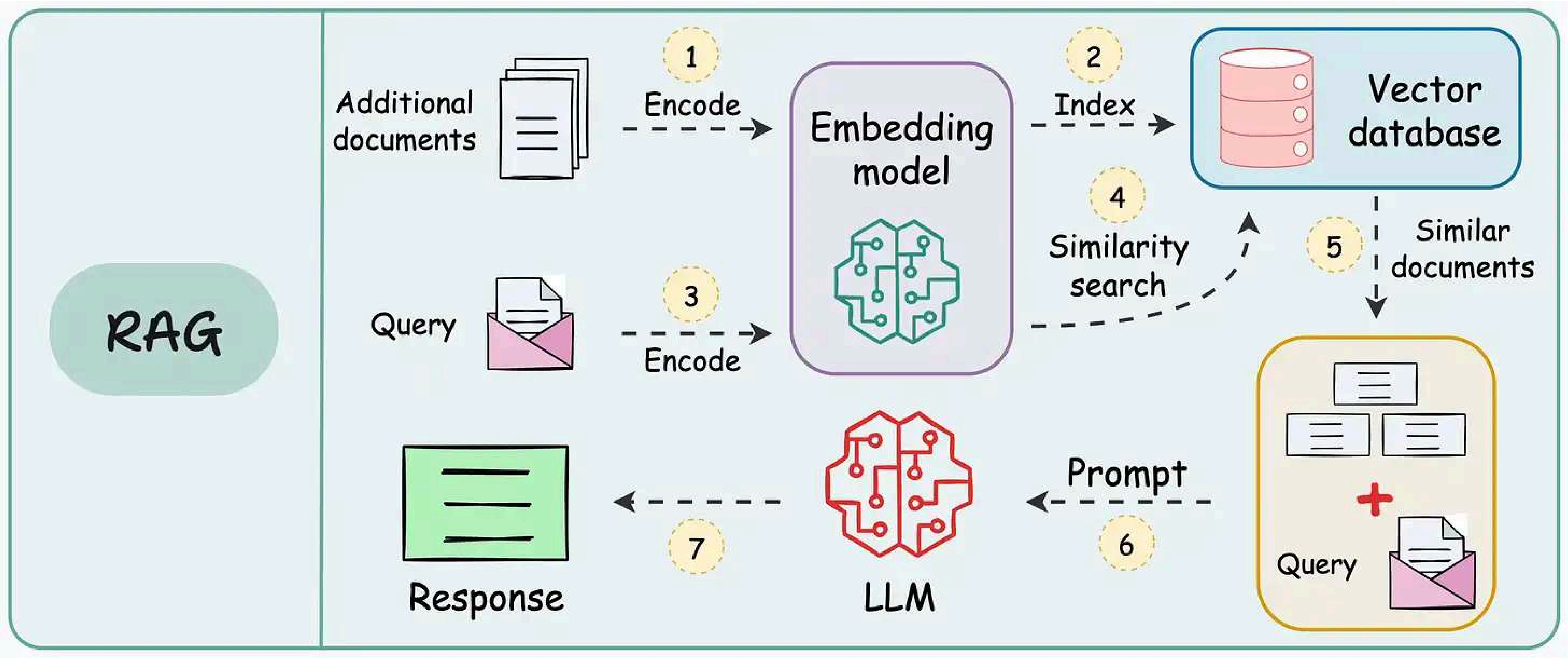
RAG (Retrieval-augmented generation)

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

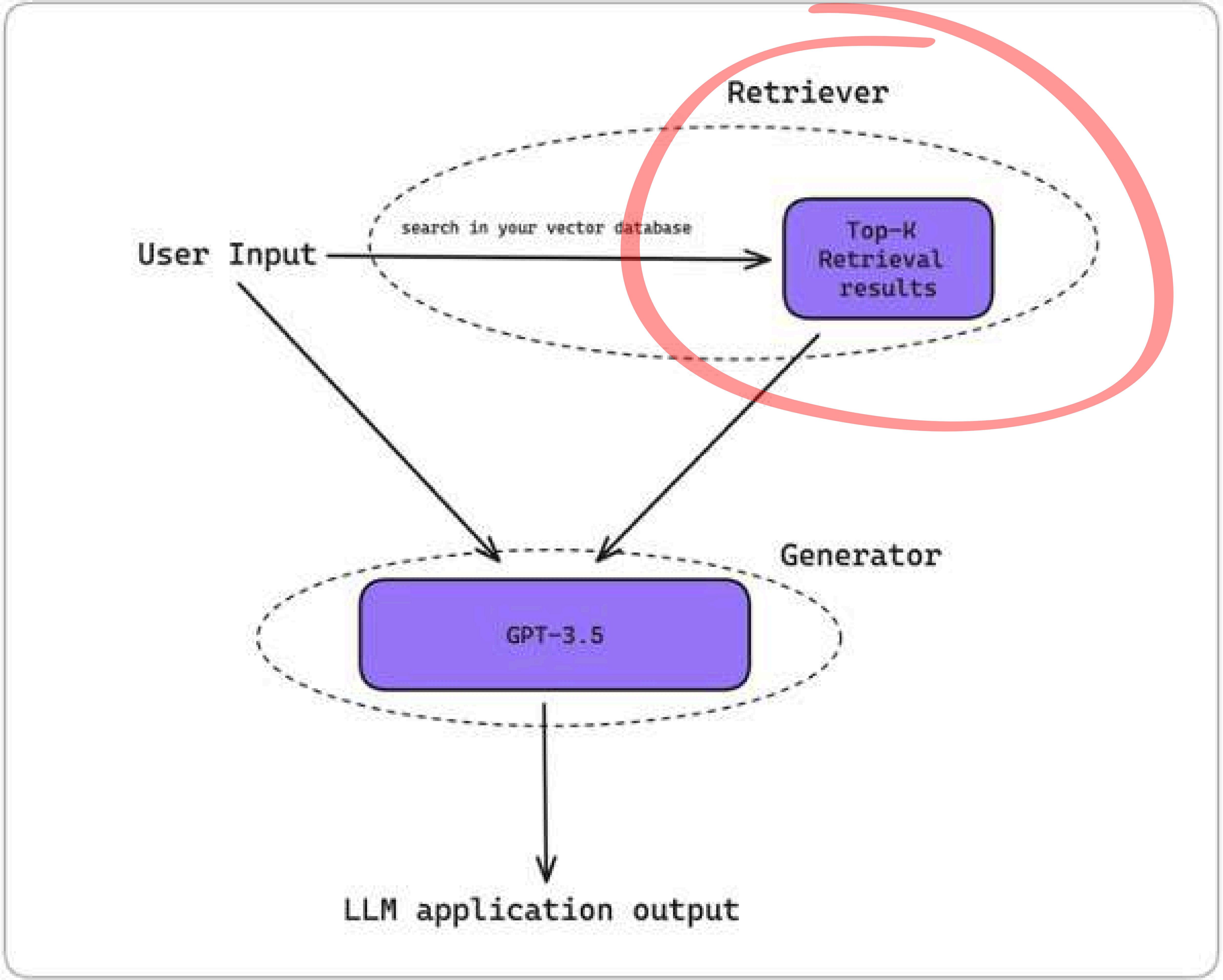
By Mohammed Arbi Nsibi


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

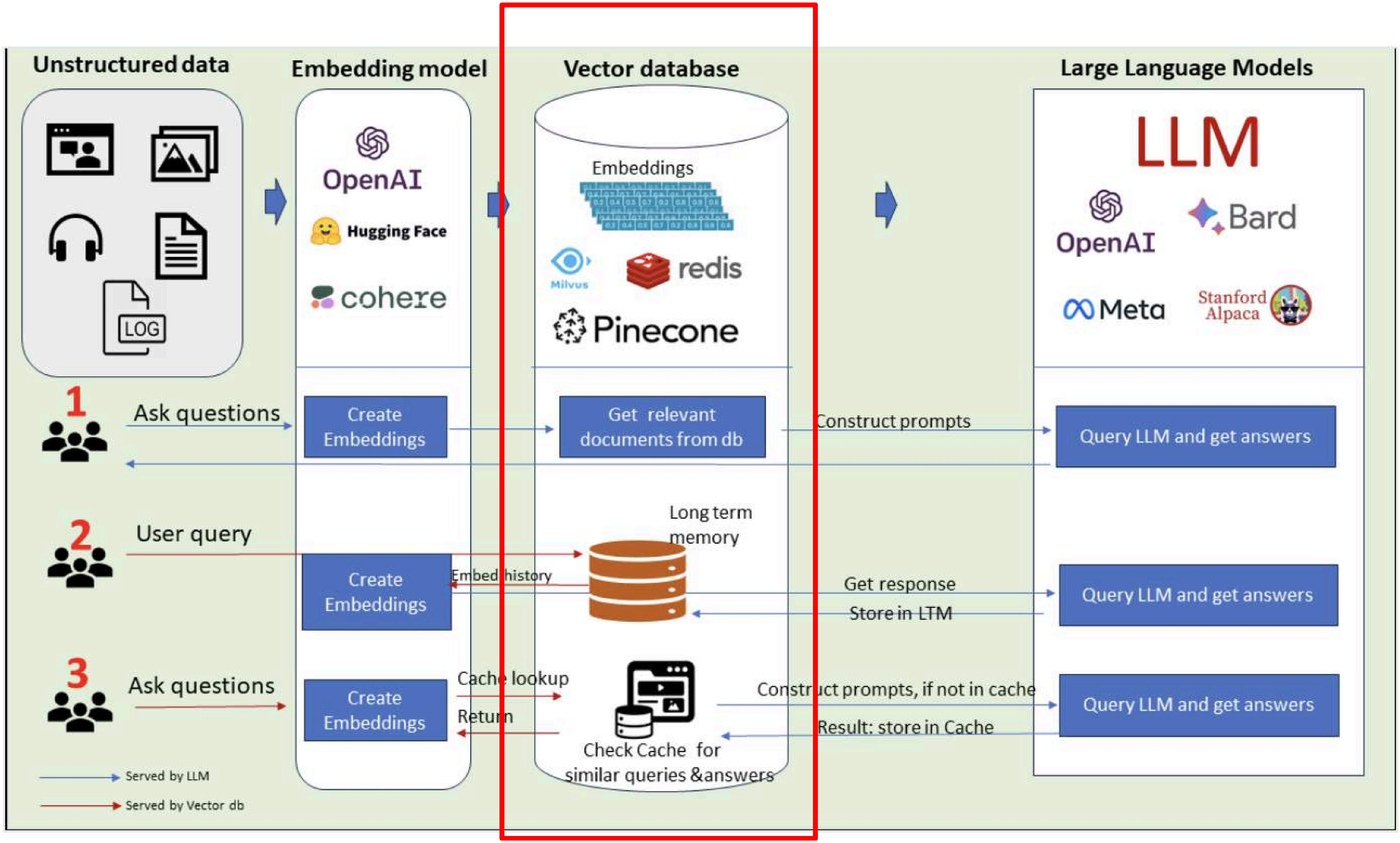
RAG (Retrieval-augmented generation)



