From Naive to Advanced RAG

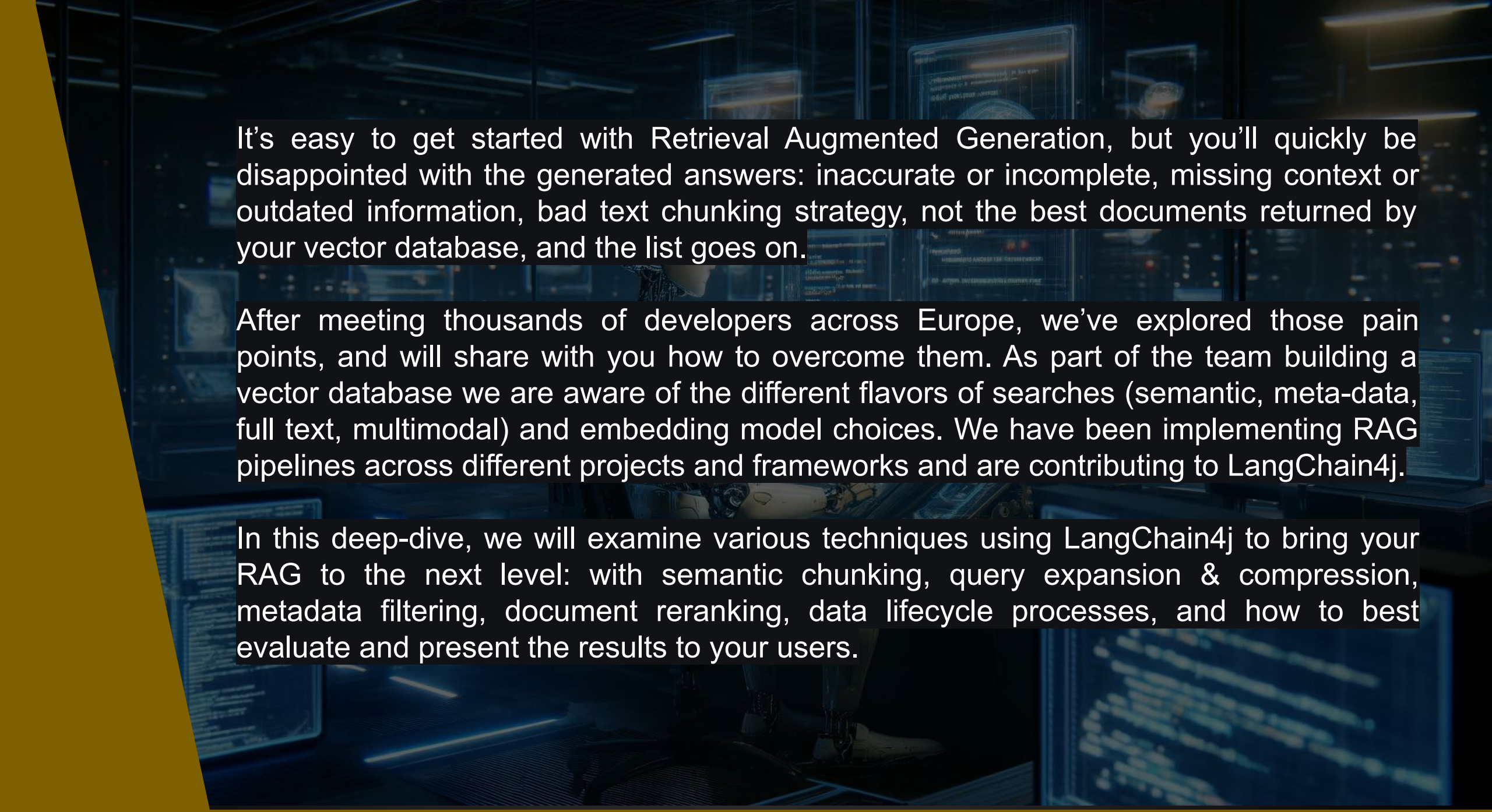
/ The Definitive Guide

CEDRICK LUNVEN
DATASTAX

GUILLAUME LAFORGE
GOOGLE

DEVOXX™

DEEP DIVE



It’s easy to get started with Retrieval Augmented Generation, but you’ll quickly be disappointed with the generated answers: inaccurate or incomplete, missing context or outdated information, bad text chunking strategy, not the best documents returned by your vector database, and the list goes on.

After meeting thousands of developers across Europe, we’ve explored those pain points, and will share with you how to overcome them. As part of the team building a vector database we are aware of the different flavors of searches (semantic, meta-data, full text, multimodal) and embedding model choices. We have been implementing RAG pipelines across different projects and frameworks and are contributing to LangChain4j.

In this deep-dive, we will examine various techniques using LangChain4j to bring your RAG to the next level: with semantic chunking, query expansion & compression, metadata filtering, document reranking, data lifecycle processes, and how to best evaluate and present the results to your users.

Guillaume Laforge

Developer Advocate @ Google Cloud

Google Cloud

- Stuff I do
  - GCP Developer Advocate, focused on Generative AI, serverless solutions & service orchestration
  - Apache Groovy founder
  - Java Champion
  - Cast Codeurs podcast
- AI
  - LangChain4j committer
  - Google Cloud Machine Learning APIs

Cédrick Lunven

Software Engineer @ DataStax

DATASTAX

- Stuff I do
  - Dev Ecosystem @DS
  - Tools (sdk, cli, plugins)
  - Dev Advocate
  - Creator of ff4j (ff4j.org)
- AI
  - CTO GoodBards
  - DataStax AI products
  - Contributor Langchain4j/SpringAI

1. Introduction
   - Generative AI and LLM
   - Prompt Engineering
   - Limitations and Why RAG
   - LangChain4j Overview

2. Naive RAG
   - Ingestion Principles
   - Query Principles

3. Advanced RAG : Ingestion
   - Loading and Parsing
   - Vectors, Embedding and Similarity
   - Introducing Vector Databases
   - Chunking
   - Embedding

Break! 15 min.

4. Advanced RAG : Query
   - **Query Preprocessing**
     - Query Preprocessing
     - Query Transformations
   - **Vector Searches**
     - Filterings and metadata
     - Projections and Sorting
   - **Question post processing**
     - Reranking
     - Recursive algorithms
     - Consolidation

5. Quality and Data Governance
   - RAG evaluation
   - Security
   - Data Lifecycle

All the code and the slides are available online

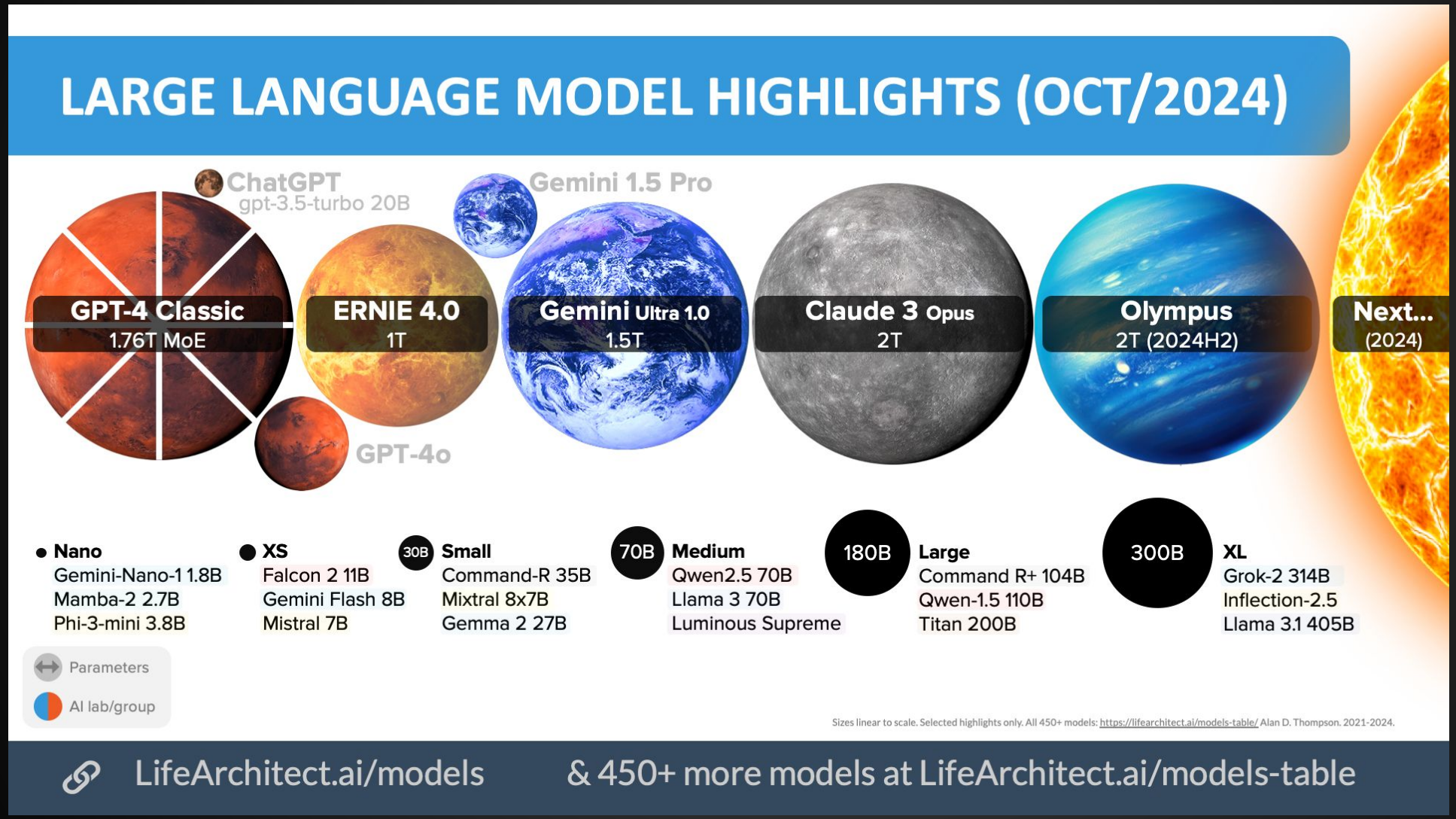
github.com/
datastaxdevs/
conference-2024-devoxx

1. Introduction

- Generative AI and LLM
- Prompt Engineering
- Limitations and Why RAG
- LangChain4j Overview

Large Language models

https://lifearchitect.ai/models/



So what are Large Language Models?

- Transformer-based neural network architecture that can **recognize**, **predict**, and **generate** human language
- Trained on huge corpuses of text, in various languages and domains
  - *Ex: PaLM 2 learned 340 billion **parameters**, and trained over 3.6 trillions of **tokens***
- Learn the **statistical relationships between words and phrases**, as well as the patterns of human language
- Can be **fine-tuned** for specific tasks or domain knowledge

Generative AI use cases 2024

Language

- Writing
- Summarization
- Ideation
- Classification
- Sentiment analysis
- Extraction
- Chat
- Search

Code

- Code generation
- Code completion
- Code chat
- Code conversion

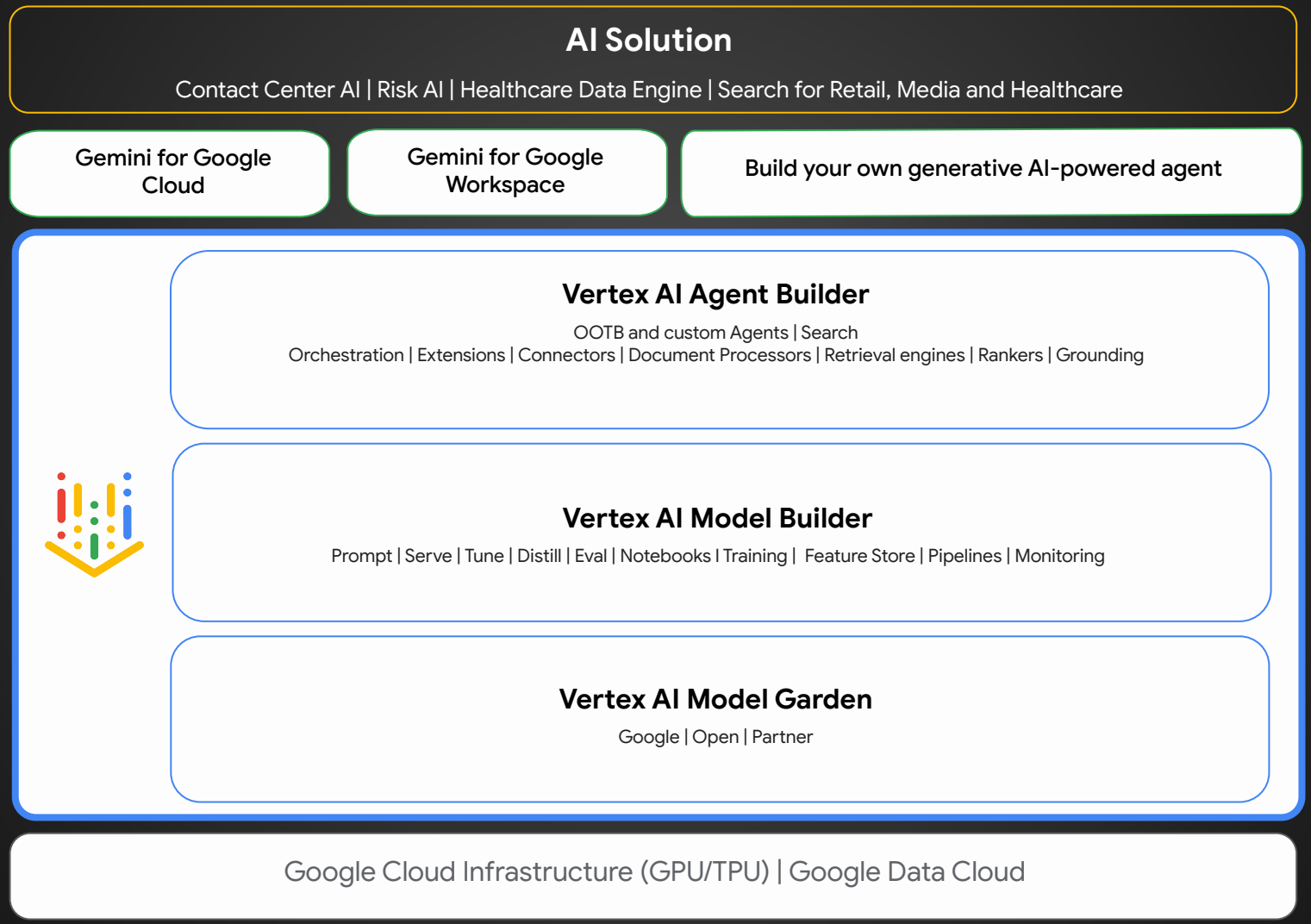
Speech

- Speech to text
- Text to speech
- Audio transcription
- Live voice streaming assistant

Vision

- Image generation
- Image editing
- Captioning
- Image Q&A
- Image search
- Video descriptions

Gemini, Imagen, Vertex AI...



How to build effective prompts ?

CONTEXT

[Roles, Persona, Audience] : You are an assistant targeting Java developers

[Objectives] : Your mission is to provide helpful answers

[Constraints] : Format, Style, Must have, Boundaries

[Question] **(inputs)** Question, Task, Entity, Completions

SAMPLE

[Techniques] One-shot Prompt, few shots prompts, check questions

Configure LLM generation parameters

TEMPERATURE

Tune the degree of randomness.

1

More Creative tasks
- Content Generation
- Can hallucinate more

More Accurate tasks
- Summarization
- Q&A

0

TOP P

Smallest set of words whose cumulative probability >= P

```
p = .8
```

java	.51
ia	.23
langchain	.11
spring	.08
…	

TOP K

The first K words ordered by their p (decreasing)

```
K = 2
```

java	.51
ia	.23
langchain	.11
spring	.08
…	

TOKENS

Size of the generated response.

PRO

- **Detailed/In-Depth**
- **Comprehensive**
- **Completion**

CONS

- **Lower Precision**
- **Processing Time**
- **Cost**
- **Repetitions**

Can we do better ?
Limitations of Prompt Engineering

LLMs...

...can be outdated *(training cut-off date)*

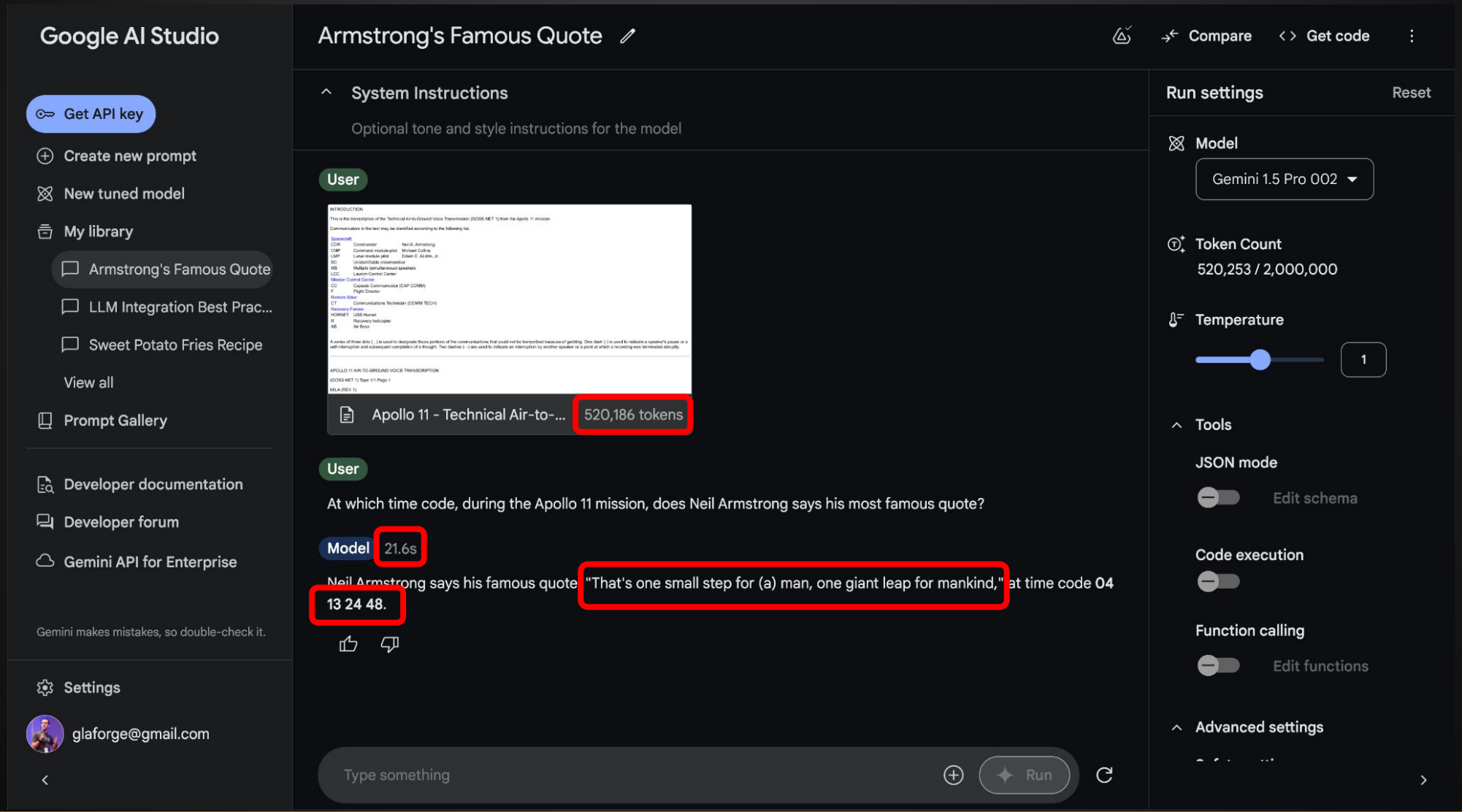
...don't know about ***your*** **data**

...aren't tuned *(hard to steer)*

...are **hallucinating** if not properly prompted

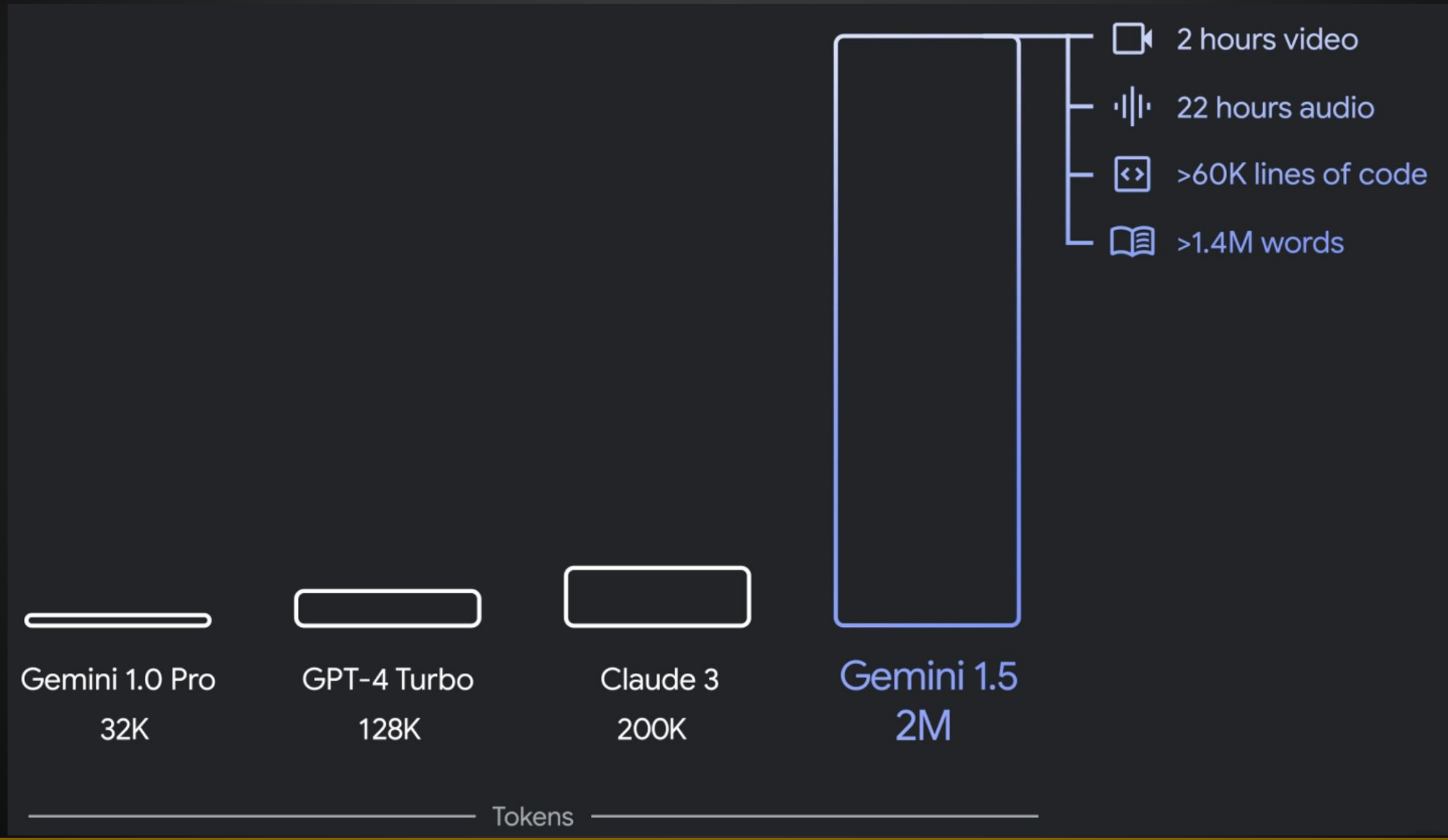
...work with limited input context windows *(can't feed all docs)*

Speaking of (large) context windows...
Gemini 1.5 Pro supports 2M input token windows



Speaking of (large) context windows...

