From words to wisdom

How LLMs and vector databases revolutionize data understanding

BoilingFrogs

Marcin Łapaj

15.03.2025

iteratec

Marcin Łapaj

Architect @ iteratec

iteratec

search history

1993 - 2000

full text search

WWW Wanderer

search history

full text search

1993 - 2000

WWW Wanderer

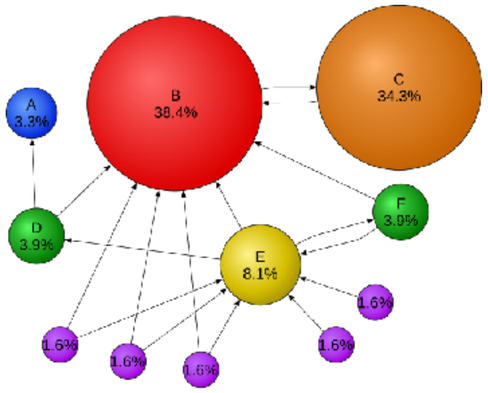
YAHOO!

ALTAVISTA
Technology, Inc.

2000 - 2012

page rank

Google™



iteratec

search history

full text search

1993 - 2000

WWW Wanderer

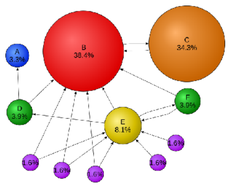
YAHOO!

ALTAVISTA Technology, Inc.

page rank

2000 - 2012

Google



2012 - 2024

things not string

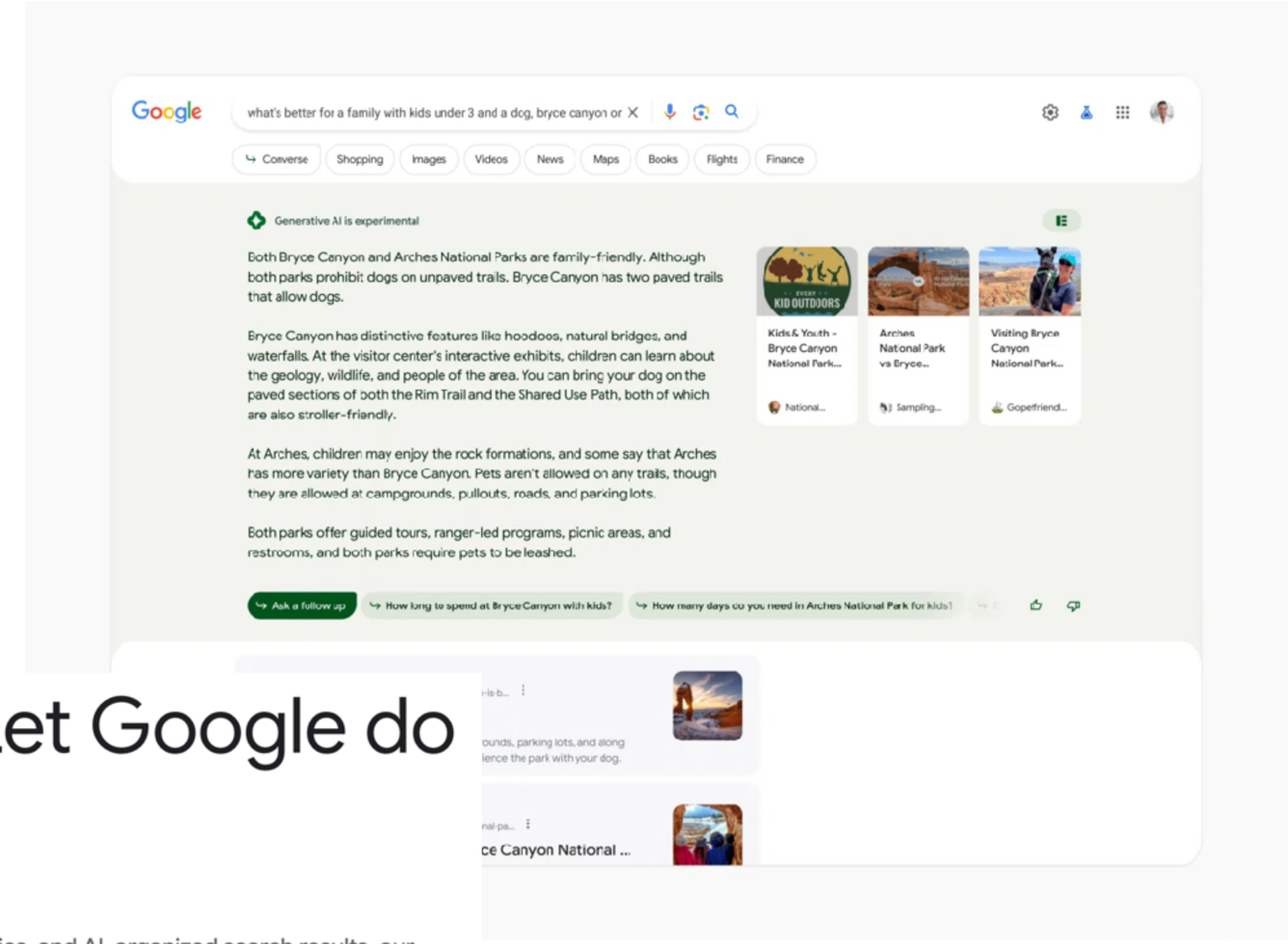
Introducing the Knowledge Graph: things, not strings

May 16, 2012 · 4 min read

A **Amit Singhal**
SVP, Engineering

iteratec

LLM + RAG search on Google



Generative AI in Search: Let Google do the searching for you

May 14, 2024
5 min read

With expanded AI Overviews, more planning and research capabilities, and AI-organized search results, our custom Gemini model can take the legwork out of searching.

