

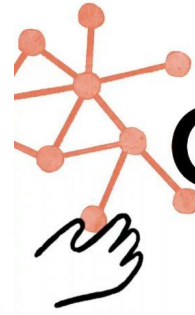
Retrieval Augmented Generation

Bach-Hoang Ngo

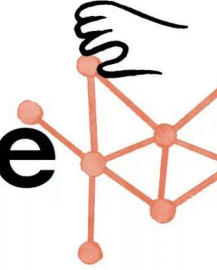
LLMs



OpenAI
GPT-4



Claude



MISTRAL
AI_



cohere



Gemini



Meta

Not up-to-date information

MODEL	DESCRIPTION	TRAINING DATA
gpt-3.5-turbo-0125	New Updated GPT 3.5 Turbo The latest GPT-3.5 Turbo model with higher accuracy at responding in requested formats and a fix for a bug which caused a text encoding issue for non-English language function calls. Returns a maximum of 4,096 output tokens. Learn more.	Up to Sep 2021
gpt-3.5-turbo	Currently points to gpt-3.5-turbo-0125.	Up to Sep 2021
gpt-3.5-turbo-1106	GPT-3.5 Turbo model with improved instruction following, JSON mode, reproducible outputs, parallel function calling, and more. Returns a maximum of 4,096 output tokens. Learn more.	Up to Sep 2021

MODEL	DESCRIPTION	TRAINING DATA
gpt-4-turbo	New GPT-4 Turbo with Vision The latest GPT-4 Turbo model with vision capabilities. Vision requests can now use JSON mode and function calling. Currently points to gpt-4-turbo-2024-04-09.	Up to Dec 2023
gpt-4-turbo-2024-04-09	GPT-4 Turbo with Vision model. Vision requests can now use JSON mode and function calling. gpt-4-turbo currently points to this version.	Up to Dec 2023
gpt-4-turbo-preview	GPT-4 Turbo preview model. Currently points to gpt-4-0125-preview.	Up to Dec 2023

Not up-to-date information

You
Who are the creators of Mamba network?

ChatGPT
The Mamba network was created by Andrey Andreev, a Russian entrepreneur who founded the company Mamba in 2003. Mamba is a popular online dating and social networking platform, particularly in Russia and Eastern Europe.

You
What is the current SOTA method for Image Classification?

ChatGPT
As of my last update in January 2022, convolutional neural networks (CNNs) were the state-of-the-art (SOTA) method for image classification tasks. Architectures like ResNet, DenseNet, and EfficientNet were among the most prominent CNN models achieving excellent performance on various image classification benchmarks such as ImageNet.

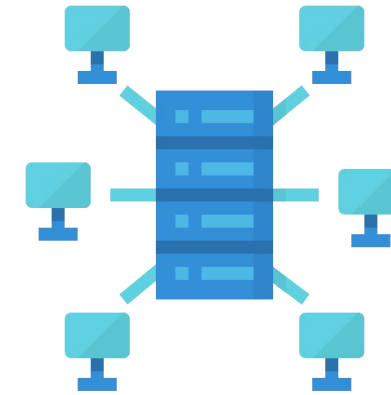
You
How many CVPR papers published in 2024

ChatGPT
I can't provide real-time data, including the number of papers published at a specific conference in 2024, as my training only includes knowledge up to January 2022. However, you can typically find this information by visiting the official website of the Conference on Computer Vision and Pattern Recognition (CVPR) or checking academic databases like IEEE Xplore or arXiv, which often host proceedings of such conferences. They usually provide information on the number of papers accepted and published for each year's conference.

Private data?



Personal data



Enterprise data



Confidential data

Retraining with more data?

Training Llama 2



2048 x

NVIDIA A100

		Time (GPU hours)	Power Consumption (W)	Carbon Emitted (tCO ₂ eq)
LLAMA 2	7B	184320	400	31.22
	13B	368640	400	62.44
	34B	1038336	350	153.90
	70B	1720320	400	291.42
Total		3311616		539.00

1.29 -> 4.10 \$ / 1h

Llama 2 – 7B ~ 240k \$

Llama 2 – 70B ~ 2.2M \$

- Risk of worse performance.
- Cannot update in real-time.
- Carbon Emission.
- Hallucination

How does the Mamba network work?

Retraining with more data?

Llama 3 numbers

	Time (GPU hours)	Power Consumption (W)	Carbon Emitted(tCO2eq)
Llama 3 8B	1.3M	700	390
Llama 3 70B	6.4M	700	1900
Total	7.7M		2290

1.29 -> 4.10 \$ / 1h

Llama 2 – 7B ~ 1.7M \$

Llama 2 – 70B ~ 8.3M \$

Hallucination

AI

You

Truyện "Dế mèn phiêu lưu ký" kể về gì



ChatGPT

"Truyện Dế Mèn Phiêu Lưu Ký" là một tác phẩm văn học nổi tiếng của nhà văn Tô Hoài. Câu chuyện xoay quanh cuộc phiêu lưu của chú dế mèn, một con vật nhỏ bé, thông minh và tinh nghịch.

Chú dế mèn bị bắt vào nhà của một người làm công việc tại cơ quan một cửa hàng kinh doanh. Tại đây, chú dế mèn gặp gỡ và kết bạn với nhiều loại động vật khác, từ chuột, bò sát đến các loại chim. Qua những cuộc trò chuyện và kinh nghiệm giao tiếp, chú dế mèn học được nhiều điều về cuộc sống, tình bạn và lòng hiếu thảo.

Cuối cùng, chú dế mèn quyết định rời khỏi nhà cửa hàng để đến một nơi mới, nơi mà anh có thể tiếp tục cuộc phiêu lưu của mình và khám phá thế giới bên ngoài. "Dế Mèn Phiêu Lưu Ký" không chỉ là một câu chuyện giải trí cho trẻ em mà còn là một tác phẩm văn học mang tính nhân văn sâu sắc, với thông điệp về tình bạn, lòng can đảm và sự khám phá cuộc sống.



AI

You

Kể tôi nghe chuyện Tấm Cám



ChatGPT

Tấm Cám là một câu chuyện dân gian nổi tiếng của Việt Nam, thường được coi là phiên bản Việt Nam của cổ tích Cinderella. Câu chuyện kể về hai chị em riêng biệt nguồn gốc: Tấm và Cám. Tấm là cô em gái hiền lành, thông minh và tốt bụng, trong khi Cám thì xấu xí, xảo trá, và ghen tuông với Tấm.

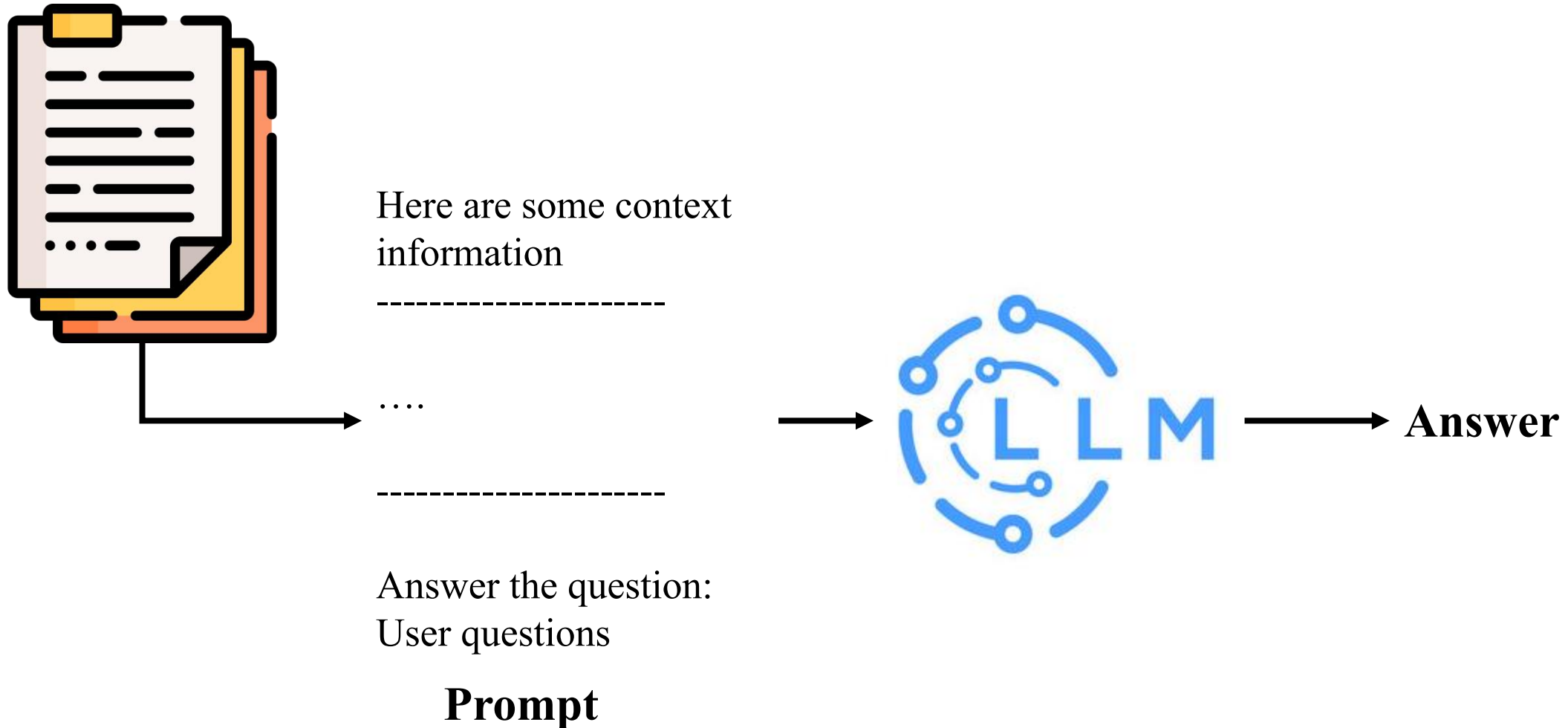
Câu chuyện bắt đầu với việc mẹ của Tấm qua đời và để lại cho cô một miếng vải. Trước khi qua đời, mẹ Tấm đã nhờ Tấm giữ miếng vải và dùng nó để gọi linh hồn của mình trở lại. Mẹ của Cám lại ghen tị và cố gắng cướp miếng vải đó. Tấm buộc phải đưa miếng vải cho Cám để tránh sự ghen tuông của nàng.

Sau đó, Cám cưới vương quốc và trở thành một phu nhân xấu xí và tàn ác. Trong khi đó, Tấm phải sống dưới mái nhà của mẹ kế, bị hãm hại và làm việc vất vả. Nhưng nhờ vào sự giúp đỡ của các loài vật và sự thấu hiểu của một ông lão nơi rừng sâu, Tấm đã có được một bộ váy để đi dự lễ hội.