Gemini 1.5 Pro supports 2M input token windows



Retrieval Augmented Generation

CONTEXT

[Roles, Persona, Audience] : You are an assistant targeting Java developers

[Objectives] : Your mission is to provide helpful answers

[Constraints] : Format, Style, Must have, Boundaries

[Question] **(inputs)** Question, Task, Entity, Completions

SAMPLE

[Techniques] One-shot prompt, few shot prompts, check questions

[Document sources] Your relevant document extracts

LangChain4j

Build GenAI Application with JAVA

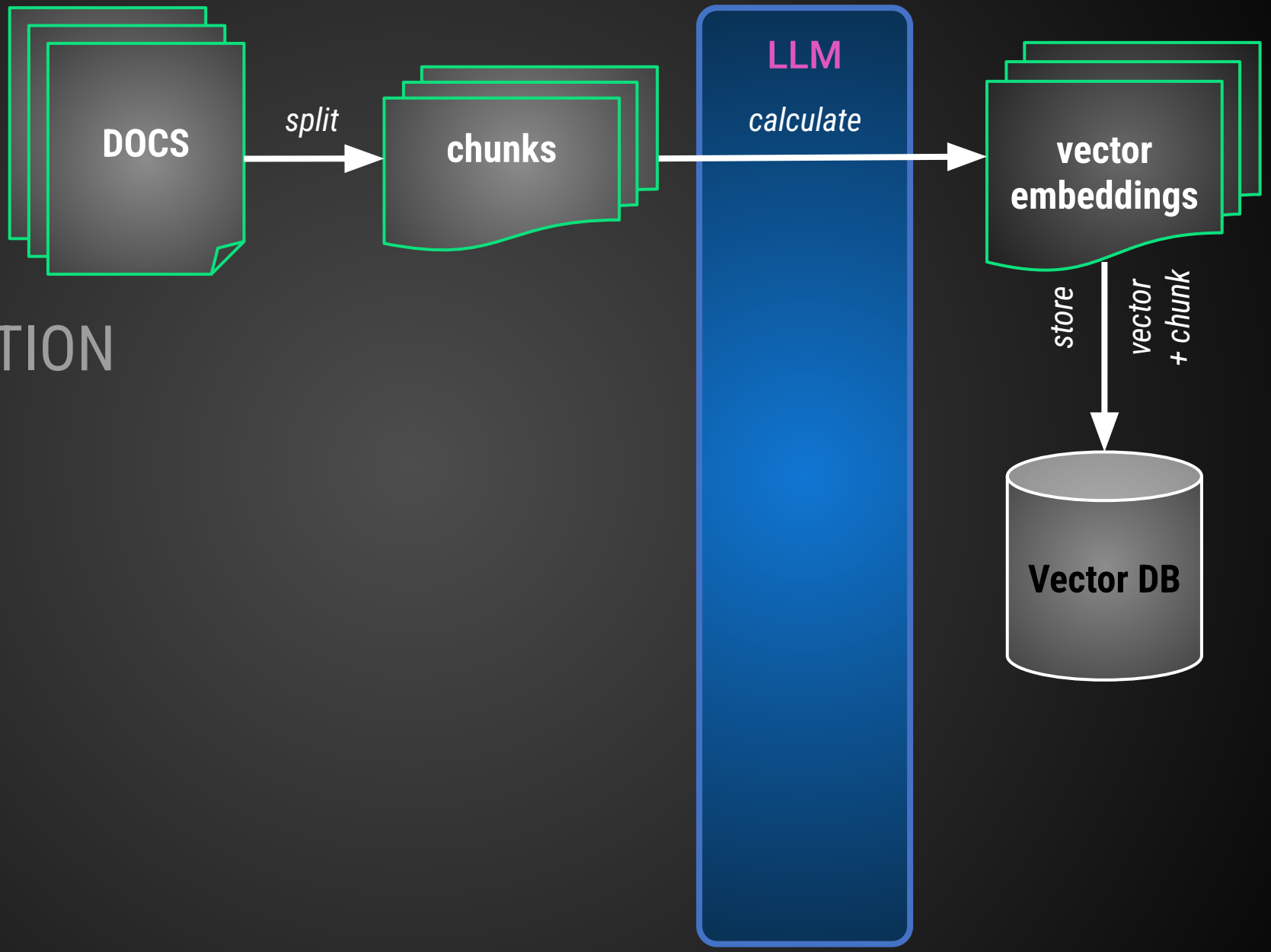
Starred 4.6k

ChatLanguage Model	Moderation Model
Language Model	Scoring Model
Image Model	Embedding Model

2. Naive RAG

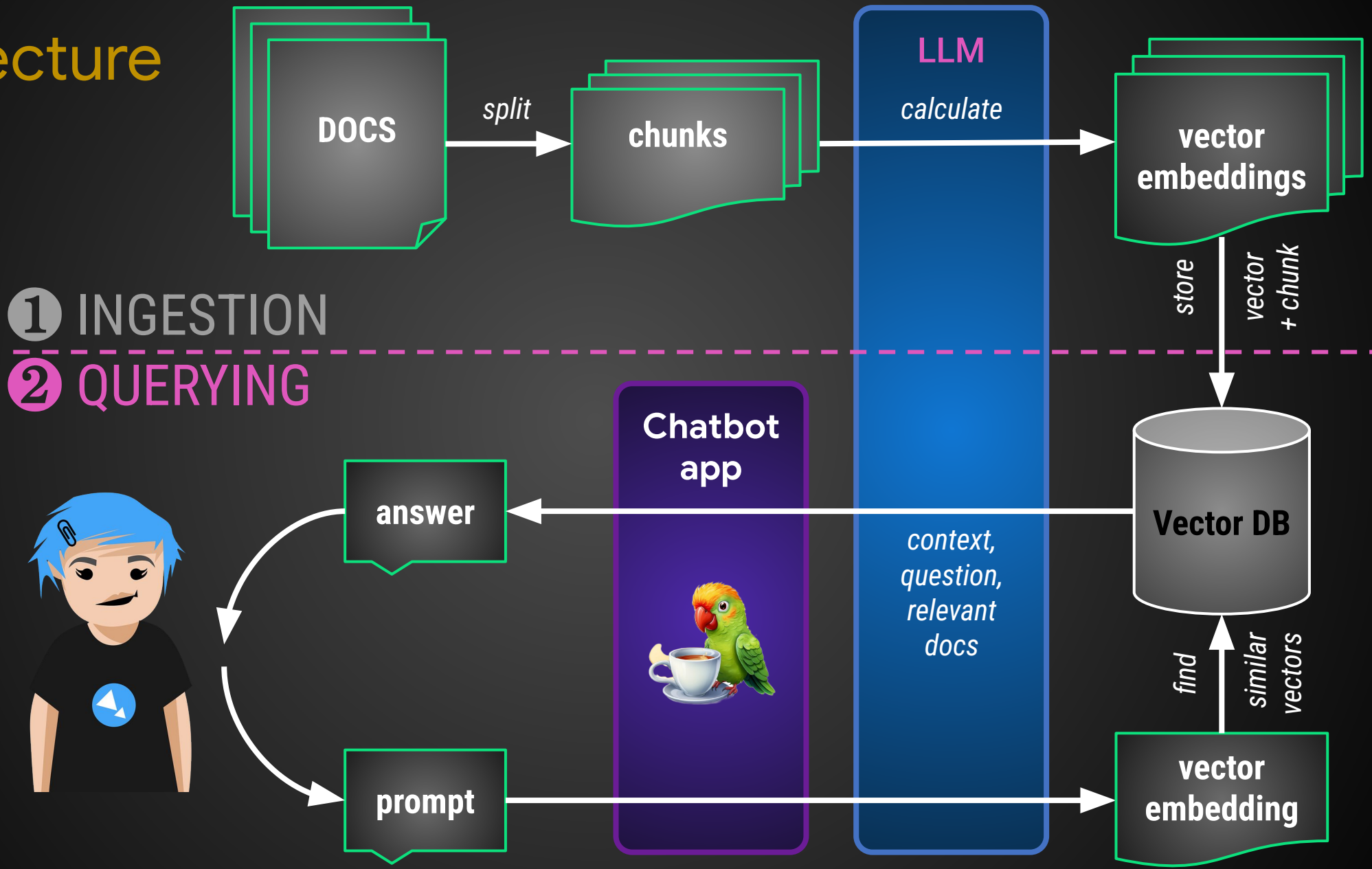
- Ingestion Principles
- Query Principles