By Mohammed Arbi Nsibi


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

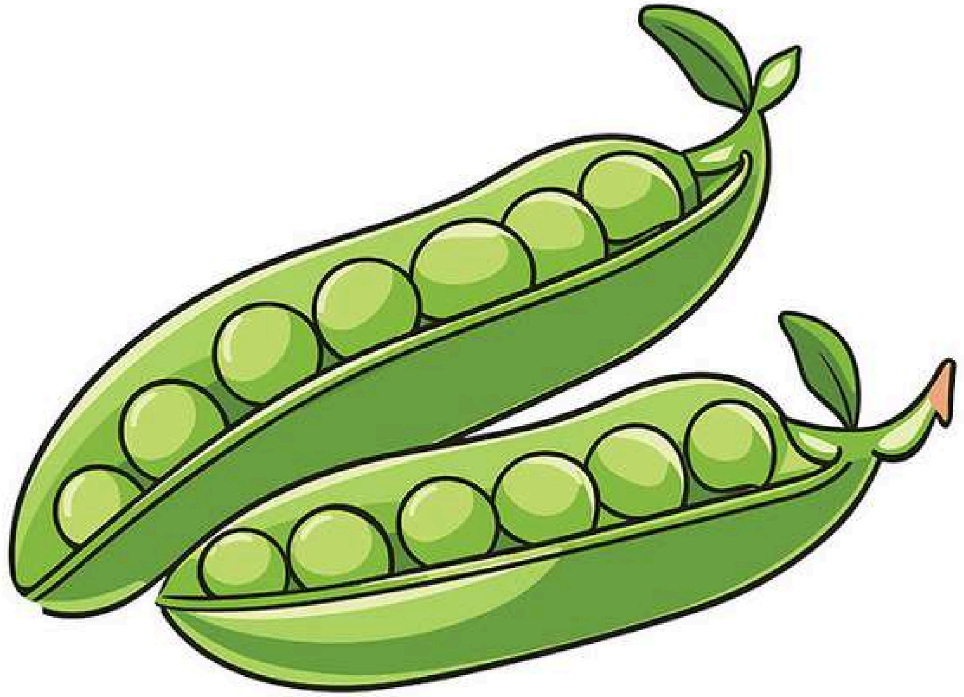
Vector Database



tea

coffee

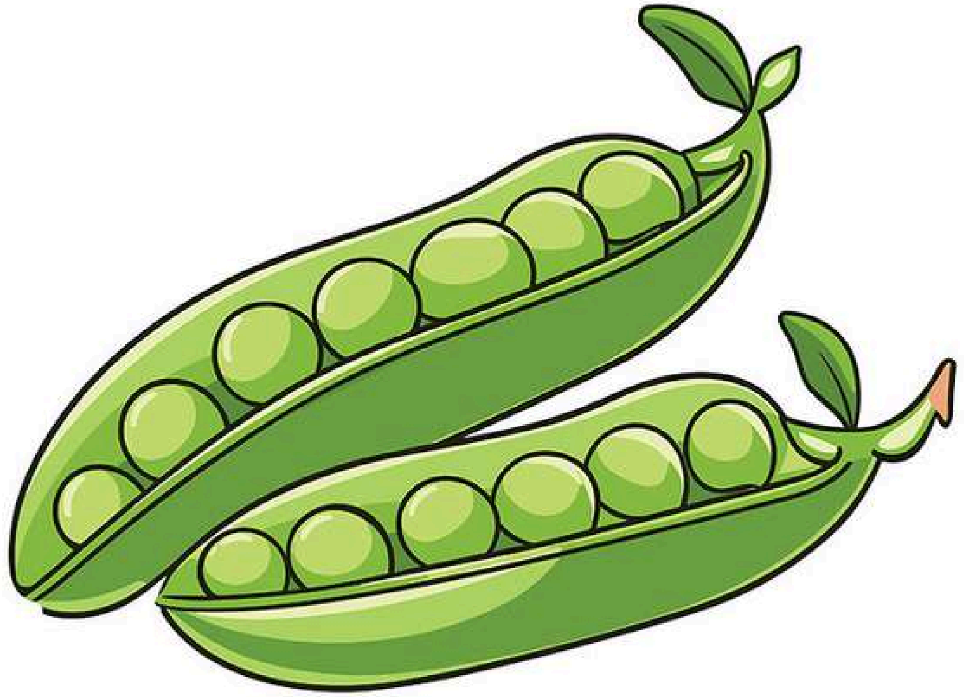
tea



pea

tea

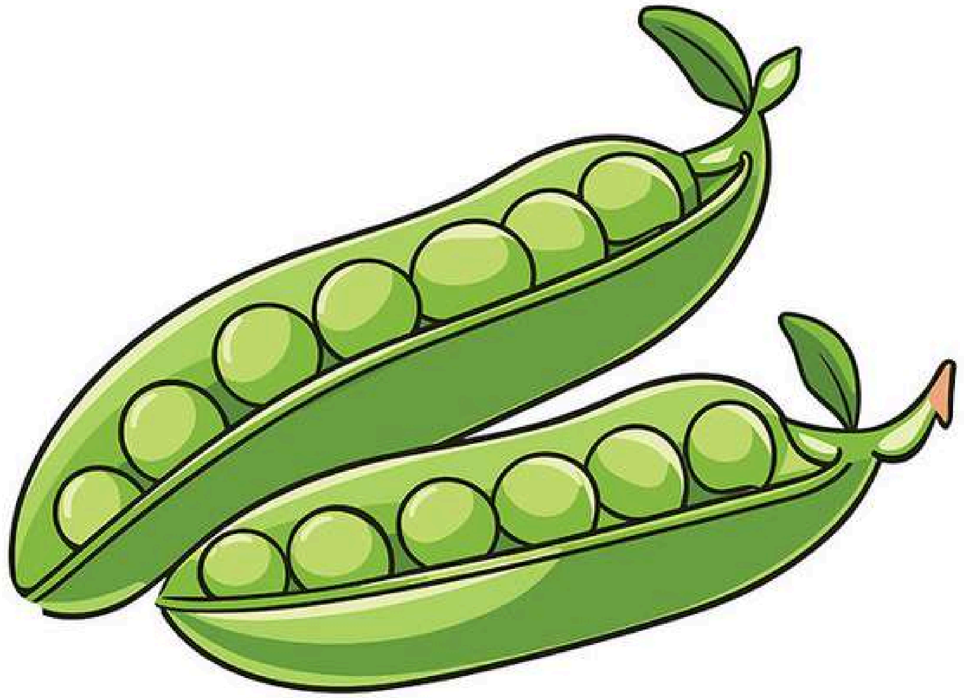
≠



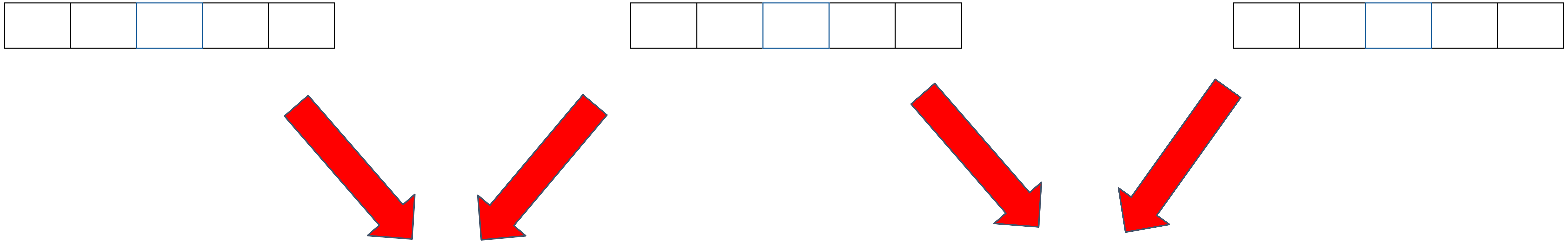
pea

tea

coffee

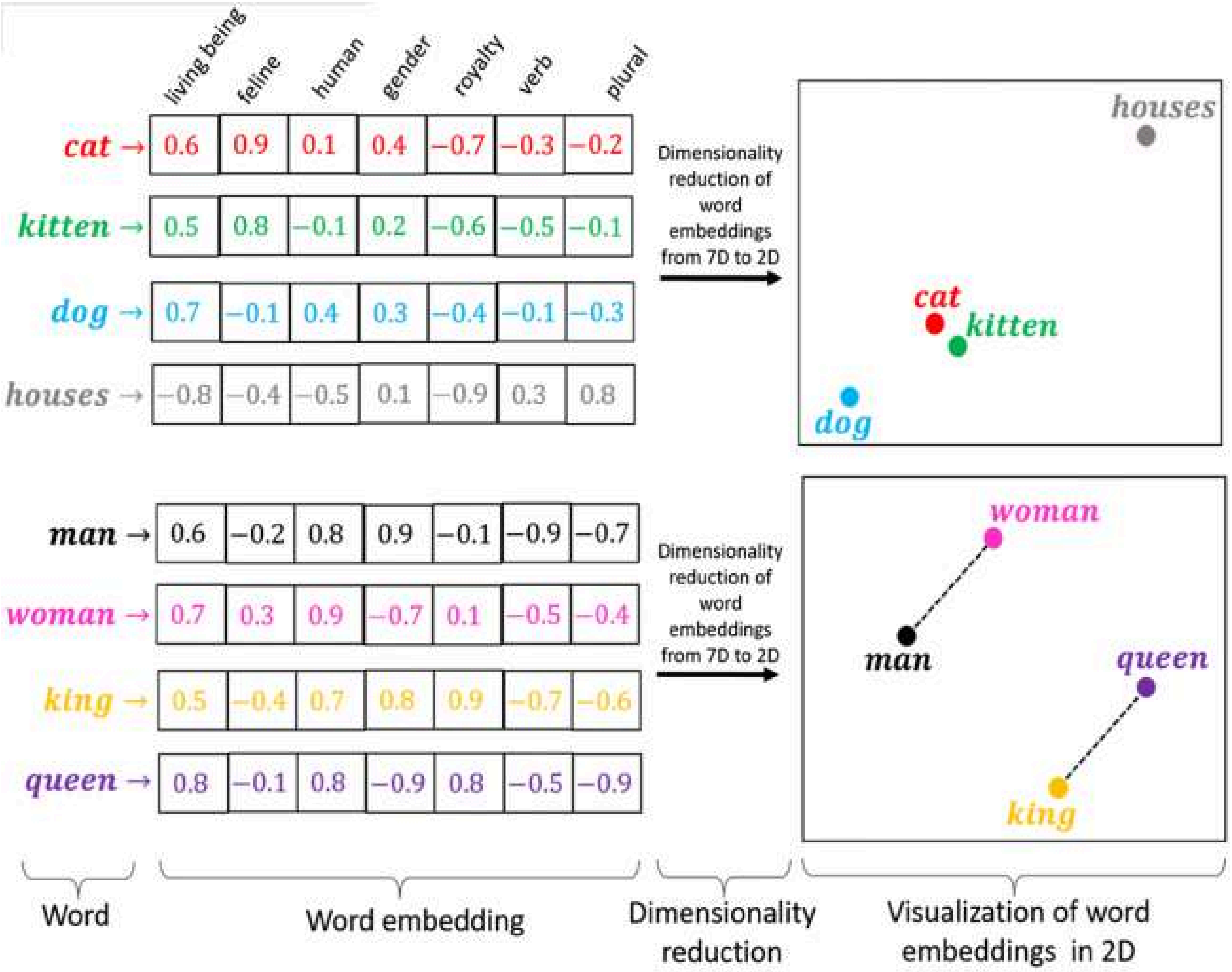


pea



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
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

Dimensionality reduction of word embeddings from 7D to 2D

Word

Word embedding

Dimensionality reduction

Visualization of word embeddings in 2D

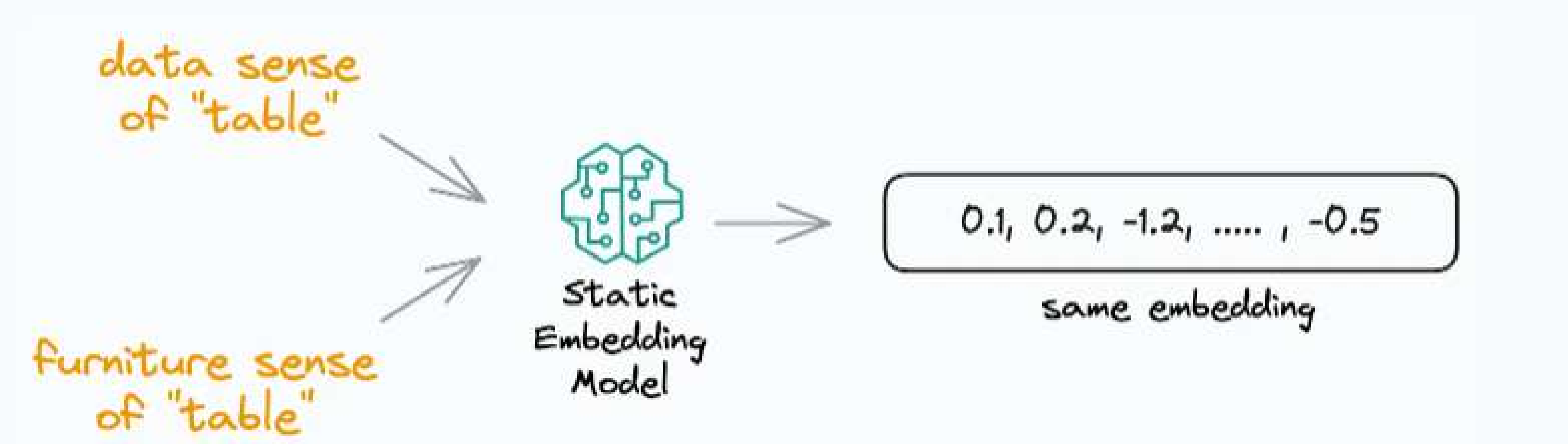
How can we get these word embedding vectors?

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.

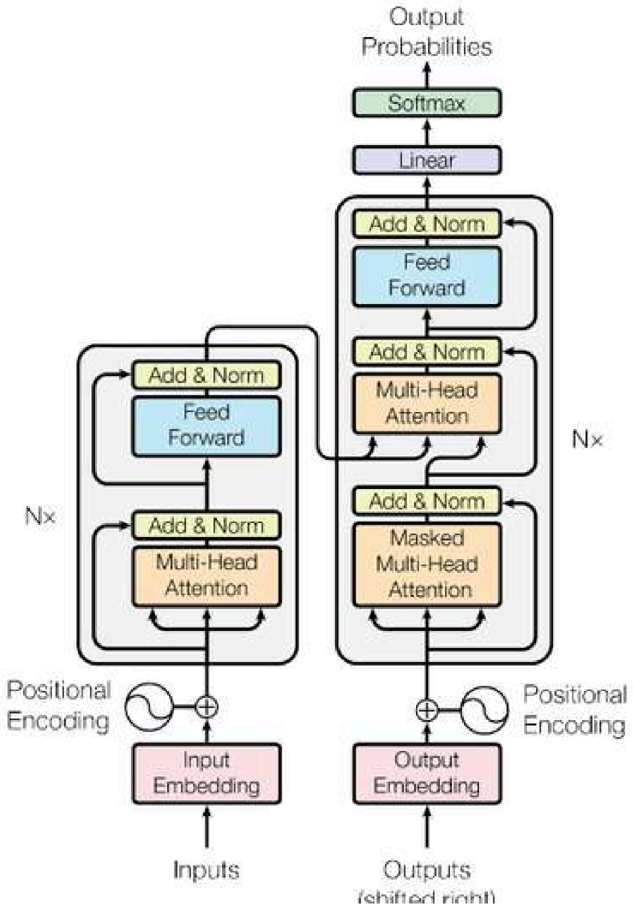


Solution ?