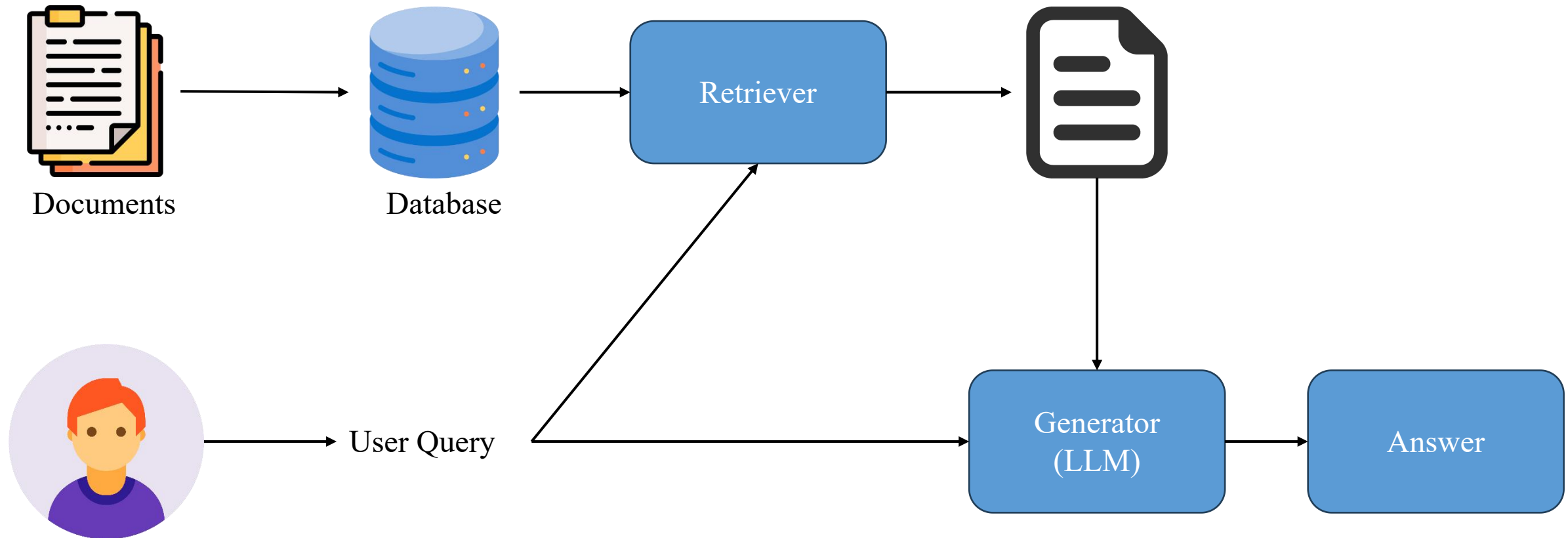
Retrieval Augmented Generation

- Overview
- Ingestion
- Retrieval
- Generation

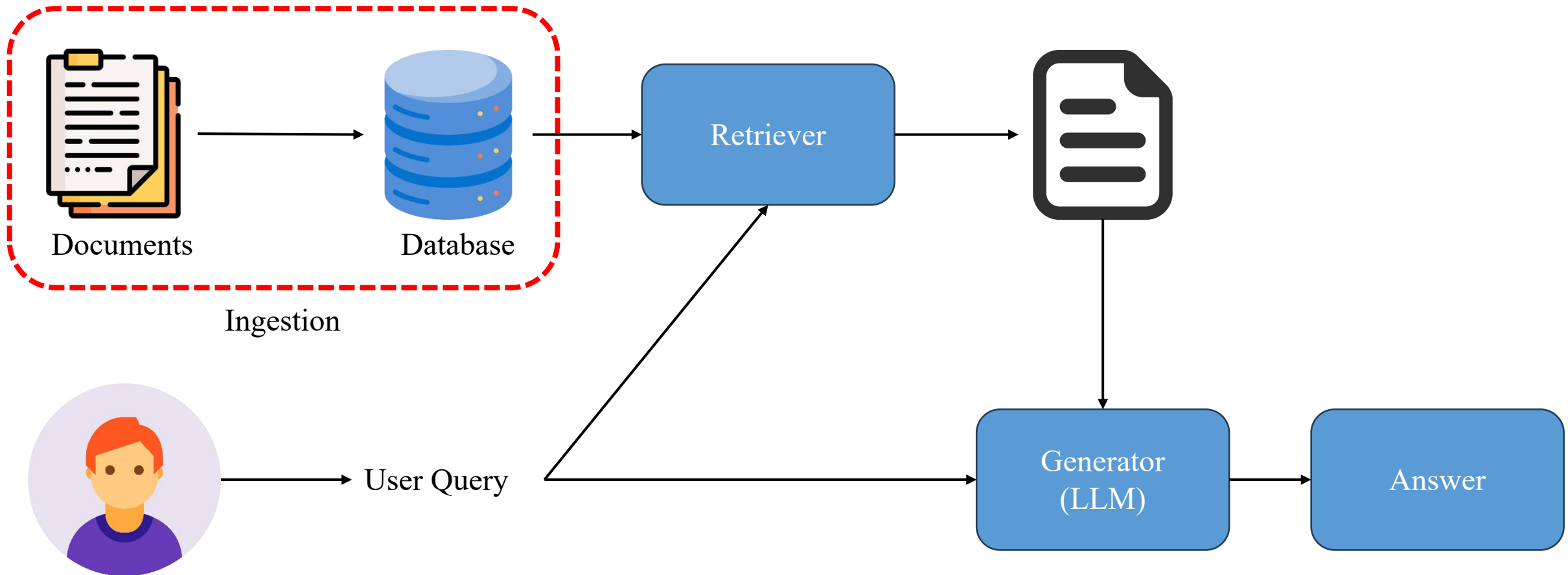
Overview



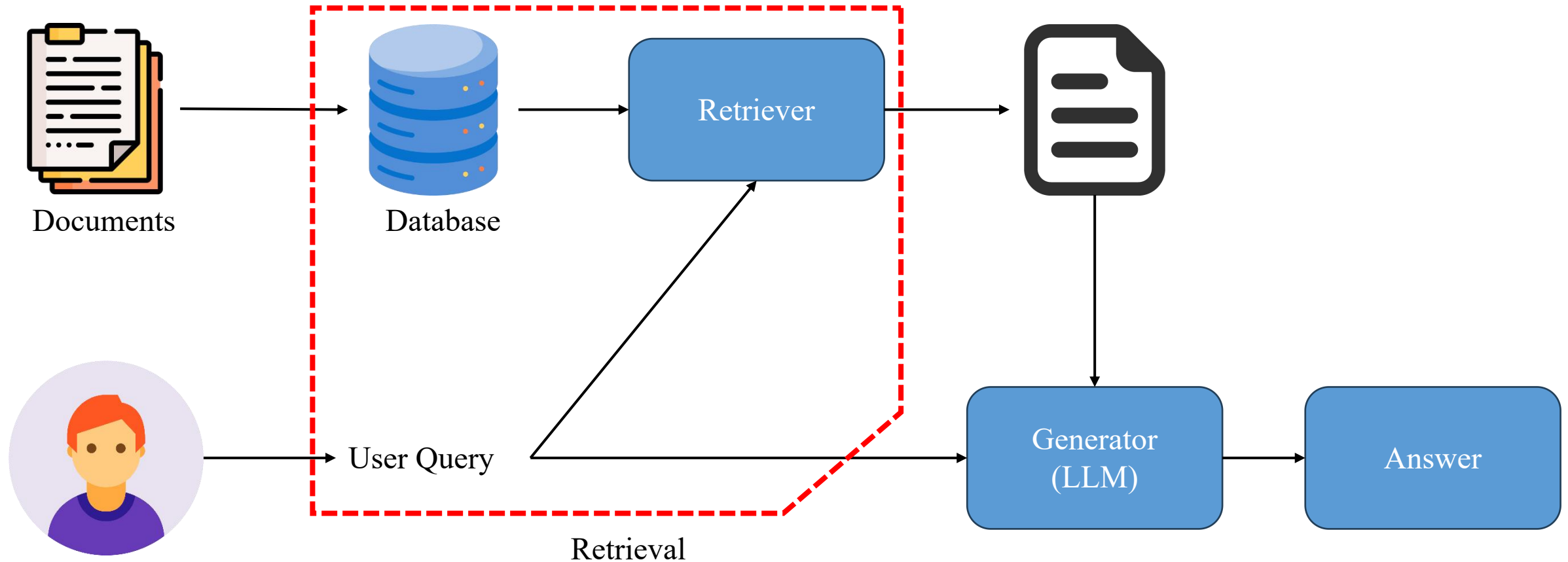
Overview



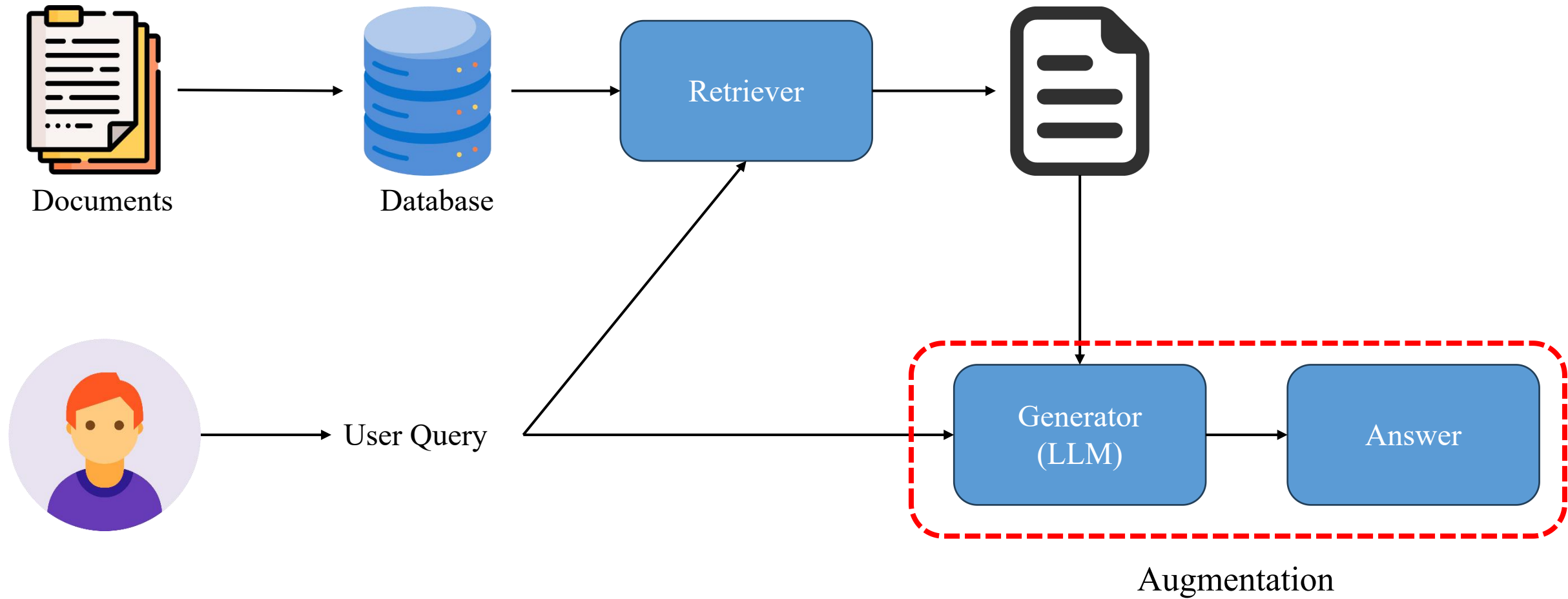
Overview



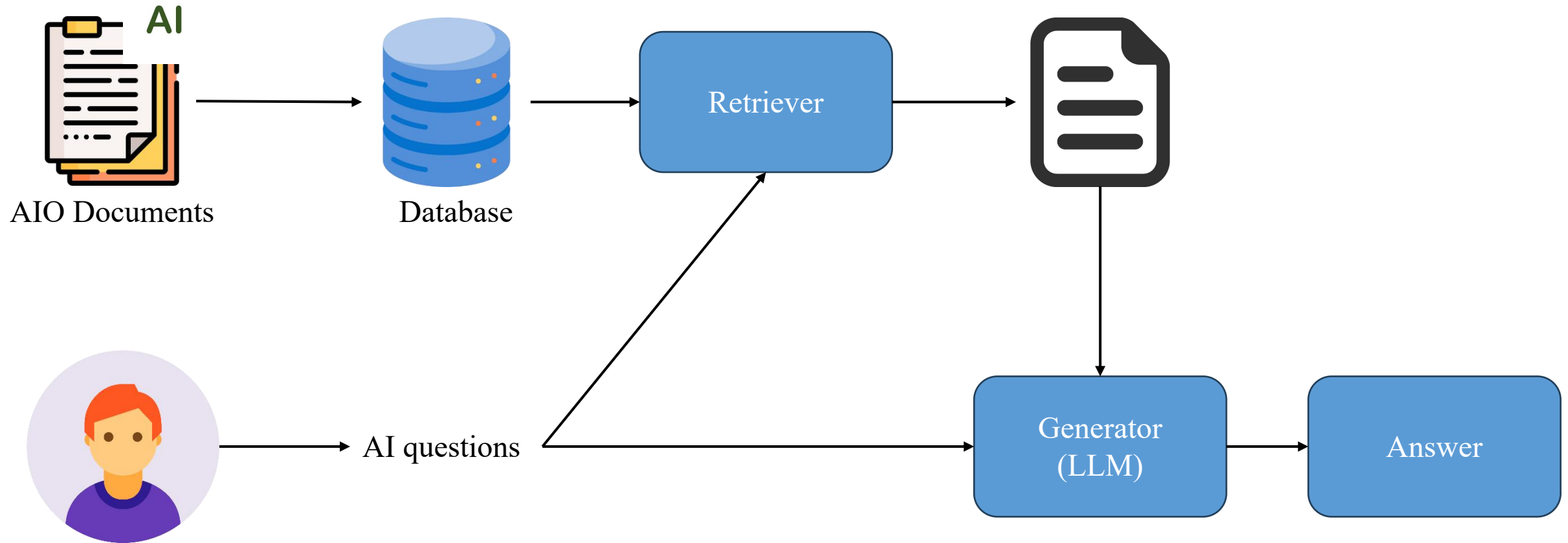
Overview



Overview



Learn by Project



Building a Chatbot from AIO Documents!!