Truth is Subjective

Inquisition Maxim

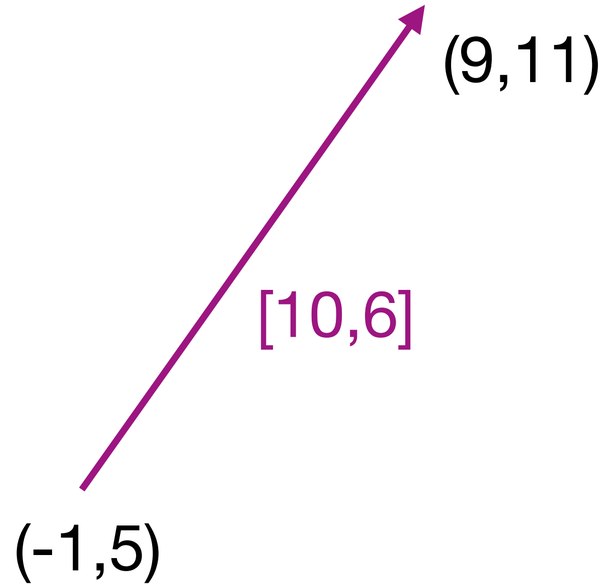
LLMs hallucinate, how to minimize that effect

what are vectors
what is text embedding
what are vector databases
let's build a rag

iteratec

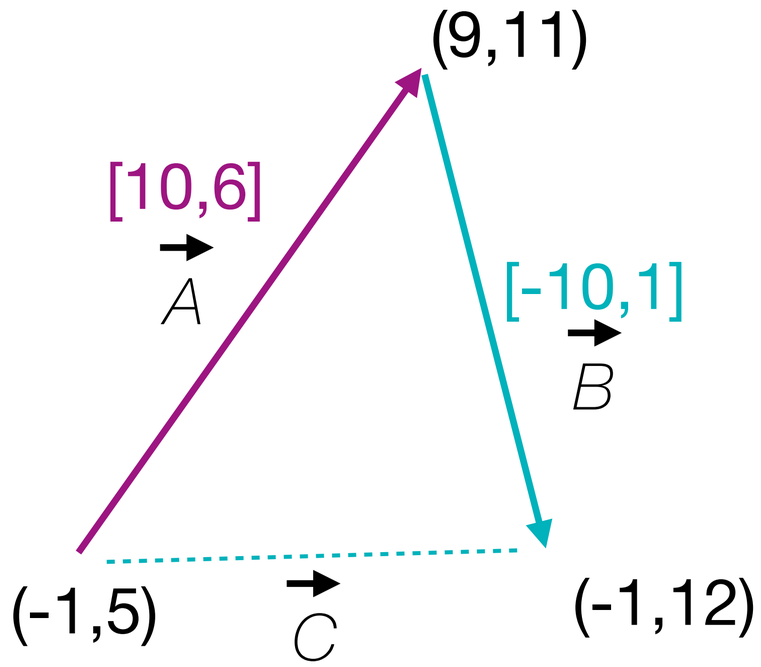
vector

properties



vector

arithmetics



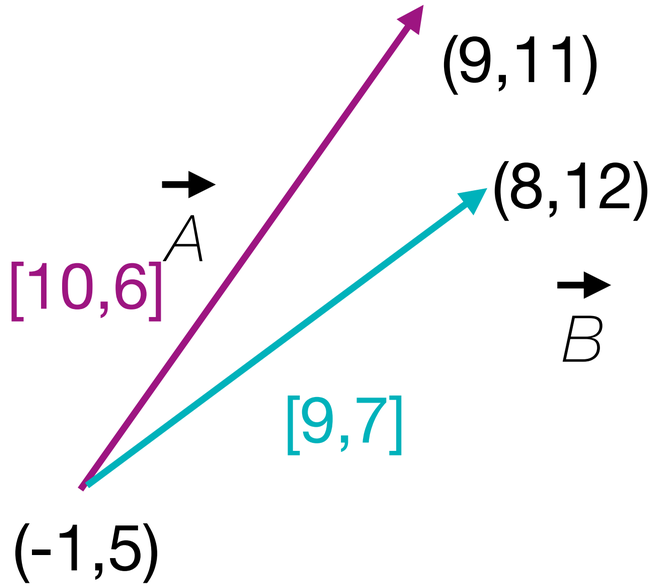
$$\vec{A} + \vec{B} = \vec{C} \quad [0,7]$$