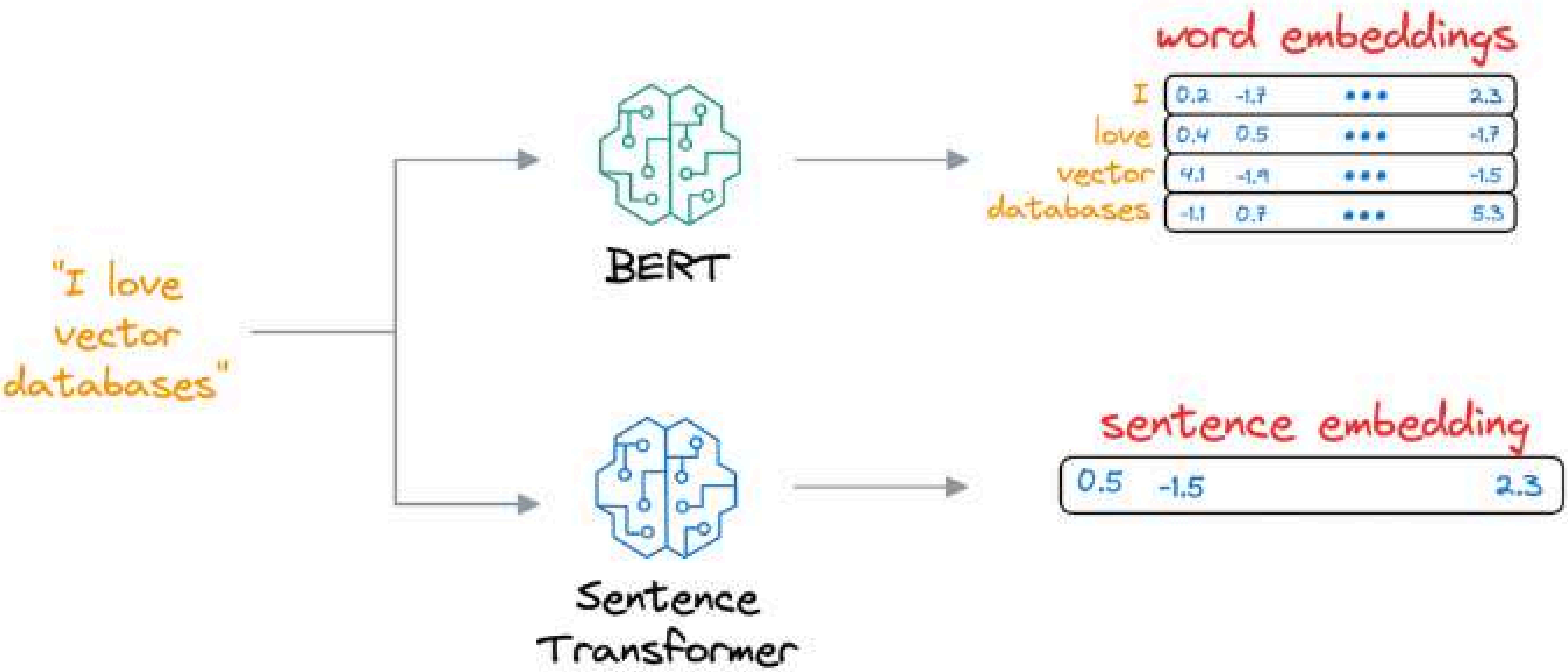
Transformer

- BERT (Bidirectional Encoder Representations from Transformers)

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



Solution ?



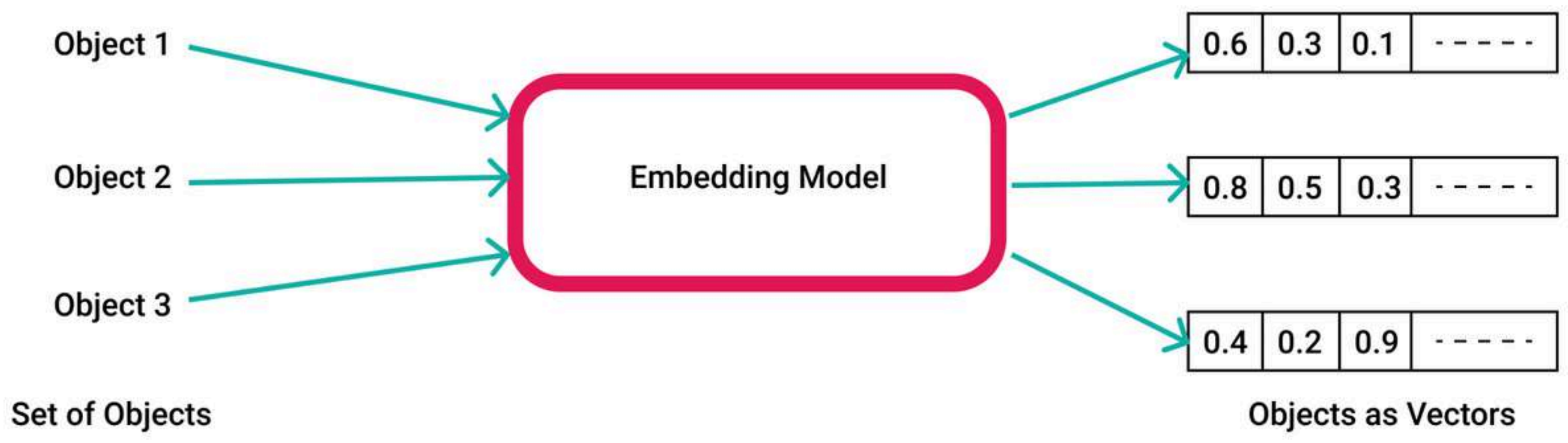
I	0.2	-1.7	•••	2.3
love	0.4	0.5	•••	-1.7
vector	4.1	-1.9	•••	-1.5
databases	-1.1	0.7	•••	5.3

0.5	-1.5	2.3

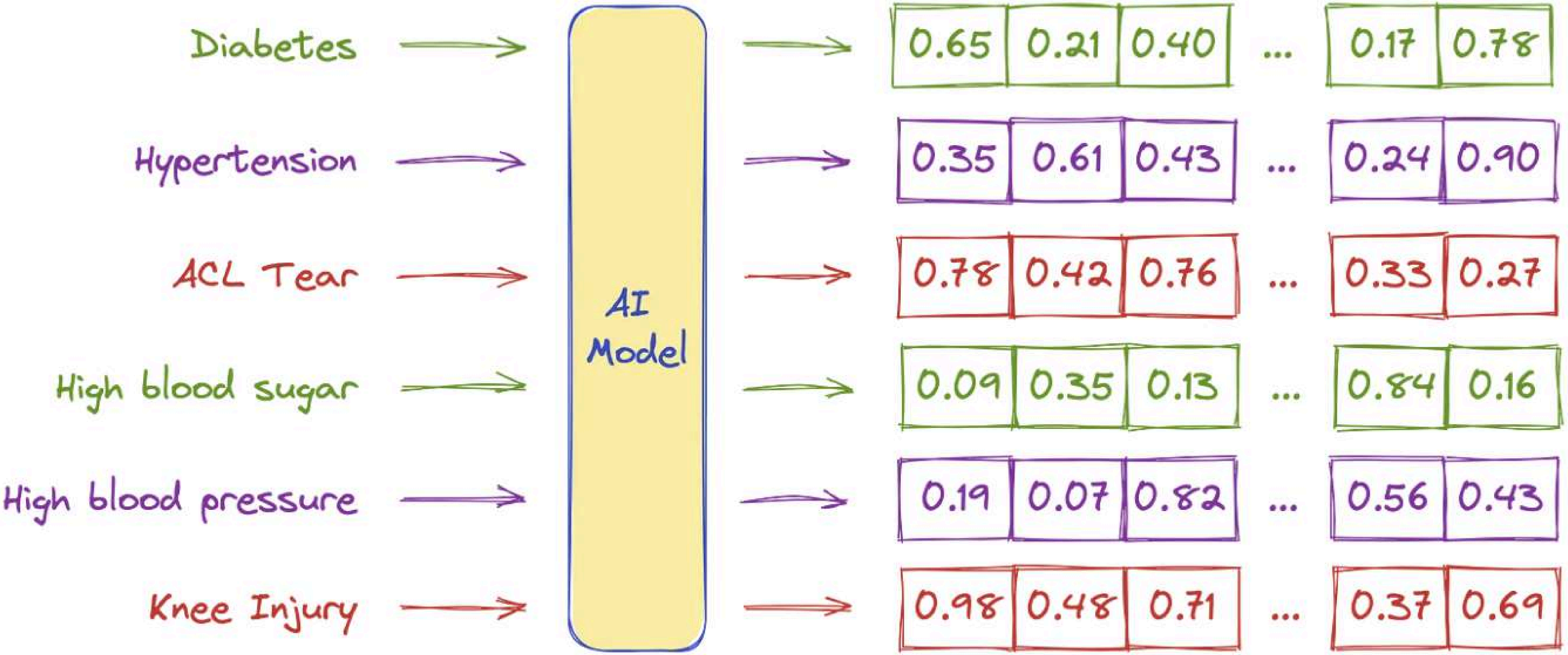
Embedding Model

- 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.