Architecture

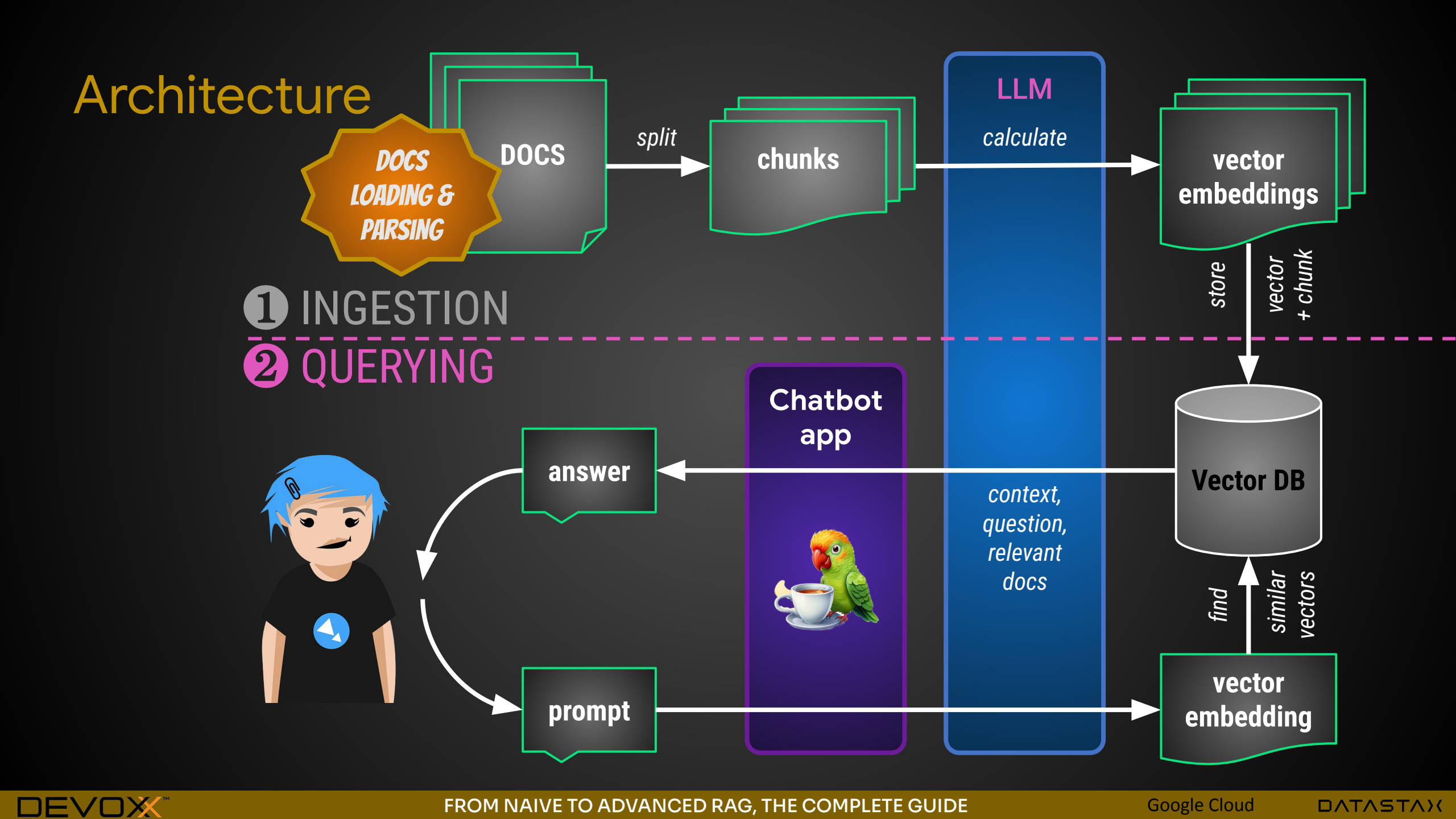


❶ INGESTION

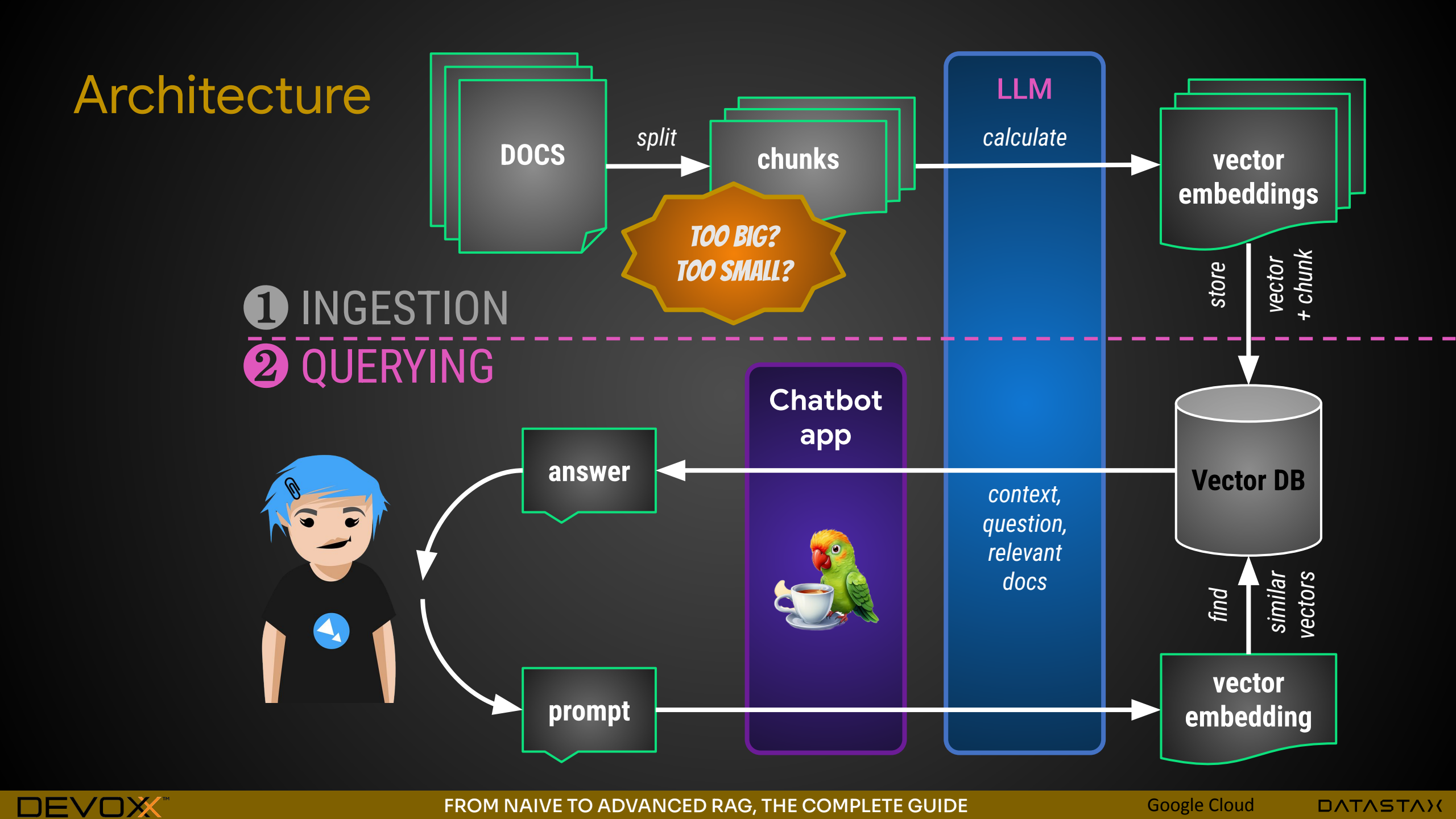
Architecture



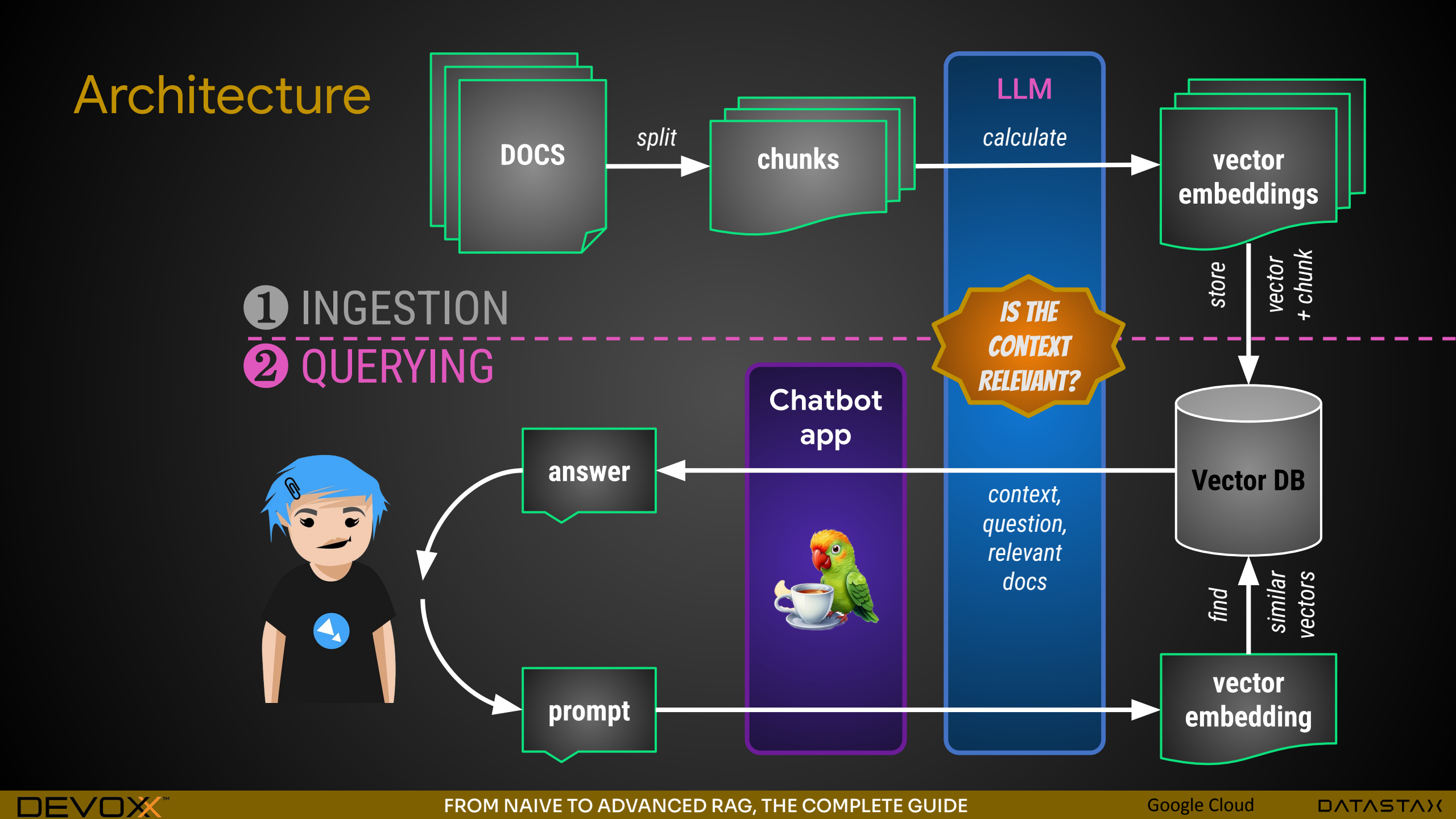
Architecture



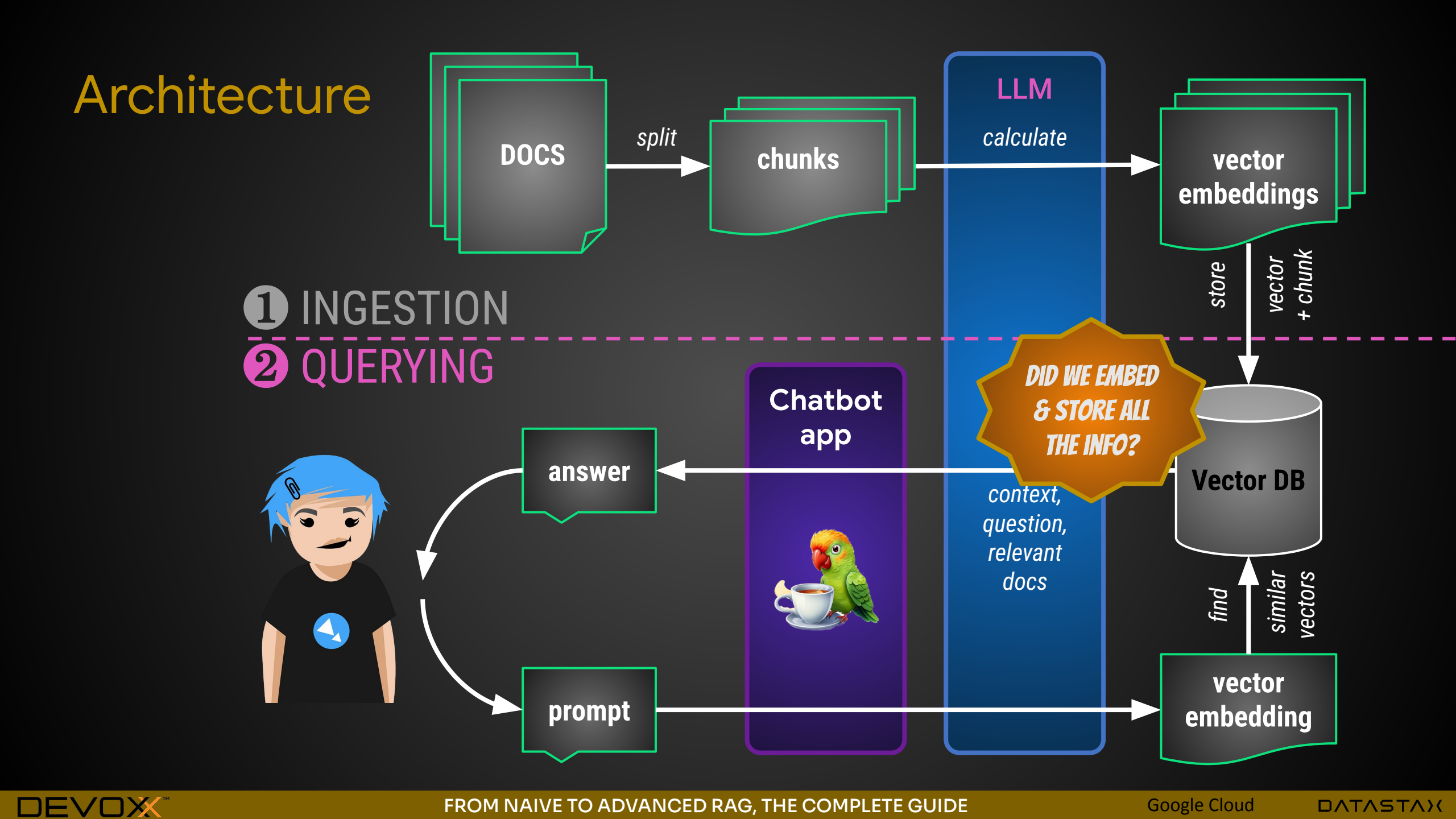
Architecture



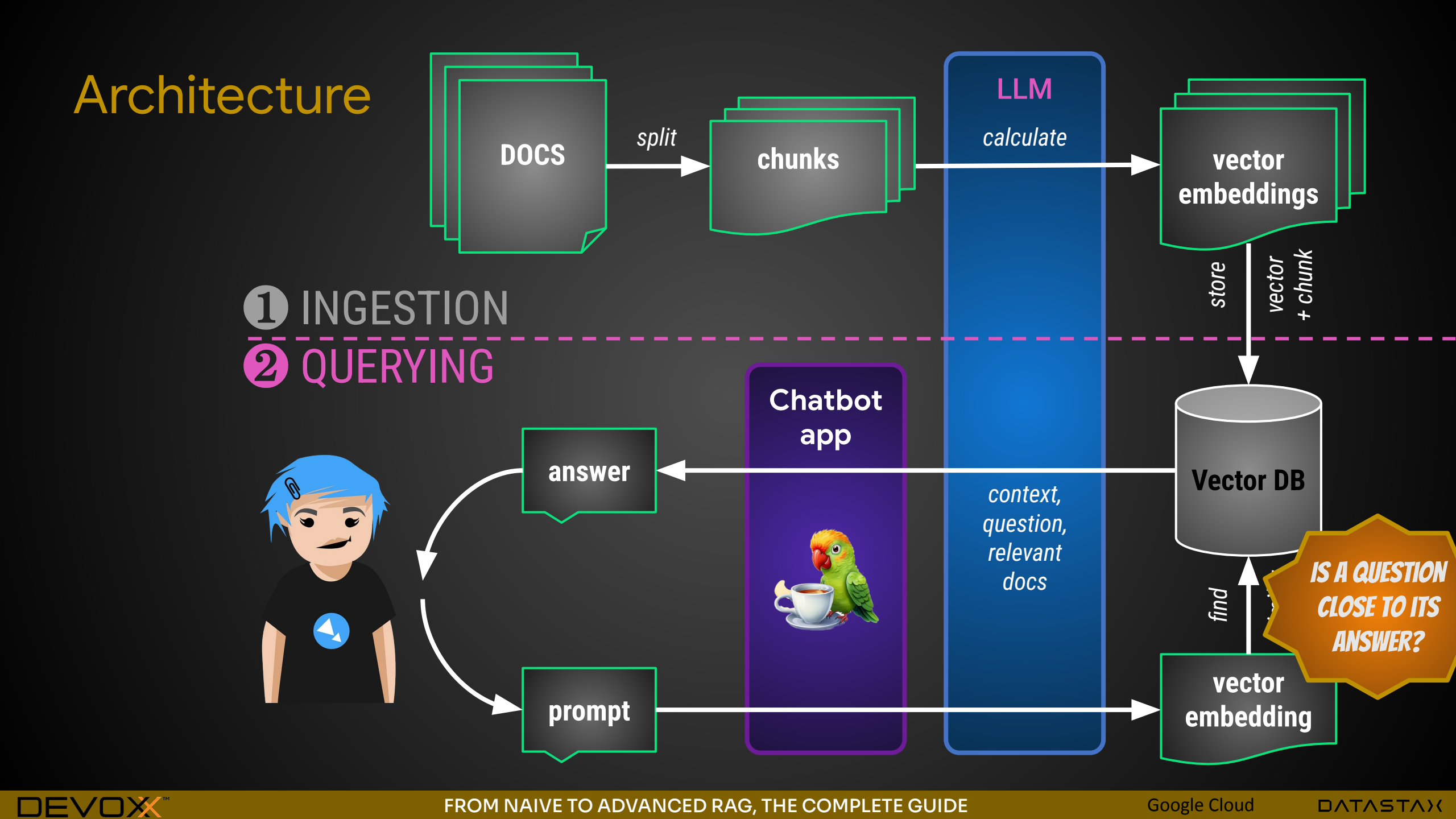
Architecture



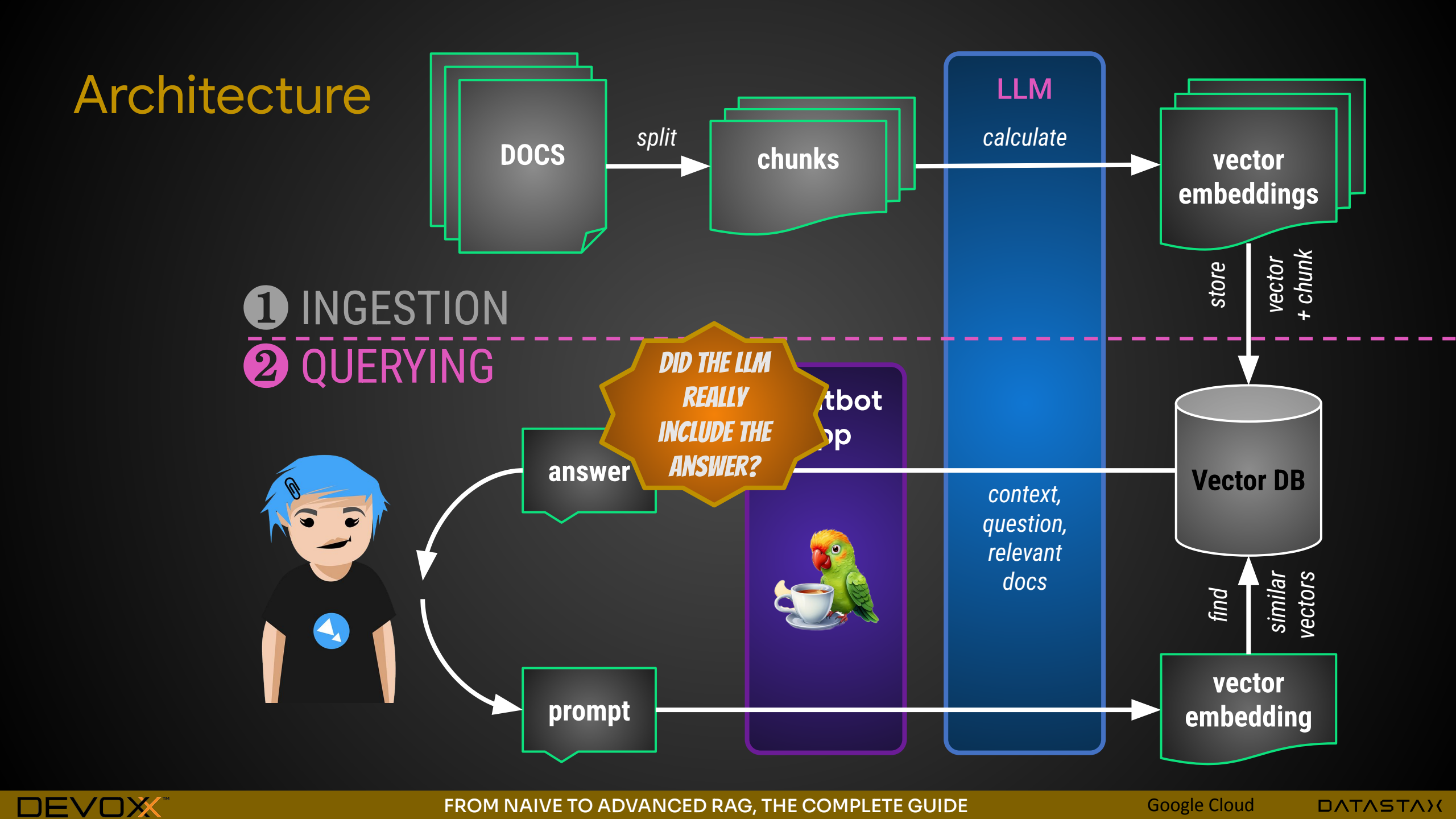
Architecture



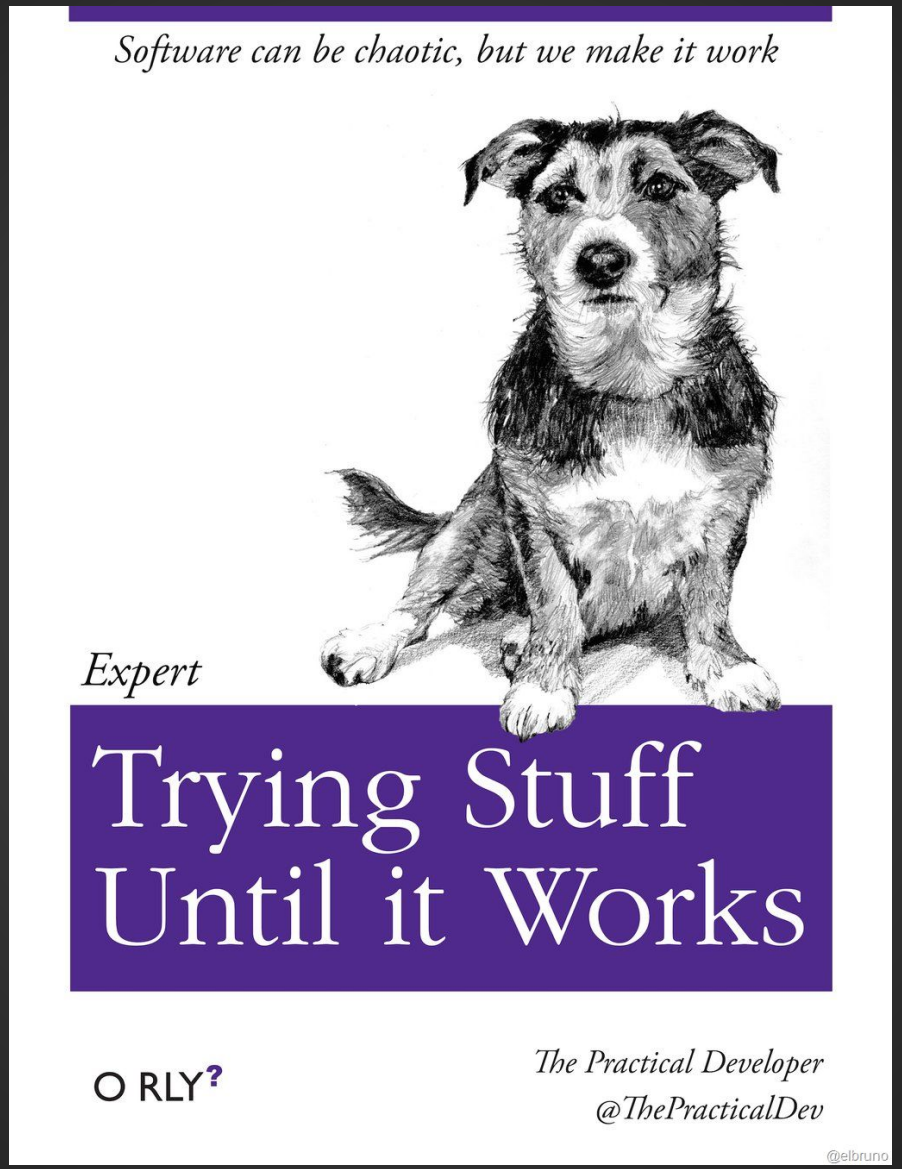
Architecture



Architecture



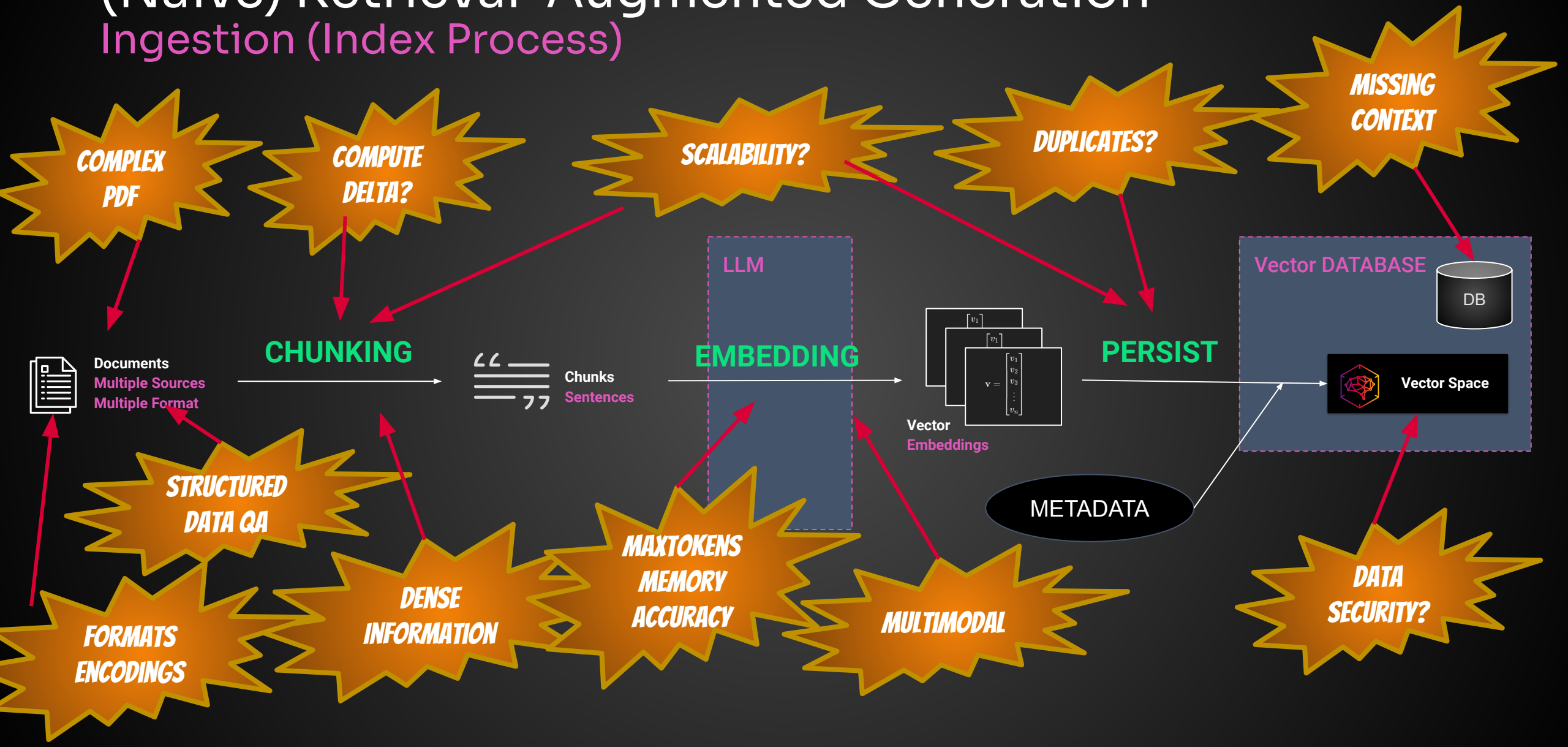
DEMO



_20_naive_rag_astra

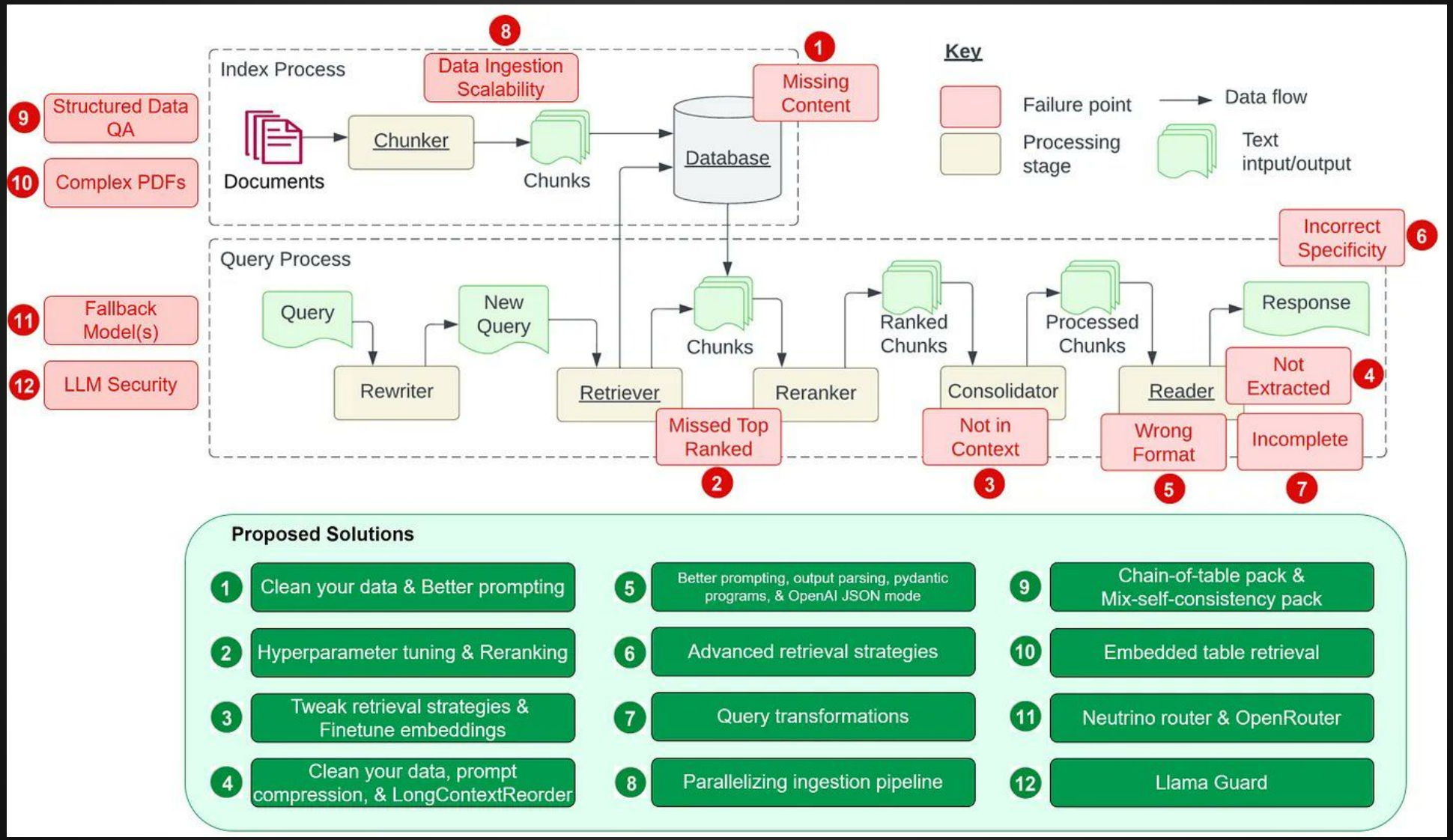
(Naive) Retrieval-Augmented Generation

Ingestion (Index Process)



The 10 pitfalls of RAG

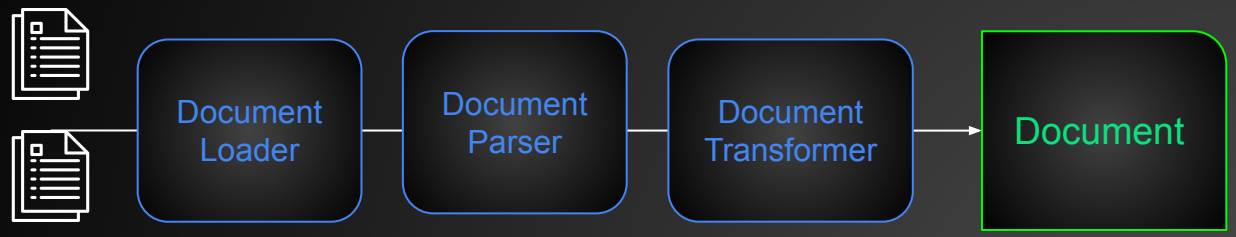
we should normalize steps



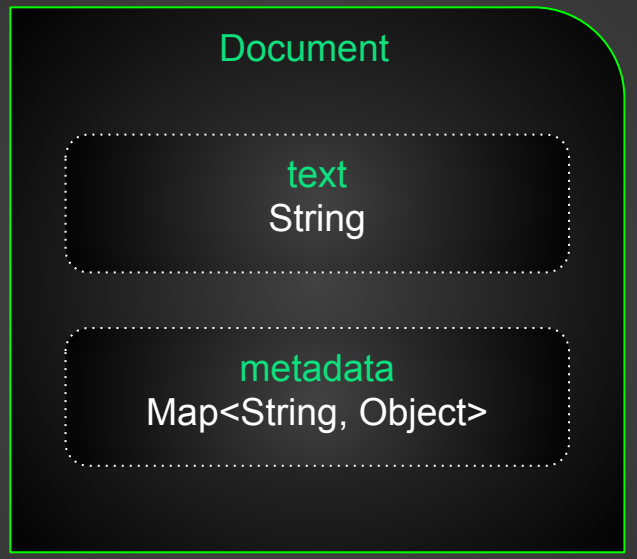
3. Advanced RAG: Ingestion

- **Loading and Parsing**
- Vectors, Embedding and Similarity
- Introducing Vector Databases
- Chunking
- Embedding

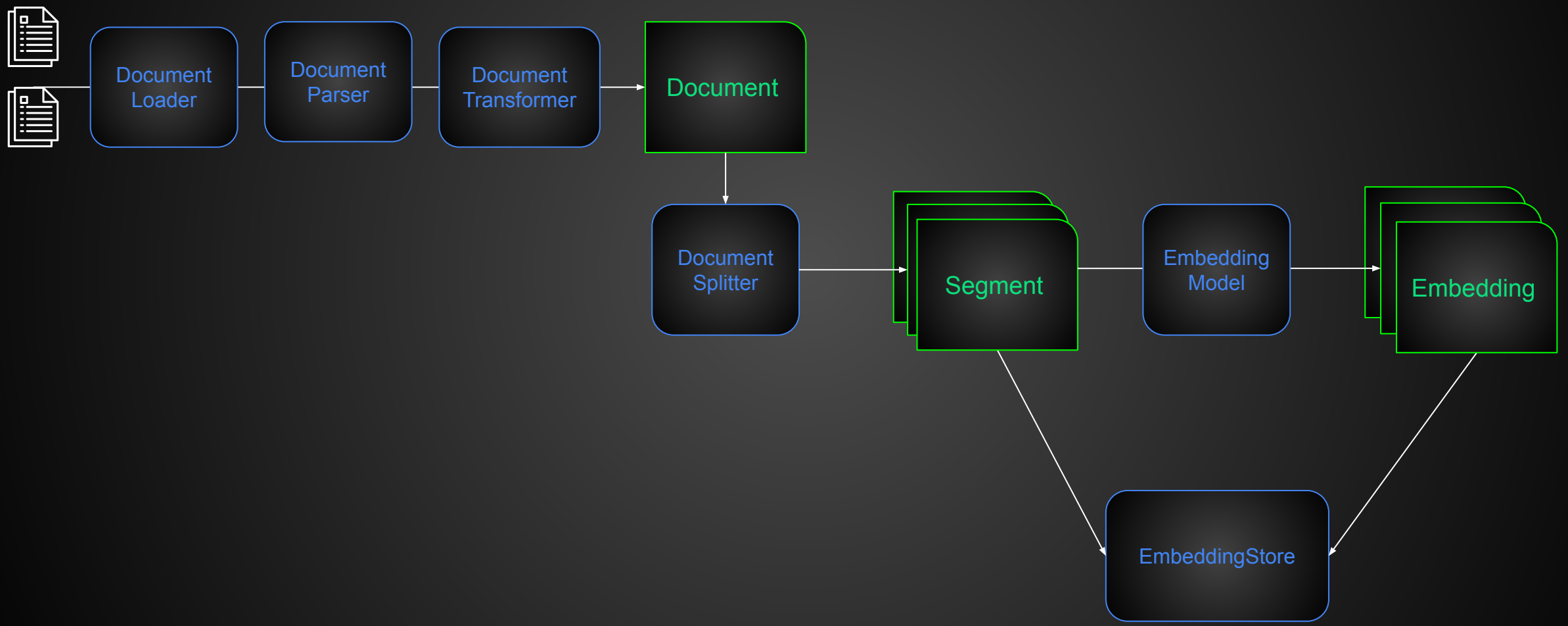
Ingestion Process



Document



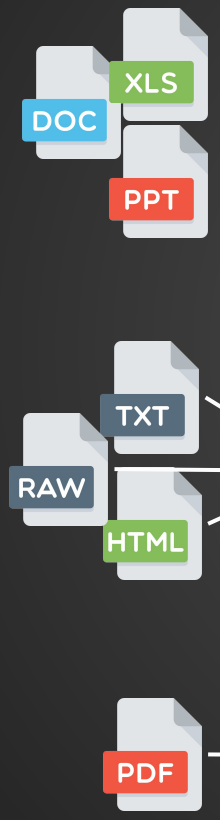
Ingestion Process



Document Ingestion (part 1)