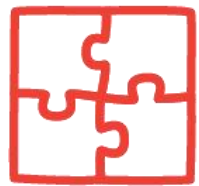
Tools



LangChain



LlamaIndex

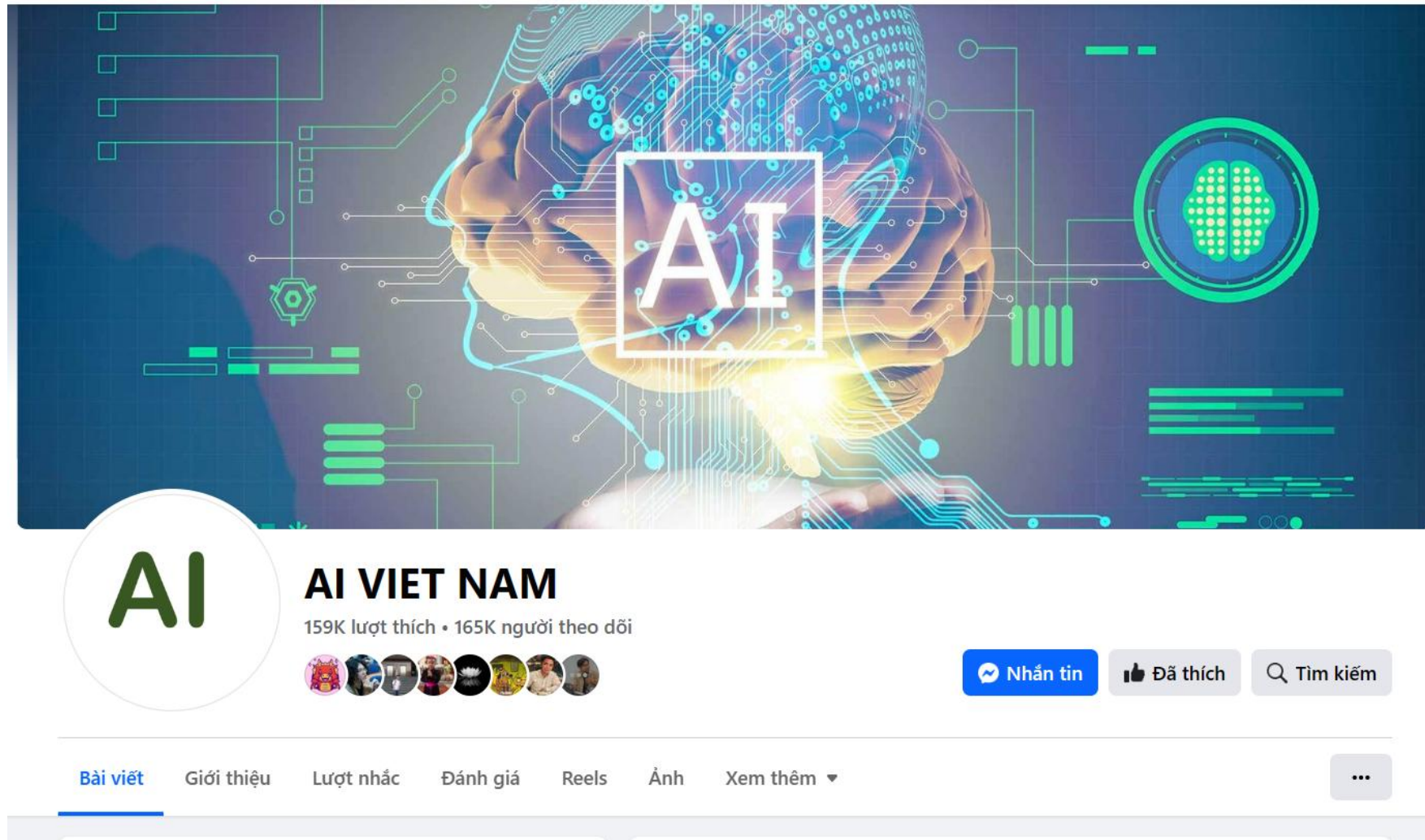


DSPy

Quantization

- Overview
- **Ingestion**
- Retrieval
- Generation

Data Collecting



Read PDF files

pypdf
BACK TO THE ROOTS

**U N S T
R U C T
U R E D**



Llama Parse

Read PDF files

```
1 import pypdf
2 filename = "/content/AIO_Documents/Documents/Mamba.pdf"
3 pdf_file = open(filename, 'rb')
4
5 reader = pypdf.PdfReader(pdf_file)
6
7 content = ""
8 for page in reader.pages:
9     text = page.extract_text()
10    content += text + "\n"
11 pdf_file.close()
```

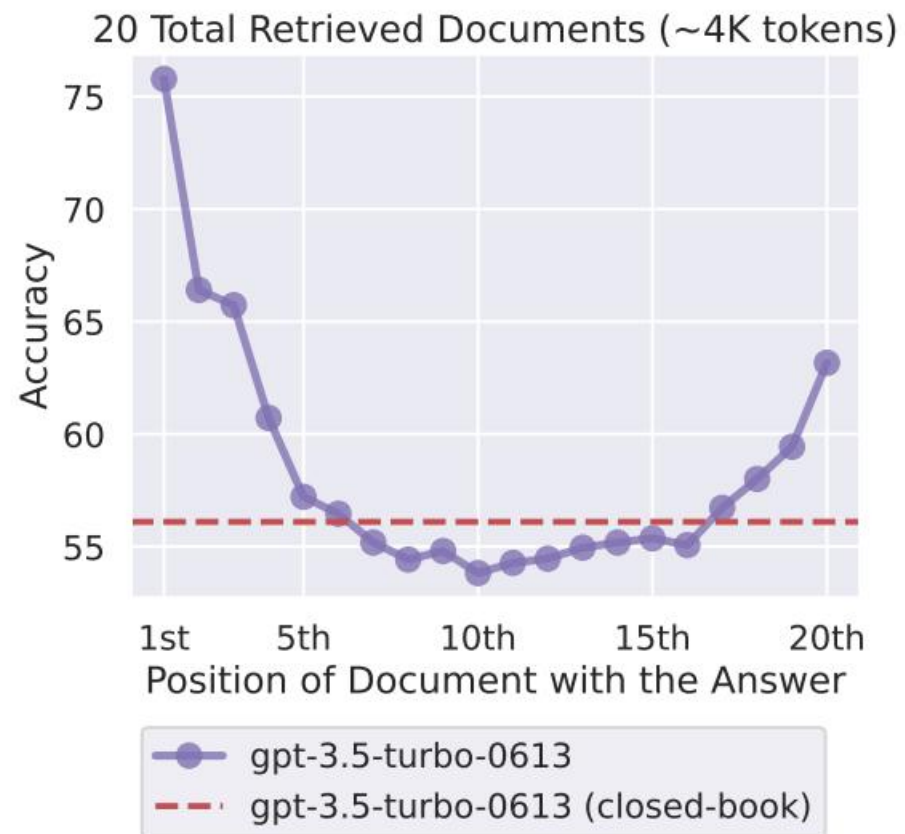
Using llama-index

```
1 from llama_index.core import SimpleDirectoryReader
2
3 documents = SimpleDirectoryReader("/content/AIO_Documents/Documents").load_data()
```

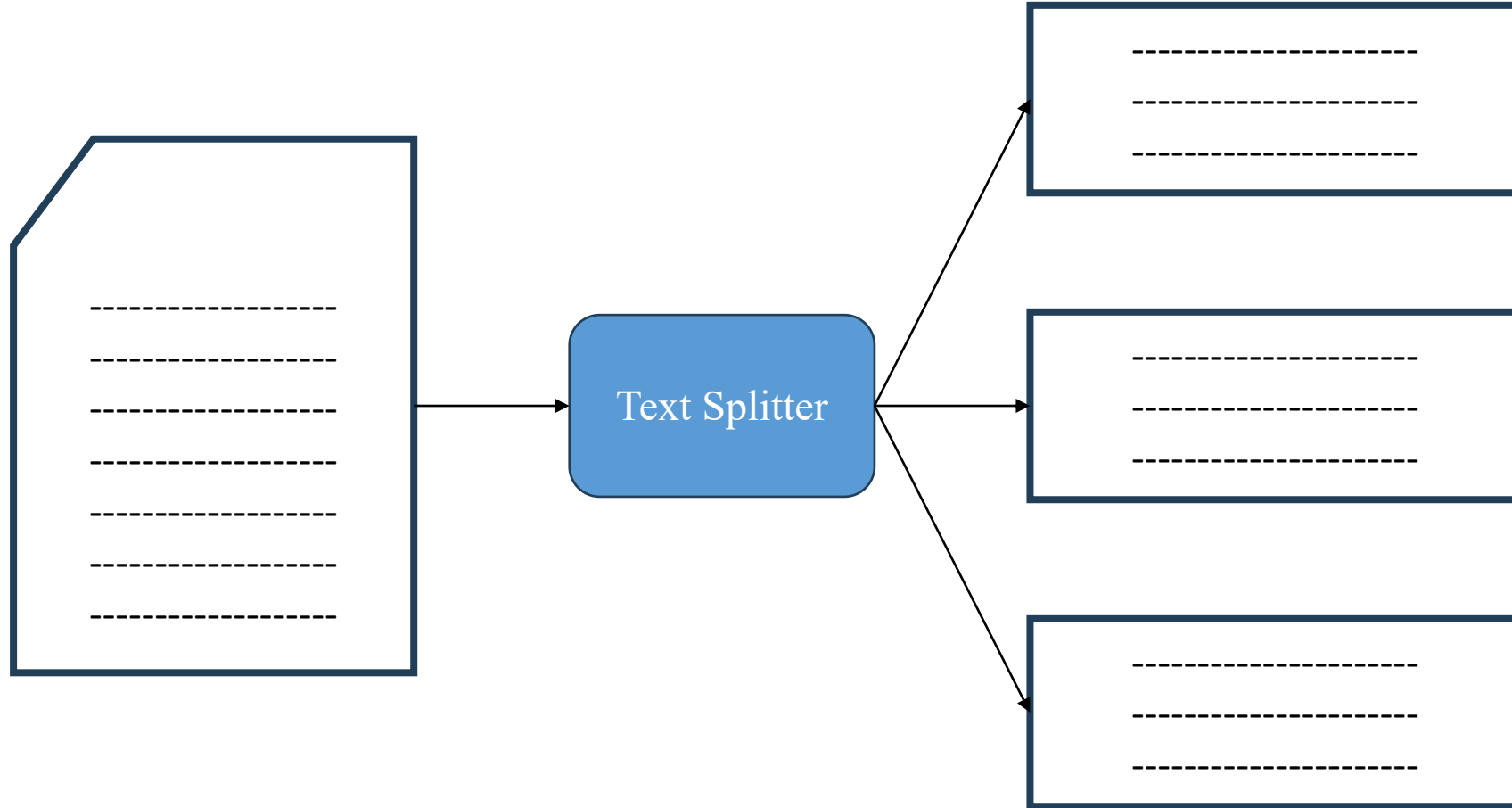
Text Splitter

Model	Context Window
gpt-4-turbo	
gpt-3.5-turbo	
Claude 3 - opus	
Mistral-7b	
Llama-7b	

Text Splitter



Text Splitter



Text Splitter

```
1 from llama_index.core.node_parser import SentenceSplitter
2 from llama_index.core import Settings, VectorStoreIndex
3
4 text_splitter = SentenceSplitter(chunk_size=512, chunk_overlap=64)
5 processed_documents = text_splitter(documents)
6
7 print("Before splitting: ", len(documents))
8 print("After splitting: ", len(processed_documents))
9
```

```
Before splitting: 318
After splitting: 597
```

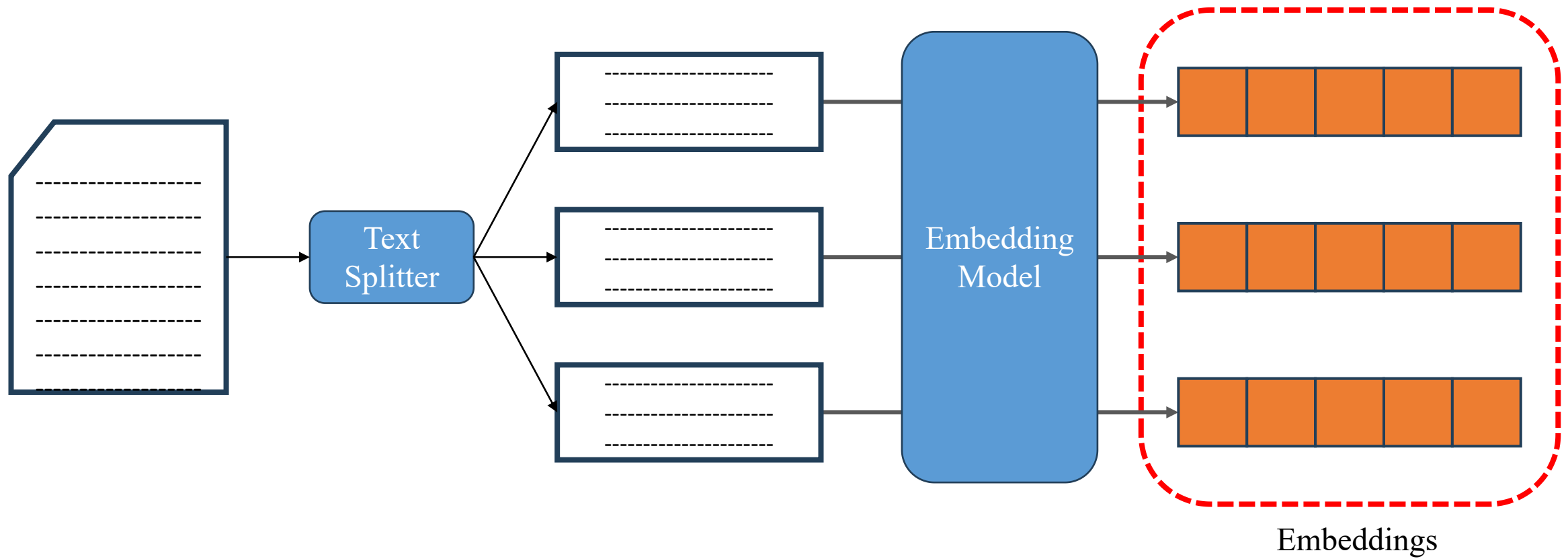
Semantic Splitter

```
1 from llama_index.core.node_parser import (  
2     SentenceSplitter,  
3     SemanticSplitterNodeParser,  
4 )  
5 from llama_index.embeddings.openai import OpenAIEmbedding  
6 from llama_index.core import SimpleDirectoryReader  
7  
8  
9 import os  
10 os.environ["OPENAI_API_KEY"] = "YOUR_API_KEY"  
11  
12 embed_model = OpenAIEmbedding()  
13 splitter = SemanticSplitterNodeParser(  
14     buffer_size=1, breakpoint_percentile_threshold=95, embed_model=embed_model  
15 )  
16  
17 nodes = splitter.get_nodes_from_documents(documents)  
18 for node in nodes:  
19     print('- ' * 100)  
20     print(node.get_content())
```

Why Embedding?