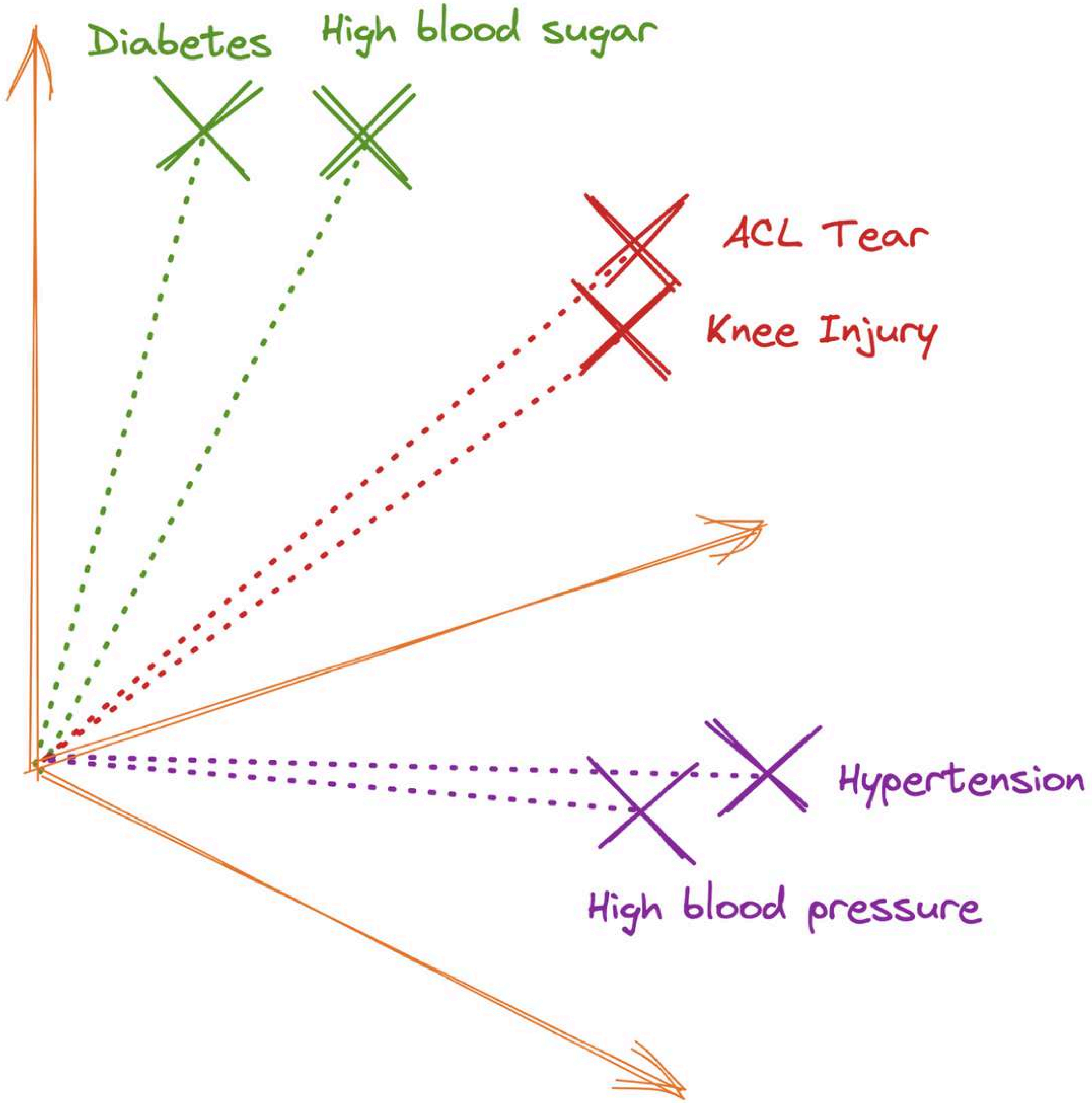


Embedding Space



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

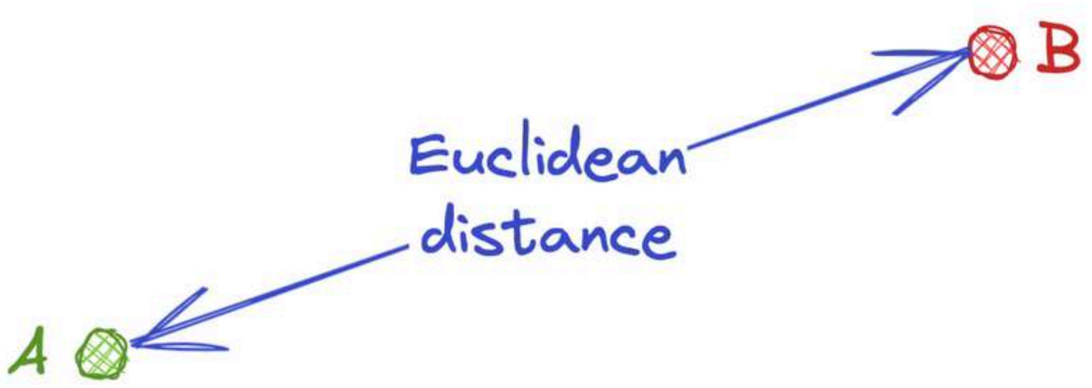
Embedding Space



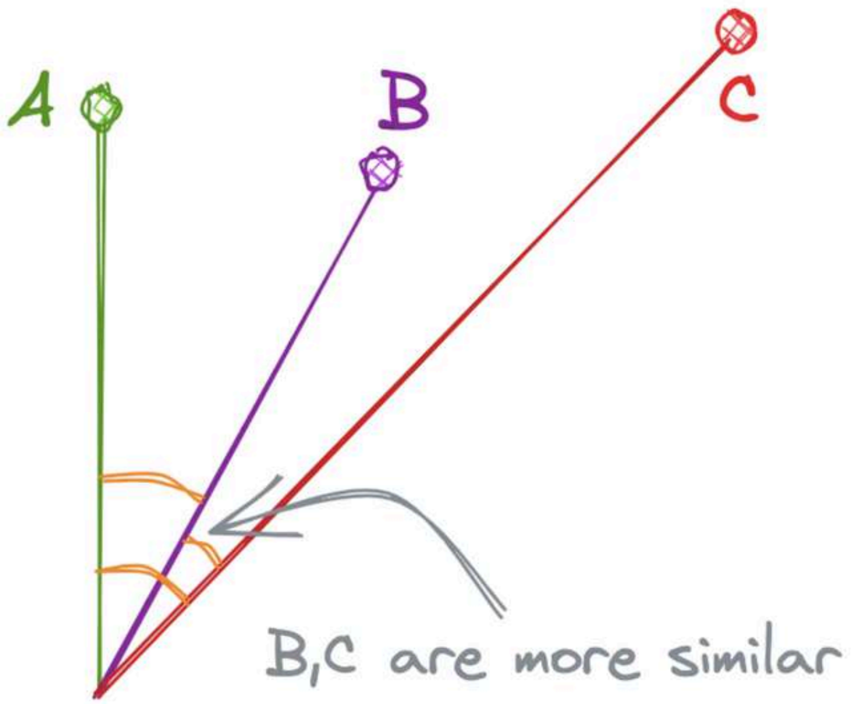
Querying a vector database

Similarity Calculation = > objective is to return the nearest neighbors

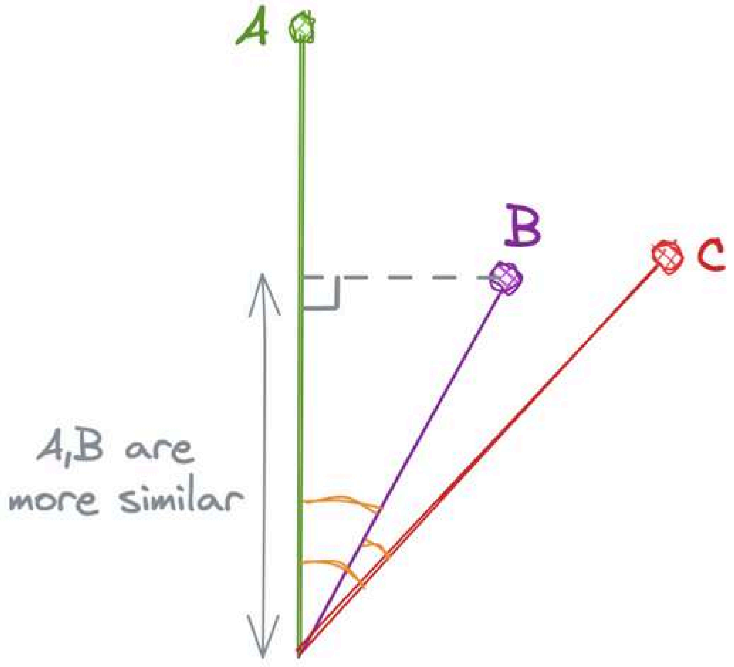
- For calculating similarity, there are several methods:
  - measuring distance - euclidean distance
  - cosine similarity or inner product



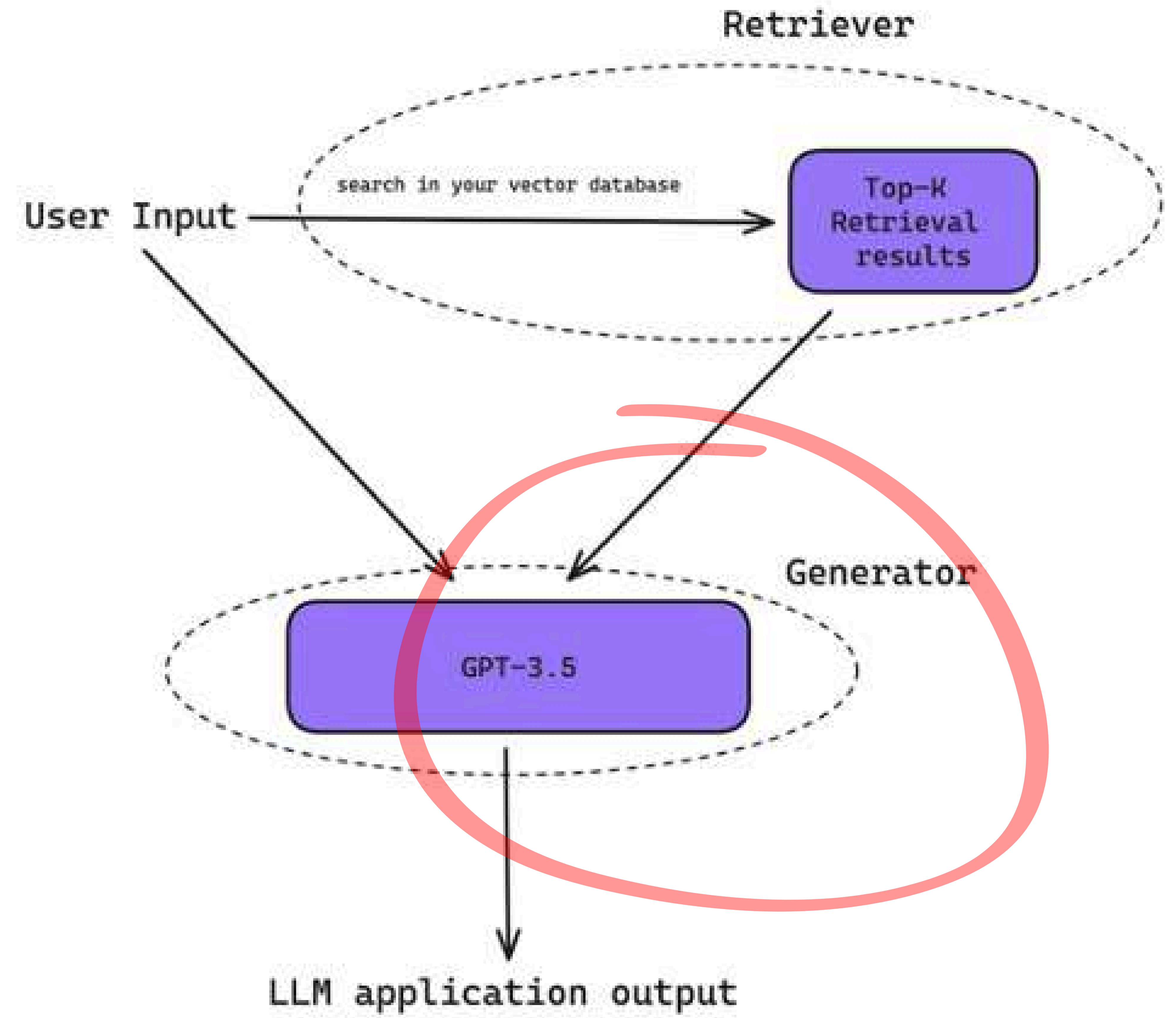
Euclidean distance



Cosine similarity



Dot Product


Retriever

search in your vector database

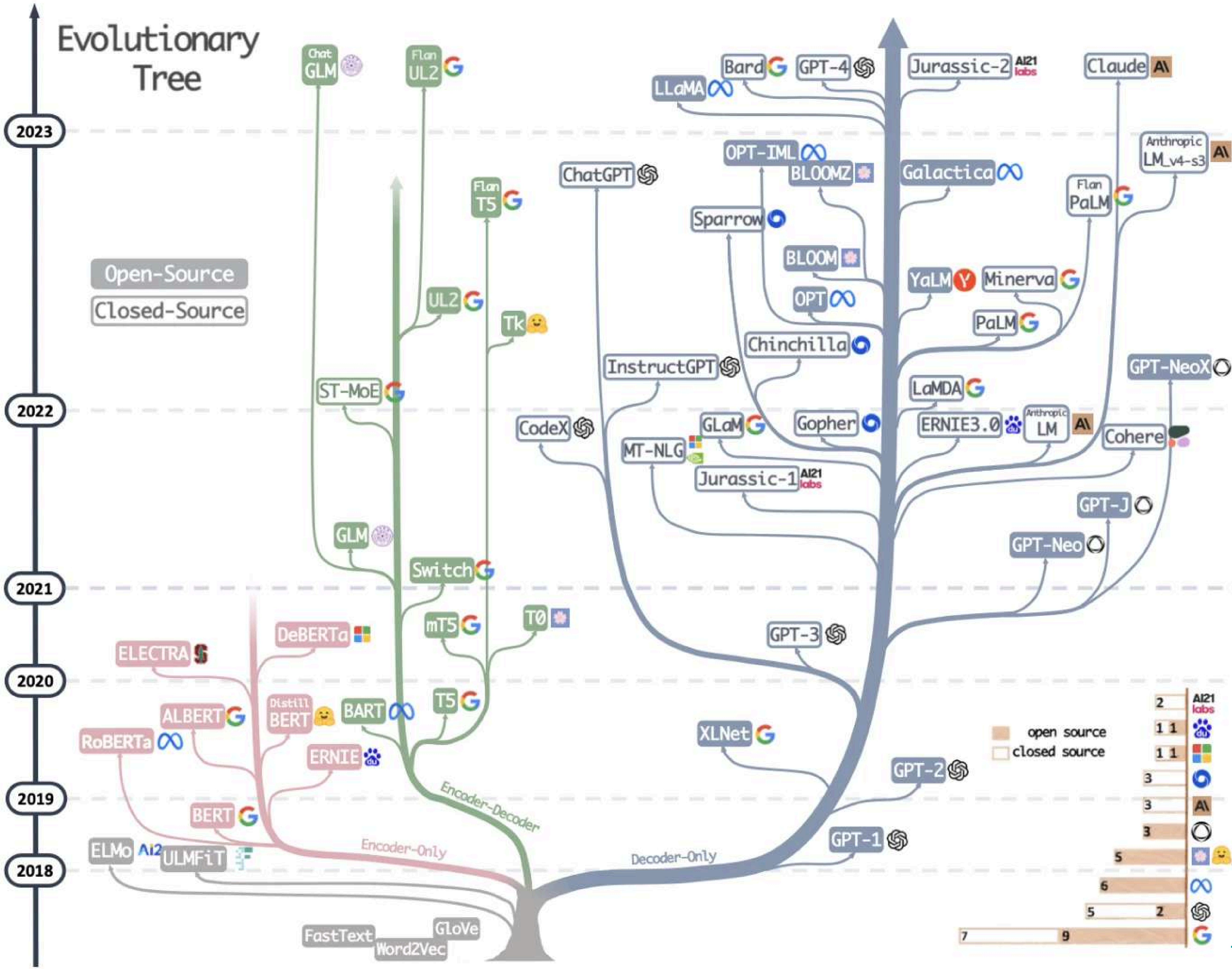
User Input

Top-K Retrieval results

Generator

GPT-3.5

LLM application output



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

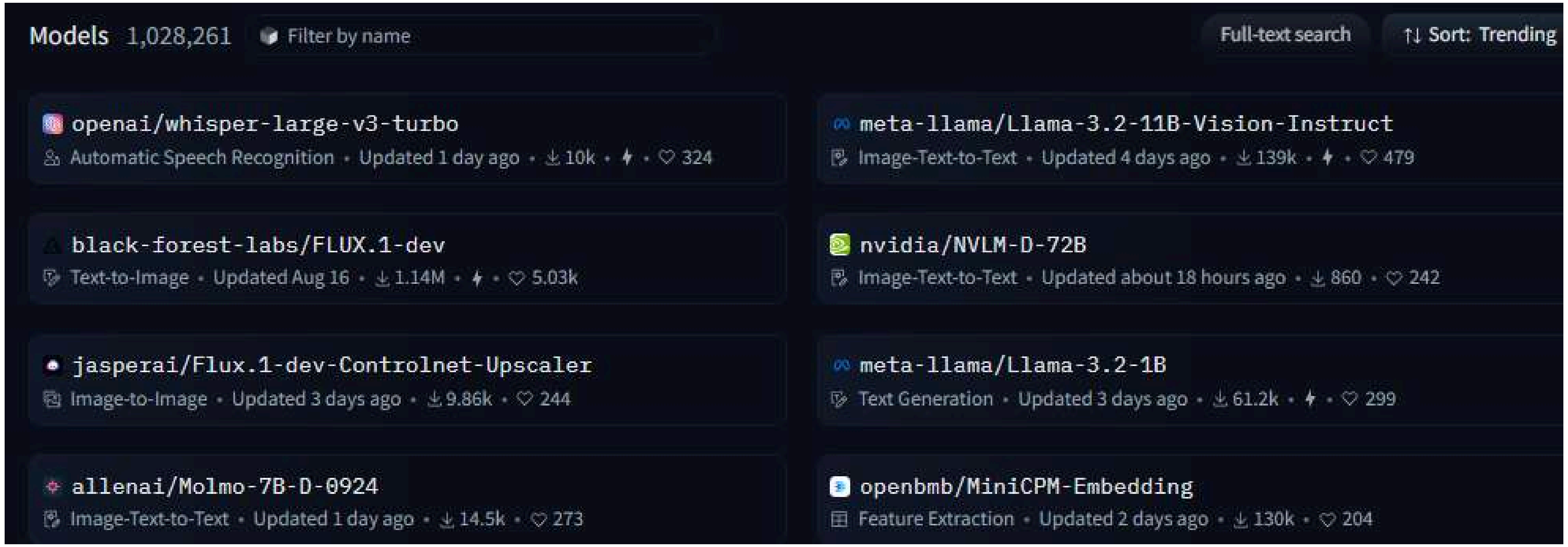
Where to find pretrained LLMs ?