$$\vec{C} - \vec{B} = \vec{A}$$

iteratec

vector

dot product

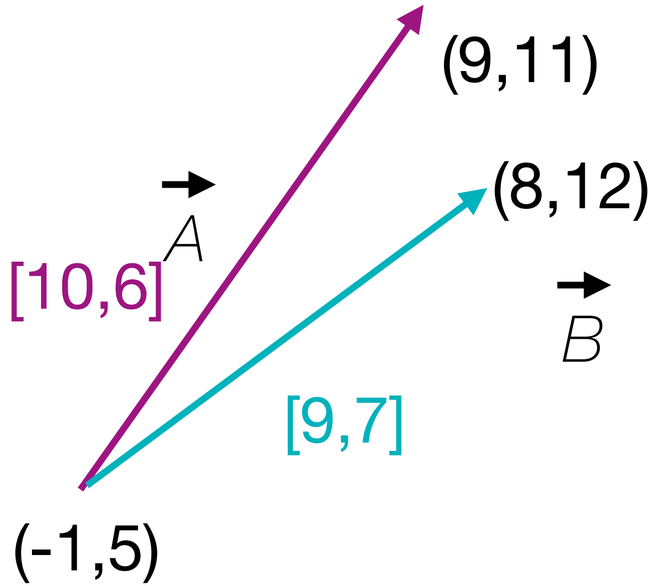


$$\vec{A} \times \vec{B} = a_1 b_1 + a_2 b_2 = 132$$

iteratec

vector

cosine similarity



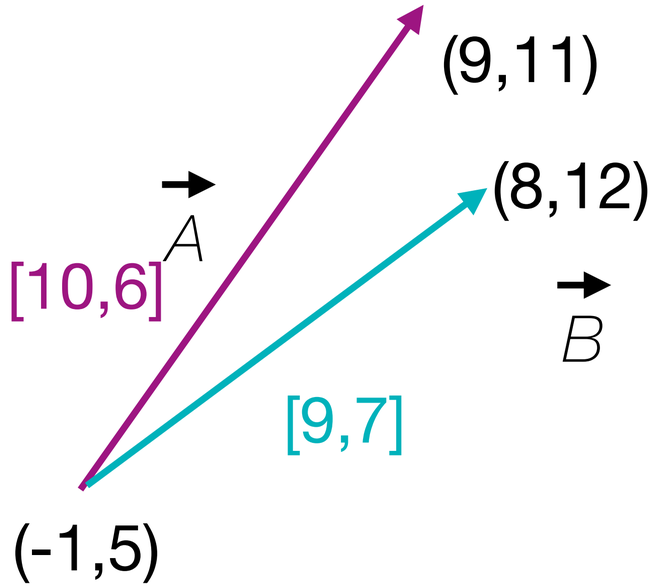
$$\cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \|\mathbf{B}\|}$$

132 / 11.66 × 11.4 = 0.99

iteratec

vector

euclidean distance



$$d(A,B) = \sqrt{(A_1 - B_1)^2 + (A_2 - B_2)^2}$$

$$d(A,B) = \sqrt{(9-10)^2 + (7-6)^2} = \sqrt{1+1} = \sqrt{2} = 1{,}41$$

iteratec

vector

dimensions

[10, 21]

dimensions = 2

[5.59, 42, 3.21]

dimensions = 3

...

[42.231, 432.23, …, 20.213]

dimensions = n

embedding

tokens

The sky is blue

The sky is blueish

embedding

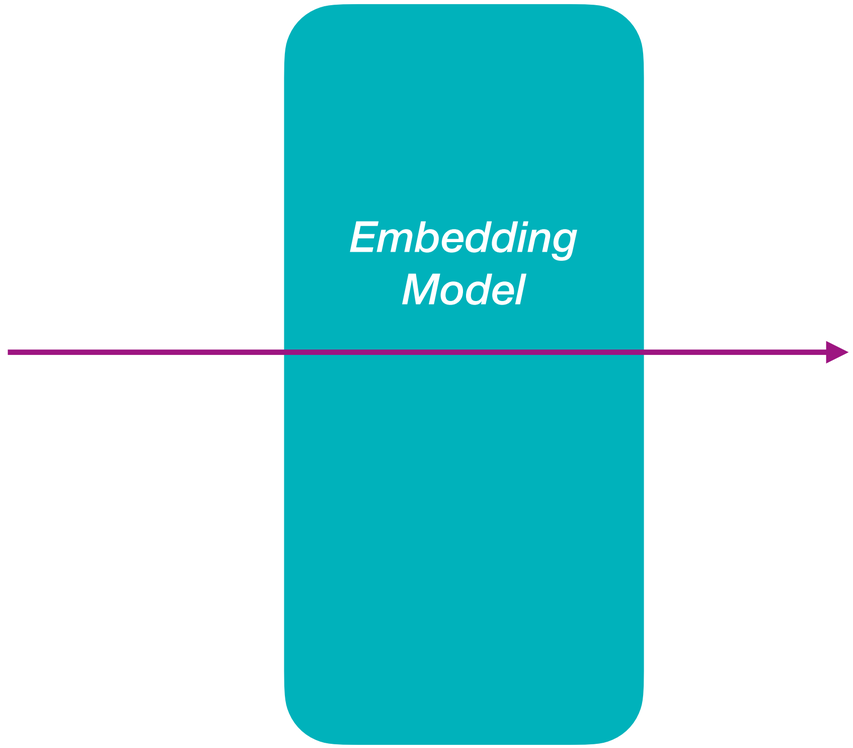
Embeddings are numerical representations, which make it easy for computers to understand the relationship.

Embeddings that are numerically similar are also semantically similar.

iteratec

embedding

In the grim darkness of the far future, there is only war.