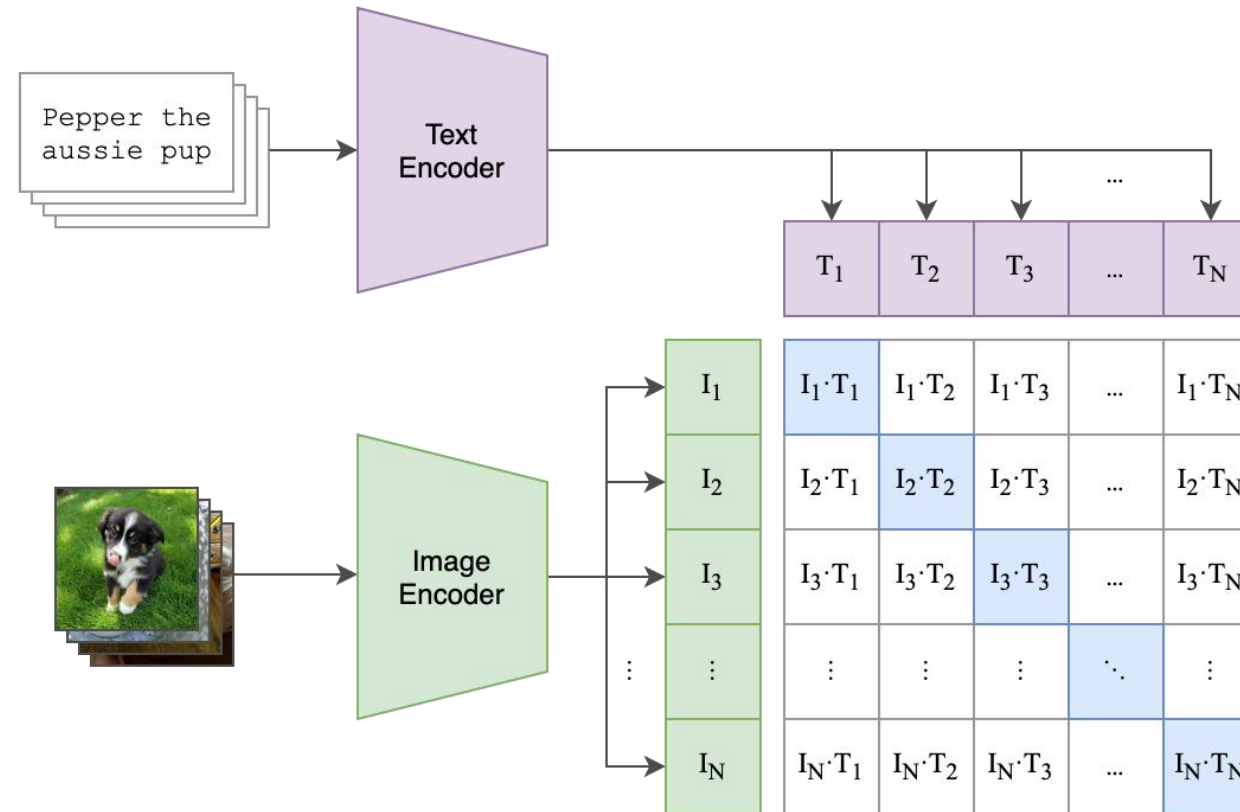
Aspect	Embedding Search	Keyword Search
Search Method	Uses vectors to represent the semantic meaning of words	Relies on exact keyword matching
Understanding Context	Understands the intent and context of queries	Focuses on the presence of specific keywords
Handling Synonyms	Can identify and match synonyms and related terms	Typically requires explicit listing of synonyms
Query Flexibility	Allows natural language queries	Best with specific, well-defined keyword queries
Computational Demand	More computationally intensive	Less computationally demanding

Embedding Models



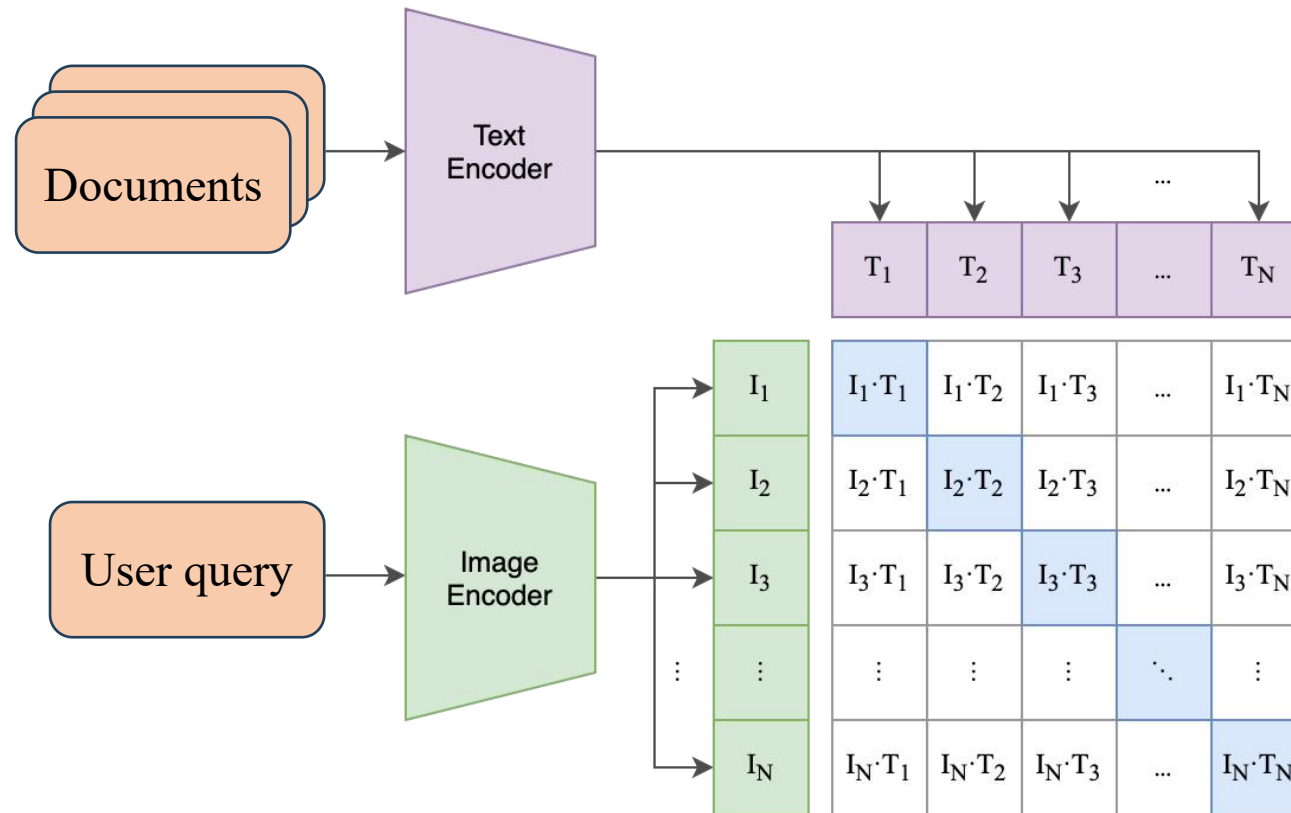
Embedding Models

(1) Contrastive pre-training



Embedding Models

(1) Contrastive pre-training



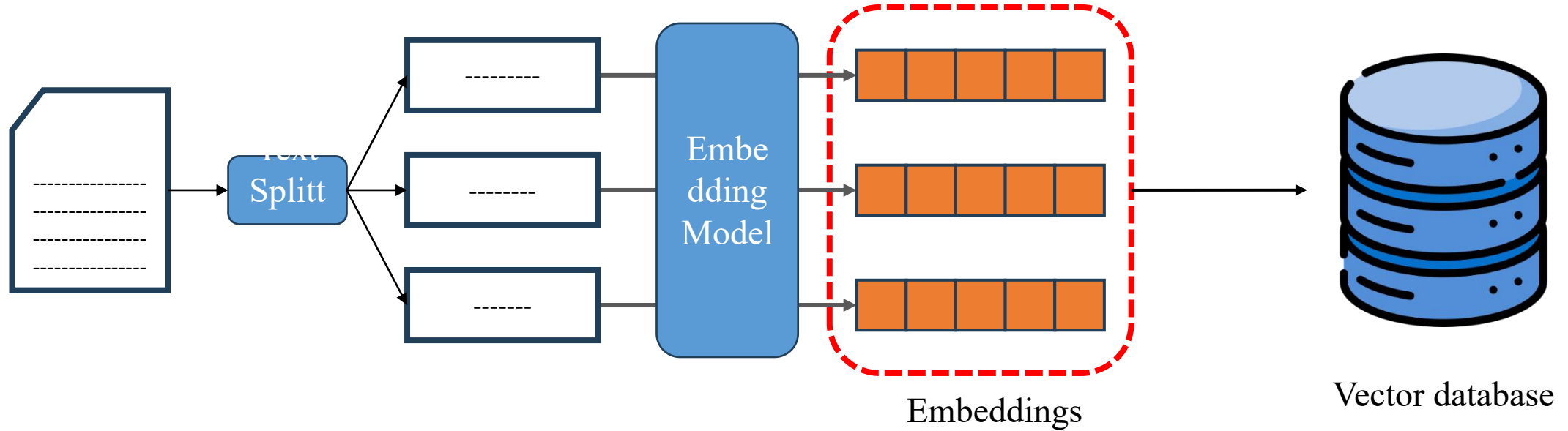
Which embedding model to use?