Hugging Face

groq®

By Mohammed Arbi Nsibi

Where to find pretrained LLMs ?



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

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

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

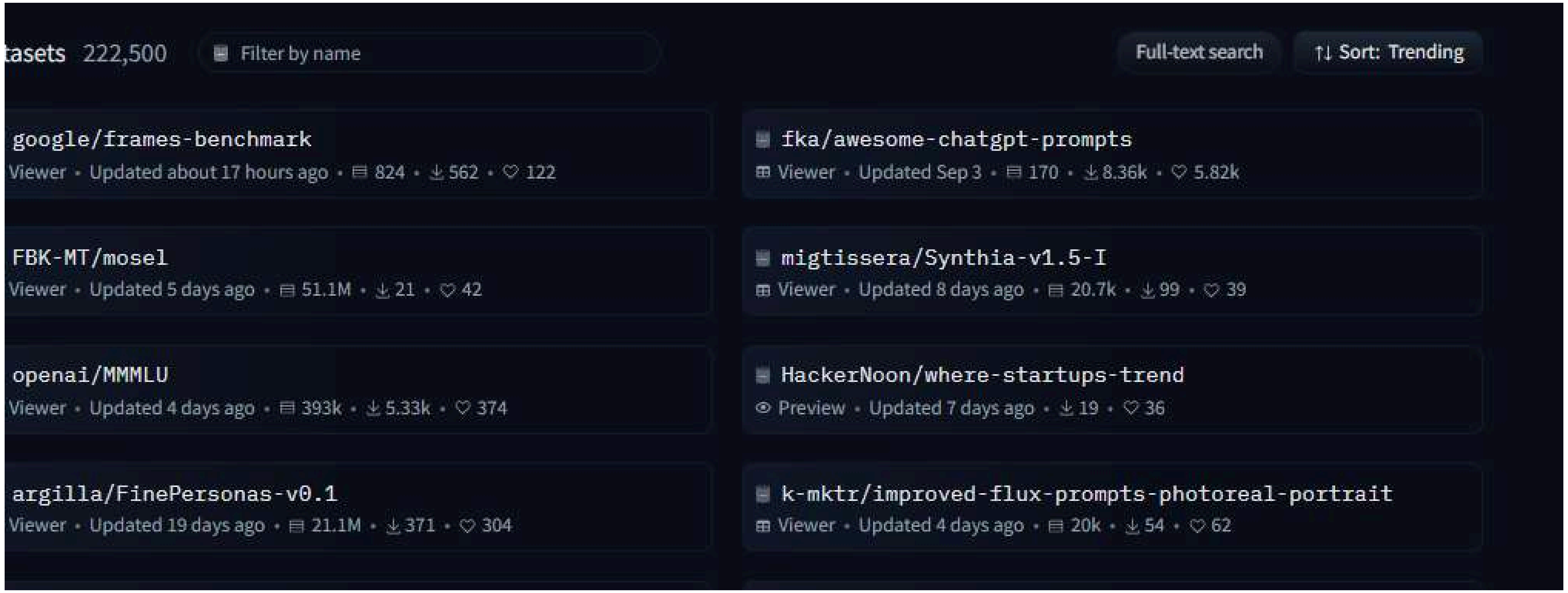
nvidia/NVLM-D-72B
Image-Text-to-Text • Updated about 18 hours ago • ⤓ 860 • ♡ 242

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

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

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

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



tasets 222,500 Filter by name Full-text search ⇅ Sort: Trending

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

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

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

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

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

HackerNoon/where-startups-trend
Preview • Updated 7 days ago • ⤓ 19 • ♡ 36

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

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

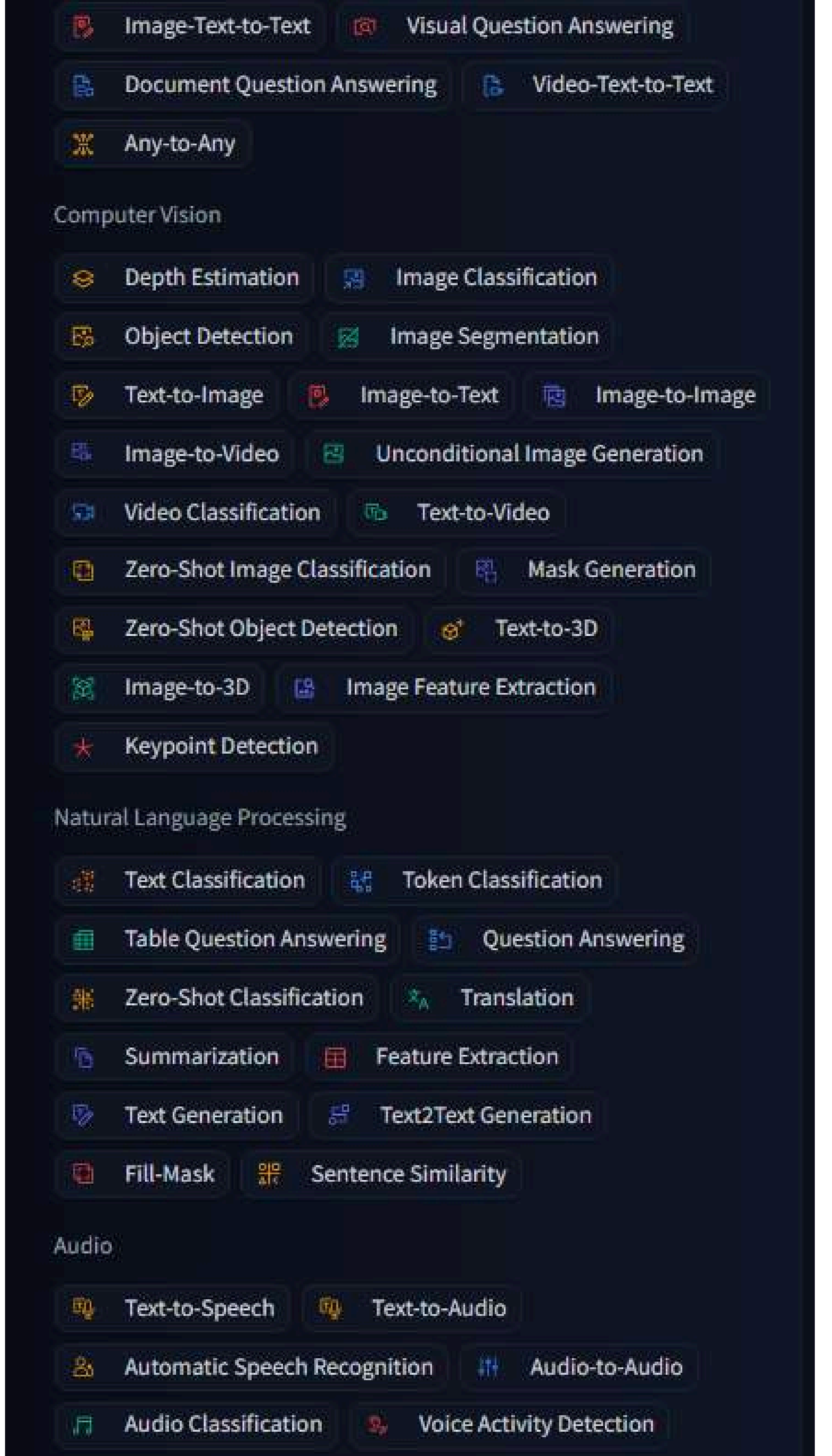


Image-Text-to-Text Visual Question Answering
Document Question Answering Video-Text-to-Text
Any-to-Any

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

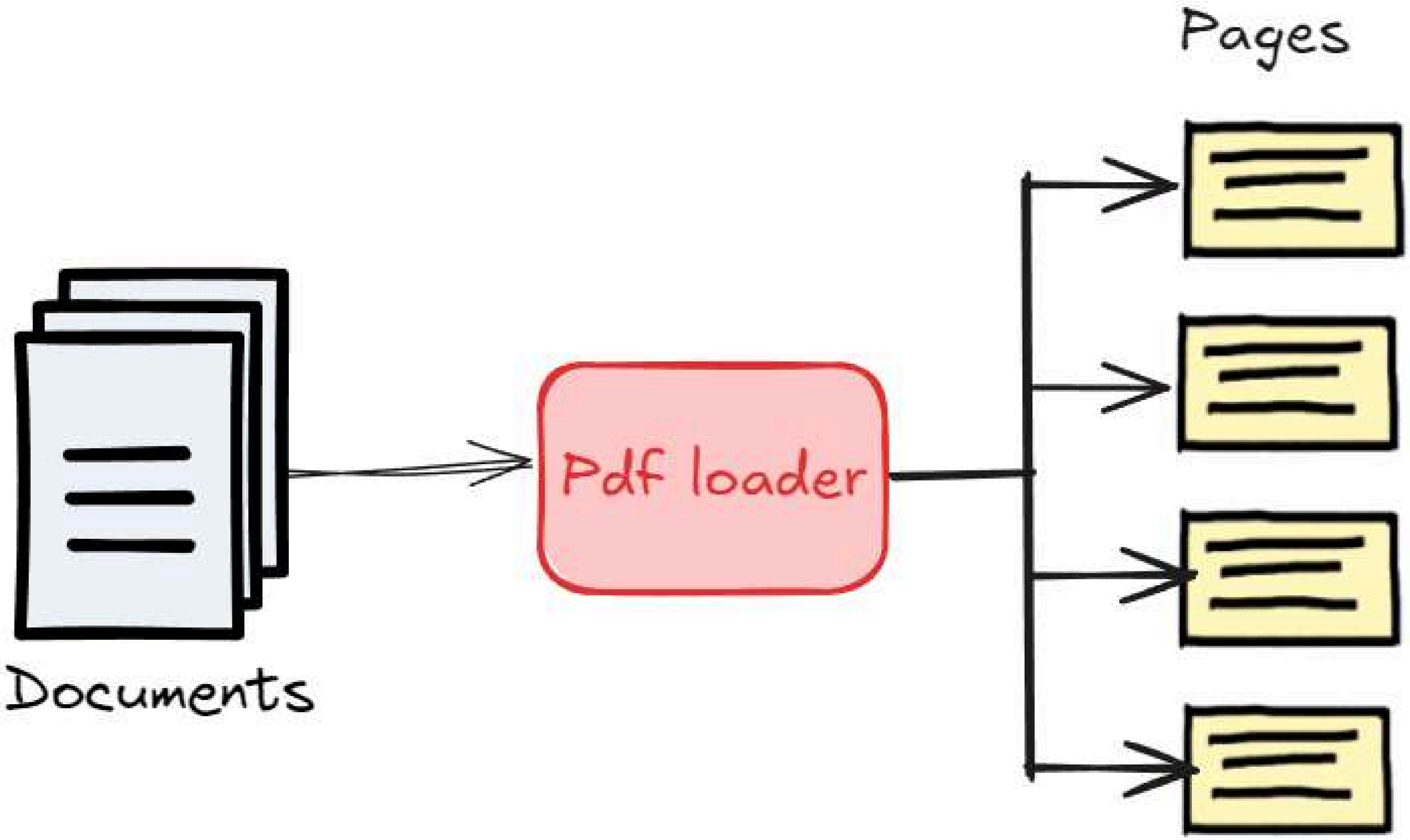
Natural Language Processing

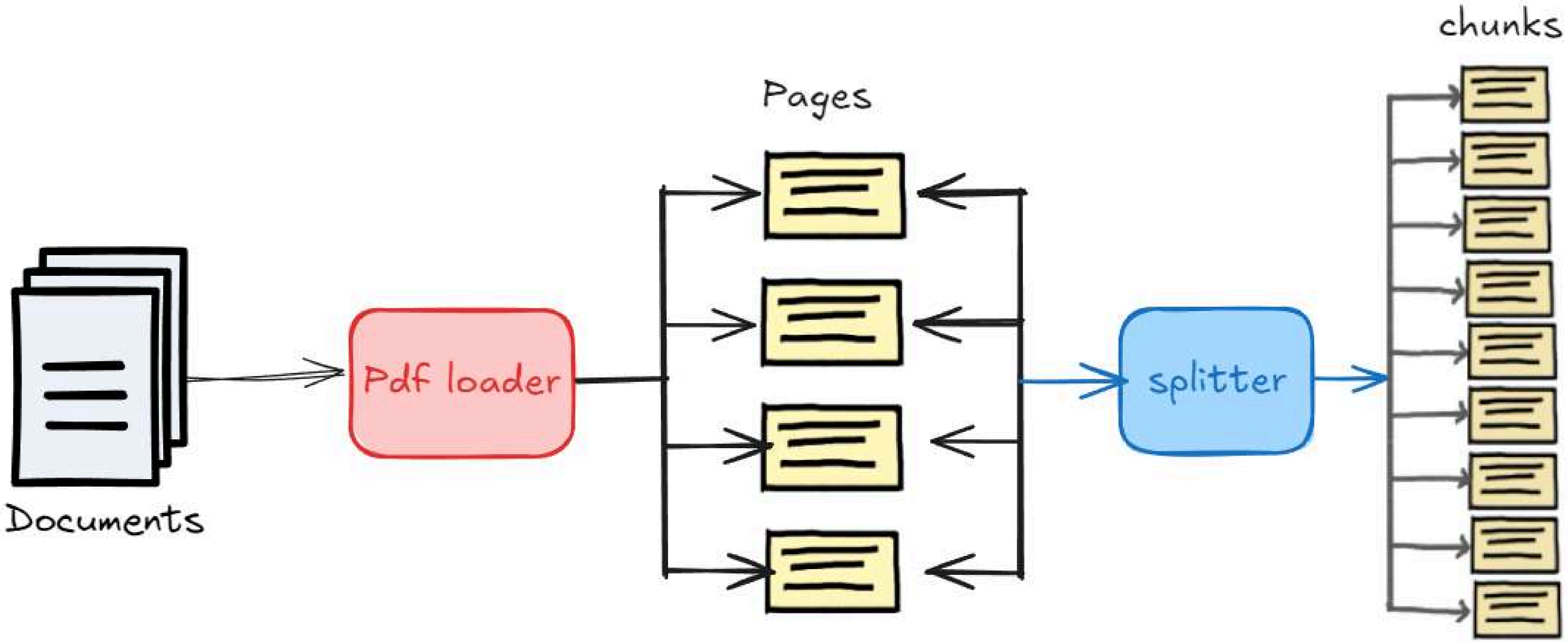
Text Classification Token Classification
Table Question Answering Question Answering
Zero-Shot Classification Translation
Summarization Feature Extraction
Text Generation Text2Text Generation
Fill-Mask Sentence Similarity