Document LOADERS
(DocumentSource)

- FileSystem
- URL
- Amazon S3
- Google Cloud Storage
- Azure Blob Storage
- Github
- Tencent
- selenium

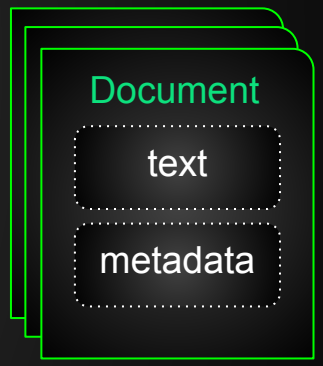


Document PARSERS
(DECODING)

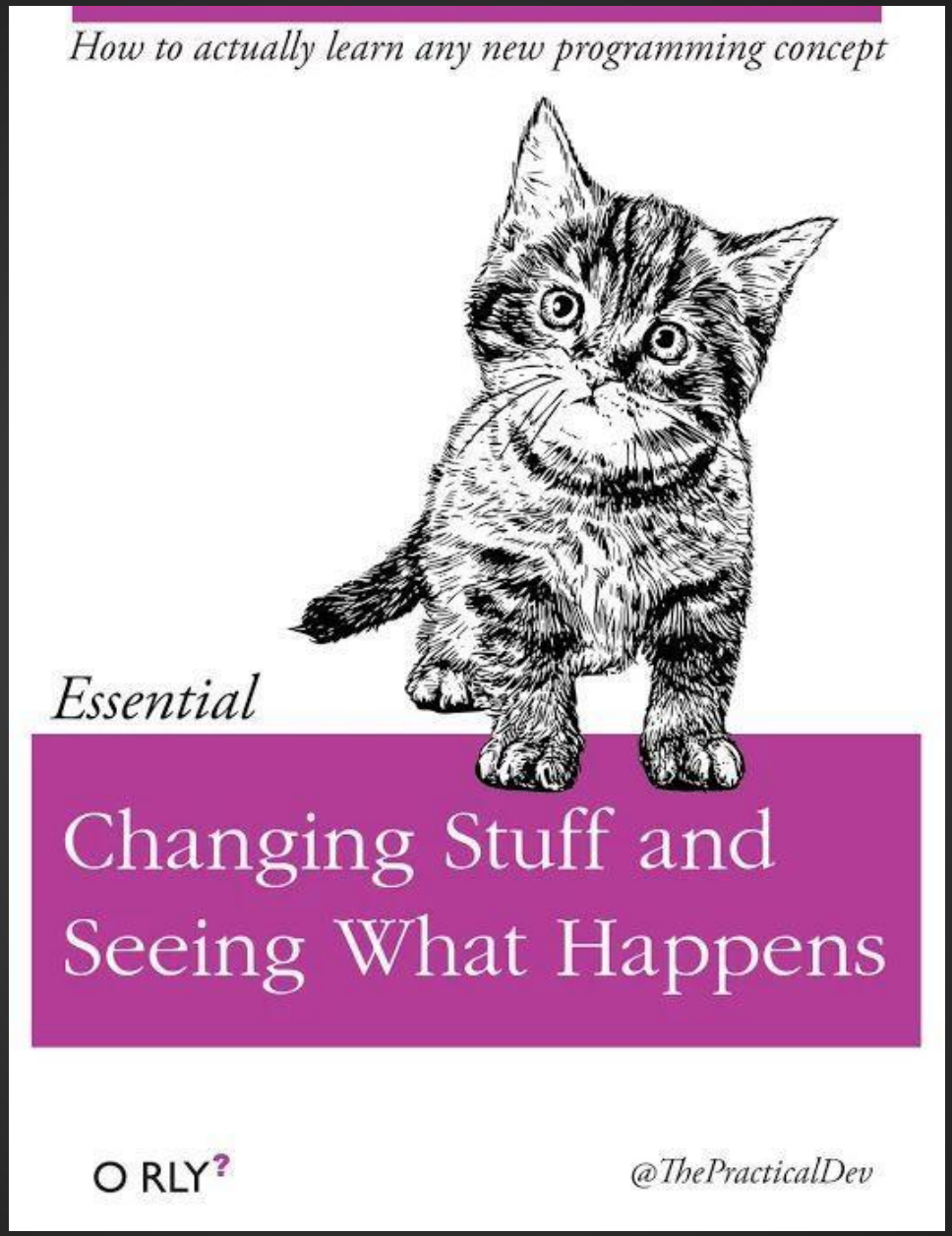
- Apache POI
- Apache TIKA
- Core JVM
- PDF BOX

Document TRANSFORMERS

- JSoup



DEMO



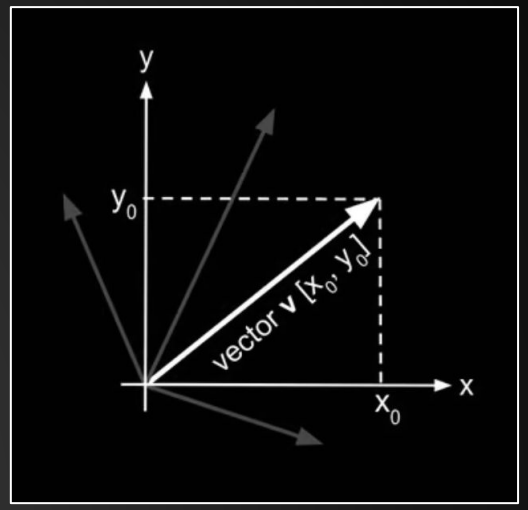
_30_loader_and_parsers

3. Advanced RAG: Ingestion

- Loading and Parsing
- **Vectors, Spaces, Similarity & Search**
- Chunking
- Embedding

Vector
Overview

- Denote a phenomenon with a direction and a length.
- Formulated as a list made of numbers (components)
- List length is the dimensionality (d)

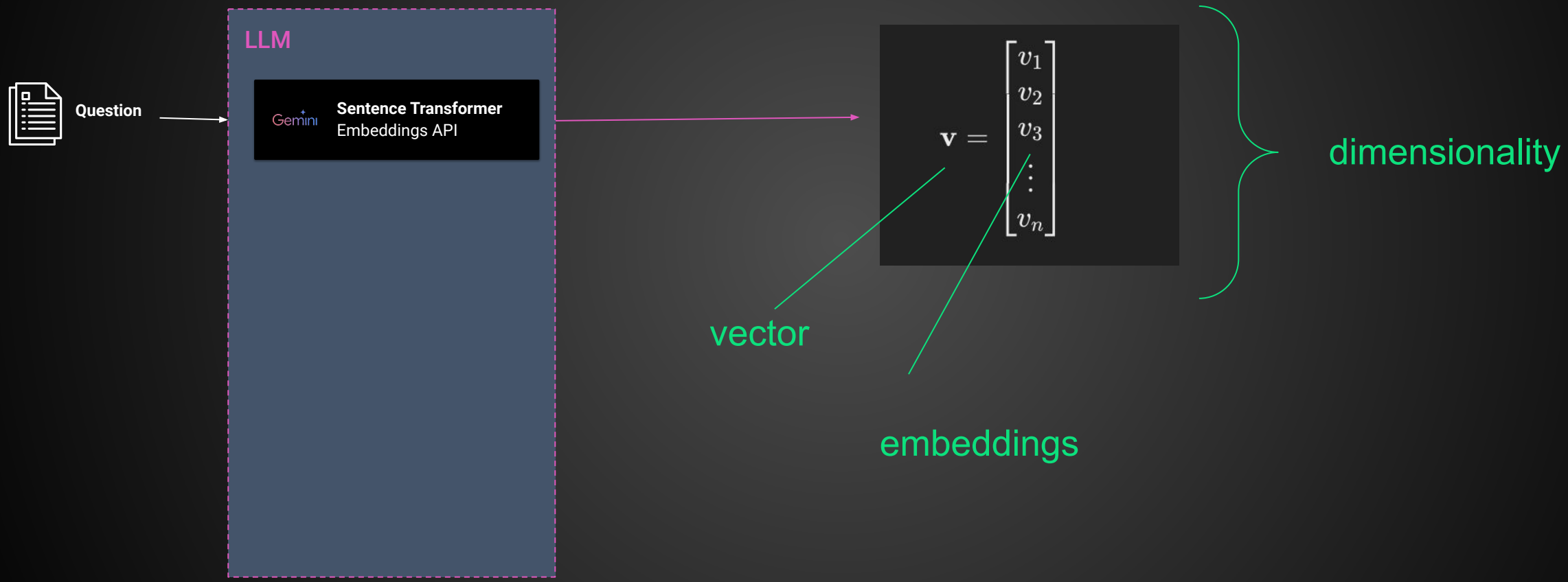


- The "length" (or norm) regardless which direction
- some meaningful notion of "rotation"
- All vectors with same **d** form a *vector space*

- *Direction* = "where the arrow points"

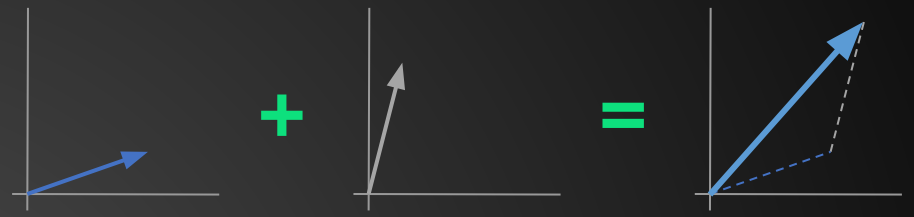
$$|\mathbf{v}| = \sqrt{\sum_i v_i^2}$$

Vector
Vectorization



Vector
Space and Dimensions

Vector spaces are "flat":

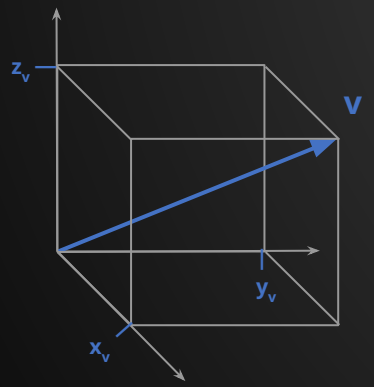


"Actual" physical spaces

d = 1: (no need to involve "vectors", no?)

d = 2: a flat plane

d = 3: the space around you



Higher dimensions

d > 3: not easily visualized

So what? Such spaces exist...

...and are pretty useful across disciplines

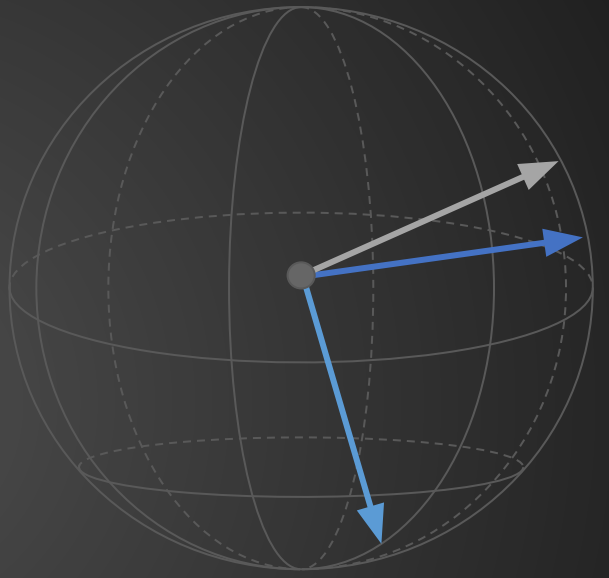
Figure 1: a 768-dimensional sphere.

$v = [v_1, v_2, v_3 \dots v_{d-1}, v_d]$

Question:

In a 3-dimensional space, the notation |v| = 1 represent a *unit vector.*

How do you call the representation of all unit vectors in a 3-dimensional space ?

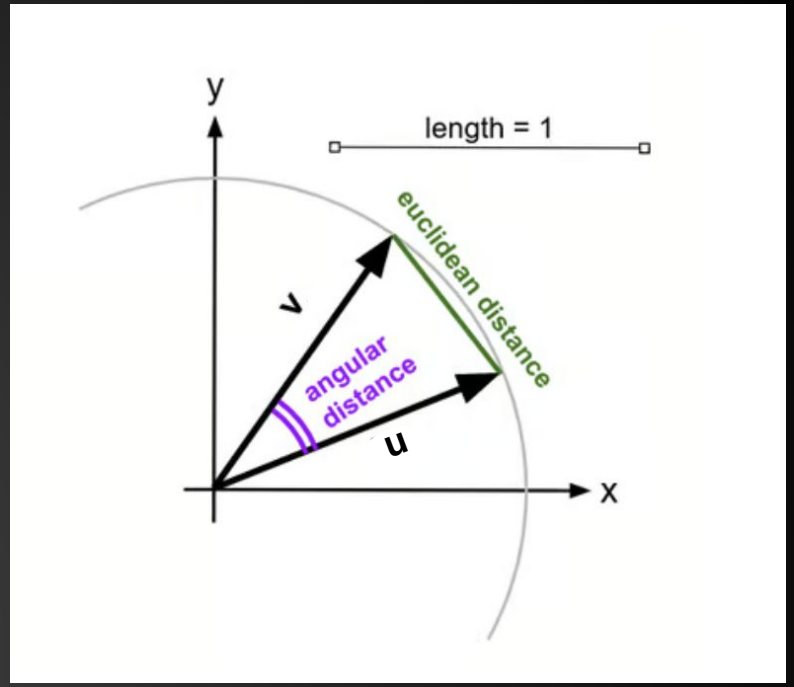


|v| = 1

The UNIT SPHERE

Vector
Similarities

A numeric way to quantify how much two vectors **v** and **u** are close to each other, computed with some formula S(*v*, *u*).



Euclidean distance (L2) * *decrease when similarity increase*

$$d(\mathbf{v}, \mathbf{u}) = \sqrt{\sum_{i=1}^{n}(v_i - u_i)^2}$$

Angular distance, cosine similarity

$$\text{cosine similarity}(\mathbf{v}, \mathbf{u}) = \frac{\mathbf{v} \cdot \mathbf{u}}{\|\mathbf{v}\|\|\mathbf{u}\|}$$

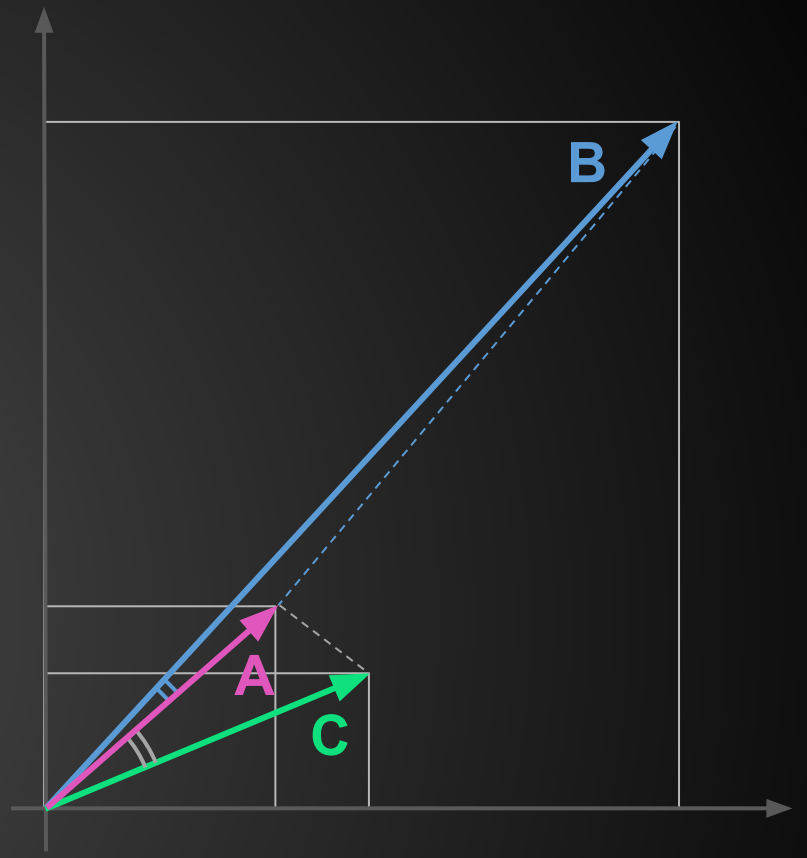
$$\mathbf{v} \cdot \mathbf{u} = \sum_{i=1}^{n} v_i u_i$$

$$\|\mathbf{v}\| = \sqrt{\sum_{i=1}^{n} v_i^2}$$
$$\|\mathbf{u}\| = \sqrt{\sum_{i=1}^{n} u_i^2}$$

Question

What is closest to **A**?

"According to Euclidean similarity, it is **C**.
With Cosine similarity, it is **B**."



DISTANCE ≠ SIMILARITY

Vector
Similarities

Similarity name	Definition (Cassandra / Astra DB)	Remarks
Euclidean	$S_{\mathrm{eucl}}(\mathbf{x},\mathbf{y}) = \frac{1}{1+\sum_i (x_i-y_i)^2} = \frac{1}{1+\delta^2_{\mathrm{eucl}}(\mathbf{x},\mathbf{y})}$	based on the Euclidean distance, $\delta_{\mathrm{eucl}}(\mathbf{x},\mathbf{y}) = \sqrt{\sum_i (x_i-y_i)^2}$
Cosine	$S_{\mathrm{cos}}(\mathbf{x},\mathbf{y}) = \frac{1+\frac{\sum_i x_i y_i}{\lvert\mathbf{x}\rvert\lvert\mathbf{y}\rvert}}{2} = \frac{1+\frac{\mathbf{x}\cdot\mathbf{y}}{\lvert\mathbf{x}\rvert\lvert\mathbf{y}\rvert}}{2} = \frac{1+S^*_{\mathrm{cos}}(\mathbf{x},\mathbf{y})}{2}$	a rescaling of the $\mathbf{S^*_{cos}}$ found on some textbooks (see later)
Dot-product	$S_{\mathrm{dot}}(\mathbf{x},\mathbf{y}) = \frac{1+\sum_i x_i y_i}{2} = \frac{1+\mathbf{x}\cdot\mathbf{y}}{2}$	rarely the right choice, except on the unit sphere!

Vector
Similarities

measure	domain	notes
Euclidean	**sphere** (all unit-norm)	Switch to Dot (same sorting, faster)
Cosine		Switch to Dot (identical, faster)
Dot-product		**OK**
Euclidean	**whole space**	**Use if the norm itself carries information**
Cosine		Normalize-on-save and switch to Dot on sphere
Dot-product		*Are you sure?*

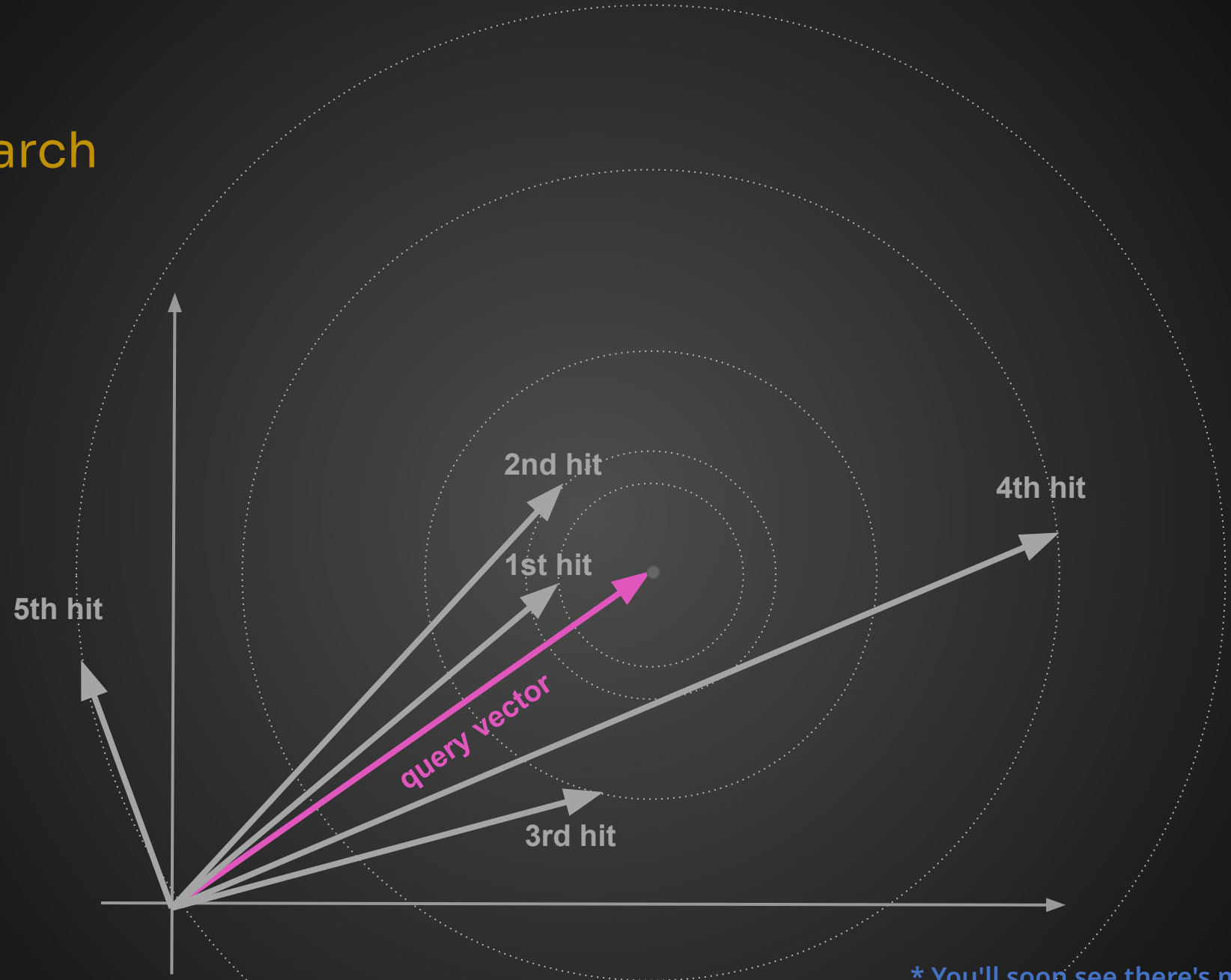
Vector Similarity != Relevance

Vectors can be similar: a query vector is similar to a passage containing the query's answer…

But **similar vectors may be irrelevant!**
(ie. they don't contain the answer)