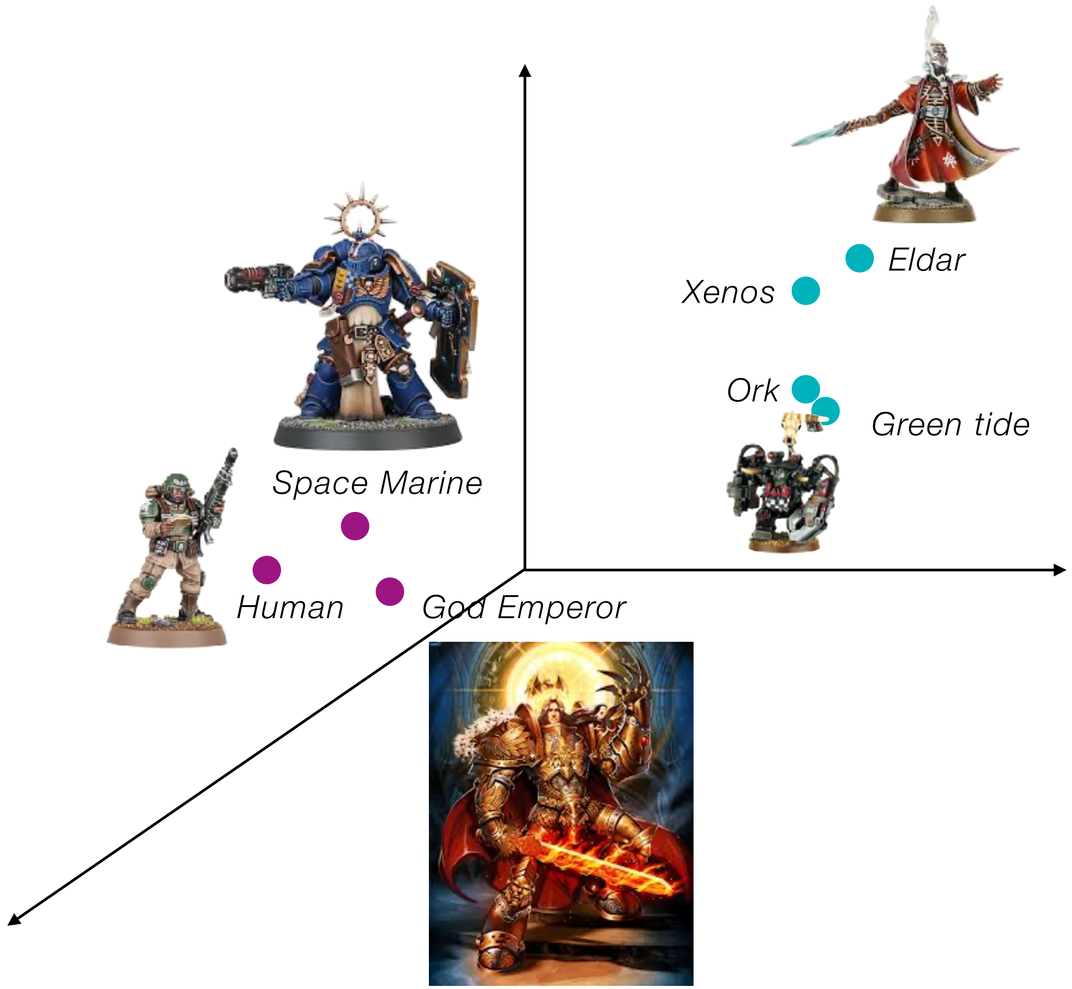
[
-0.0006528618978336453,
-0.03336622938513756,
-0.004293263889849186,
-0.017323806881904602,
-0.0009665339603088796,
0.013140465132892132
...
-0.021770961582660675,
0.006146245170384645 5,
-0.024811107665300 37,
-0.034019481390714645,
0.020276013761758804,
0.010891761630773544,
-0.009465908631682396,
0.006306418217718601,
0.005037596914917231
]

embedding

	living being	human	living at home	feline
Cat	0.92	0.13	0.65	0.8
Human	0.88	0.97	0.79	0.05
Tiger	0.9	0.13	-0.83	0.7

iteratec

embedding



suffer not the unclean to live

Black Templar pray

iteratec

embedding

models

word2vec	*GloVe*	*BERT*	*...*
2013	*2014*	*2018*	

embedding

properties


```
vector('Poland') + vector('capital') = vector('Warszawa')

vector('empress') - vector('emperor') = vector('woman')
```