Audio

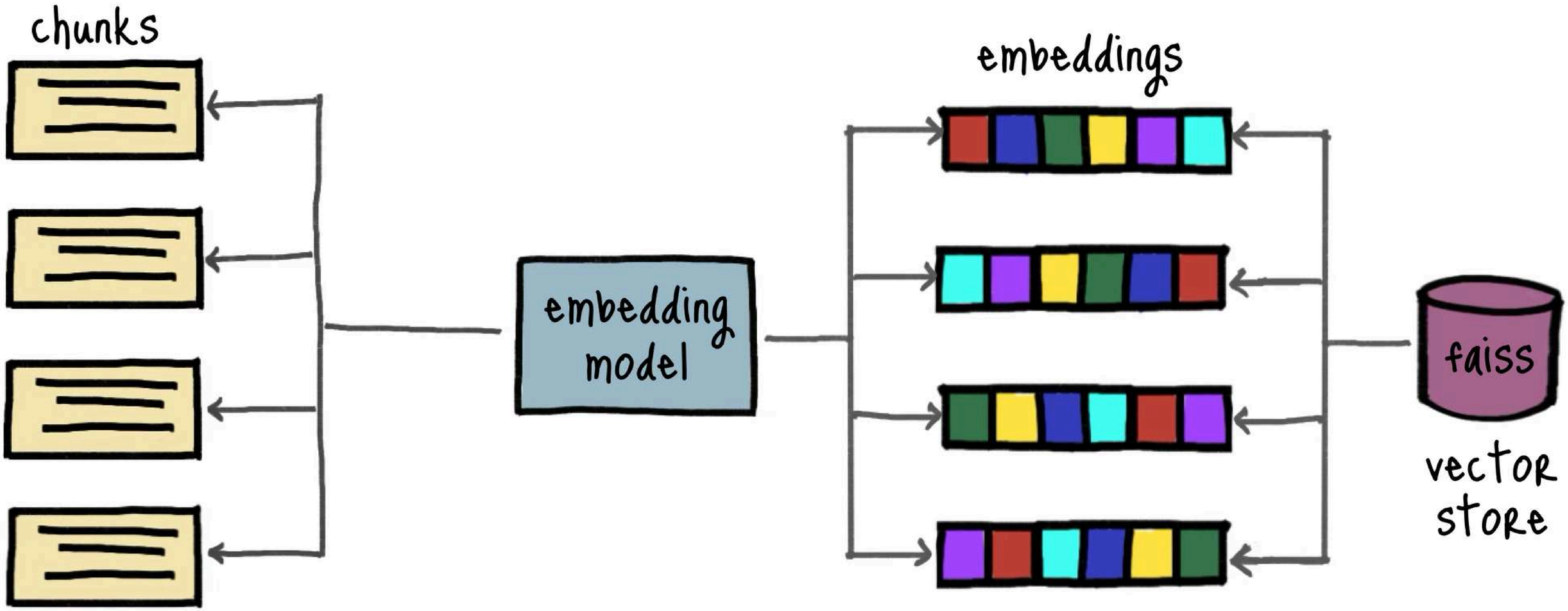
Text-to-Speech Text-to-Audio
Automatic Speech Recognition Audio-to-Audio
Audio Classification Voice Activity Detection

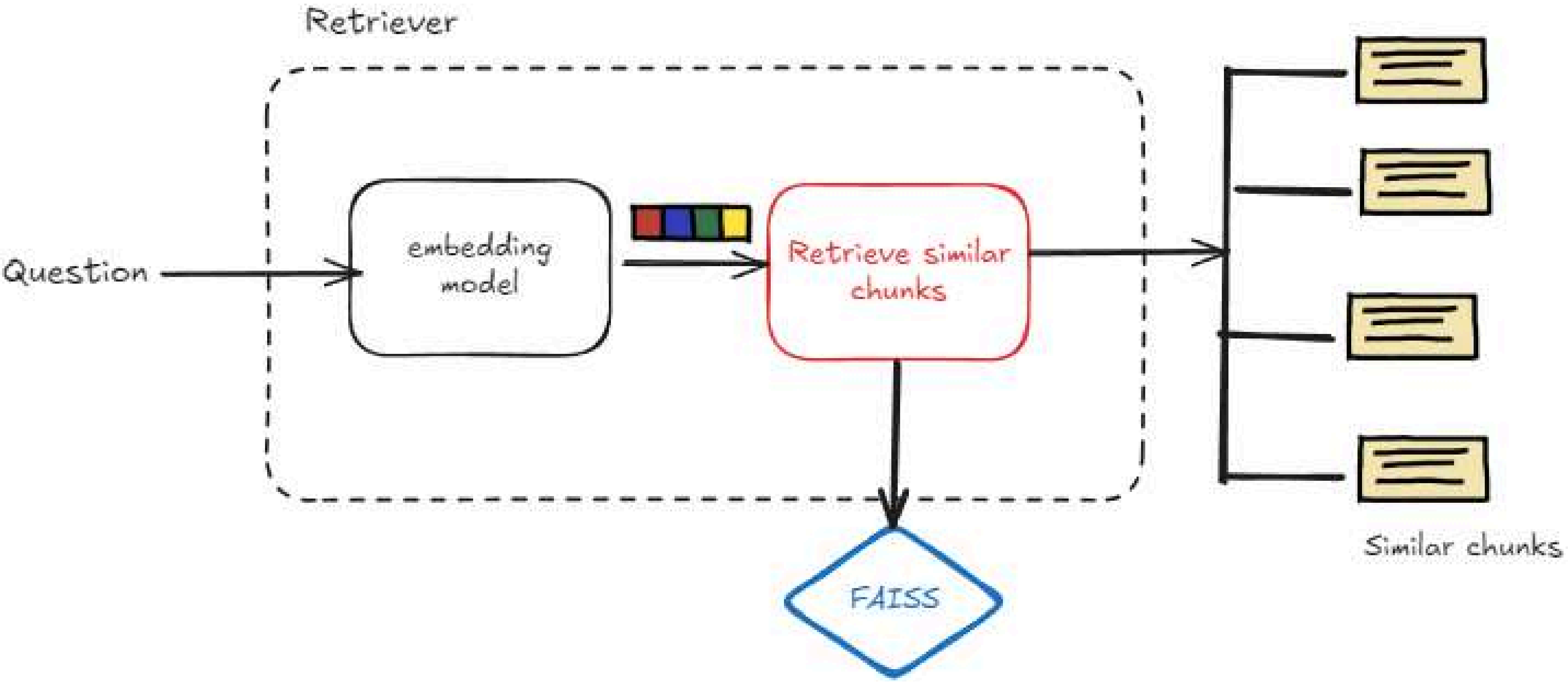
RAG architecture



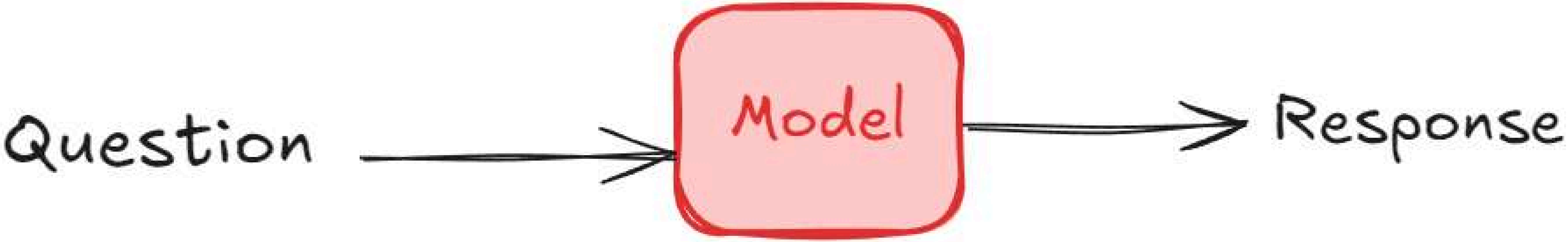


By Mohammed Arbi Nsibi


chunks
embedding model
embeddings
faiss
vector store



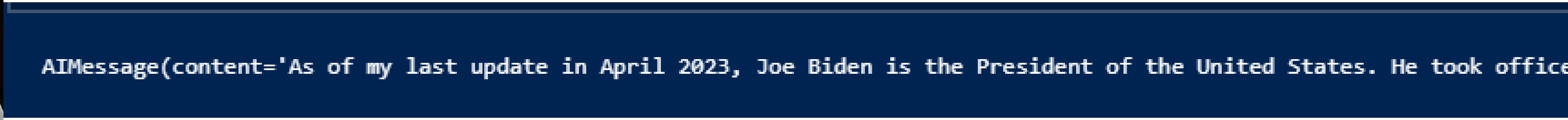
By Mohammed Arbi Nsibi



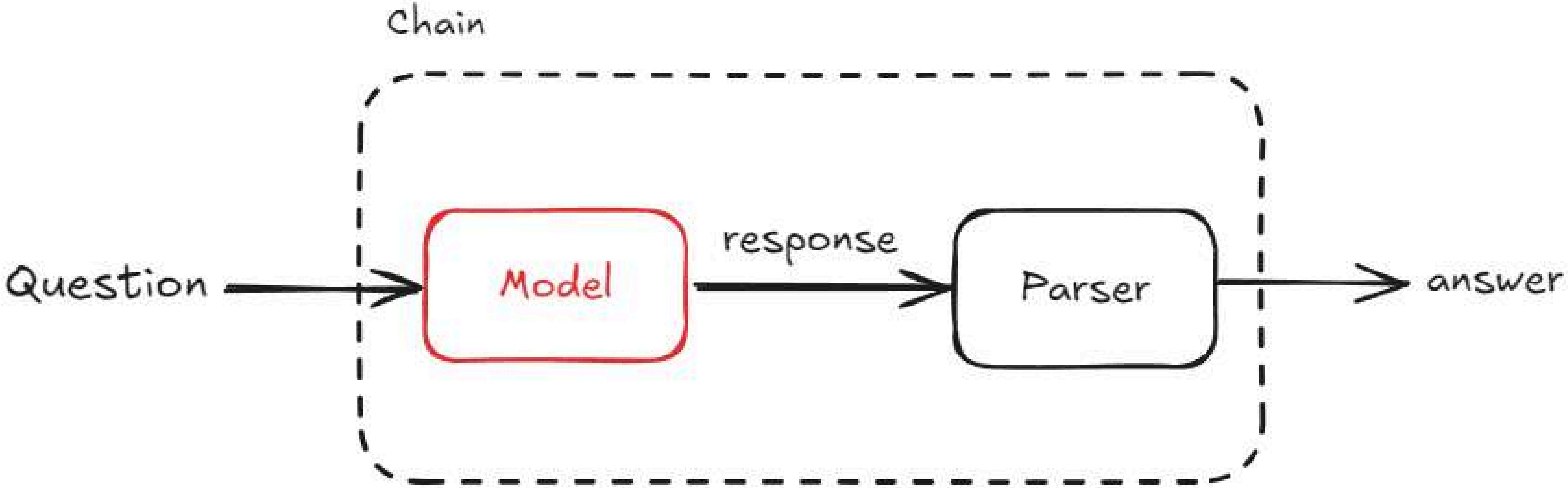
Question

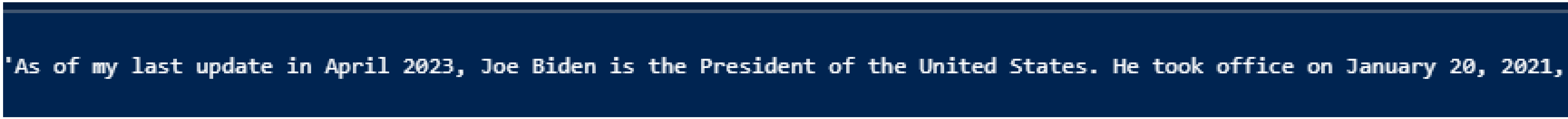
Model

Response

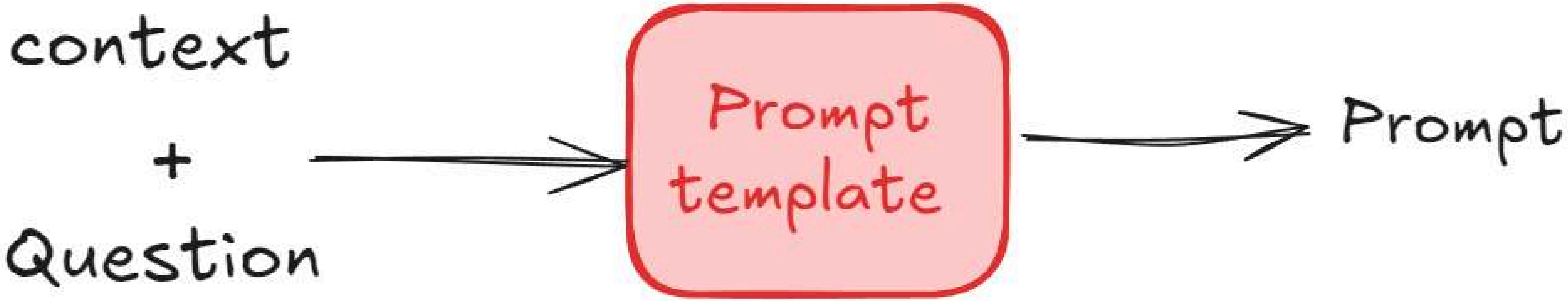


```
AIMessage(content='As of my last update in April 2023, Joe Biden is the President of the United States. He took office
```