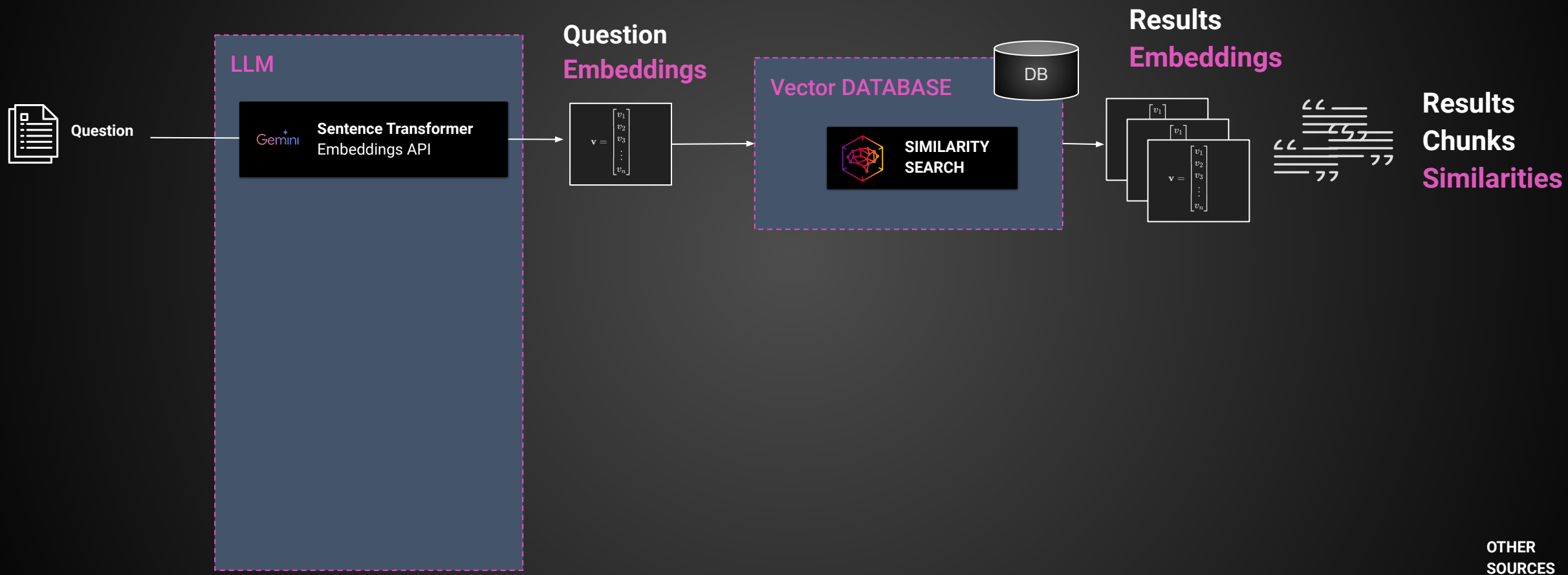
⇒ Importance of scoring, with (Re)Ranking APIs

Vector
Vector Search

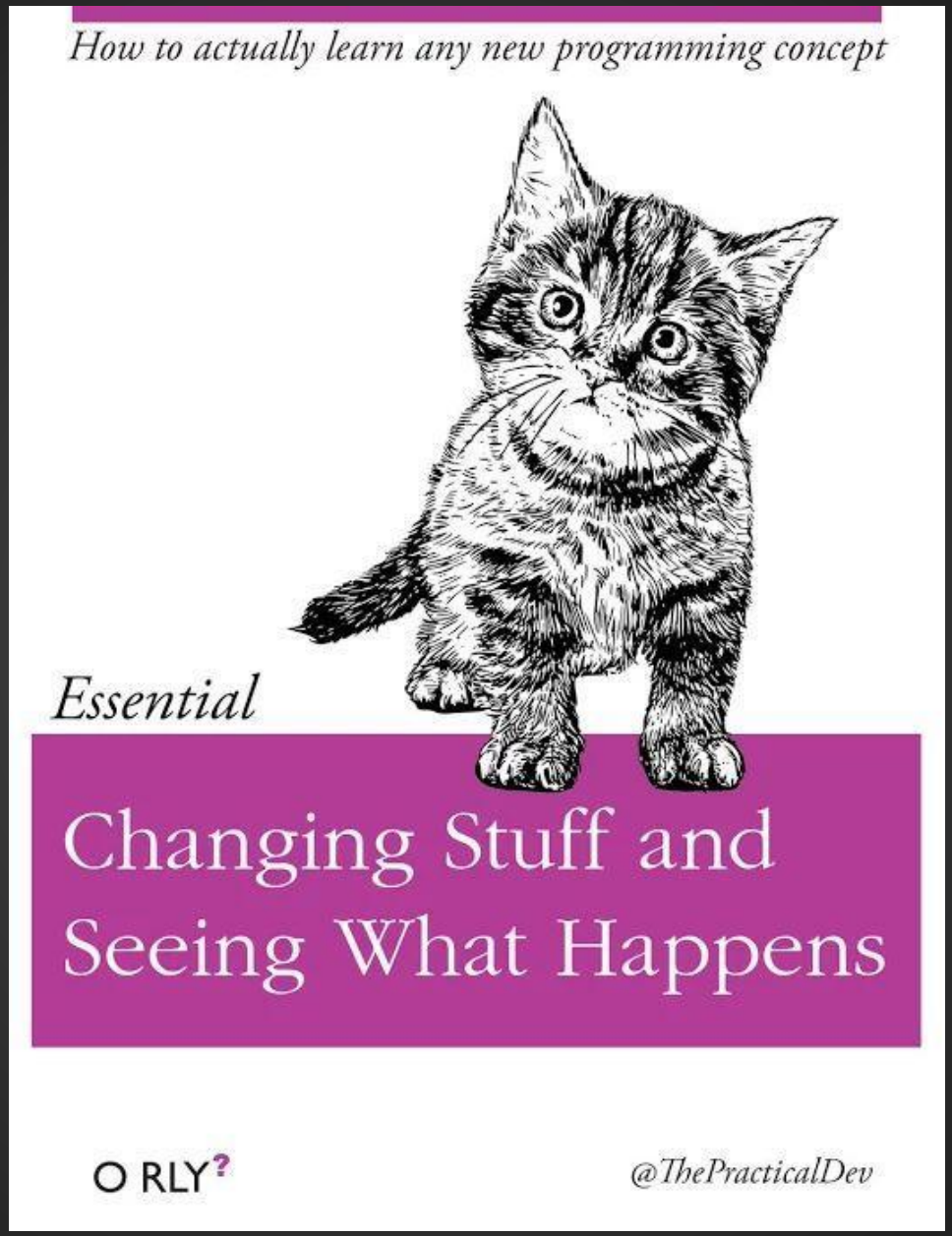


* You'll soon see there's more to say . . .

Vector
Vector Search



DEMO

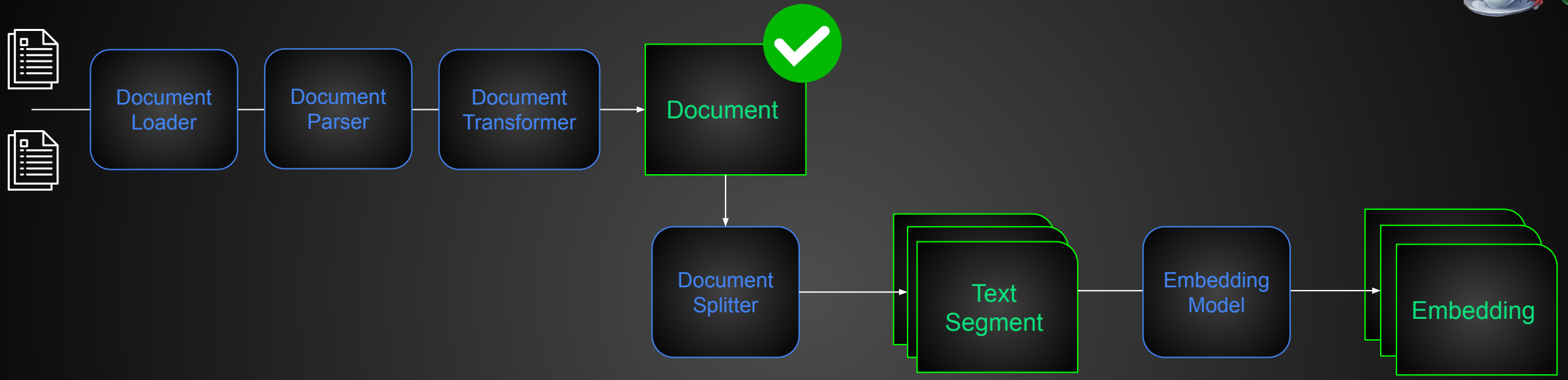


_31_vectors
_32_vectors_similarity

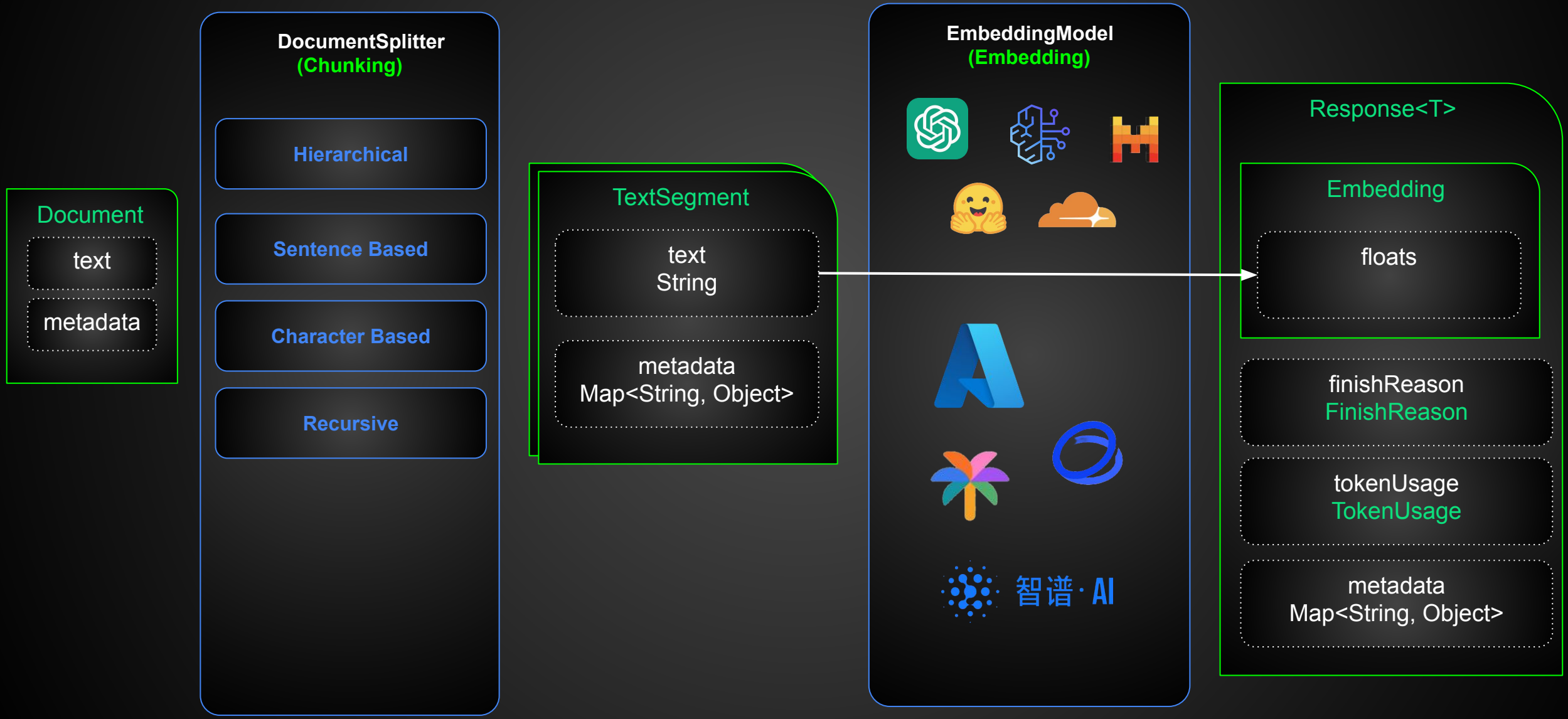
3. Advanced RAG: Ingestion

- Loading and Parsing
- Vectors, Spaces, Similarity & Search
- **Chunking**
- Embedding
- Vector Database

Ingestion Process



Document Ingestion (part 2)



Chunking techniques

- **Hierarchical chunking**
- **Context expansion splitting**
  parent/child, sliding window
- **Hypothetical Questions**
  generate relevant questions
- **Contextual retrieval**
  recent article from Anthropic
- **Semantic chunking**

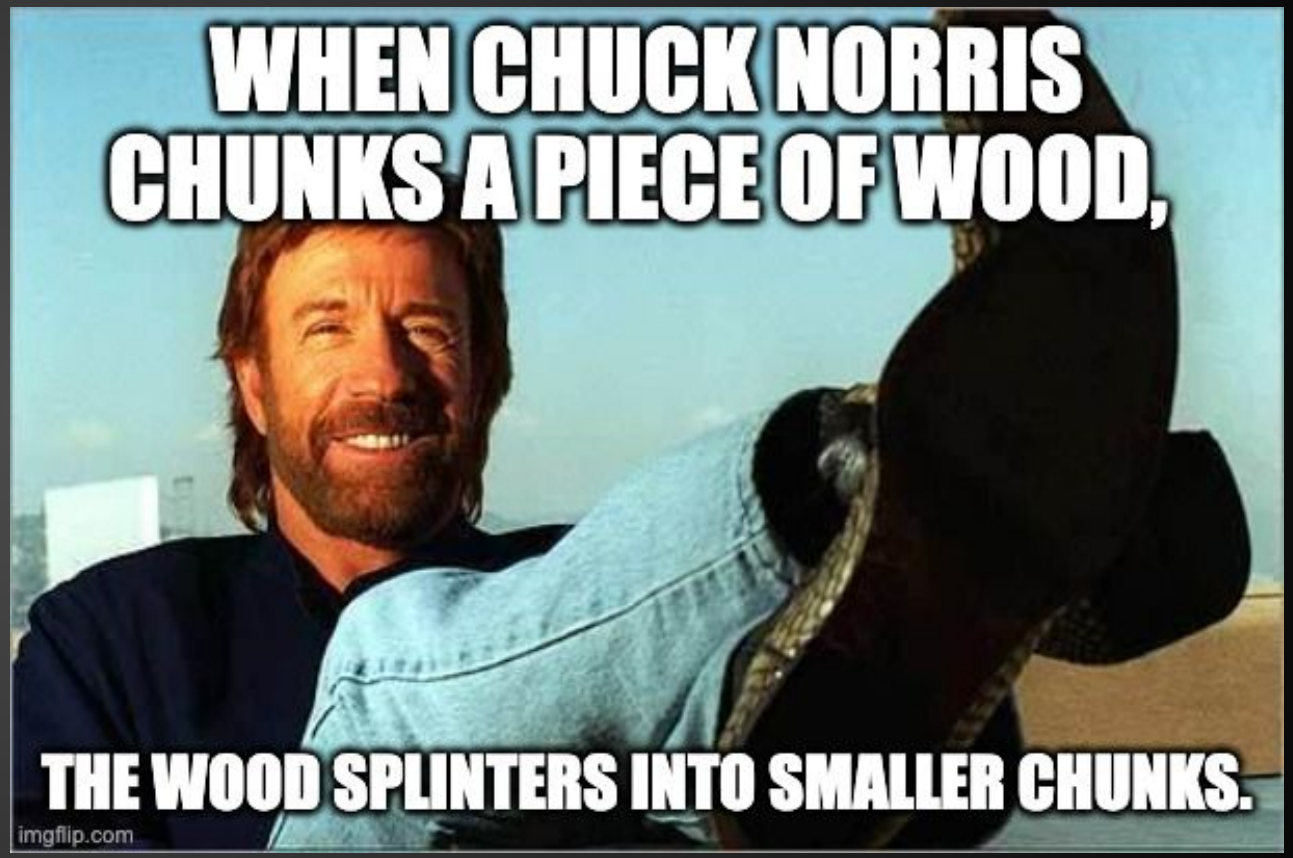


Illustration with a Wikipedia article about Berlin

WIKIPEDIA
The Free Encyclopedia

Search Wikipedia | Search

Donate | Create account | Log in | •••

Berlin

261 languages

Article | Talk | Read | Edit | View history | Tools

From Wikipedia, the free encyclopedia

Coordinates: 52°31′12″N 13°24′18″E

Contents hide

This article is about the capital city of Germany. For other uses, see Berlin (disambiguation).

Berlin[a] is the capital and largest city of Germany, both by area and by population.[11] Its more than 3.85 million inhabitants[12] make it the European Union's most populous city, as measured by population within city limits.[13] The city is also one of the states of Germany, and is the third smallest state in the country in terms of area. Berlin is surrounded by the state of Brandenburg, and Brandenburg's capital Potsdam is nearby. The urban area of Berlin has a population of over 4.5 million and is therefore the most populous urban area in Germany.[5][14] The Berlin-Brandenburg capital region has around 6.2 million inhabitants and is Germany's second-largest metropolitan region after the Rhine-Ruhr region, and the sixth-biggest metropolitan region by GDP in the European Union.[15]

Berlin was built along the banks of the Spree river, which flows into the Havel in the western borough of Spandau. The city incorporates lakes in the western and southeastern boroughs, the largest of which is Müggelsee. About one-third of the city's area is composed of forests, parks and gardens, rivers, canals, and lakes.[16]

First documented in the 13th century[10] and at the crossing of two important historic trade routes,[17] Berlin was designated the capital of the Margraviate of Brandenburg (1417–1701), Kingdom of Prussia

Berlin

Capital city, state and municipality

Spree river, Museum Island, Berlin TV Tower and Berlin Palace in Mitte

Victory Column

Charlottenburg Palace

Rotes Rathaus

Brandenburg Gate

Reichstag

Appearance hide

Text
- Small
- Standard
- Large

Width
- Standard
- Wide

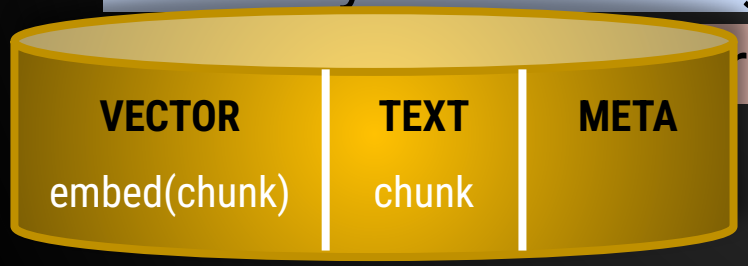
Color (beta)
- Automatic
- Light
- Dark

Raw text

Berlin is the capital and largest city of Germany, both by area and by population. Its more than 3.85 million inhabitants make it the European Union's most populous city, as measured by population within city limits. The city is also one of the states of Germany, and is the third smallest state in the country in terms of area. Berlin is surrounded by the state of Brandenburg, and Brandenburg's capital Potsdam is nearby. The urban area of Berlin has a population of over 4.5 million and is therefore the most populous urban area in Germany. The Berlin-Brandenburg capital region has around 6.2 million inhabitants and is Germany's second-largest metropolitan region after the Rhine-Ruhr region, and the sixth-biggest metropolitan region by GDP in the European Union.

Naive chunking (~100 characters)

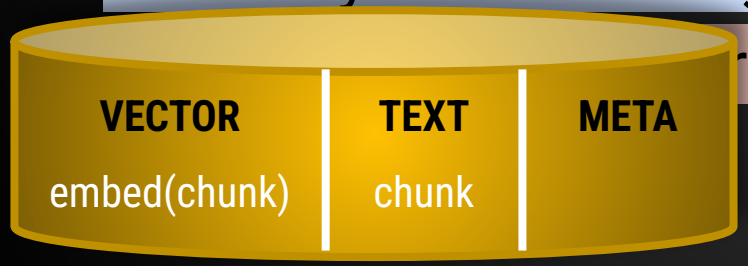
Berlin is the capital and largest city of Germany, both by area and by population. Its more than 3.85 million inhabitants make it the European Union's most populous city, as measured by population within city limits. The city is also one of the states of Germany, and is the third smallest state in the country in terms of area. Berlin is surrounded by the state of Brandenburg, and Brandenburg's capital Potsdam is nearby. The urban area of Berlin has a population of over 4.5 million and is therefore the most populous urban area in Germany. The Berlin-Brandenburg capital region has around 6.2 million inhabitants and is Germany's second-largest metropolitan region after the Rhine-Ruhr region, and ropolitan region by GDP in the European Union.



_34_chunking_defaults

Naive chunking with overlap (~120 chars + 20 overlap)

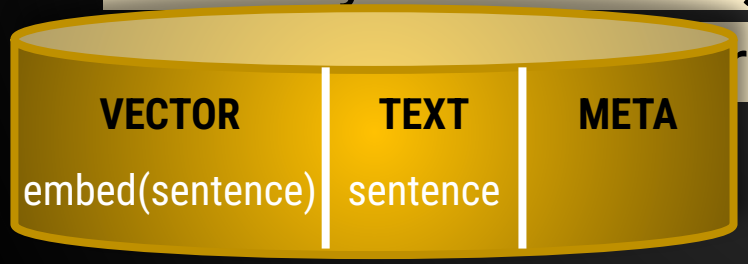
Berlin is the capital and largest city of Germany, both by area and by population. Its more than 3.85 million inhabitants make it the European Union's most populous city, as measured by population within city limits. The city is also one of the states of Germany, and is the third smallest state in the country in terms of area. Berlin is surrounded by the state of Brandenburg, and Brandenburg's capital Potsdam is nearby. The urban area of Berlin has a population of over 4.5 million and is therefore the most populous urban area in Germany. The Berlin-Brandenburg capital region has around 6.2 million inhabitants and is Germany's second-largest metropolitan region after the Rhine-Ruhr region, and ropolitan region by GDP in the European Union.



_34_chunking_defaults

Chunking by sentence

Berlin is the capital and largest city of Germany, both by area and by population. Its more than 3.85 million inhabitants make it the European Union's most populous city, as measured by population within city limits. The city is also one of the states of Germany, and is the third smallest state in the country in terms of area. Berlin is surrounded by the state of Brandenburg, and Brandenburg's capital Potsdam is nearby. The urban area of Berlin has a population of over 4.5 million and is therefore the most populous urban area in Germany. The Berlin-Brandenburg capital region has around 6.2 million inhabitants and is Germany's second-largest metropolitan region after the Rhine-Ruhr region, and ropolitan region by GDP in the European Union.

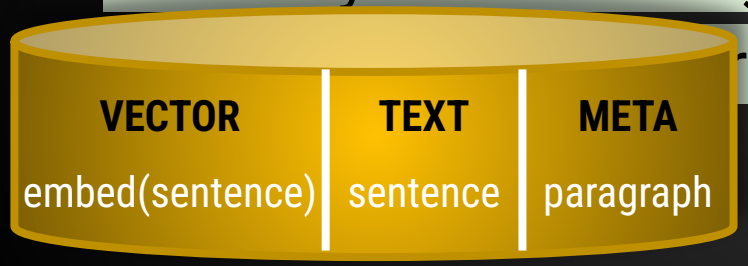


_34_chunking_defaults

Parent (context) / child (embedding) chunking

Berlin is the capital and largest city of Germany, both by area and by population. Its more than 3.85 million inhabitants make it the European Union's most populous city, as measured by population within city limits. The city is also one of the states of Germany, and is the third smalles tate in the country in terms of area. Berlin is surrounded by the state of Bra andenburg's capital Potsdam is nearby. The urban area of B tion of over 4.5 million and is therefore the most populous urb y. The Berlin-Brandenburg capital region has around 6.2 million inhabitants and is Germany's second-largest metropolitan region after the Rhine-Ruhr region, and ropolitan region by GDP in the European Union.