embedding

resources

Hugging Face

https://huggingface.co/

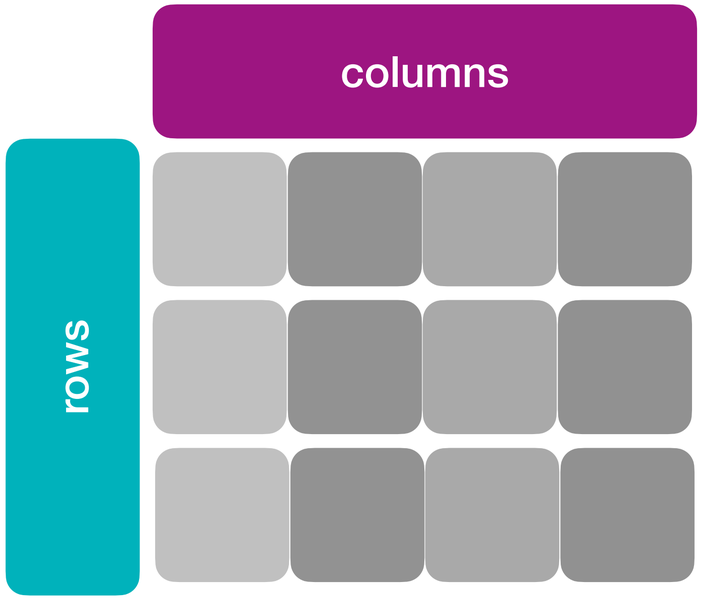
https://ollama.com/

iteratec

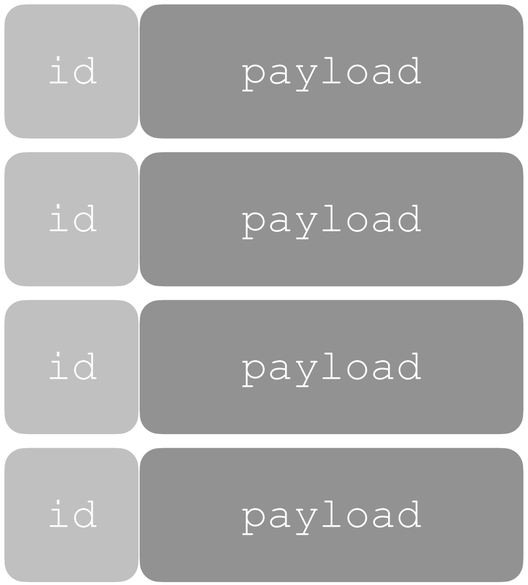
the database

Vector DB

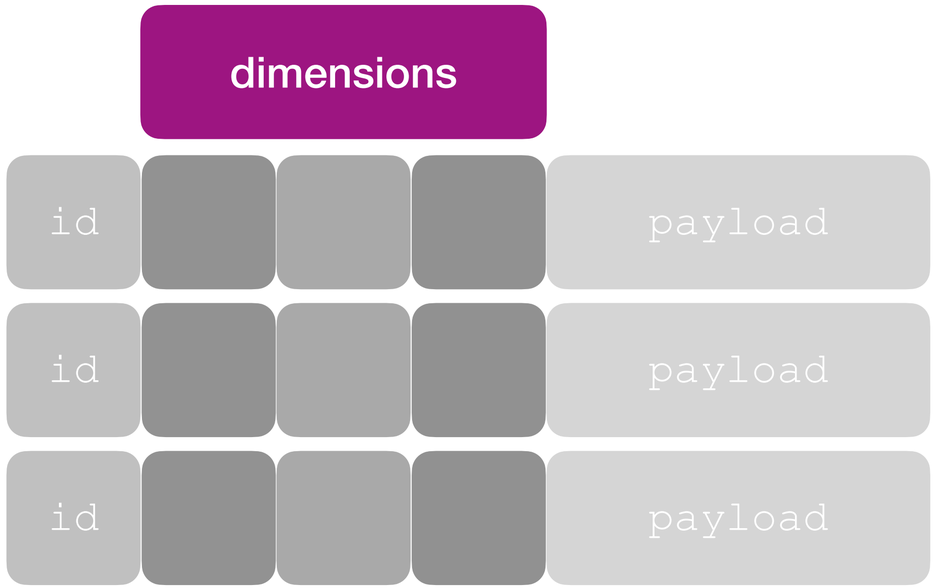
relation



document



vector

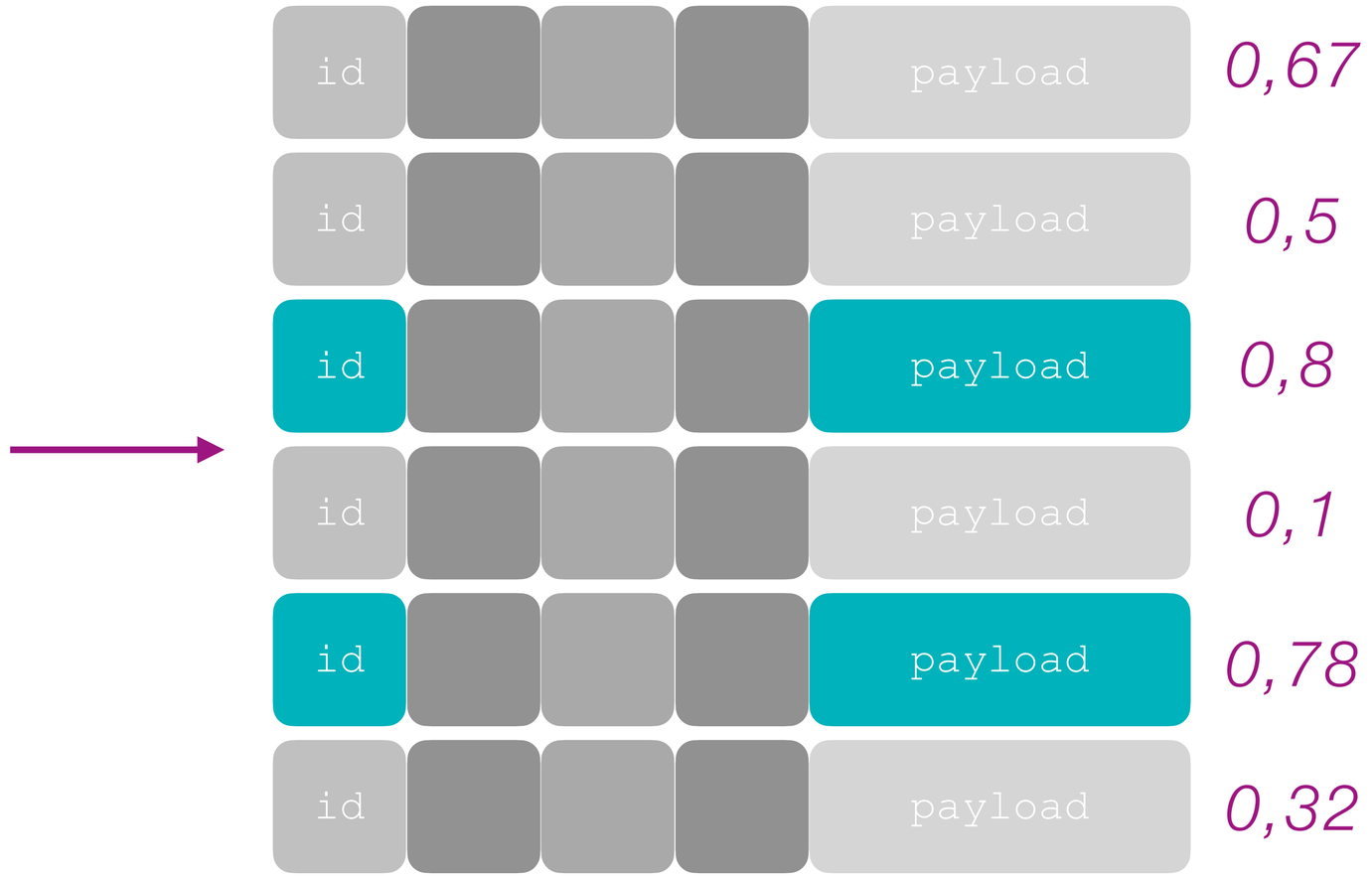


the database

similarity search

query vector

[1, 0,6, 3 ...]