Embedding Models

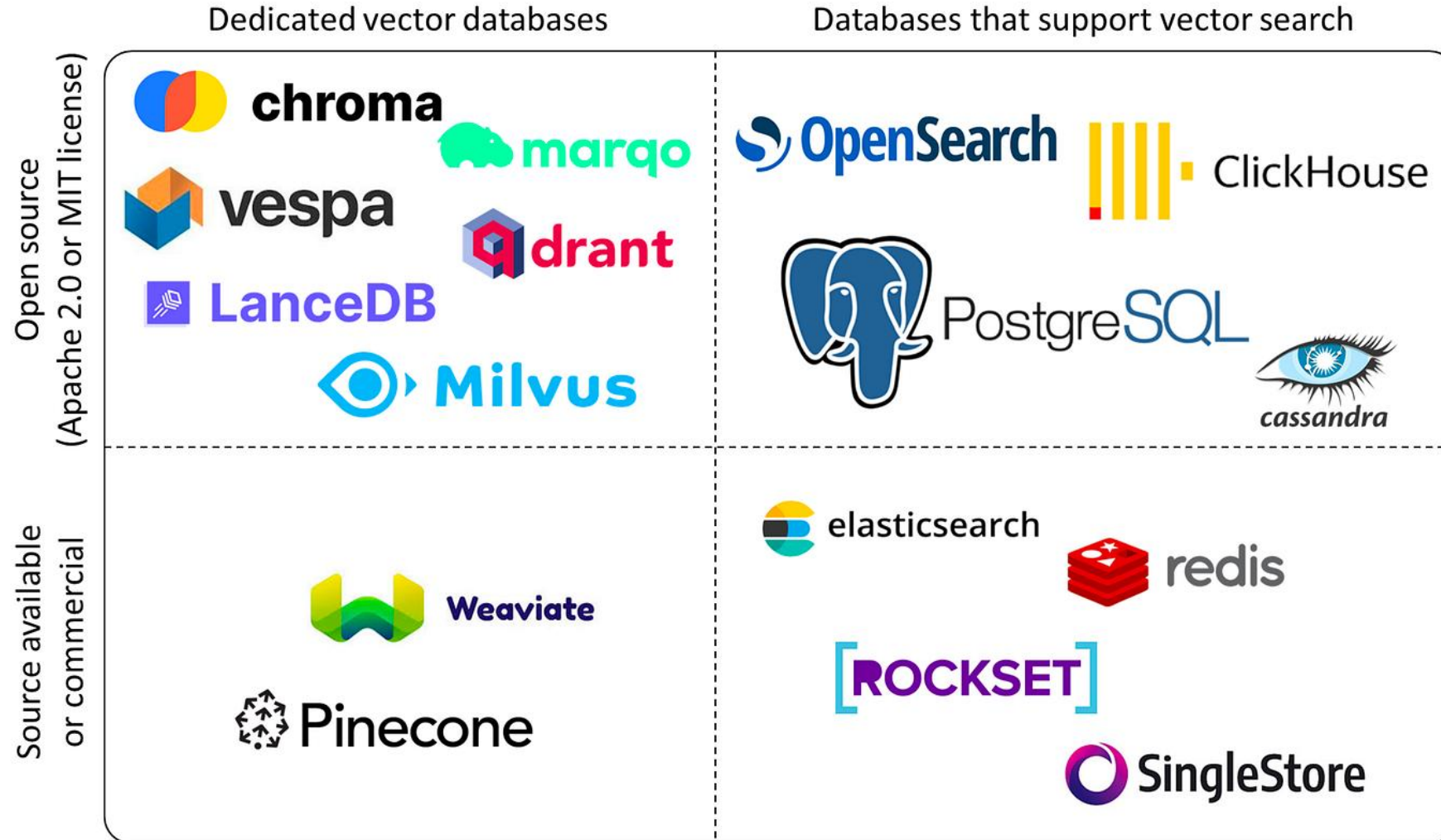
[MTEB Leaderboard - a Hugging Face Space by mteb](#)

Rank ▲	Model ▲	Model Size (Million Parameters) ▲	Memory Usage (GB, fp32) ▲	Embedding Dimensions ▲	Max Tokens ▲	Average (56 datasets) ▲	Classification Average (12 datasets) ▲	Clustering Average (11 datasets) ▲
1	SFR-Embedding-Mistral	7111	26.49	4096	32768	67.56	78.33	51.67
2	voyage-lite-02-instruct	1220	4.54	1024	4000	67.13	79.25	52.42
3	GritLM-7B	7242	26.98	4096	32768	66.76	79.46	50.61
4	e5-mistral-7b-instruct	7111	26.49	4096	32768	66.63	78.47	50.26
5	google-gecko.text-embedding-p	1200	4.47	768	2048	66.31	81.17	47.48
6	GritLM-8x7B	46703	173.98	4096	32768	65.66	78.53	50.14
7	LLM2Vec-Mistral-7B-Instruct-v					64.8	76.63	45.54
8	echo-mistral-7b-instruct-last	7111	26.49	4096	32768	64.68	77.43	46.32
9	mxbai-embed-large-v1	335	1.25	1024	512	64.68	75.64	46.71

Vector database



Vector database



Vector database

[Vector DB Comparison \(superlinked.com\)](https://superlinked.com)

Vector DB Comparison

by Superlinked | Last Updated : Today

Search

Get insights

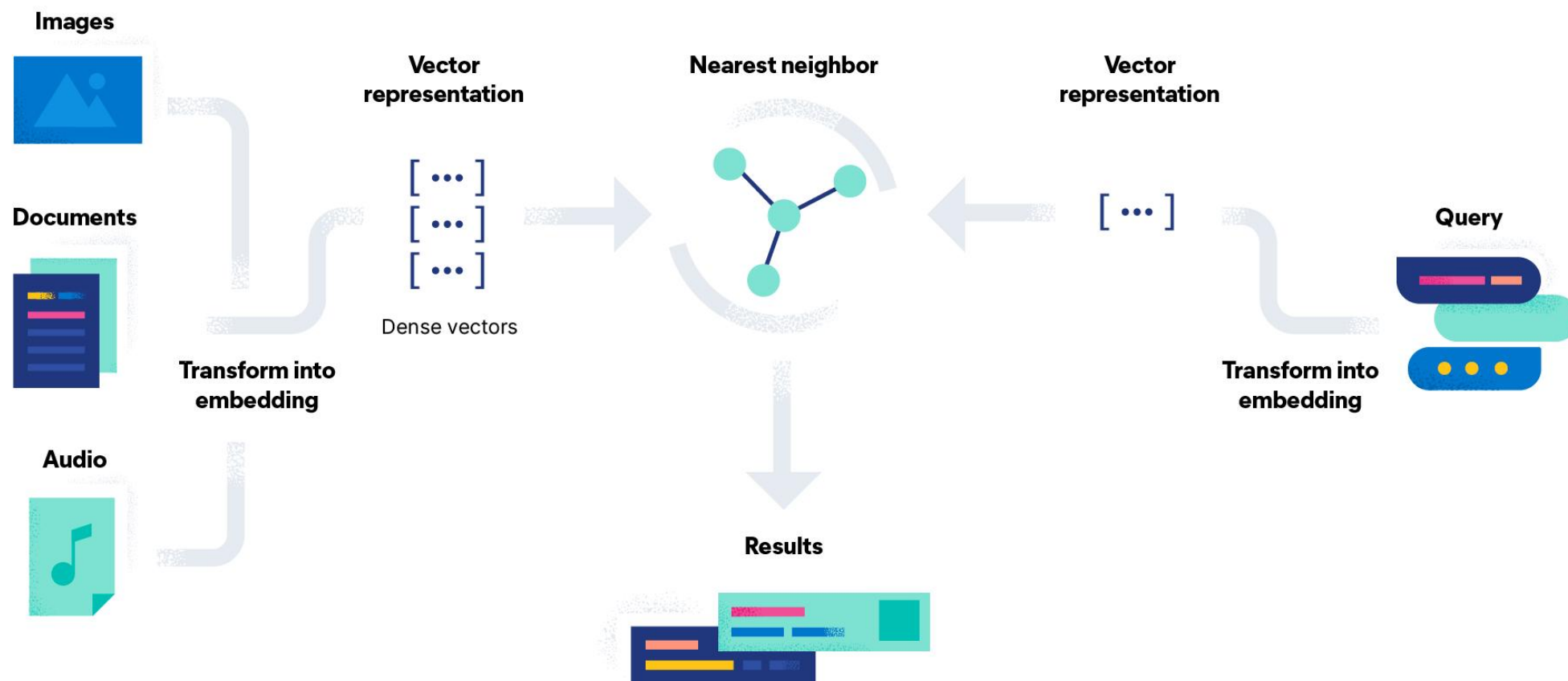
Give us a star

Vendor	About				Search							
	OSS	License	Dev Lang	VSS Launch	Filters	Hybrid Search	Facets	Geo Search	Multi-Vector	Sparse	BM25	
KDB.AI	✗	Proprietary	python	2023	✓	✓		-	-	✓	✓	
LanceDB	✓	Apache-2.0	rust	2023	✓	✓	-	-	✗	-	✓	
Marqo	✓	Apache-2.0	python	2022	✓	✓	-	-	✓	✗	✓	
Meilisearch	✓	MIT	rust	2023	✓	✓	✓	✓	✓	✗		
Milvus	✓	Apache-2.0	go c++	2019	✓	✓	-	✗	✓	✓	✓	
MongoDB Atlas	✗	GNU AGPL ...	c++ java	2023	✓	✓	✓	✓	✓	✗	✓	
MyScale	✓	Apache-2.0	c++	2023	✓	✓	✓	✓	-	✗	✗	
Neo4j	✗	Pro...	java scala	2023	✓	✓	✗	✓	✓	✗	✗	
Nuclia DB	✓	AGPLv3	rust python	2021	✓	✓	✓	✗	-	✗	✓	
OpenSearch	✓	Apache-2.0	java	2021	✓	✓	✓	✓	✓	✓	✓	