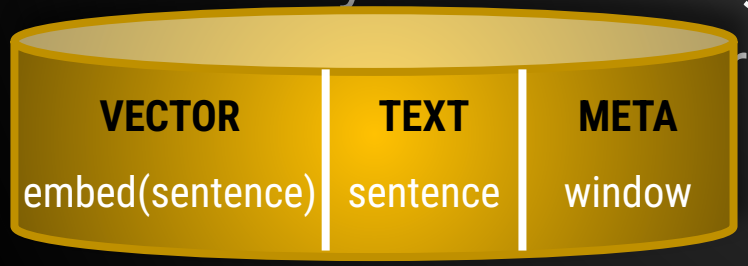
Embed sentence, but return context



Sentence sliding window chunking

Berlin is the capital and largest city of Germany, both by area and by population. Its more than 3.85 million inhabitants make it the European Union's most populous city, as measured by popu tion within city limits. The city is also one of the states of Germany, and is the d smallest state in the country in terms of area. Berlin is surrounded by t nburg, and Brandenburg's capital Potsdam is nearby. The u has a population of over 4.5 million and is therefore the most populous urban area in Germany. The Berlin-Brandenburg capital region has around 6.2 million inhabitants and is Germany's second-largest metropolitan region after the Rhine-Ruhr region, and opolitan region by GDP in the European Union.

Embed sentence, but return context



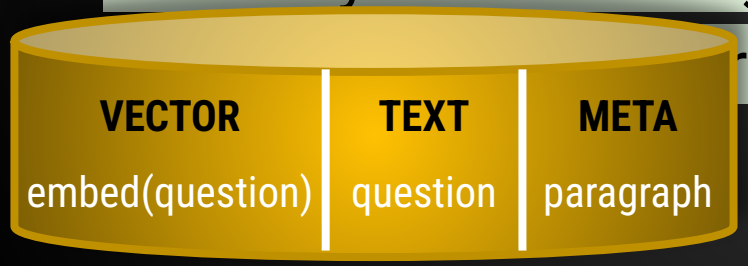
_36_chunks_with_wider_context

Hypothetical Questions

Berlin is the capital and largest city of German
Its more than 3.85 million inhabitants make it
populous city, as measured by population with
of the states of Germany, and is the third
of area. Berlin is surrounded by the state of
capital Potsdam is nearby. The urban area of
million and is therefore the most populous urb
Berlin-Brandenburg capital region has around 6.2 million inhabitants and is
Germany's second-largest metropolitan region after the Rhine-Ruhr region, and
ropolitan region by GDP in the European Union.

Embedding questions:

- *What is the capital and largest city of Germany?*
- *What is the population of Berlin?*
- *Which state is Berlin located in?*
- *What is the name of the state surrounding Berlin?*
- *What is the name of the capital of the state surrounding Berlin?*



_37_hypothetical_questions_embedding

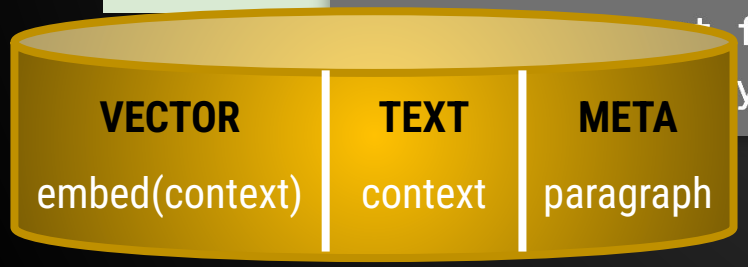
Contextual Retrieval

Berlin is the capital and largest city of Germany, both by area and by population. Its more than 3.85 million inhabitants make it the European Union's most populous city, as measured by population within limits. The city is also one of the states of Germany, and is the third sm terms of area. urg's capital l over 4.5 million a Berlin-B d is Germa Ruhr region, and nion.

```
<document>
{{WHOLE_DOCUMENT}}
</document>
Here is the chunk we want to situate within th
<chunk>
{{CHUNK_CONTENT}}
</chunk>
Please give a short succinct context to situate this chunk within the
for the purposes of improving search retrieval of the
y with the succinct context and nothing else.
```

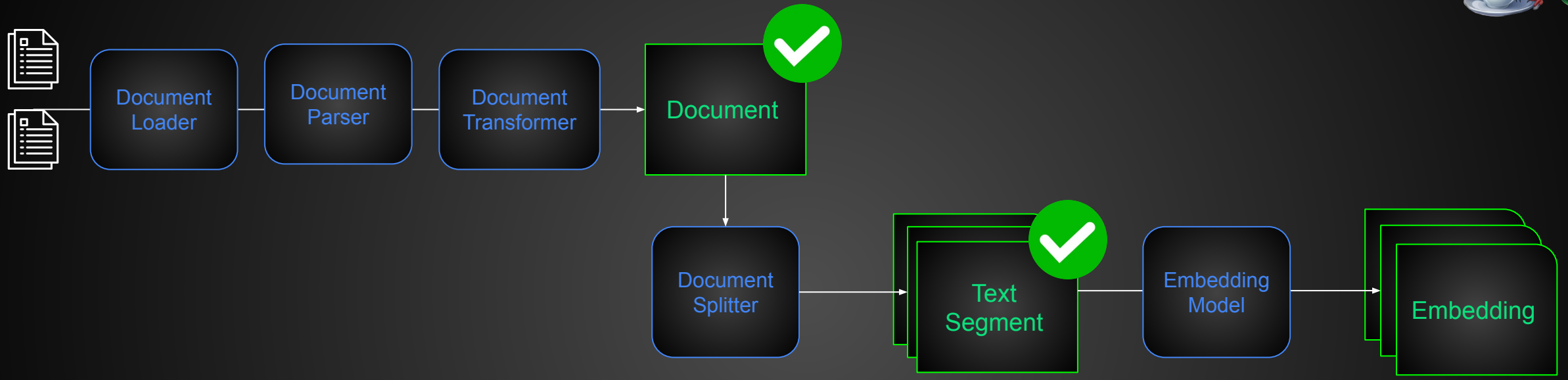
Embed chunk *"in context"*:

Berlin's population within its city limits is the largest in the European Union.



_38_contextual_retrieval

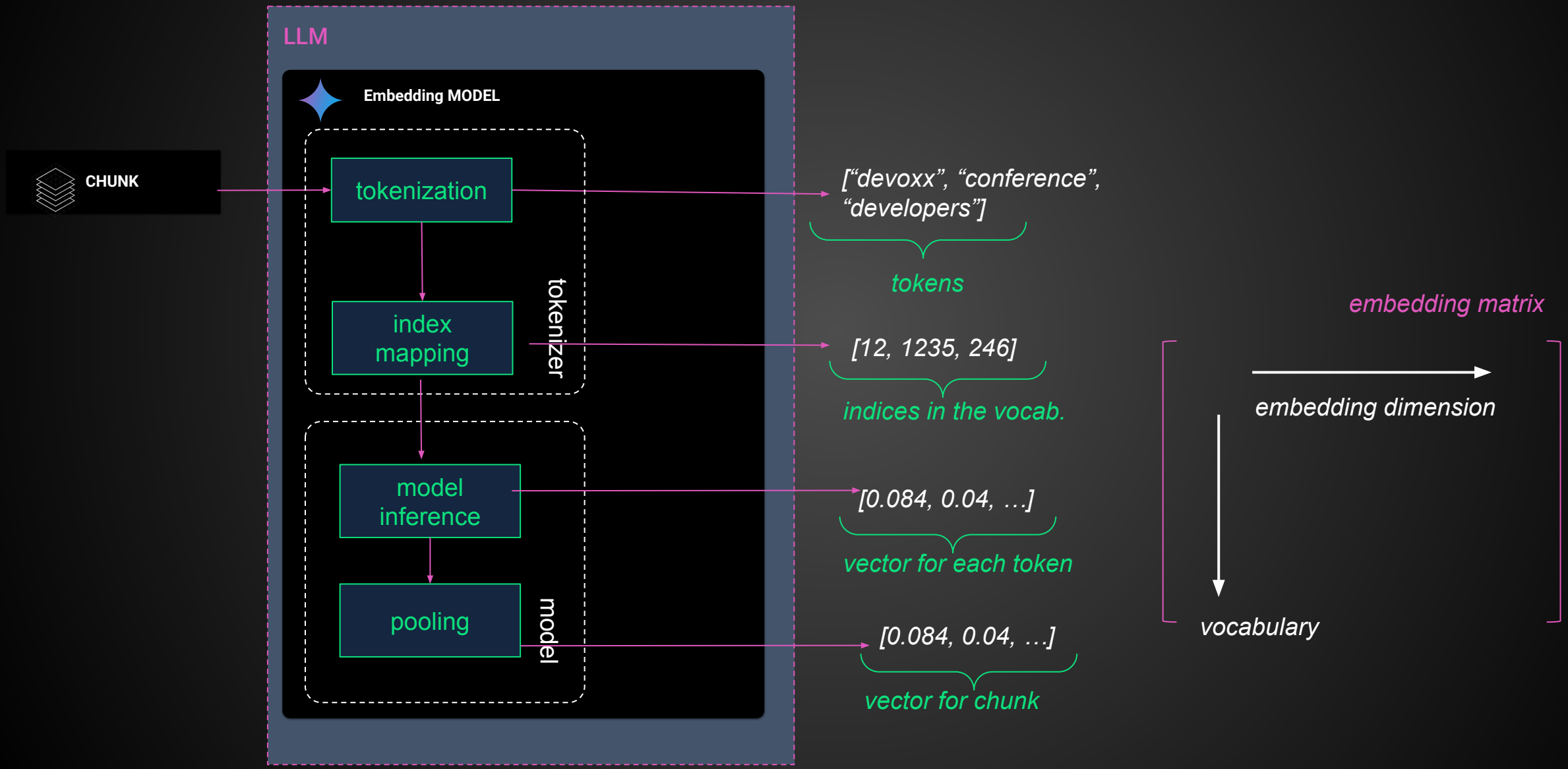
Ingestion Process



3. Advanced RAG: Ingestion

- Loading and Parsing
- Vectors, Spaces, Similarity & Search
- Chunking
- **Embedding**
- Vector Database

Embedding

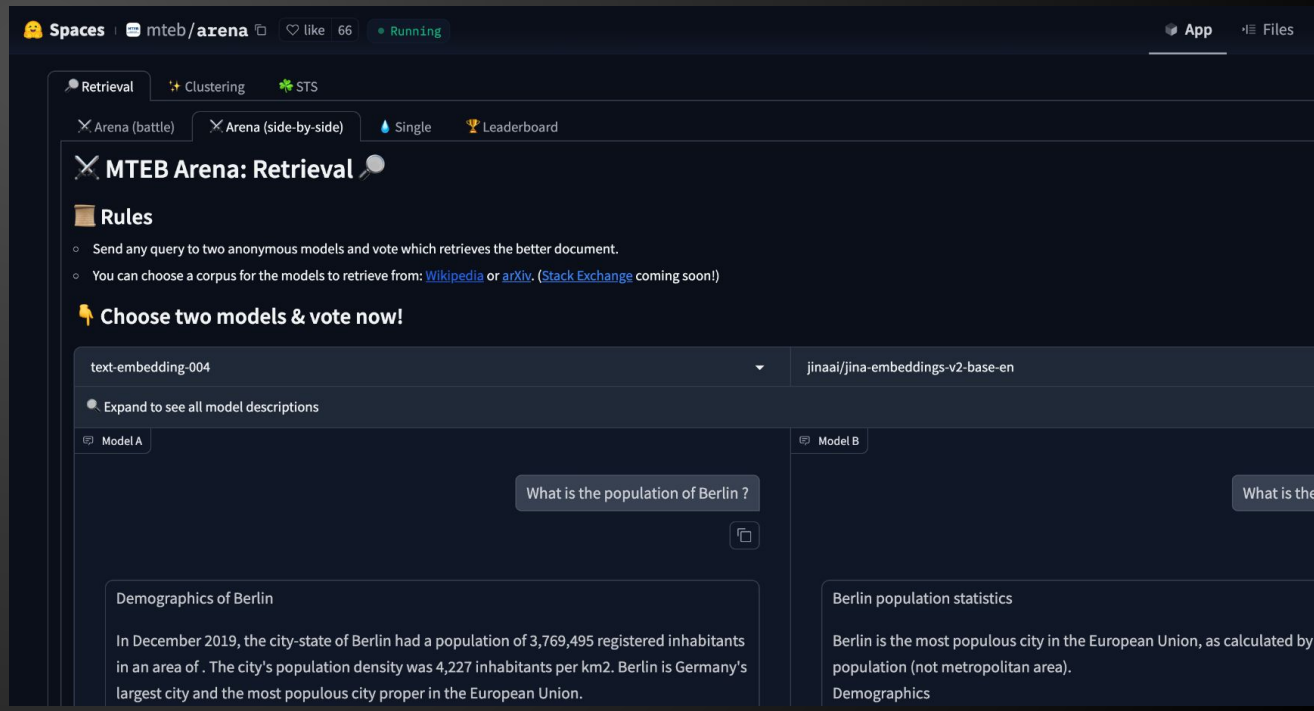

LLM
Embedding MODEL
CHUNK
tokenization
index mapping
tokenizer
model inference
pooling
model
["devoxx", "conference", "developers"]
tokens
[12, 1235, 246]
indices in the vocab.
[0.084, 0.04, …]
vector for each token
[0.084, 0.04, …]
vector for chunk
embedding matrix
embedding dimension
vocabulary

How to choose your embedding models

https://huggingface.co/spaces/mteb/leaderboard

- Task Requirements (use case, retrieval works great)
- Language (second tab)
- Domain Specificity (~fine-tuning)
- Model Size (resources, latency)
- Cost (self hosted vs API)
- Dimensionality
- Multi Modal Requirements
- Your existing stack/provider

https://huggingface.co/spaces/mteb/arena



Matryoshka Embeddings

Matryoshka Representation Learning (MRL)

dim=1024

dim=512

dim=256

TRUNCATE

embedding matrix

embedding dimension

vocabulary

Colbert Embeddings

High Density DataSet

Pre-Processing

- (Optional) Fine-tune BERT on your document corpus
- Use the generic checkpoint provided by ColBERT project
- Generate contextualized embeddings vectors E[d] for each of your documents (one per token)

Query Time

- Generate embeddings vectors for your query (one per token)

$$E[d] = \{\mathbf{e}_t \mid t \in T[d]\}$$

$E[d]$ is the set of embedding vectors for document d.

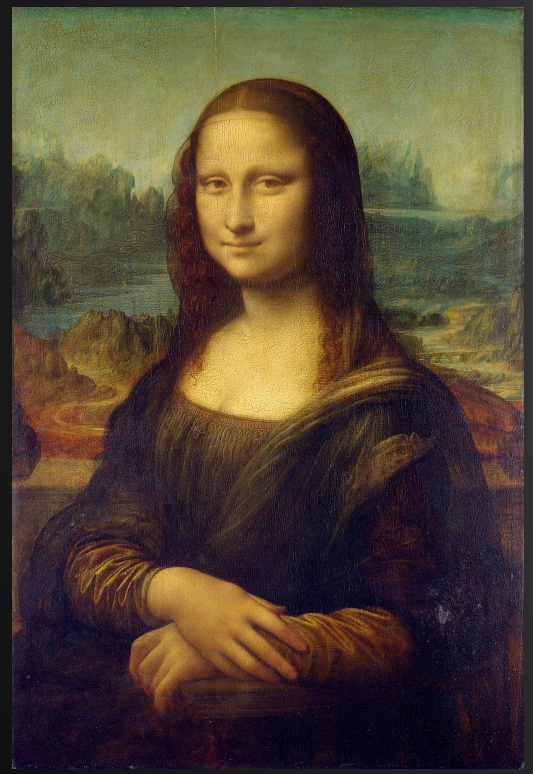
$T[d]$ is the set of tokens in document d.

$\mathbf{e}_t$ is the embedding vector corresponding to token t.

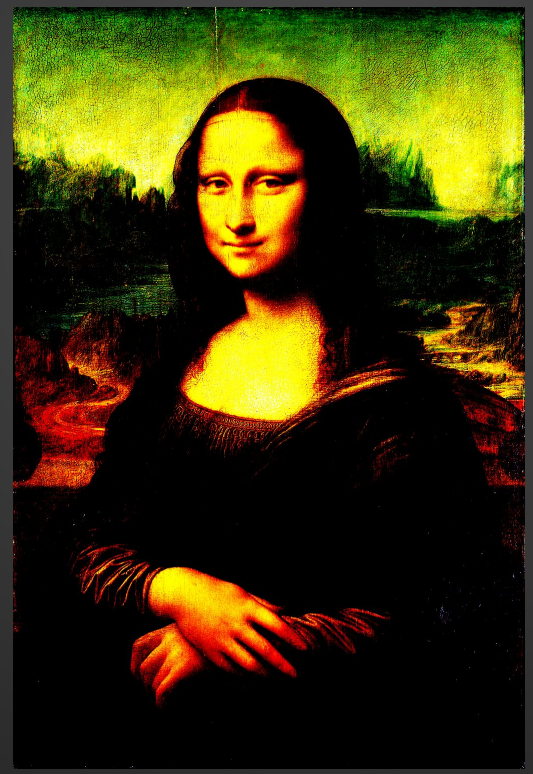
$$\text{Score} = \sum_{v \in E[q]} \max_{x \in E[d]} \text{sim}(v, x)$$

Embedding (scalar) quantization

- Instead of using float32, use smaller types (float16/bfloat16, int8, binary...)
- Distribution is ignored (PQ covers later)