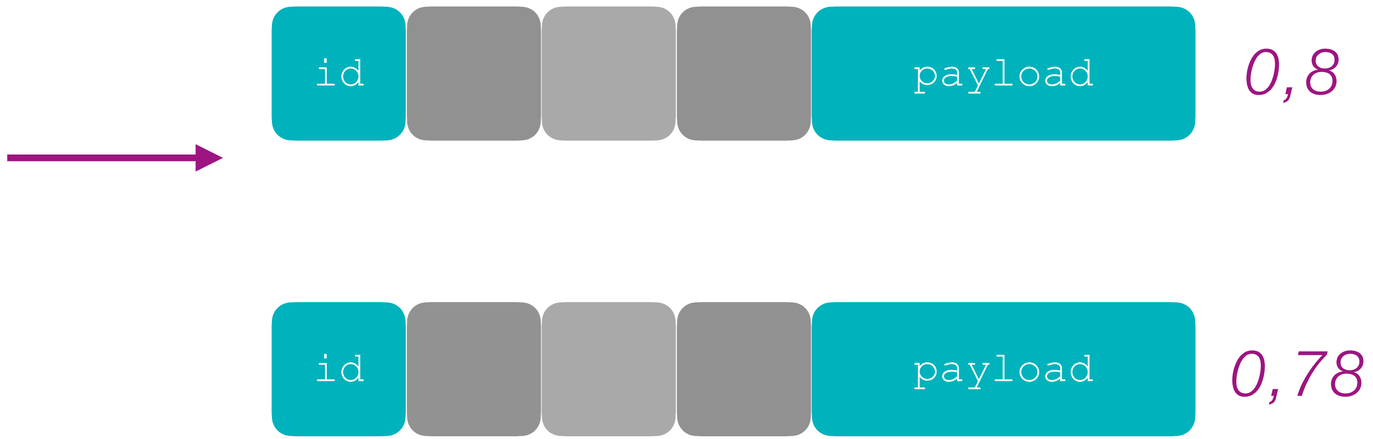


the database

similarity search

query vector

[1, 0,6, 3 ...]



the database

similarity search

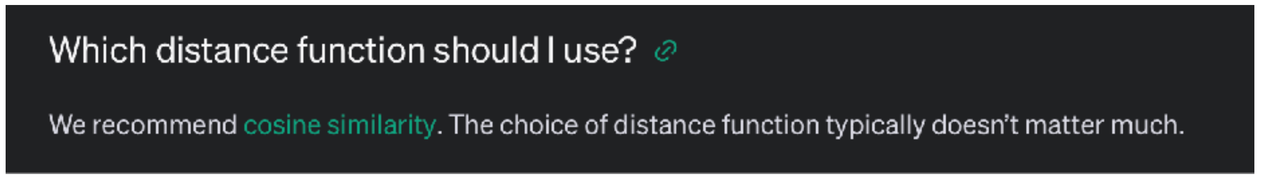
embedding model for Data and Query must be the same

match the distance function you use to the one used to train the vector embedding model

source Oracle

Dot Product | Cosine Similarity | Euclidean Distance

Manhattan Distance | Hammering Distance | Jaccard Distance



or pick one and see what happens

source OpenAI

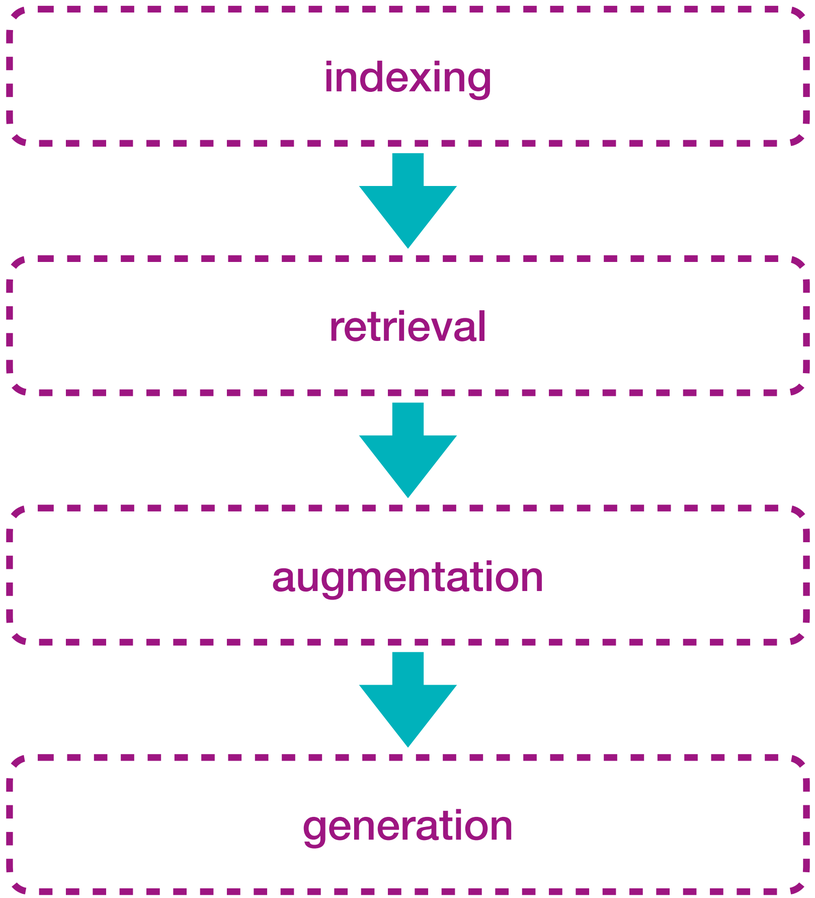
Vector DBs

Elasticsearch

...