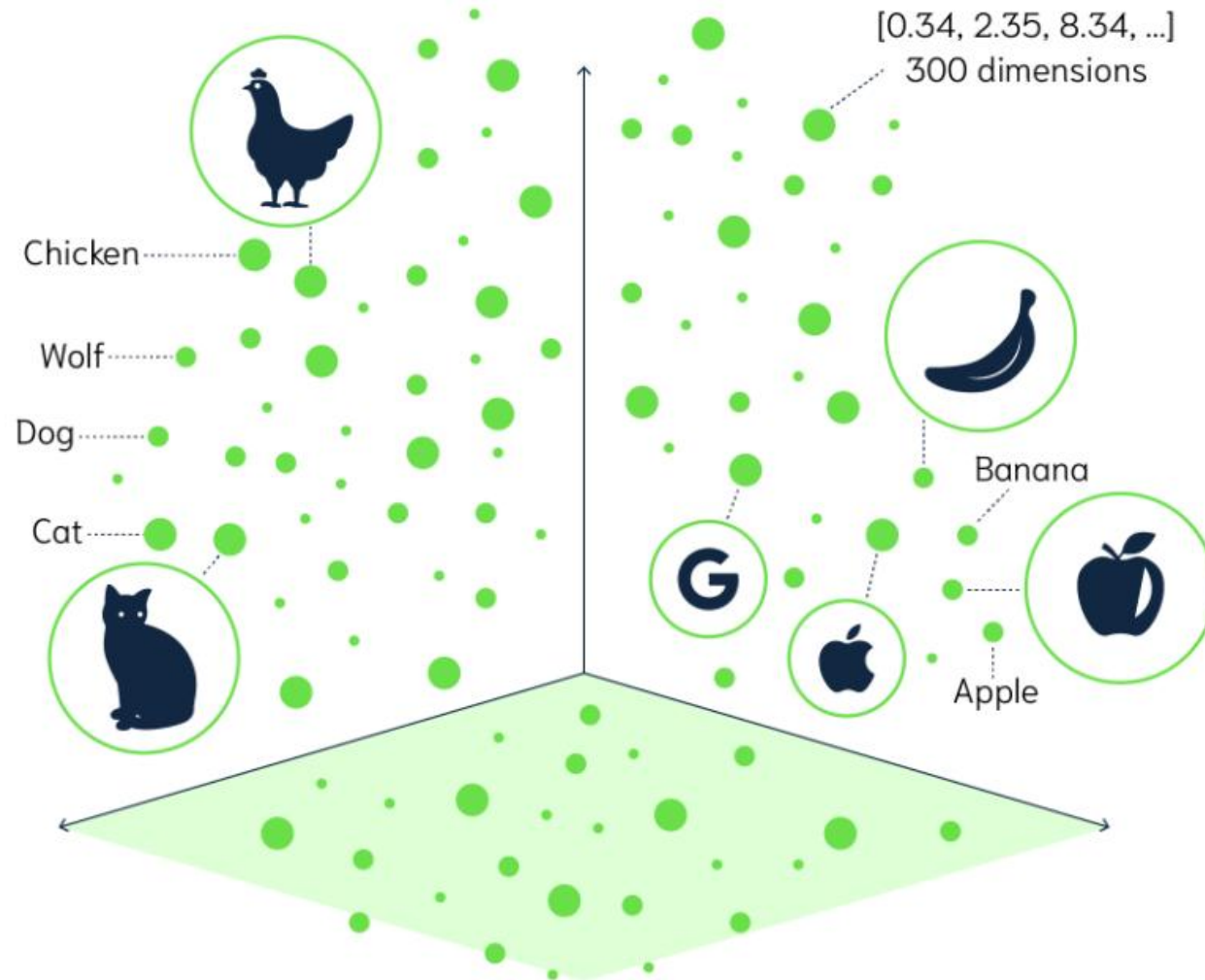
Quantization

- Overview
- Ingestion
- **Retrieval**
- Generation

Vector Search

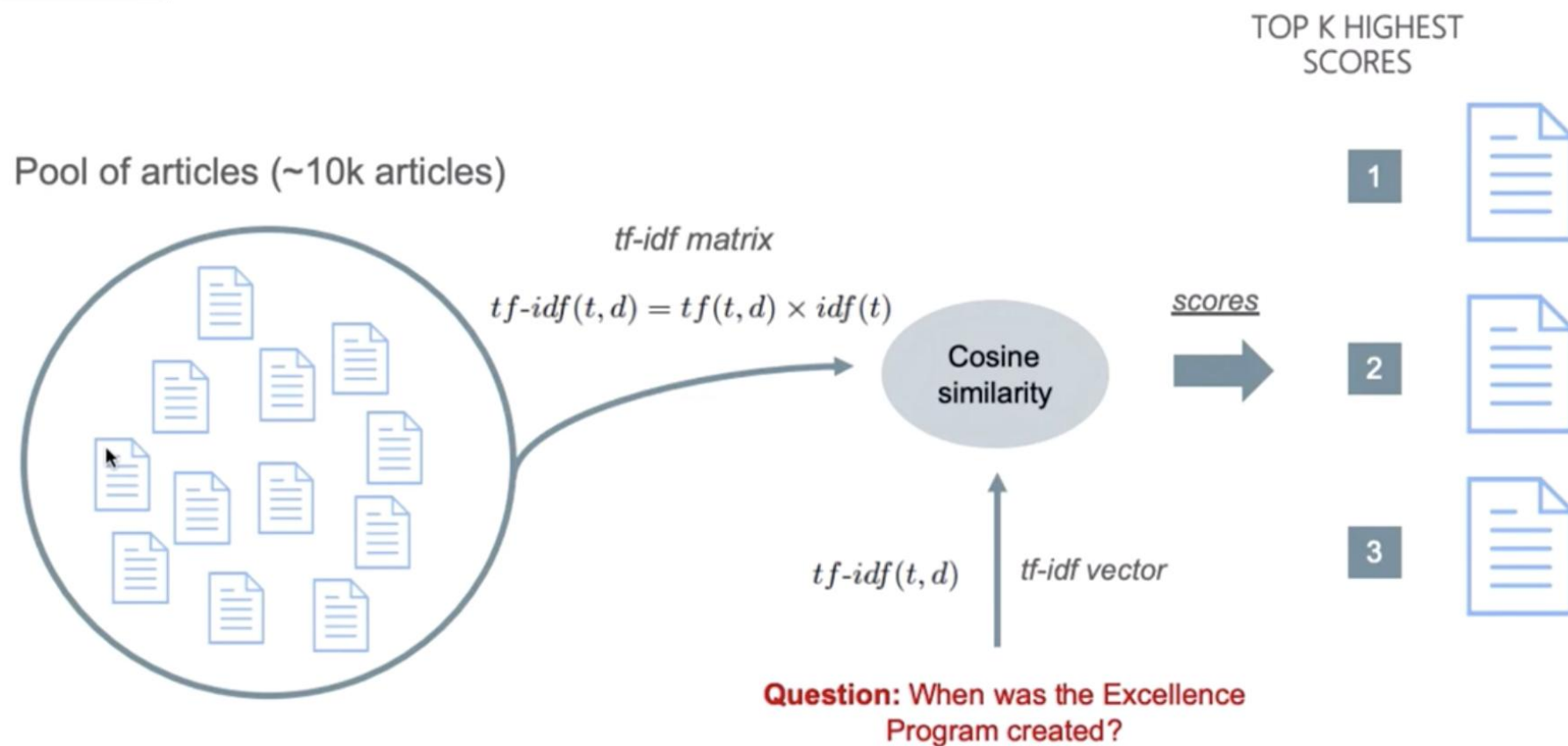


Vector Search

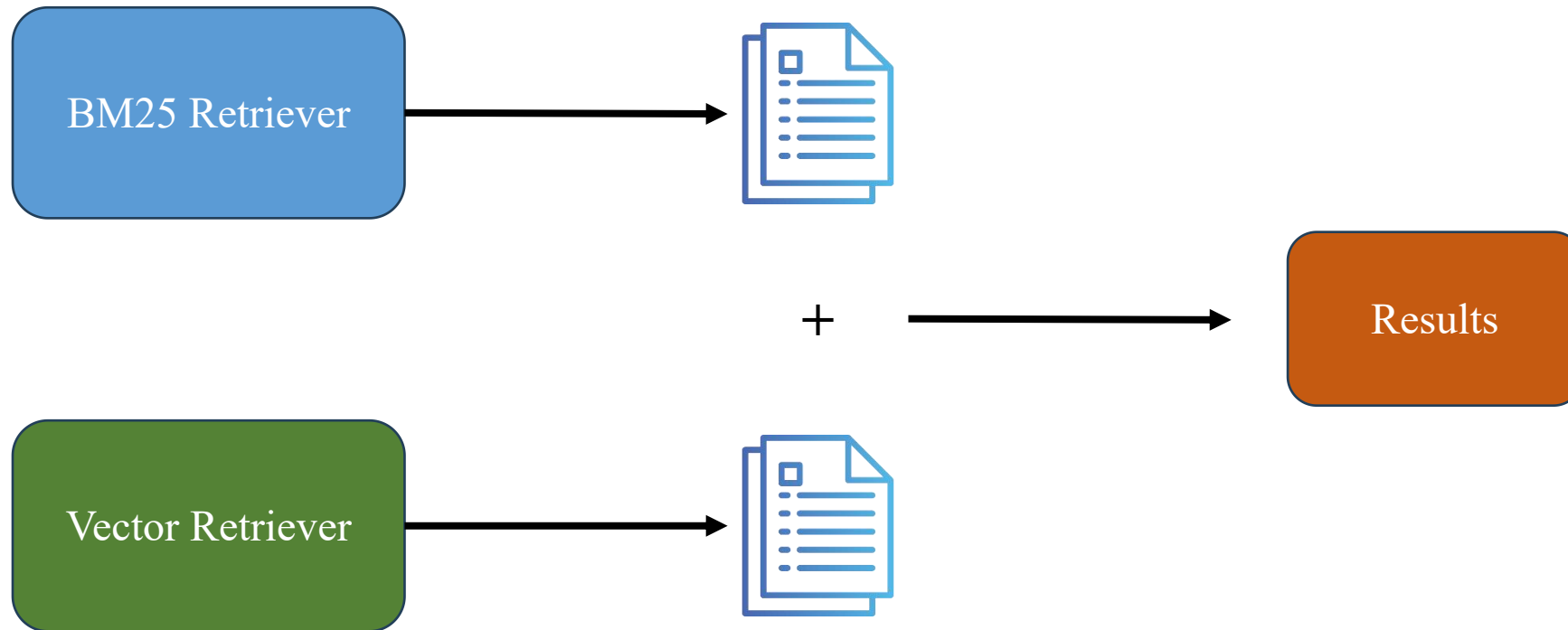


Keyword Search

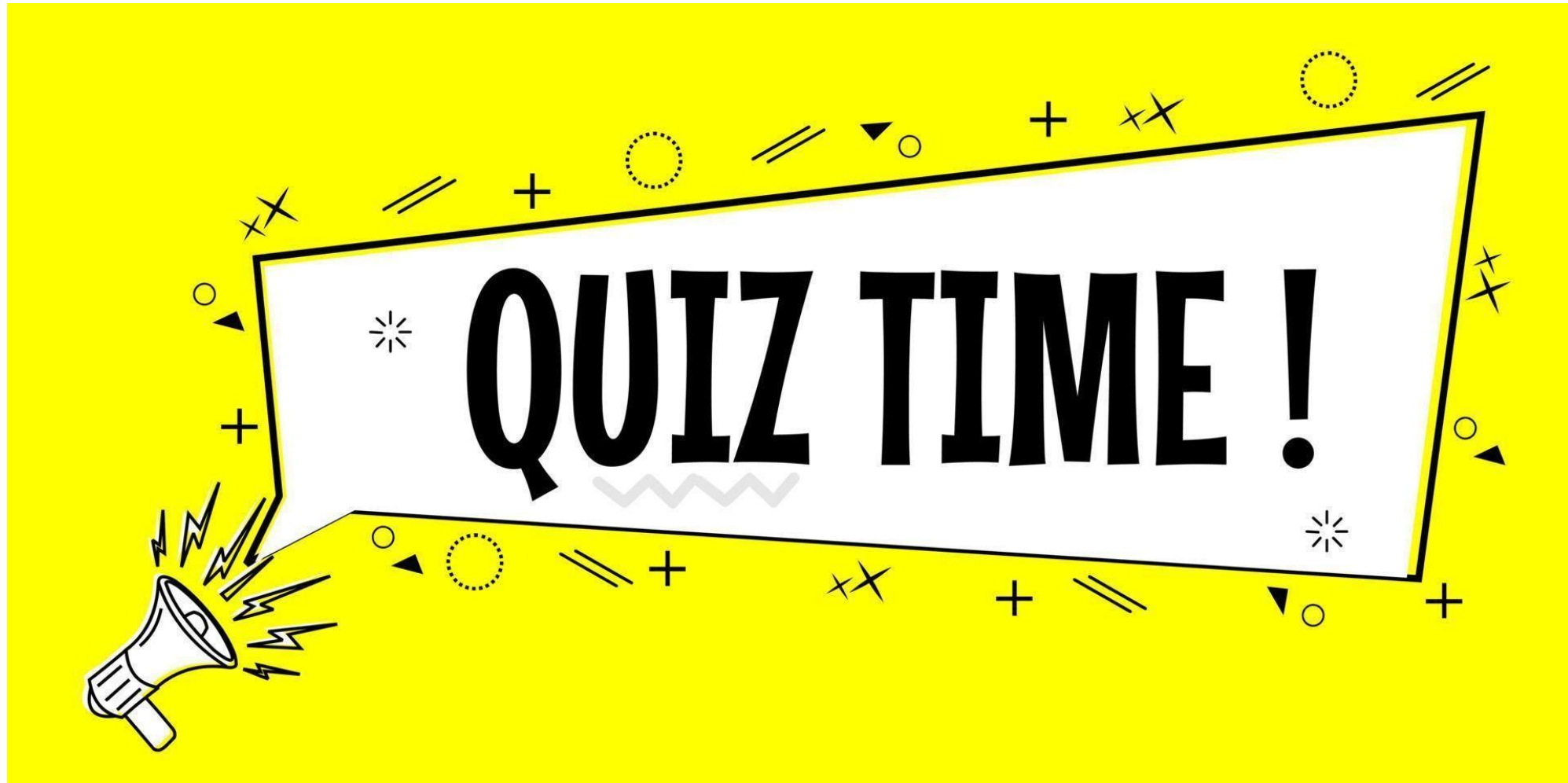
RETRIEVER



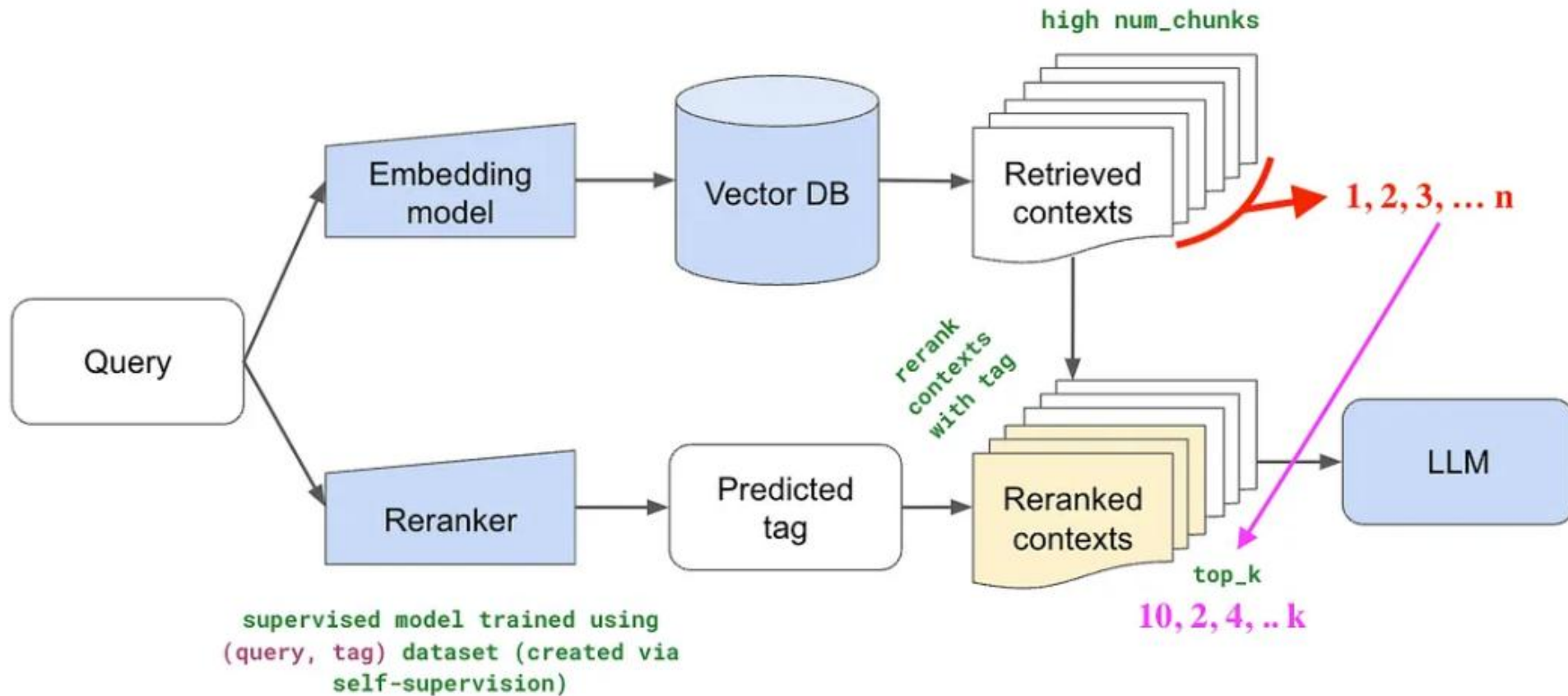
Hybrid Search



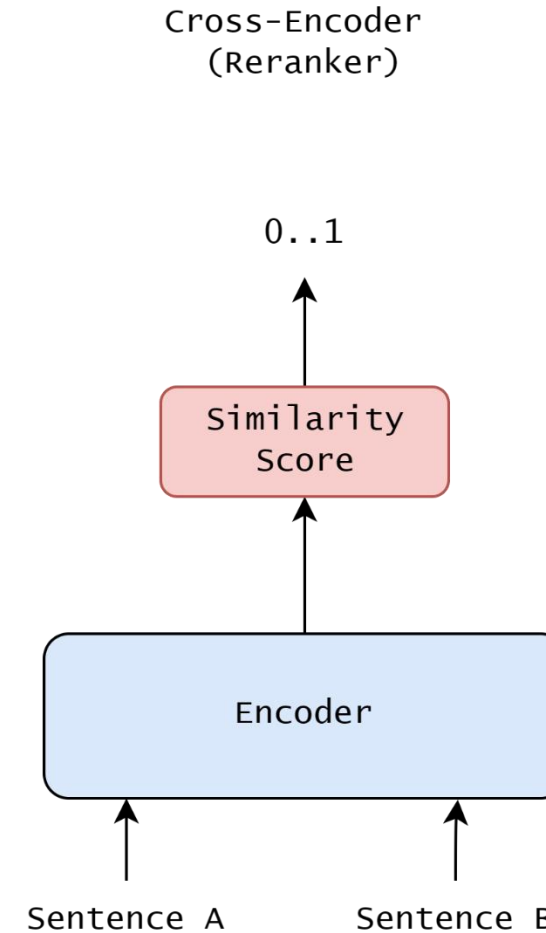
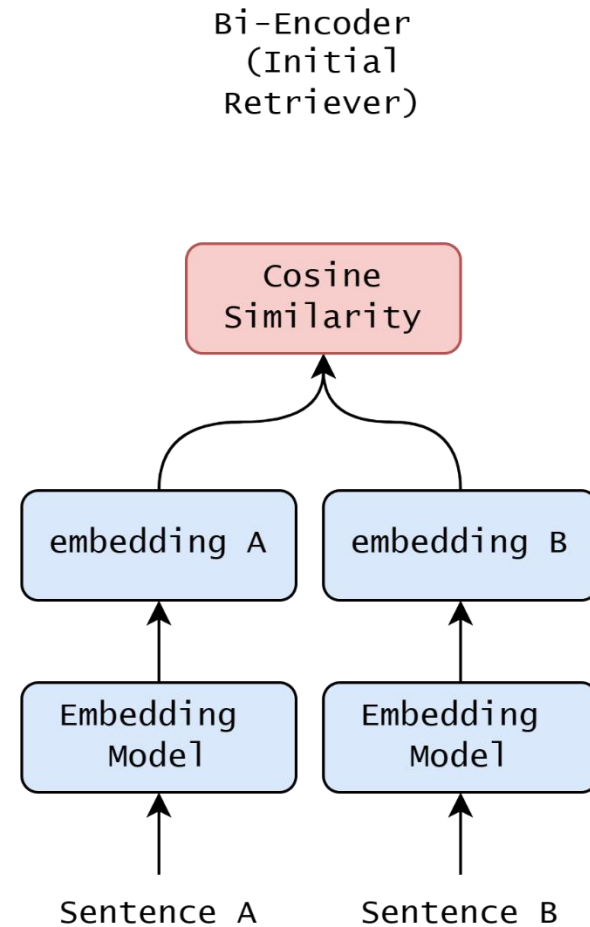
QUIZ TIME!!!



Reranking