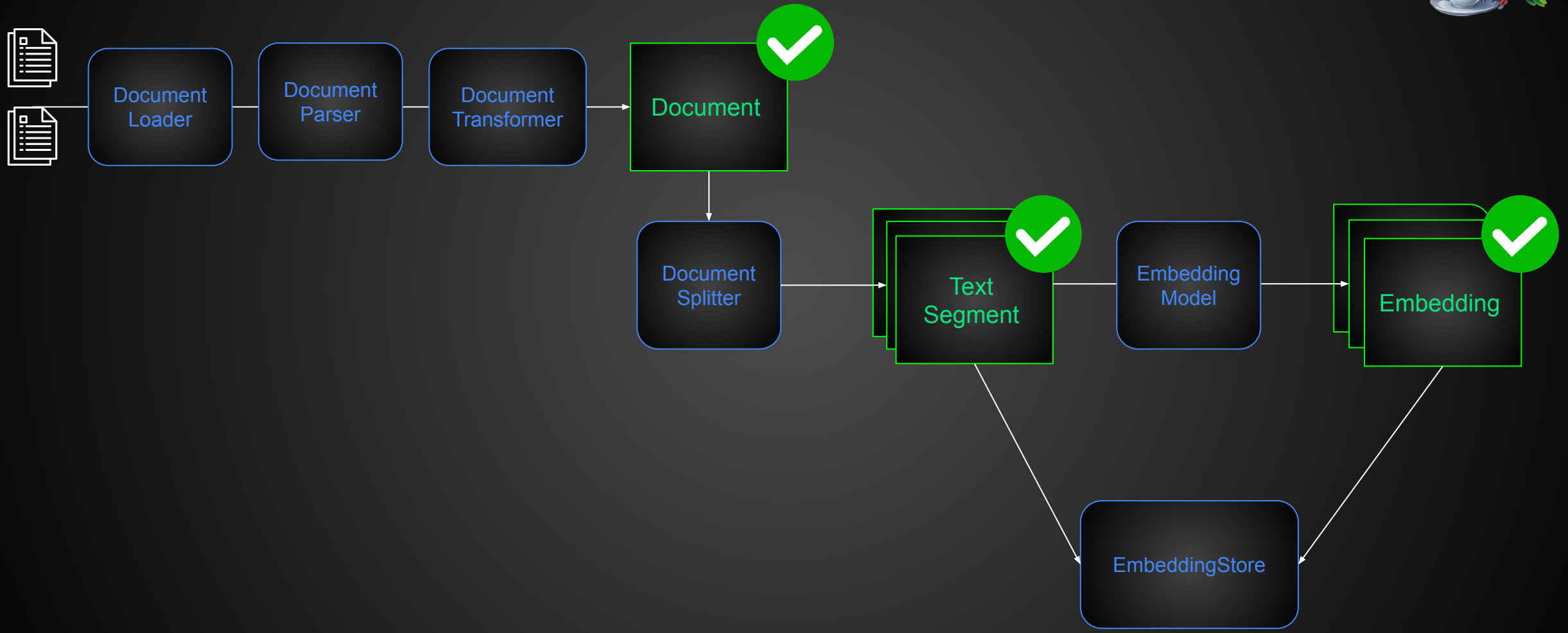
float32



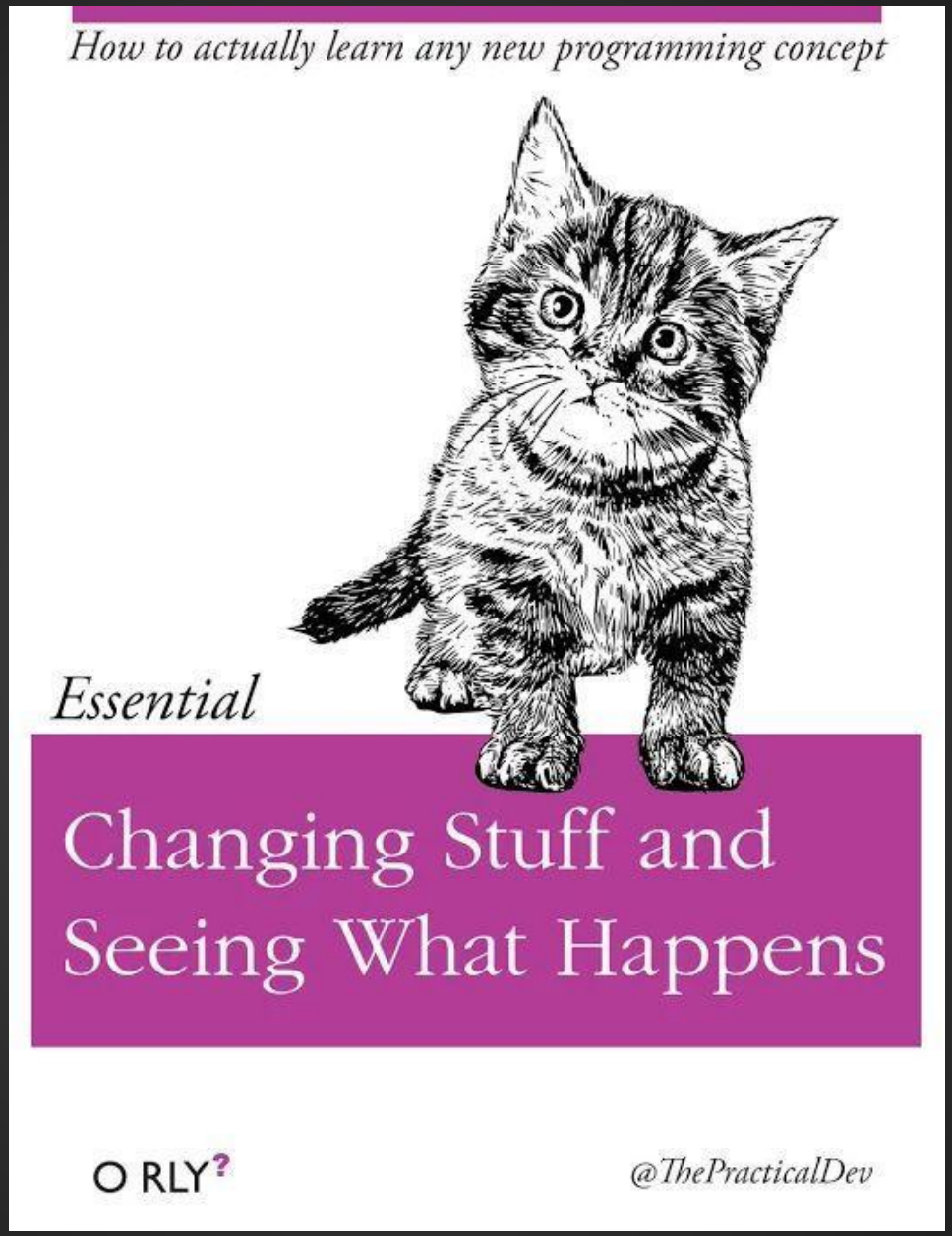
int8

binary / ternary

Ingestion Process

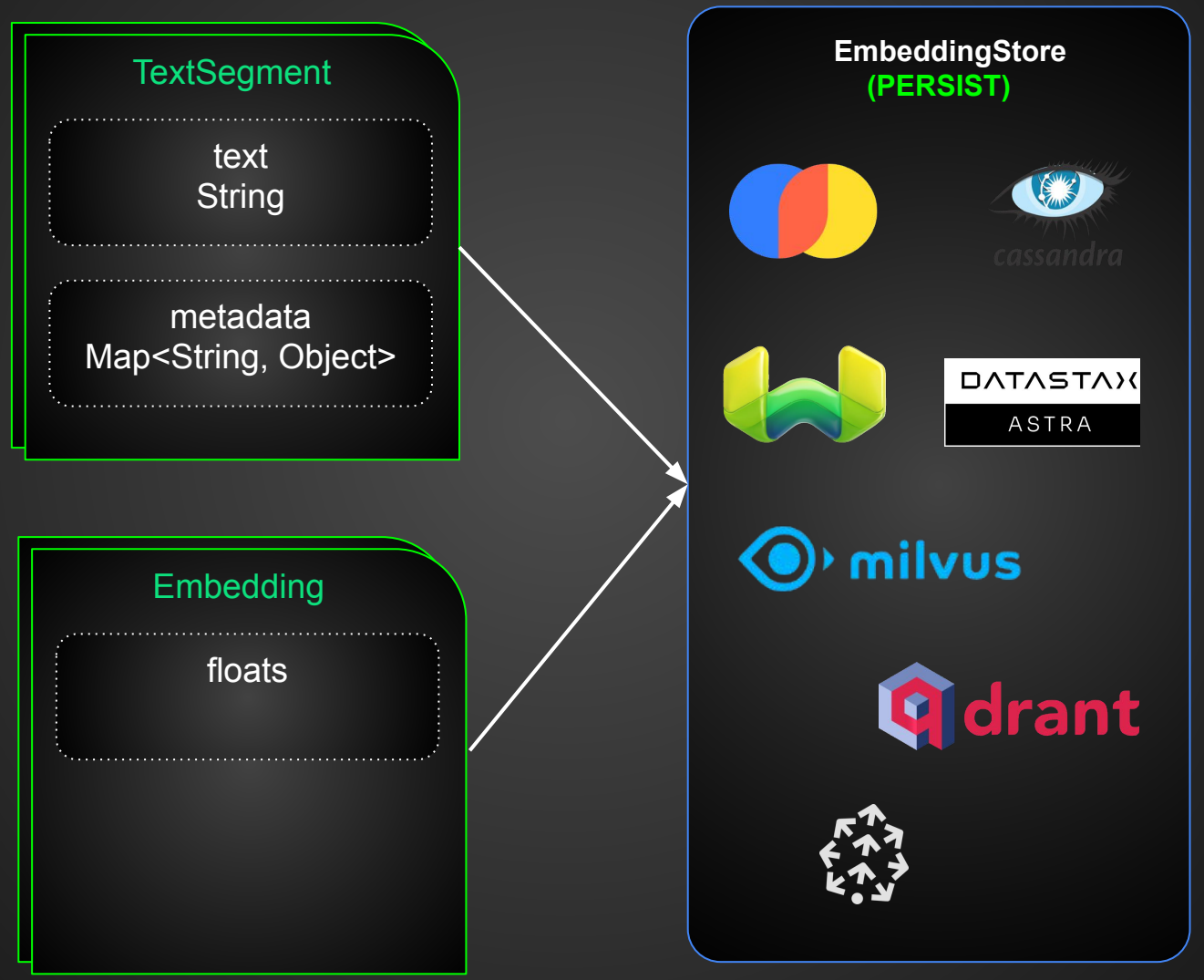


DEMO

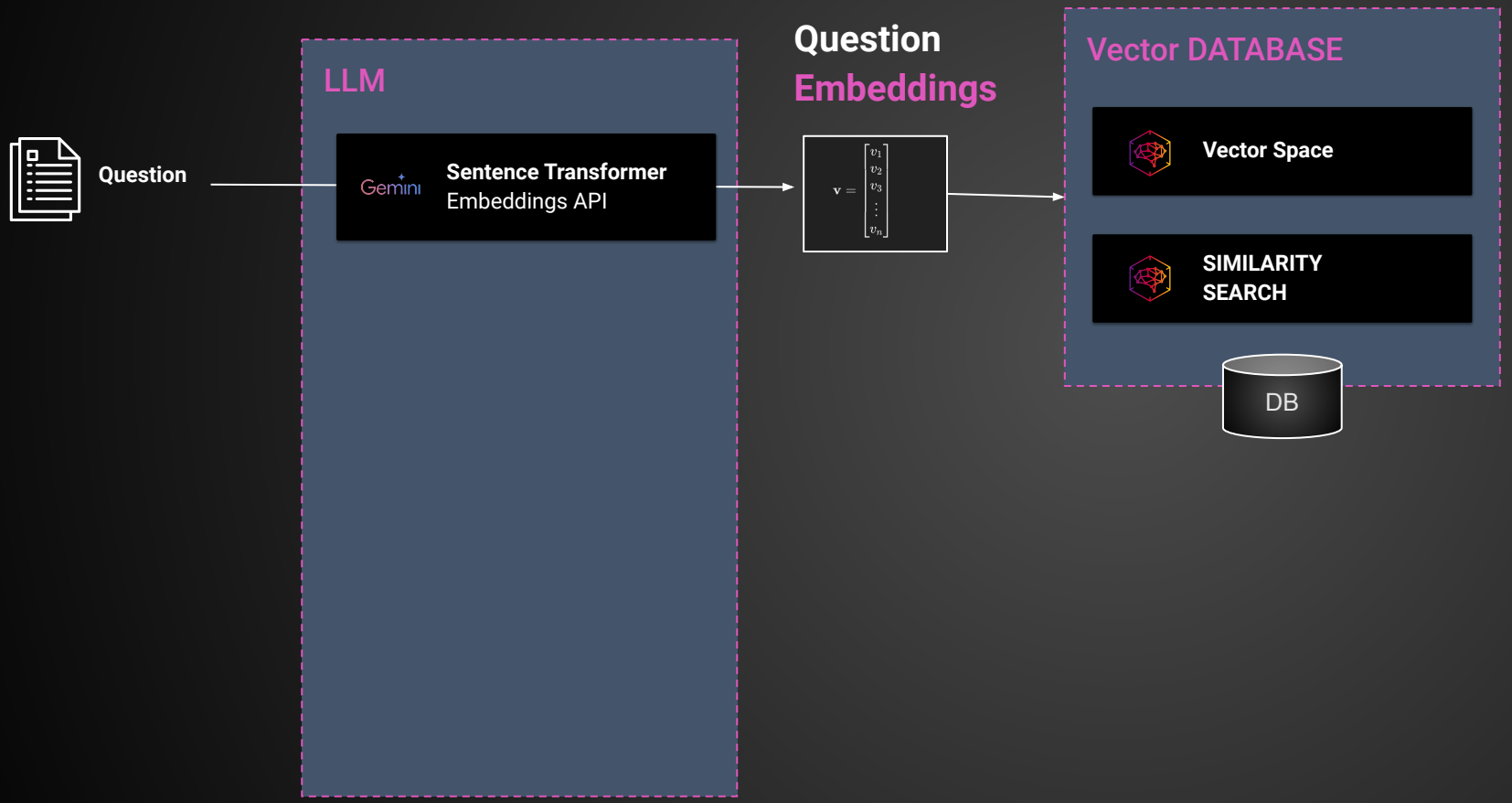


_33_1_embeddings_tokenizers
_33_2_embedding_multilingual
_33_3_embedding_retrieval_task
_33_4_embedding_matryoshka

Document Ingestion (part 3)



Vector Databases
Overview



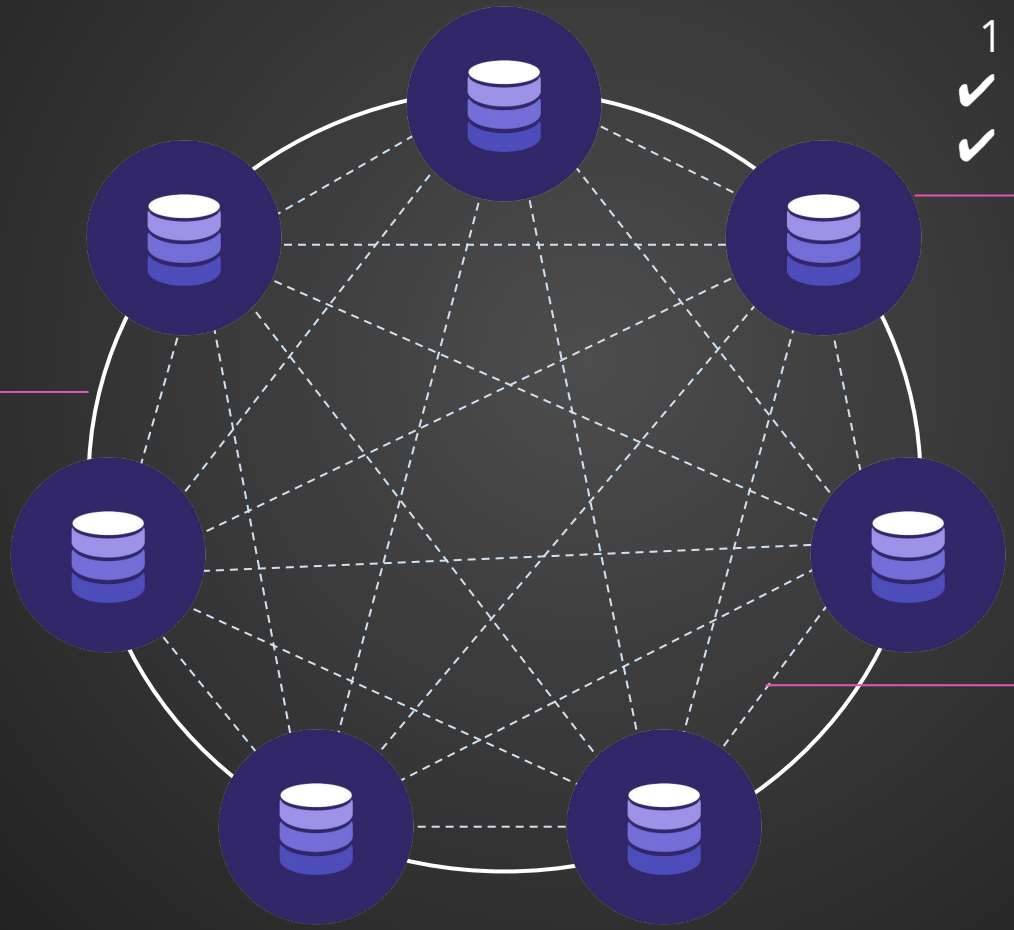
OTHER
SOURCES

Apache Cassandra™

Undisputed Leader for Scale and Reliability

Apache Cassandra™

NoSQL Distributed database



1 Installation = 1 NODE

- ✔ Capacity = ~ 2-4TB
- ✔ Throughput = LOTS Tx/sec/core

DataCenter (DC) | Ring

Communication:

- ✔ Gossiping
- ✔ No Master (peer-to-peer)

Apache Cassandra™

NoSQL Distributed database

High Availability

Always On

Every second of downtime translates into lost revenue

Linear Scalability

Hyper Scalability

Millions of operations per day, hour, or second

Low Latency

Faster Pace

Every millisecond of latency has consequence

Global Distribution

Data Everywhere

On-premises, hybrid, multi-cloud, centralized, or edge

Cassandra 5 as a Vector database

New Type

- New Vector type introduced

```
CREATE TABLE IF NOT EXISTS vsearch.products (
  id int PRIMARY KEY,
  name TEXT,
  description TEXT,
  item_vector VECTOR<FLOAT, 5> //5-dimensional embedding
);
```

Cassandra 5 as a Vector database

SAI Secondary indices

```
CREATE CUSTOM INDEX IF NOT EXISTS ann_index
ON vsearch.products(item_vector)
USING 'StorageAttachedIndex';
```

Cassandra 5 as a Vector database

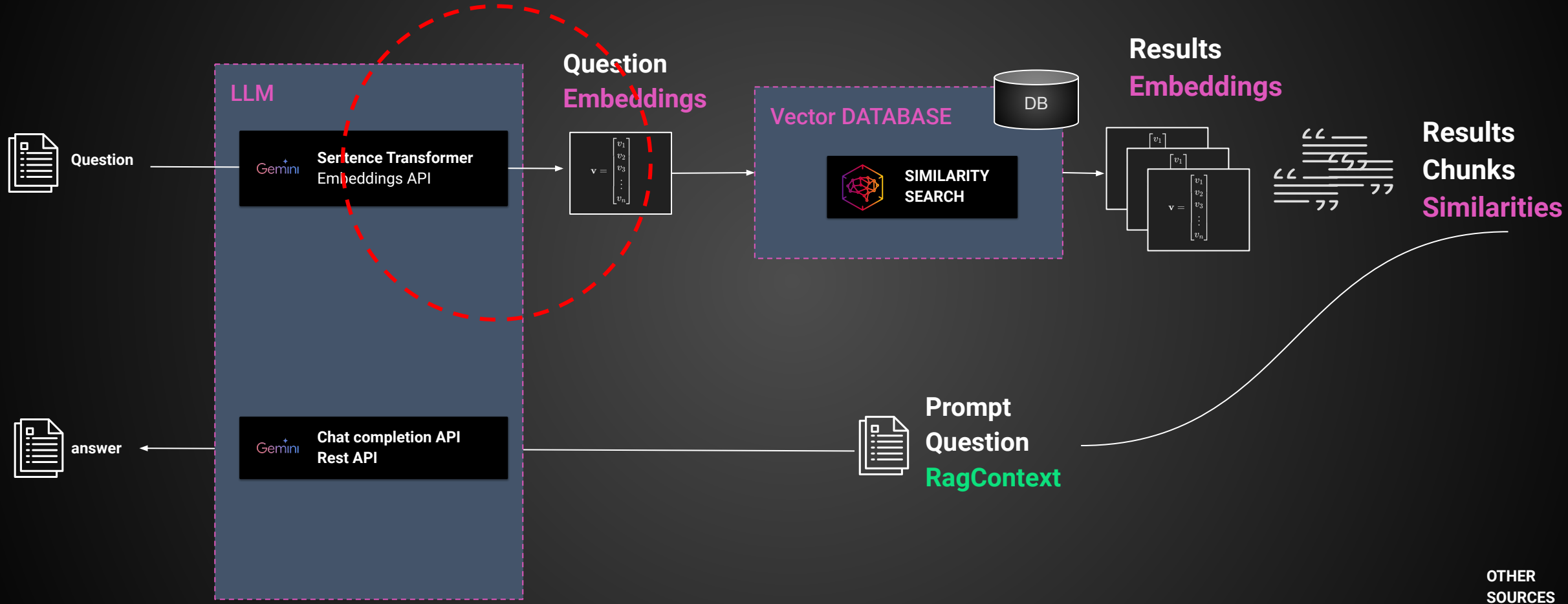
Sample Neighbour Search

```
SELECT * FROM vsearch.products
ORDER BY item_vector ANN OF [0.15, 0.1, 0.1, 0.35, 0.55]
LIMIT 1;
```

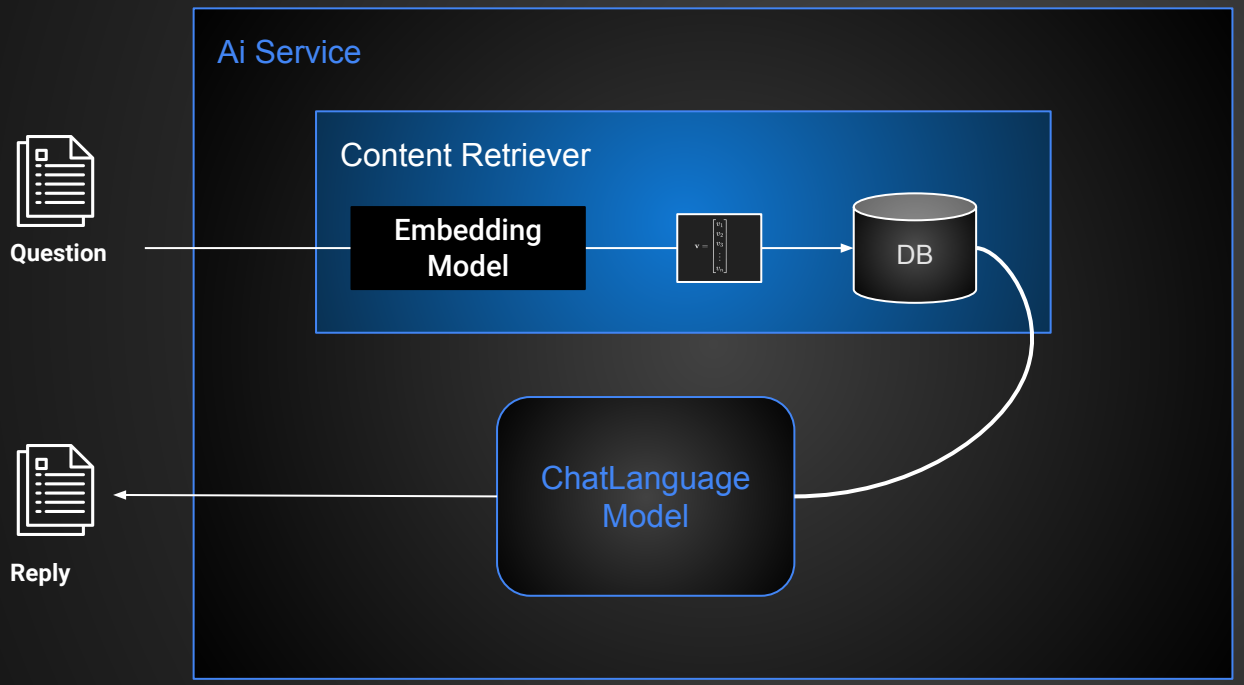
4. Advanced RAG: Query

- **Query Preprocessing**
  - Query Preprocessing
  - Query Transformations
- **Vector Searches**
  - Filterings and metadata
  - Projections and Sorting
- **Question post processing**
  - Reranking
  - Recursive algorithms
  - Consolidation

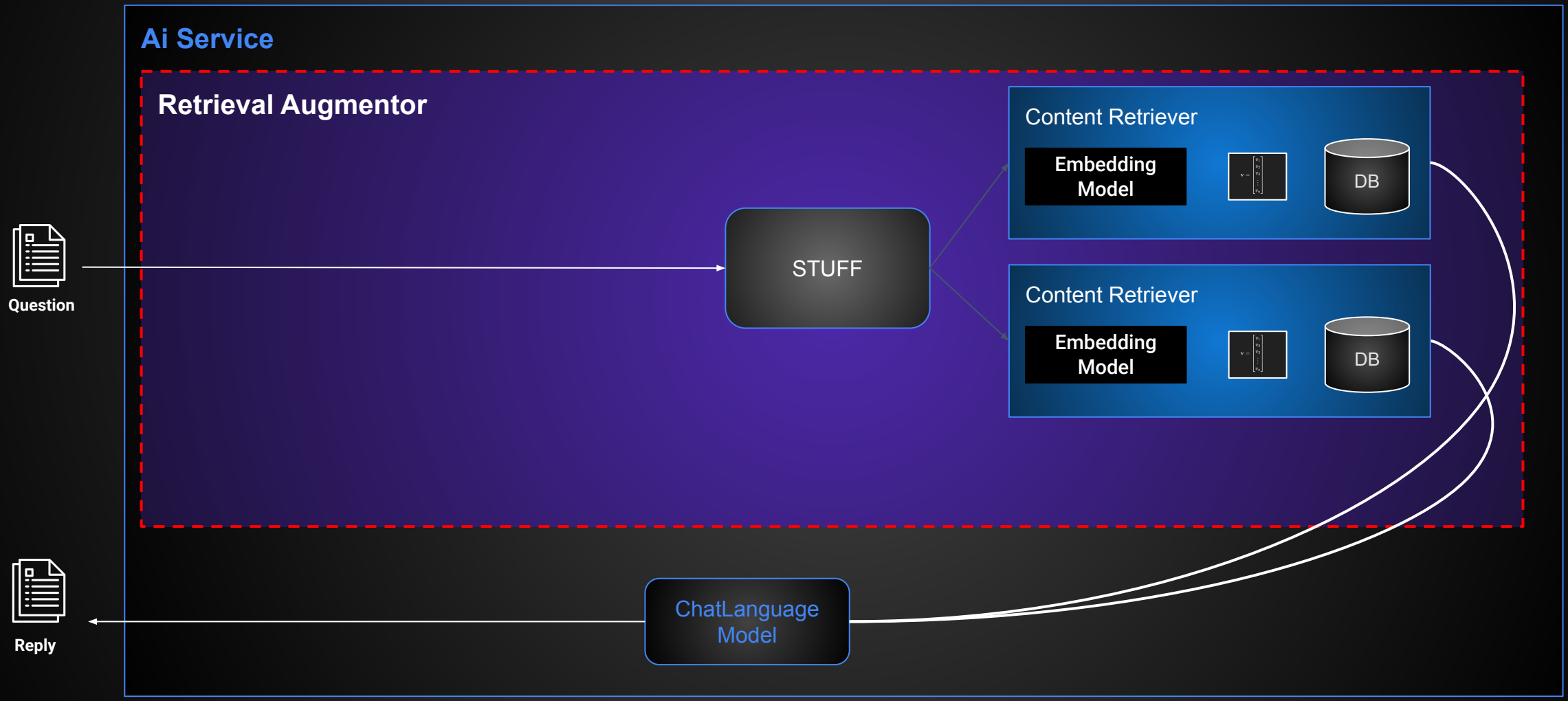
Query Preprocessing
Overview



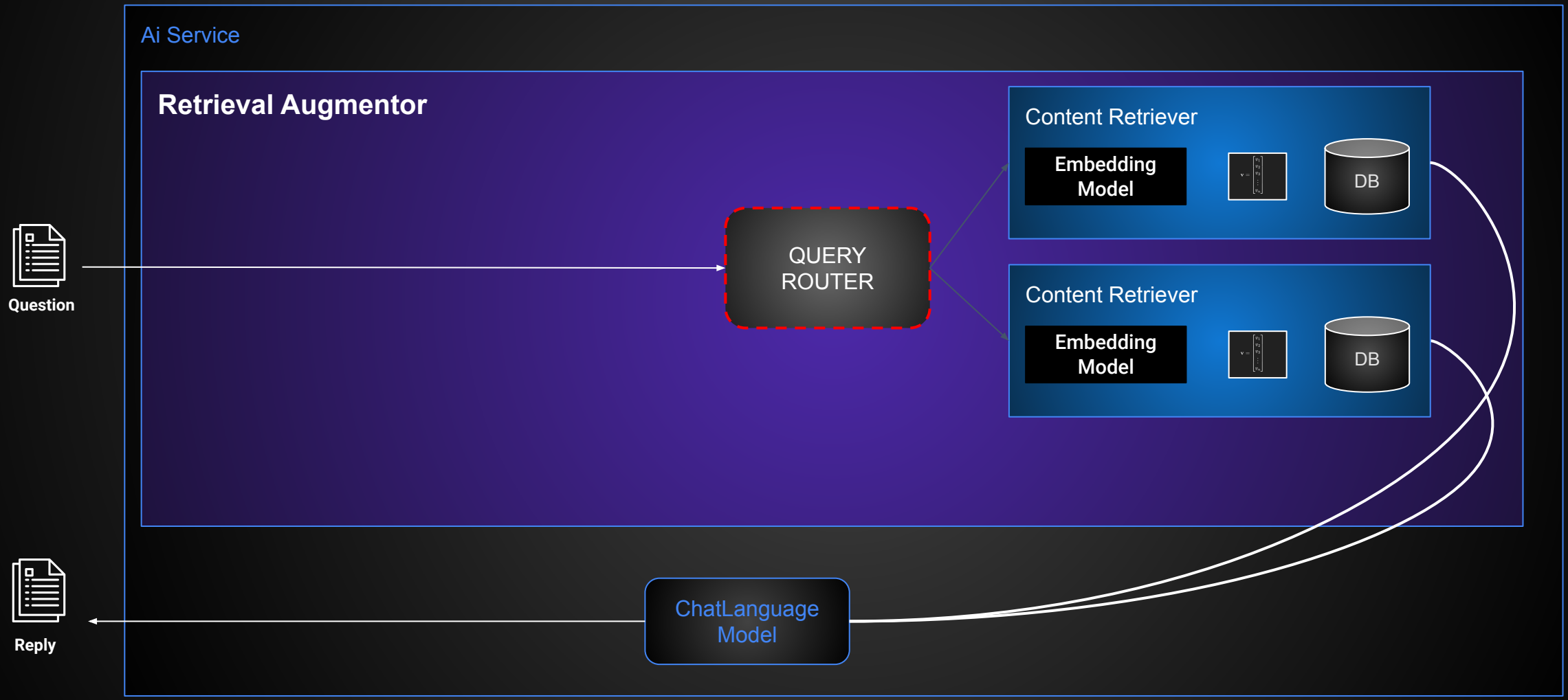
Query Preprocessing
Extending the Content Retriever