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



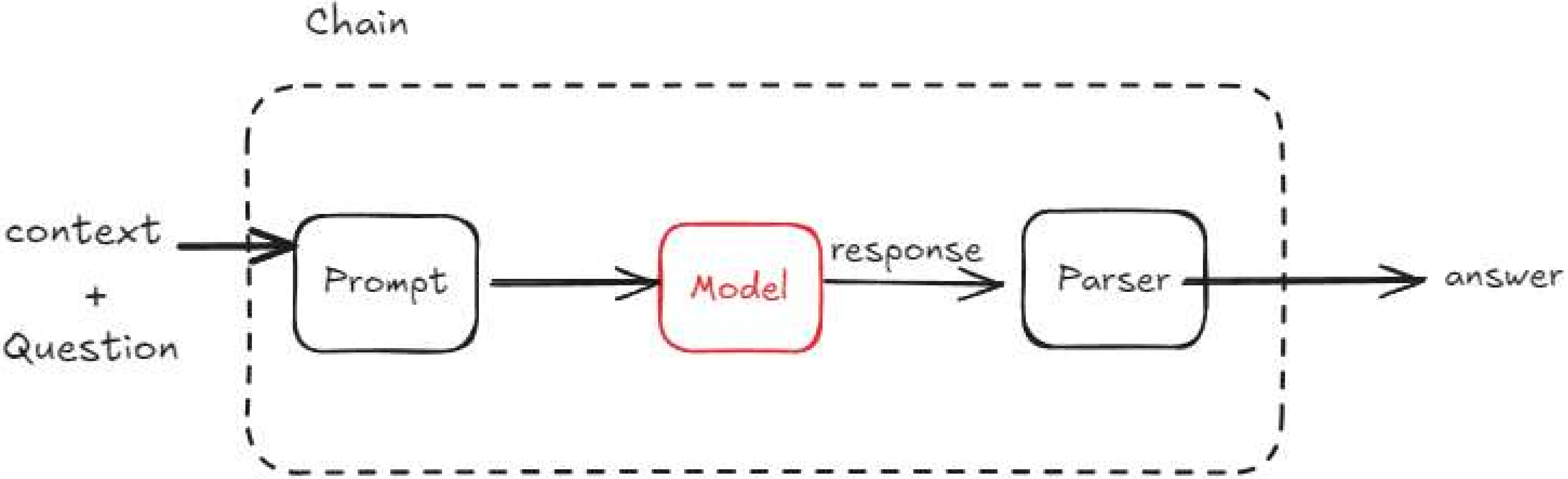
context

+

Question

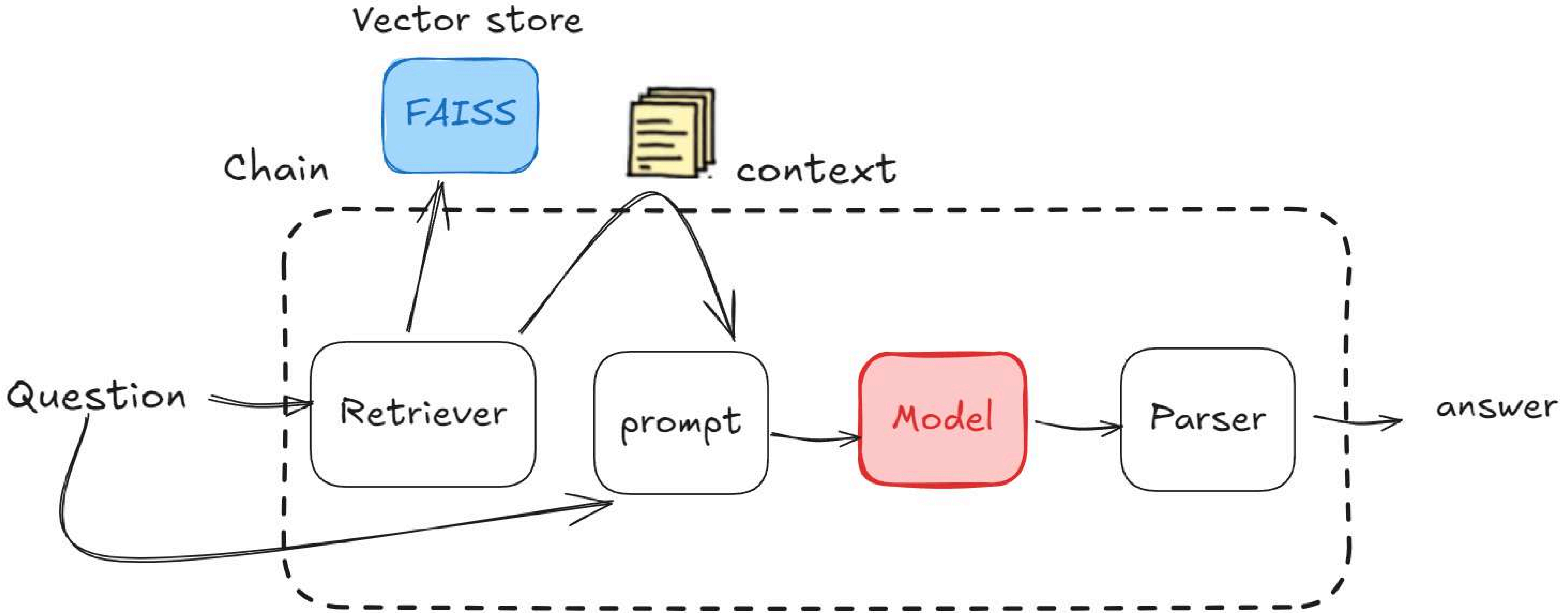
Prompt template

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.

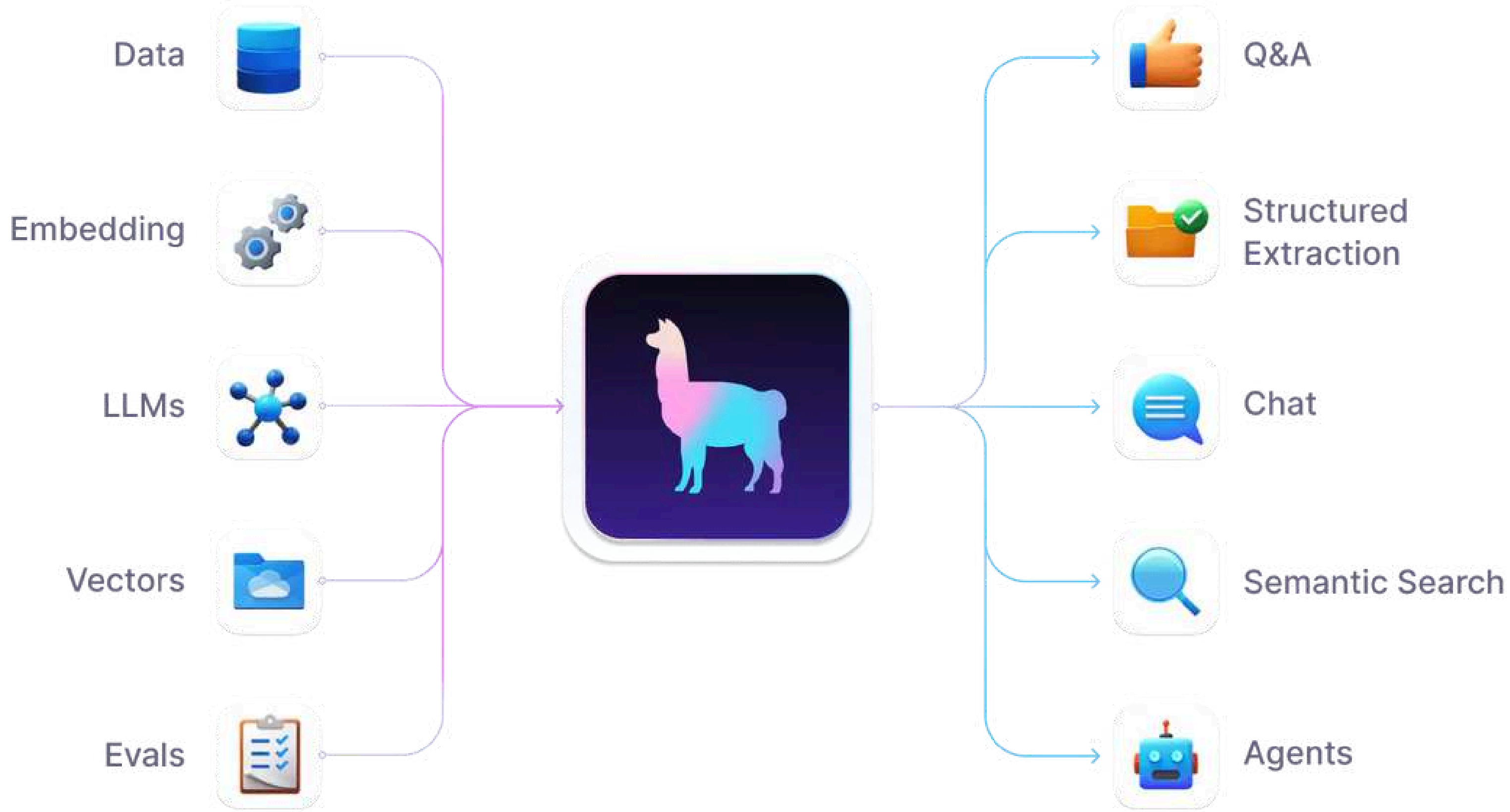
- It is based on LCEL (LangChain Expression Language)(build, compose, or manage sequences of operations)

- Use-cases including chatbots, RAG, document summarization and synthetic data generation.

🦜🔗 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

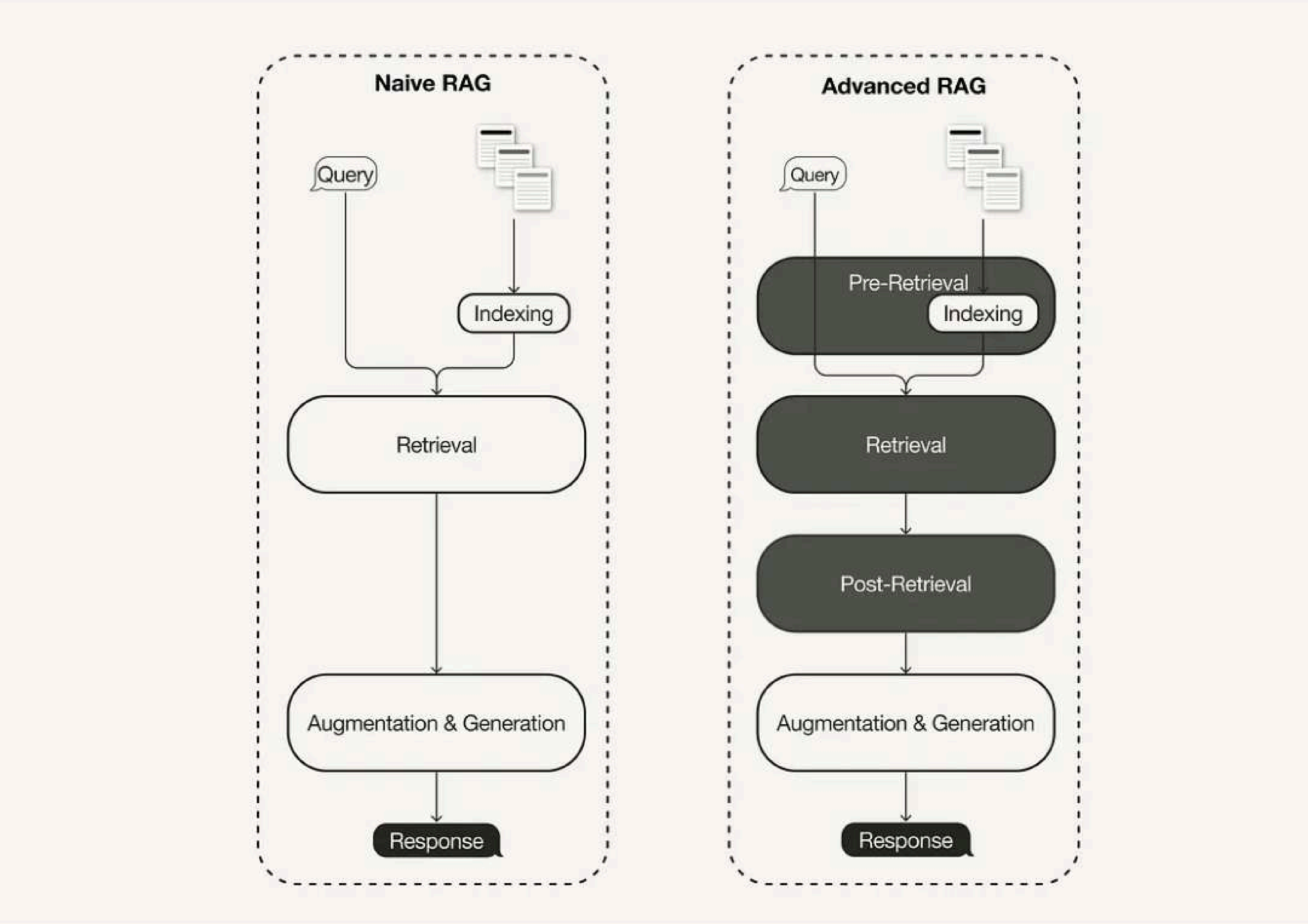
always has been
wait so creating chatbots is that easy ?