Reranking - Model based



Reranking – LLM based

Microsoft **LLMLingua**: Compressing Prompts for Accelerated Inference of Large Language Models

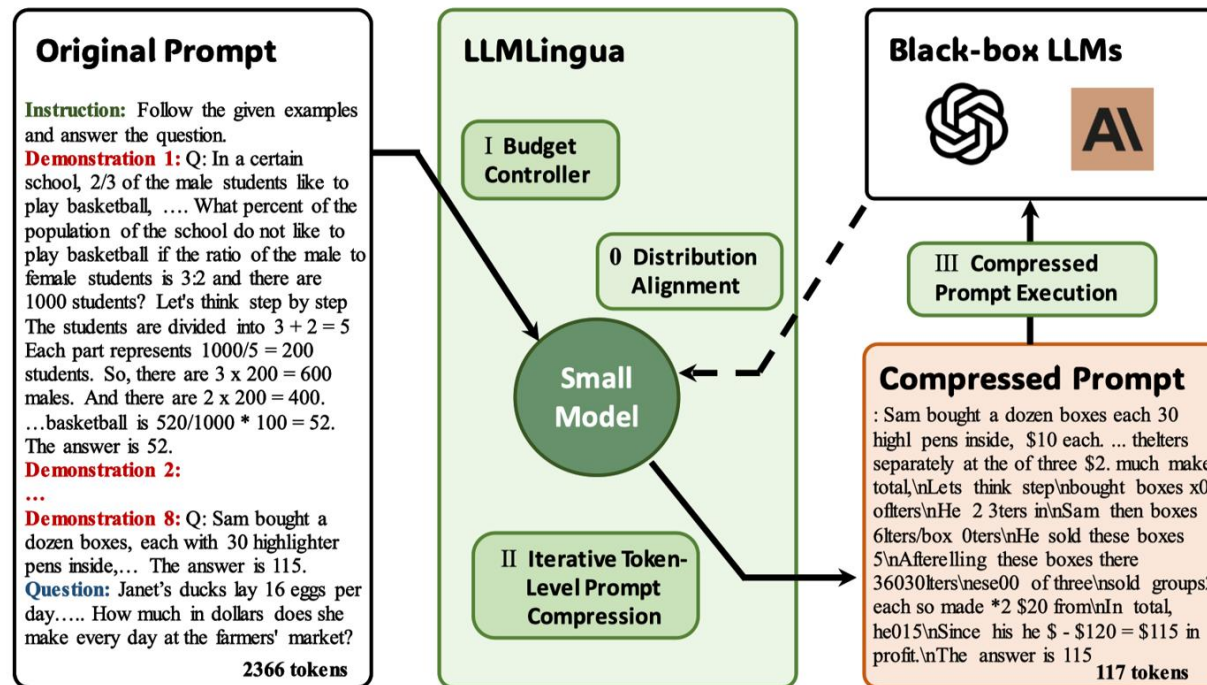


Figure 1: Framework of the proposed approach LLMLingua.

<https://aka.ms/LLMLingua>

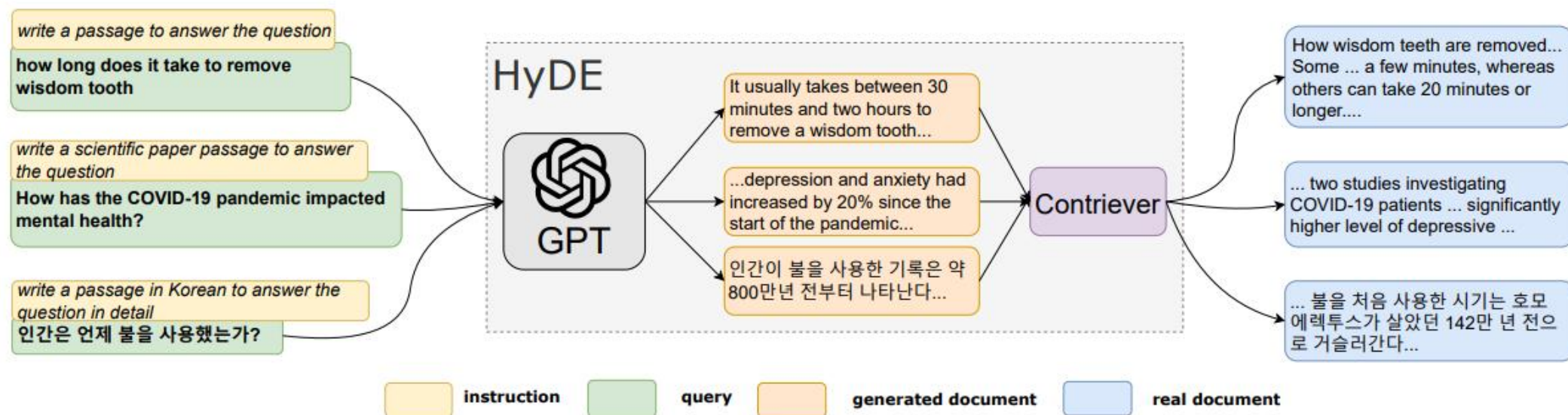
Now you can use
LLMLingua!



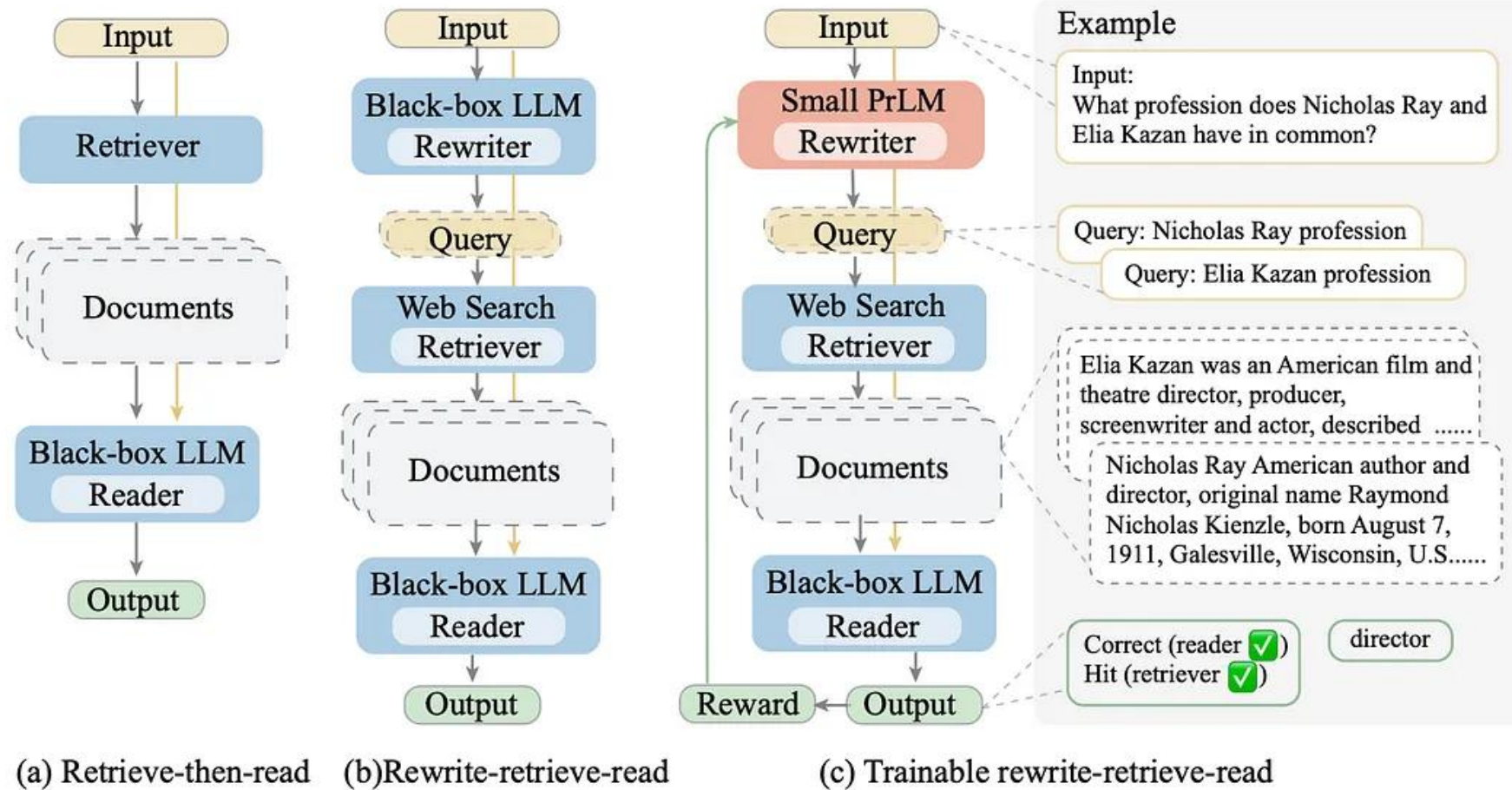
A simple and efficient method to compress prompt up to **20x**.