RAG

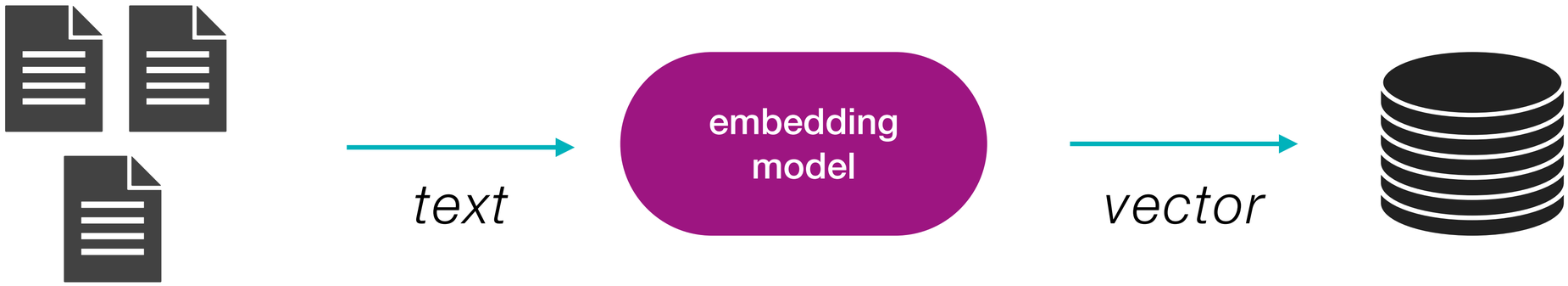
Retrieval augmented generation



iteratec

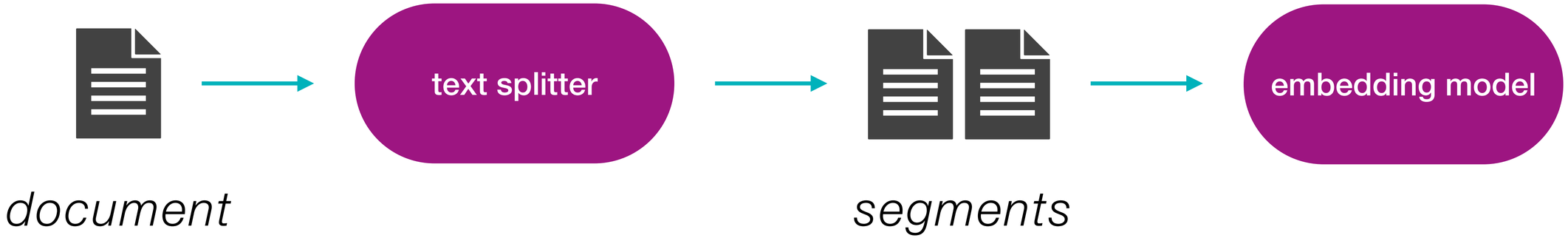
RAG

indexing



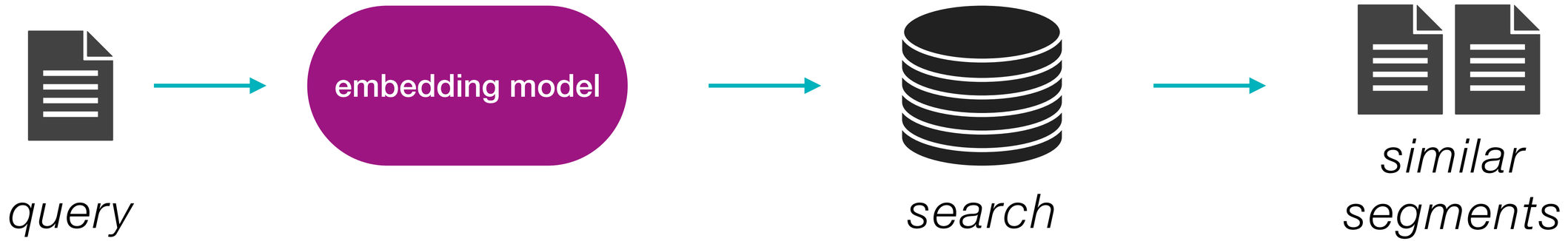
RAG

indexing



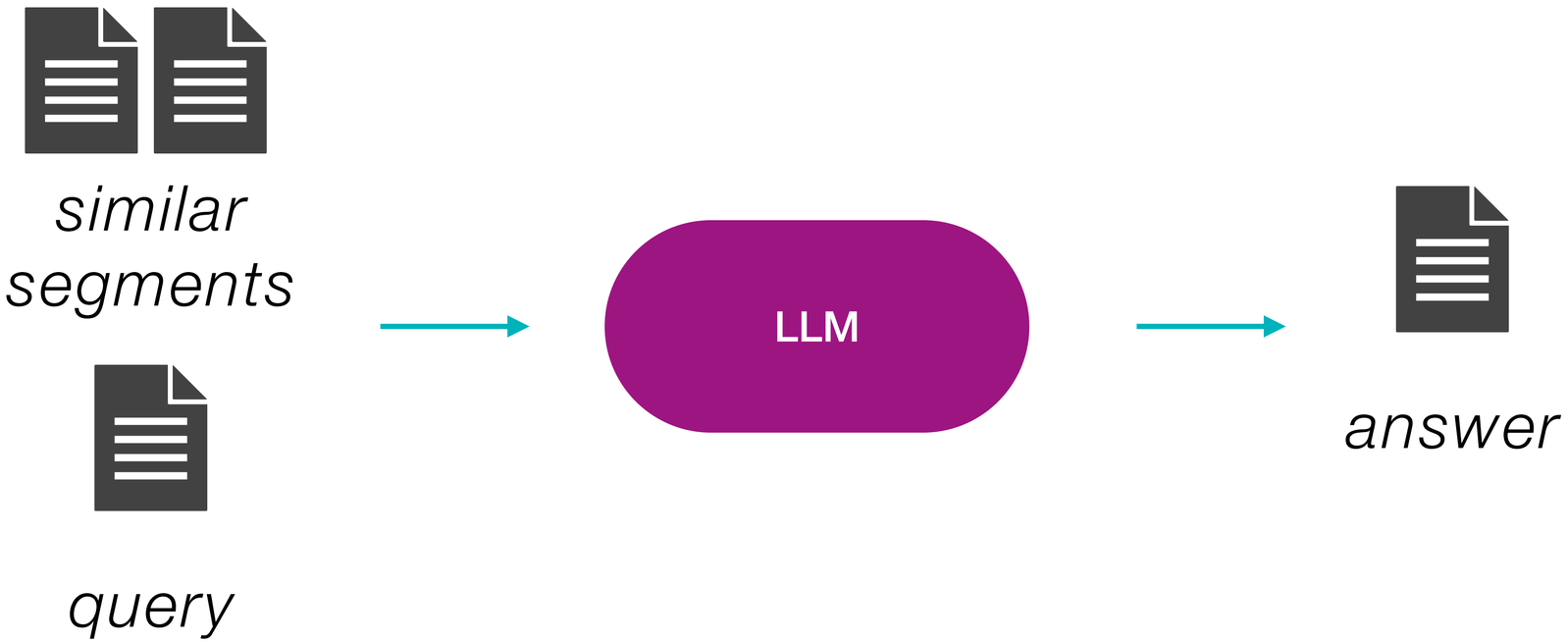
RAG