THEY SAID U CREATED A CHATBOT?
I DID
WHAT DID IT COST ?
NOTHING, IT'S FREE

Naive RAG vs Advanced RAG

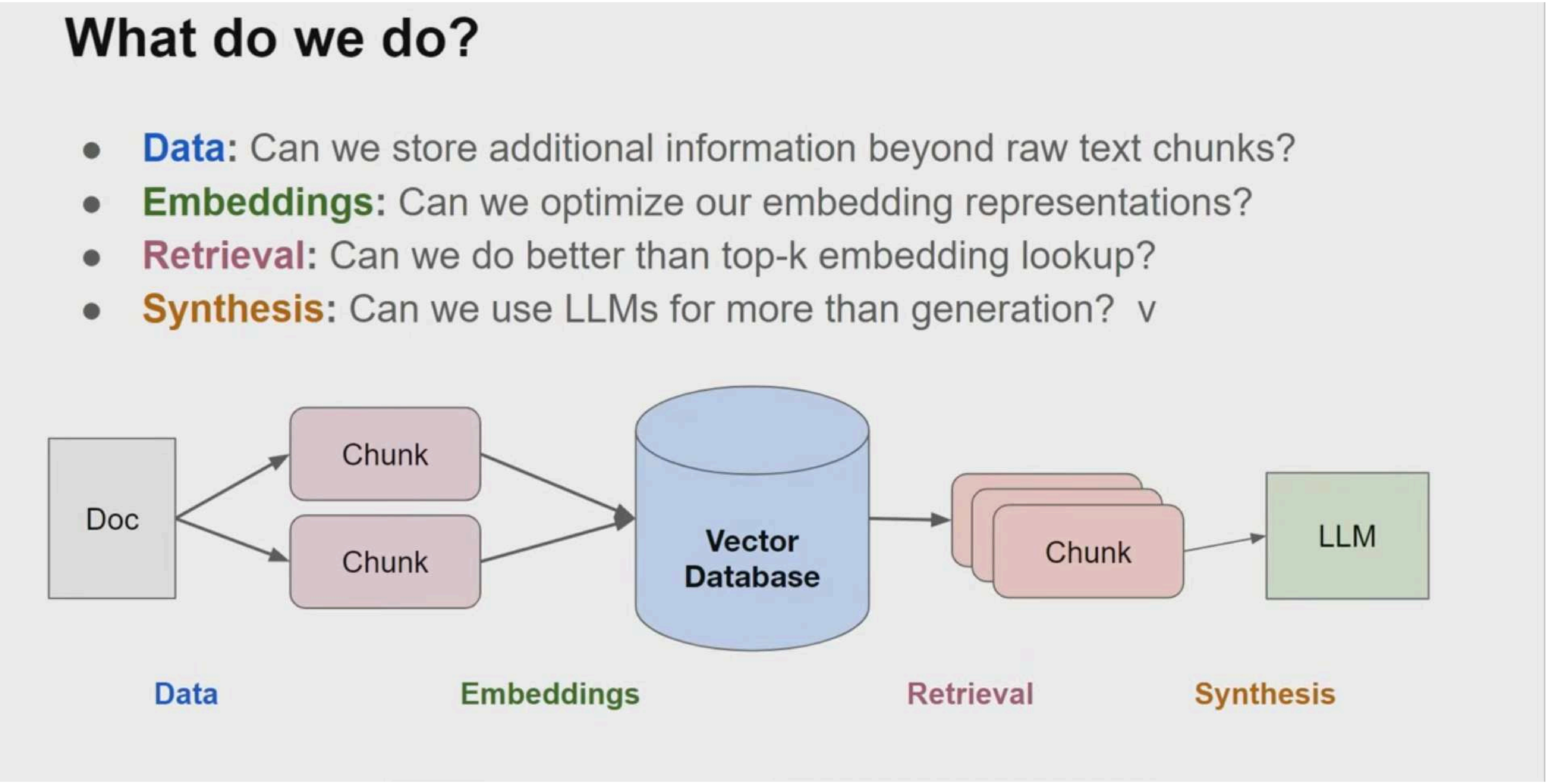
- There are many implementation to further improve performance of Naive RAG.
- Advanced RAG has evolved as a new paradigm with targeted enhancements to address some of the limitations of the naive RAG paradigm.

- Advanced RAG techniques can be categorized into
  - pre-retrieval optimization,
  - retrieval optimization, and
  - post-retrieval optimization



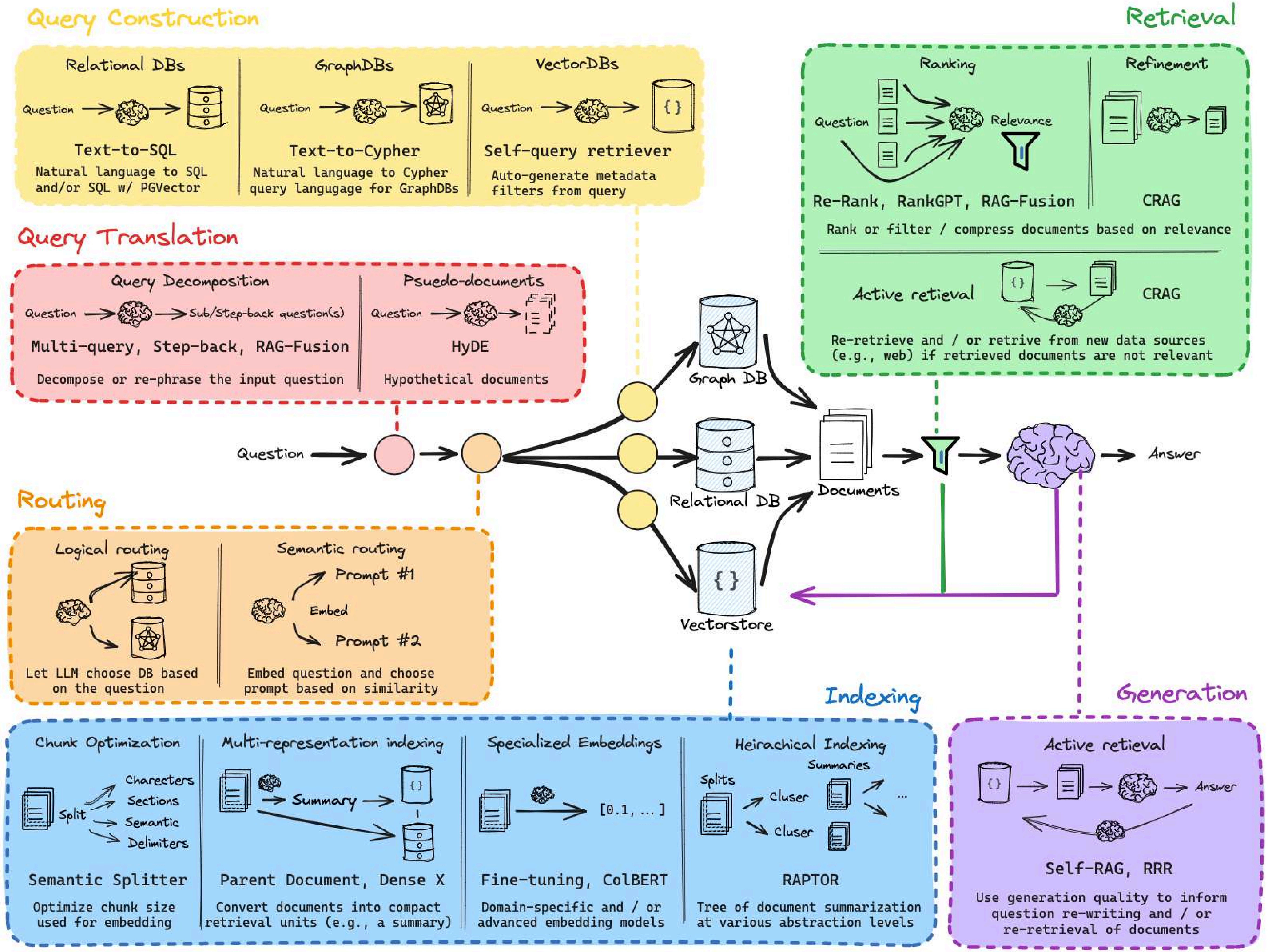
Difference between Naive and Advanced RAG (Image by the author, inspired by [1])

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? v



Text-to-SQL

Natural language to SQL and/or SQL w/ PGVector

Text-to-Cypher

Natural language to Cypher query langugage for GraphDBs

Self-query retriever

Auto-generate metadata filters from query

Multi-query, Step-back, RAG-Fusion

Decompose or re-phrase the input question

HyDE

Hypothetical documents

Let LLM choose DB based on the question

Embed question and choose prompt based on similarity

Re-Rank, RankGPT, RAG-Fusion

CRAG

Rank or filter / compress documents based on relevance

CRAG

Re-retrieve and / or retrive from new data sources (e.g., web) if retrieved documents are not relevant

Semantic Splitter

Optimize chunk size used for embedding

Parent Document, Dense X

Convert documents into compact retrieval units (e.g., a summary)

Fine-tuning, ColBERT

Domain-specific and / or advanced embedding models

RAPTOR

Tree of document summarization at various abstraction levels

Self-RAG, RRR

Use generation quality to inform question re-writing and / or re-retrieval of documents

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

THE WHITE HOUSE
WASHINGTON
LLM hallucinations are now illegal

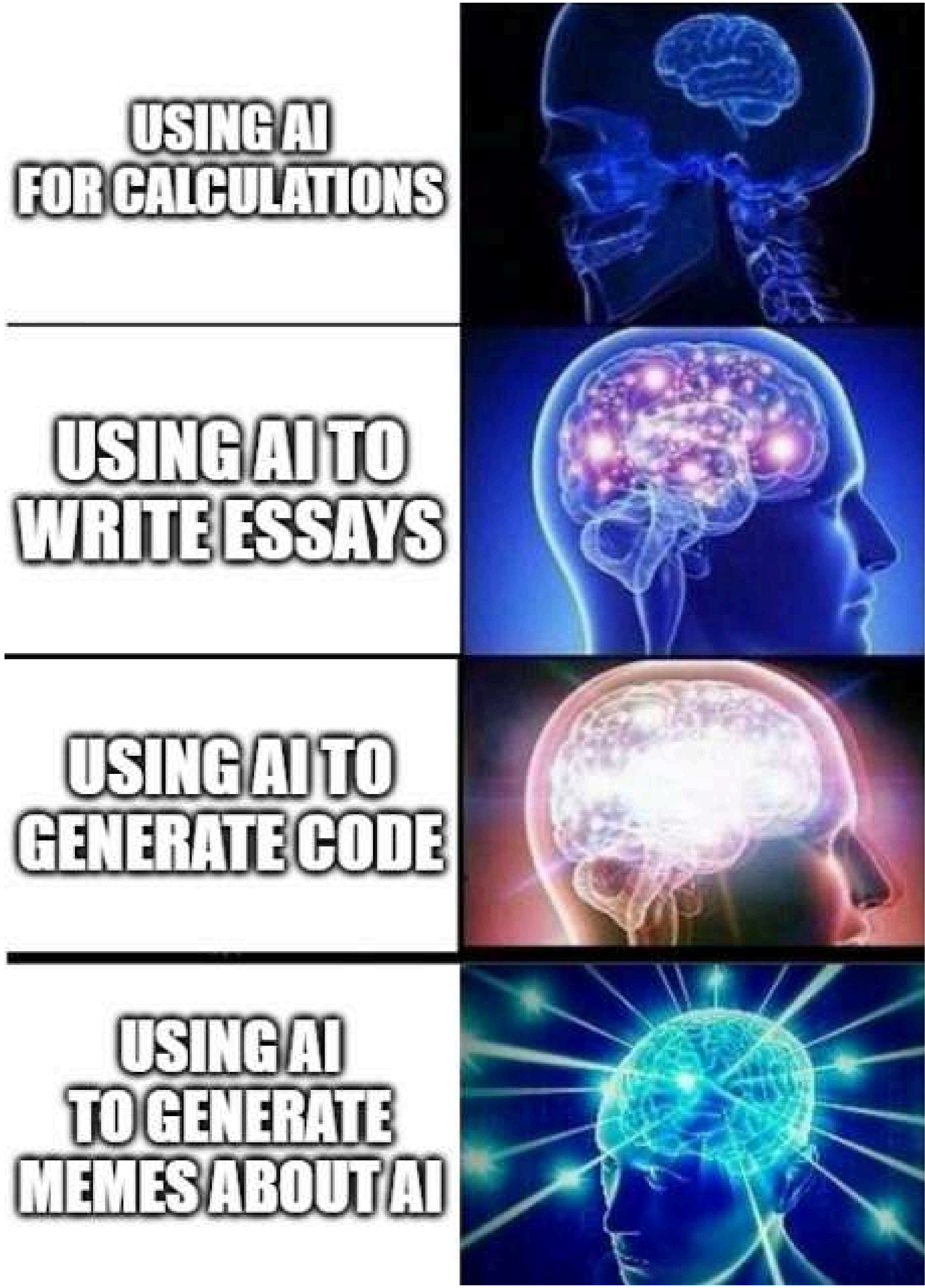
Resources

- 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!!


USING AI FOR CALCULATIONS
USING AI TO WRITE ESSAYS
USING AI TO GENERATE CODE
USING AI TO GENERATE MEMES ABOUT AI

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

QUIZ TIME 😁

By Mohammed Arbi Nsibi

😁