- 💰 **Saving cost**, not only prompt, but also the generation length;
- 📄 **Support longer contexts**;
- ⚖️ **Robustness**, no need any training for the LLMs;
- 🔍 **Keeping** the original prompt knowledge like ICL, reasoning, etc.
- 📄 **KV-Cache compression**, speedup inference;
- 🦜 **GPT-4 can recovery all key information from compressed prompt.**

Query Rewriting



Query Rewriting



Quantization

- Overview
- Ingestion
- Retrieval
- **Generation**

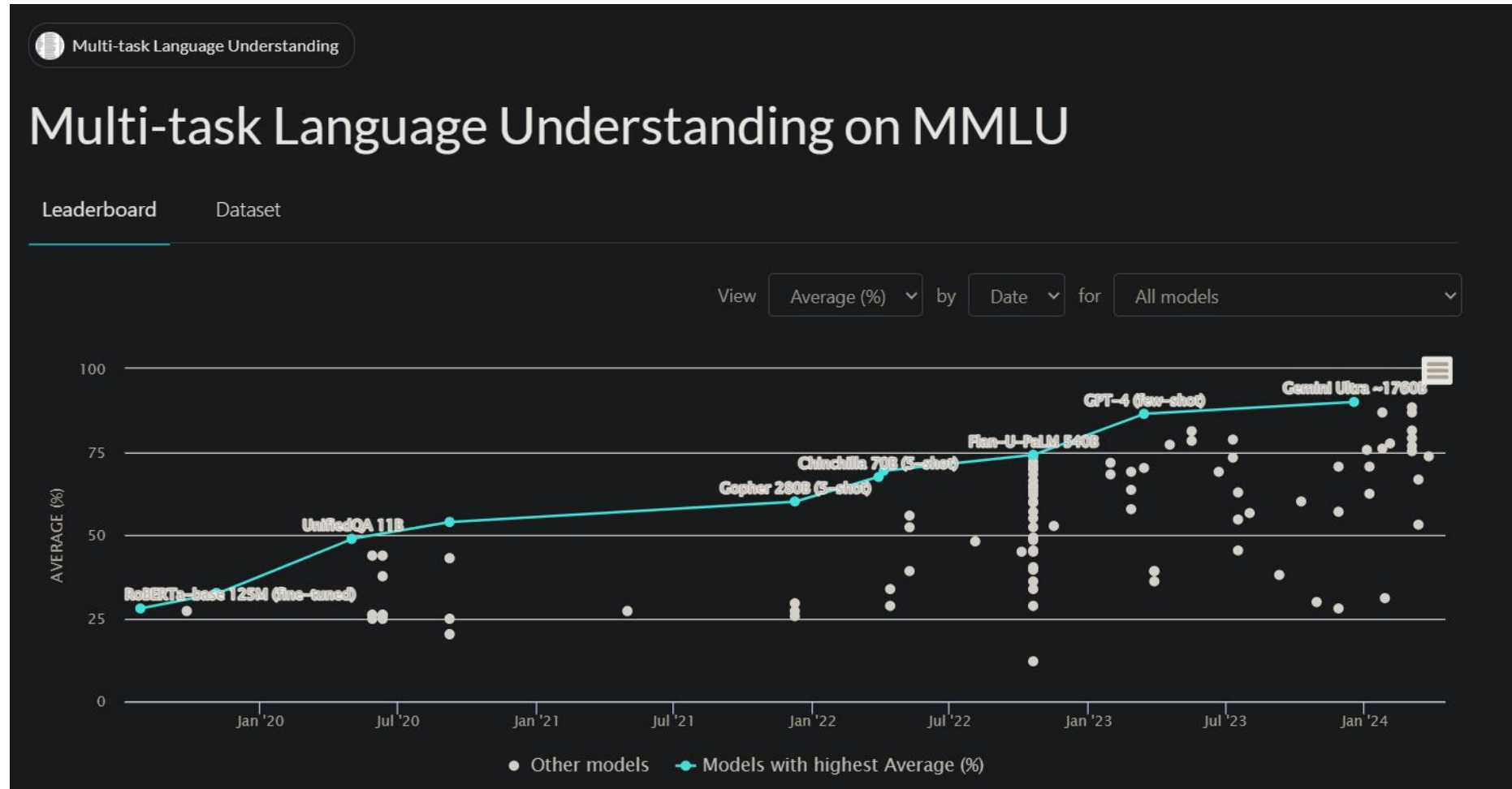
LLM Benchmarks

	Meta Llama 3 70B
MMLU 5-shot	82.0
GPQA 0-shot	39.5
HumanEval 0-shot	81.7
GSM-8K 8-shot, CoT	93.0
MATH 4-shot, CoT	50.4

BENCHMARK	DESCRIPTION	Gemma 7B
MMLU 5-shot, top-1	Representation of questions in 57 subjects (incl. STEM, humanities and others)	64.3
BBH -	Diverse set of challenging tasks requiring multi-step reasoning	55.1
DROP -	Reading comprehension (F1 Score)	-
HellaSwag 0-shot	Commonsense reasoning for everyday tasks	81.2
GSM8K maj@1	Basic arithmetic manipulations (incl. Grade School math problems)	46.4
MATH 4-shot	Challenging math problems (incl. algebra, geometry, pre-calculus, and others)	24.3
HumanEval pass@1	Python code generation	32.3
Natural2Code -	Python code generation. New held out dataset HumanEval-like, not leaked on the web.	-

	Gemini Ultra	Gemini Pro	GPT-4
MMLU Multiple-choice questions in 57 subjects (professional & academic) (Hendrycks et al., 2021a)	90.04% CoT@32*	79.13% CoT@8*	87.29% CoT@32 (via API**)
	83.7% 5-shot	71.8% 5-shot	86.4% 5-shot (reported)
GSM8K Grade-school math (Cobbe et al., 2021)	94.4% Maj1@32	86.5% Maj1@32	92.0% SFT & 5-shot CoT
MATH Math problems across 5 difficulty levels & 7 subdisciplines (Hendrycks et al., 2021b)	53.2% 4-shot	32.6% 4-shot	52.9% 4-shot (via API**)
			50.3% (Zheng et al., 2023)
BIG-Bench-Hard Subset of hard BIG-bench tasks written as CoT problems (Srivastava et al., 2022)	83.6% 3-shot	75.0% 3-shot	83.1% 3-shot (via API**)
HumanEval Python coding tasks (Chen et al., 2021)	74.4% 0-shot (IT)	67.7% 0-shot (IT)	67.0% 0-shot (reported)
Natural2Code Python code generation. (New held-out set with no leakage on web)	74.9% 0-shot	69.6% 0-shot	73.9% 0-shot (via API**)
DROP Reading comprehension & arithmetic. (metric: F1-score) (Dua et al., 2019)	82.4 Variable shots	74.1 Variable shots	80.9 3-shot (reported)
HellaSwag (validation set) Common-sense multiple choice questions (Zellers et al., 2019)	87.8% 10-shot	84.7% 10-shot	95.3% 10-shot (reported)

LLM Benchmarks - MMLU



GSM8K + MATH

Problem: Beth bakes 4, 2 dozen batches of cookies in a week. If these cookies are shared amongst 16 people equally, how many cookies does each person consume?

Solution: Beth bakes 4 2 dozen batches of cookies for a total of $4 \times 2 = \ll 4 \times 2 = 8 \gg$ 8 dozen cookies

There are 12 cookies in a dozen and she makes 8 dozen cookies for a total of $12 \times 8 = \ll 12 \times 8 = 96 \gg$ 96 cookies

She splits the 96 cookies equally amongst 16 people so they each eat $96 / 16 = \ll 96 / 16 = 6 \gg$ 6 cookies

Final Answer: 6

Problem: Mrs. Lim milks her cows twice a day. Yesterday morning, she got 68 gallons of milk and in the evening, she got 82 gallons. This morning, she got 18 gallons fewer than she had yesterday morning. After selling some gallons of milk in the afternoon, Mrs. Lim has only 24 gallons left. How much was her revenue for the milk if each gallon costs \$3.50?

Mrs. Lim got 68 gallons - 18 gallons = $\ll 68 - 18 = 50 \gg$ 50 gallons this morning.

So she was able to get a total of 68 gallons + 82 gallons + 50 gallons = $\ll 68 + 82 + 50 = 200 \gg$ 200 gallons.

She was able to sell 200 gallons - 24 gallons = $\ll 200 - 24 = 176 \gg$ 176 gallons.

Thus, her total revenue for the milk is $\$3.50/\text{gallon} \times 176 \text{ gallons} = \$\ll 3.50 \times 176 = 616 \gg$ 616.

Final Answer: 616

HumanEval

```
def incr_list(l: list):  
    """Return list with elements incremented by 1.  
    >>> incr_list([1, 2, 3])  
    [2, 3, 4]  
    >>> incr_list([5, 3, 5, 2, 3, 3, 9, 0, 123])  
    [6, 4, 6, 3, 4, 4, 10, 1, 124]  
    """
```

```
    return [i + 1 for i in l]
```

```
def solution(lst):  
    """Given a non-empty list of integers, return the sum of all of the odd elements  
    that are in even positions.
```

Examples

```
solution([5, 8, 7, 1]) ==>12  
solution([3, 3, 3, 3, 3]) ==>9  
solution([30, 13, 24, 321]) ==>0  
"""
```

```
    return sum(lst[i] for i in range(0, len(lst)) if i % 2 == 0 and lst[i] % 2 == 1)
```


Chatbot Arena

[Chat with Open Large Language Models \(lmsys.org\)](https://lmsys.org)

Question: If the FED buys bonds in the secondary market
(A) the money supply will increase. (B) the money supply will decrease. (C) the money supply will not be affected. (D) the discount rate would be affected. Output the answer only.

Assistant A:

The Federal Reserve buys bonds in the secondary market to increase the money supply.

Assistant B:

(A) the money supply will increase.

Follow-up Question: How does it affect my daily life? Give 3 examples.

Assistant A:

The Federal Reserve buys bonds in the secondary market to increase the money supply. This affects your daily life by:

1. Increasing the money supply.
2. Increasing the money supply and lowering the discount rate.
3. Increasing the money supply and lowering the discount rate and increasing the money supply.

Assistant B:

When the Federal Reserve buys bonds in the secondary market, it can have several effects on the economy that can impact your daily life. Here are three examples:

1. Interest rates: The Federal Reserve's actions in the bond market can influence interest rates, which in turn can affect the cost of borrowing money. For example, if the Federal Reserve buys bonds and drives up demand for them, interest rates on things like mortgages, car loans, and credit cards may go up, making it more expensive to borrow money.
2. Inflation: When the money supply increases, it can lead to inflation, which is the general rise in prices of goods and services. This can impact your daily life by making it more expensive to buy things like groceries, gas, and other necessities.
3. Employment: The Federal Reserve's actions in the bond market can also have an impact on employment. For example, if the Federal Reserve's actions lead to economic growth and increased demand for goods and services, it can create more job opportunities and boost employment.

GPT-4 Judgment:

Assistant A provided an incorrect response to the user's question about how the Federal Reserve buying bonds in the secondary market affects daily life. The answer given is repetitive and lacks clear examples of how the action impacts daily life.

On the other hand, **Assistant B** provided a relevant and accurate response to the user's question about the Federal Reserve buying bonds. The answer includes three clear examples of how the action impacts daily life, such as interest rates, inflation, and employment.

Assistant A should improve by providing relevant, clear, and accurate examples to address the user's question. **Assistant B's** response is more helpful, relevant, and detailed.

VMLU (Vietnamese)

Leaderboard of fine-tuned models

#	MODEL	CREATOR	ACCESS	BASE MODEL	EVALUATION DATE	STEM	SOCIAL SCIENCE	HUMANITIES	OTHERS	AVG
1	VNPTAI.O-14B	VNPT AI	Private	Qwen1.5-14B-Chat	11/03/2024	51.64	61.75	58.09	54.51	55.83
2	SeaLLM-7B-v2.5	DAMO Academy	Private	llama-2-7b	09/04/2024	49.35	60.66	55.95	49.05	53.30
3	Vistral-7B-Chat	UONLP x Ontocord	Weight	Mistral-7B-v0.1	16/01/2024	43.32	57.02	55.12	48.01	50.07
4	SeaLLM-7b-v2	DAMO Academy	Weight	llama-2-7b	15/02/2024	39.95	52.02	49.38	45.27	45.79
5	bloomz-7b1	BigScience	Weight	Bloom-7b1	08/01/2024	32.63	45.73	41.85	39.89	38.87
6	T-Llama	FPTU HCM	Weight	llama-2-7b	18/03/2024	32.2	43.15	40.31	36.57	37.28
7	vbd-llama2-7b-50b...	Vin BigData	Weight	llama-2-7b	08/01/2024	31.45	40.34	40.24	39.62	36.98