retrieval



RAG

augmentation & generation



demo

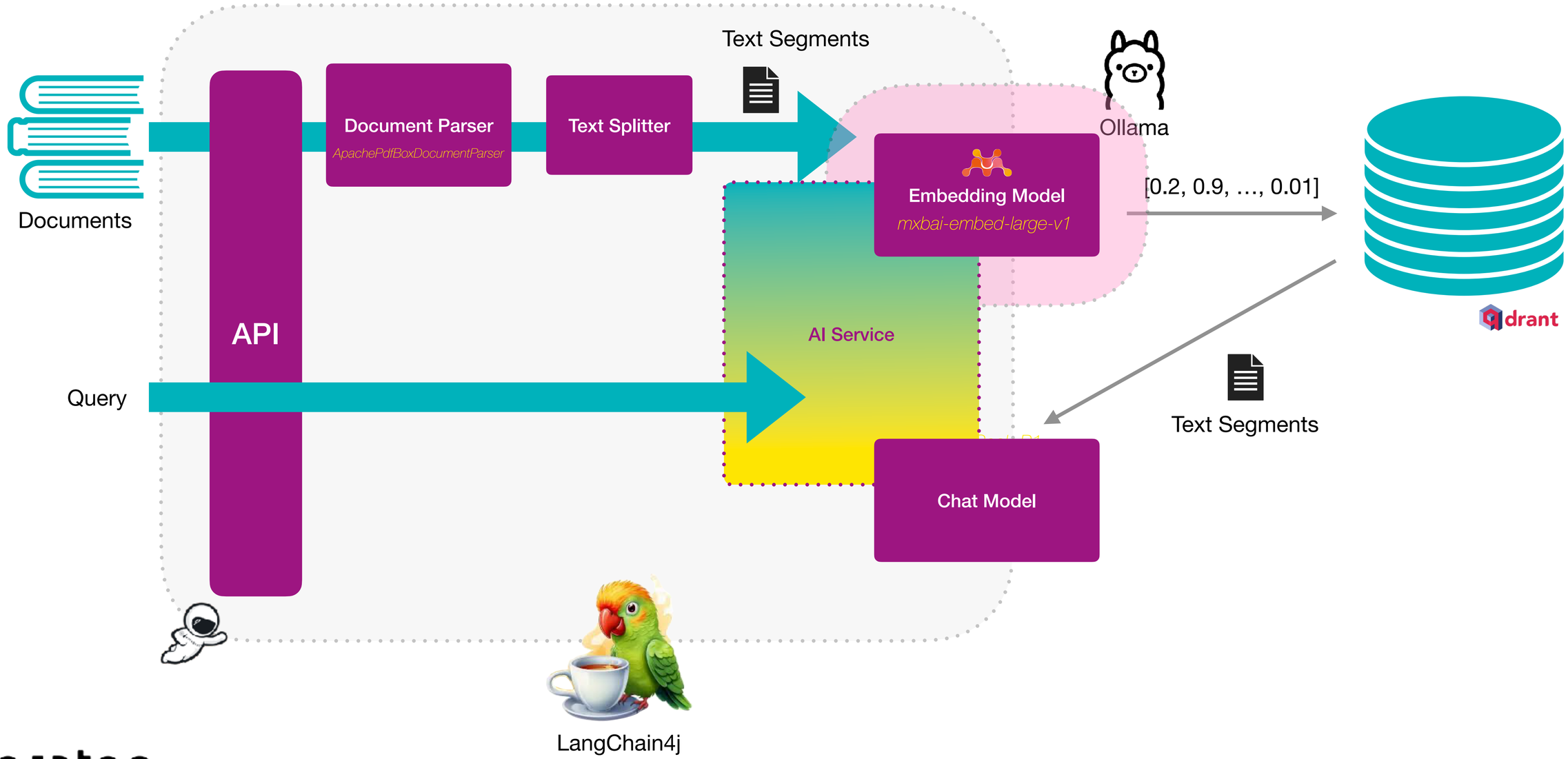
https://docs.langchain4j.dev/

DeepSeek R1

mxbai-embed-large-v1

RAG

Demo architecture



Thank you

The Emperor Protects

Oceń moją prelekcję

BoilingFrogs

iteratec