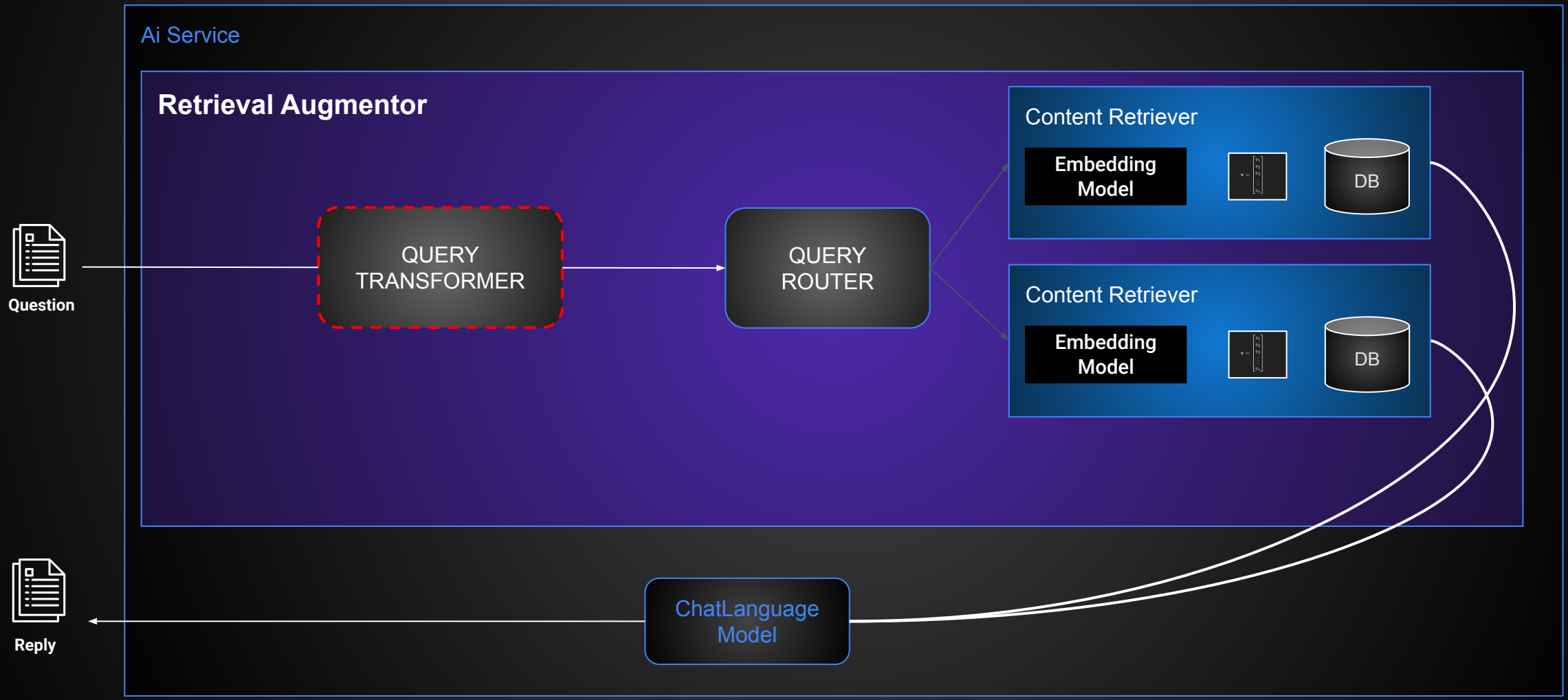
Query Preprocessing
Retrieval Augmentor



Query Preprocessing
Query Router



Query Preprocessing
Query Transformer (compression, HyDE)

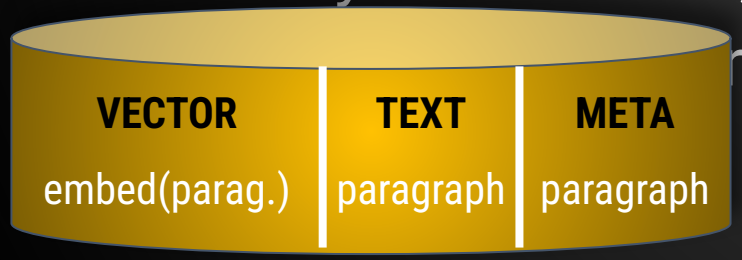


HyDE — Hypothetical Document Embedding

Berlin is the capital and largest city of Germany, both by area and by population. Its more than 3.85 million inhabitants make it the European Union's most populous city, as measured by ulation within city limits. The city is also one of the states of German country in terms of area. Berlin is surrour Brandenburg's capital Potsdam is near ulation of over 4.5 million and is therefore any. The Berlin-Brandenburg capi itants and is Germany's second-largest metropolitan region after the Rhine-Ruhr region, and ropolitan region by GDP in the European Unio n.

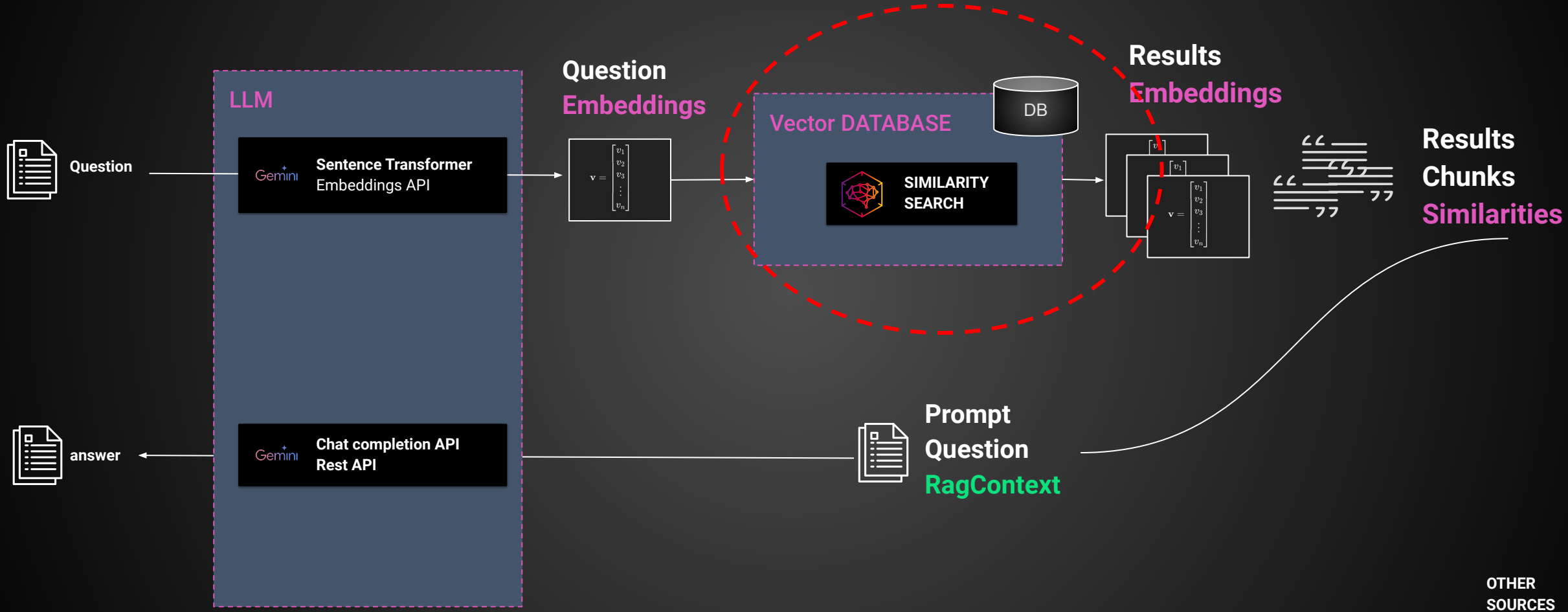
User query:
What is the population of Berlin?

Hypothetical answer (provided by LLM):
There are 3 million inhabitants in Berlin



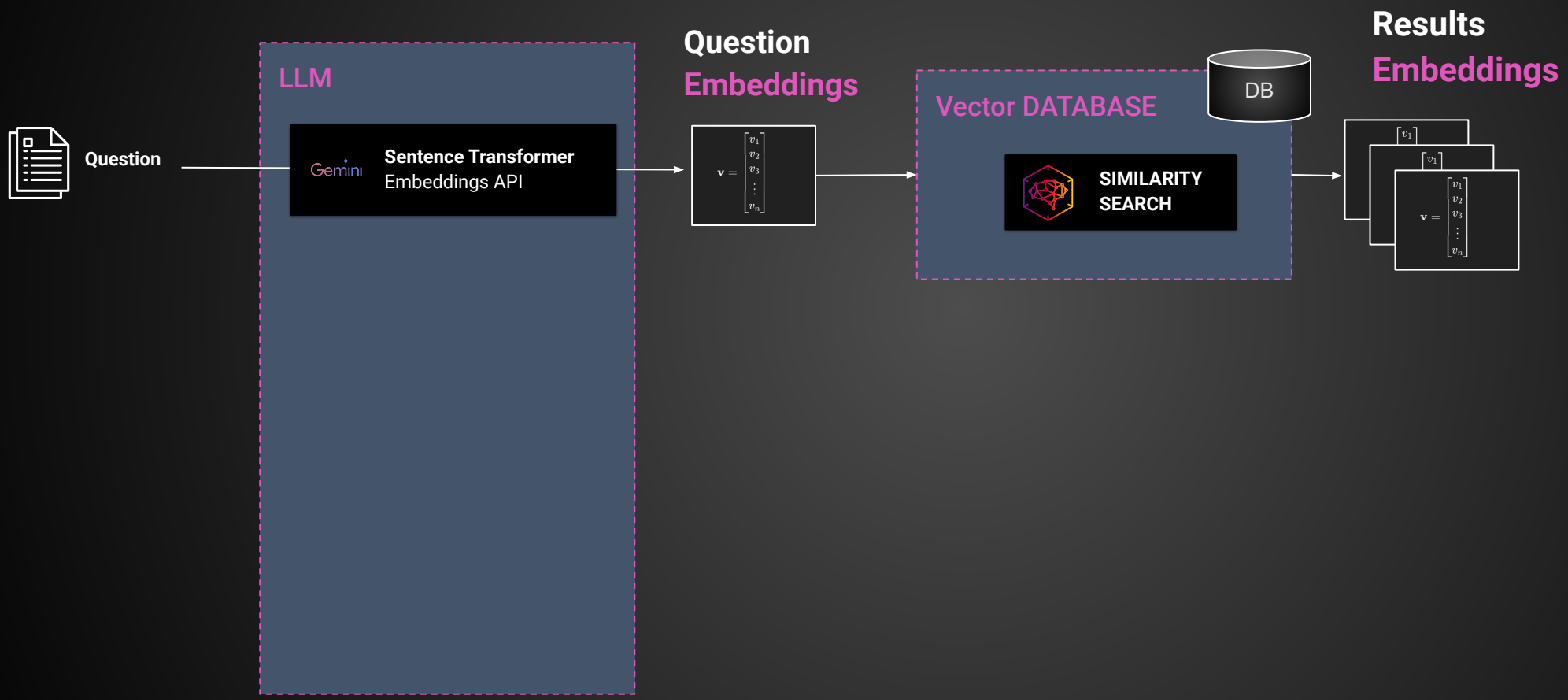
_47_hypothetical_document_embedding

Vector Search
Overview



Vector Databases

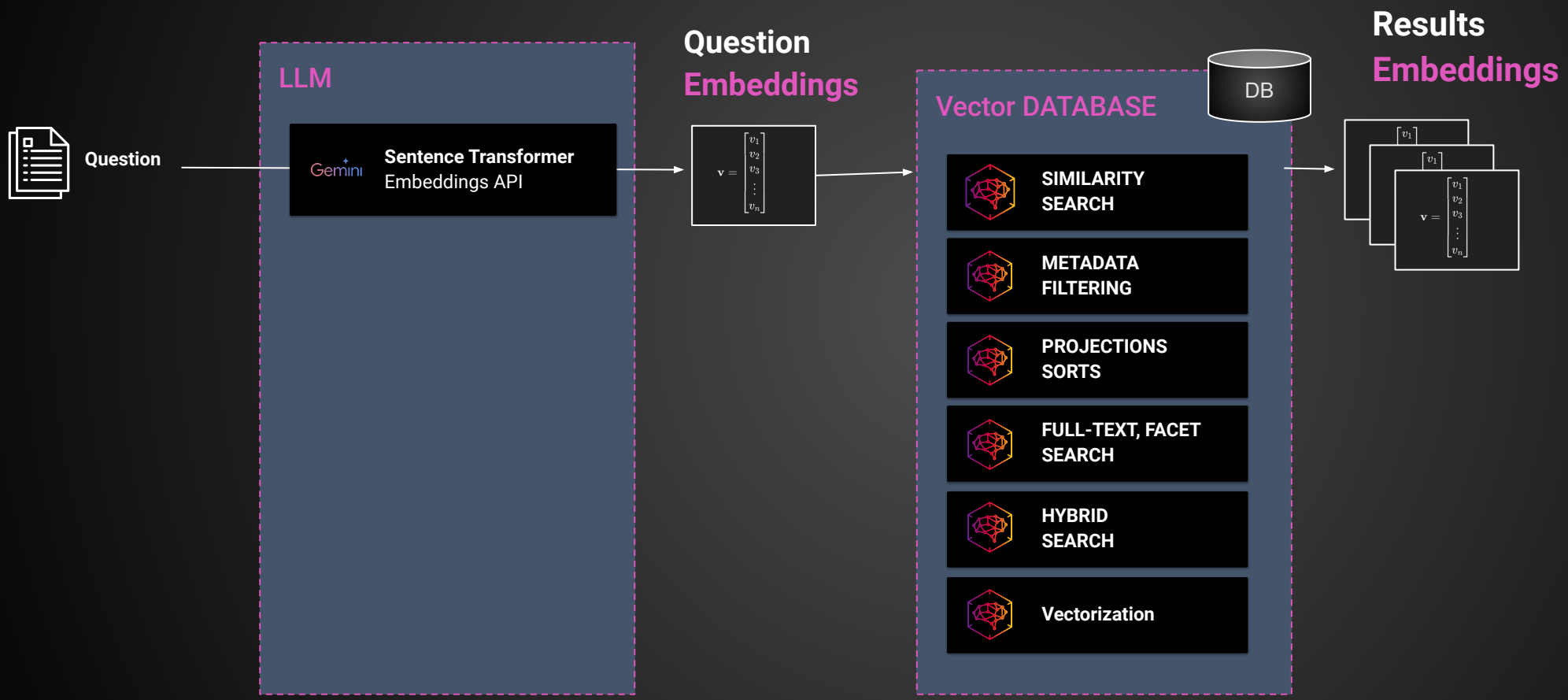
What you think...



OTHER
SOURCES

Vector Databases

But... in reality



OTHER SOURCES

Vector Databases
Vector Database Features

https://superlinked.com/vector-db-comparison

Vector DB Comparison

by Superlinked | Last Updated : Today

Search | Get insights | Give us a star

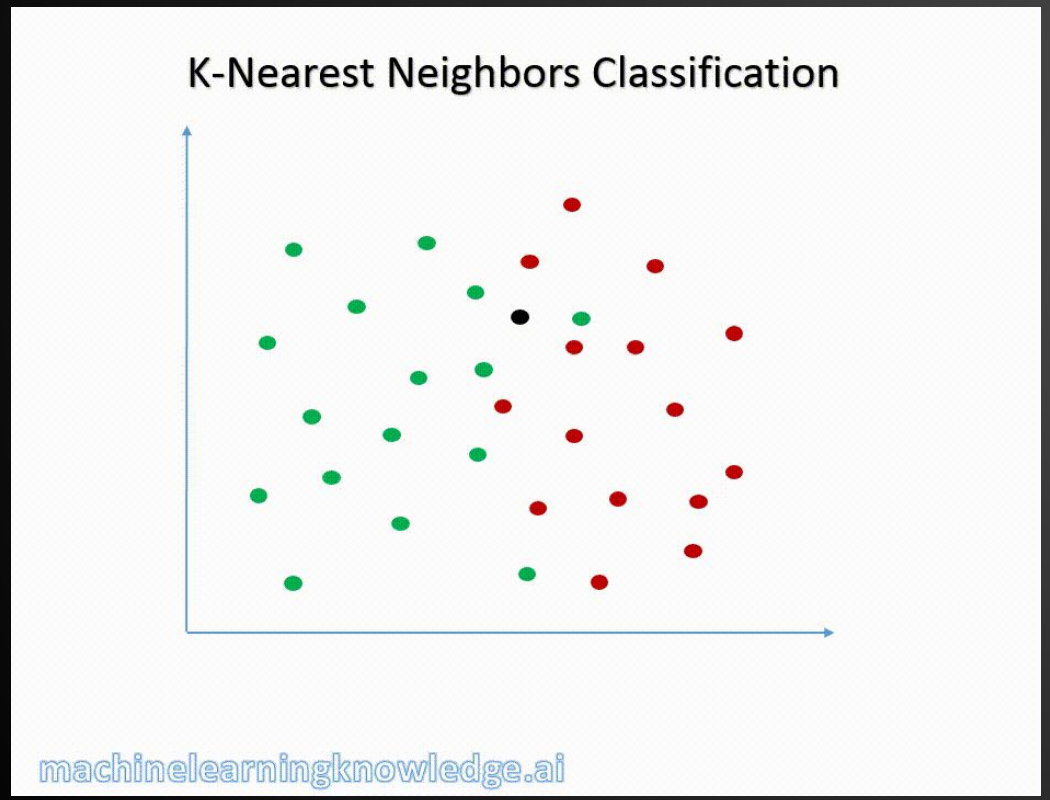
Vendor	About				Search							
	OSS	License	Dev Lang	VSS Launch	Filters	Hybrid Search	Facets	Geo Search	Multi-Vector	Sparse	BM25	Full-Text
Activeloop Deep Lake	✓	MPL 2.0	python c++	2023	✓	-	-	✗	✓	■	✗	■
Anari AI	✗	Proprietary	-	2023	-	✗	-	✗	-	-	✗	-
Apache Cassandra	✓	Apache-2.0	java	2023	✓	✓	-	-	-	✗	-	✗
Apache Solr	✓	Apache-2.0	java	2022	✓	✓	✓	✓	✓	-	✓	✓
ApertureDB	-	-	-	-	-	-	-	-	-	-	-	-
Azure AI Search	✗	Proprietary	c++	2023	✓	✓	✓	✓	✓	✗	✓	✓
Chroma	✓	Apache-2.0	python	2022	✓	✗	-	-	✗	✗	✗	✗
ClickHouse	✓	Apache 2.0	c++	2022	✓	✗	✓	✓	✓	✗	✗	✓
CrateDB	✓	Apache 2.0	java	2023	✓	-	✓	✓	✓	-	✓	✓
DataStax Astra DB	✓	Proprietary	java go	2023	✓	✓	✓	✓	✗	✗	✗	✓
Elasticsearch	✗	Elastic Lice...	java	2021	✓	✓	✓	✓	✓	✓	✓	✓

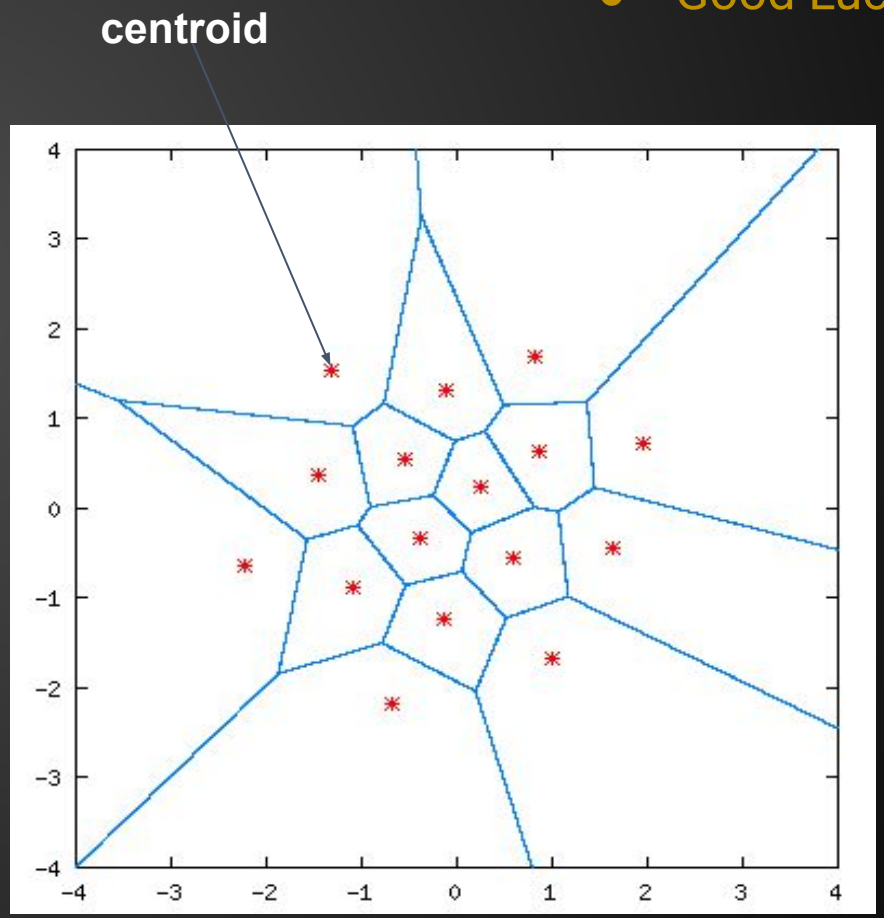
Vector Search: KNN

Inverted File Index (IVF)

https://machinelearningknowledge.ai/k-nearest-neighbor-classification-simple-explanation-beginners/

- Exhaustive
- Linear Complexity
- 1000+ dimensions
- Millions of Vector
- Good Luck !



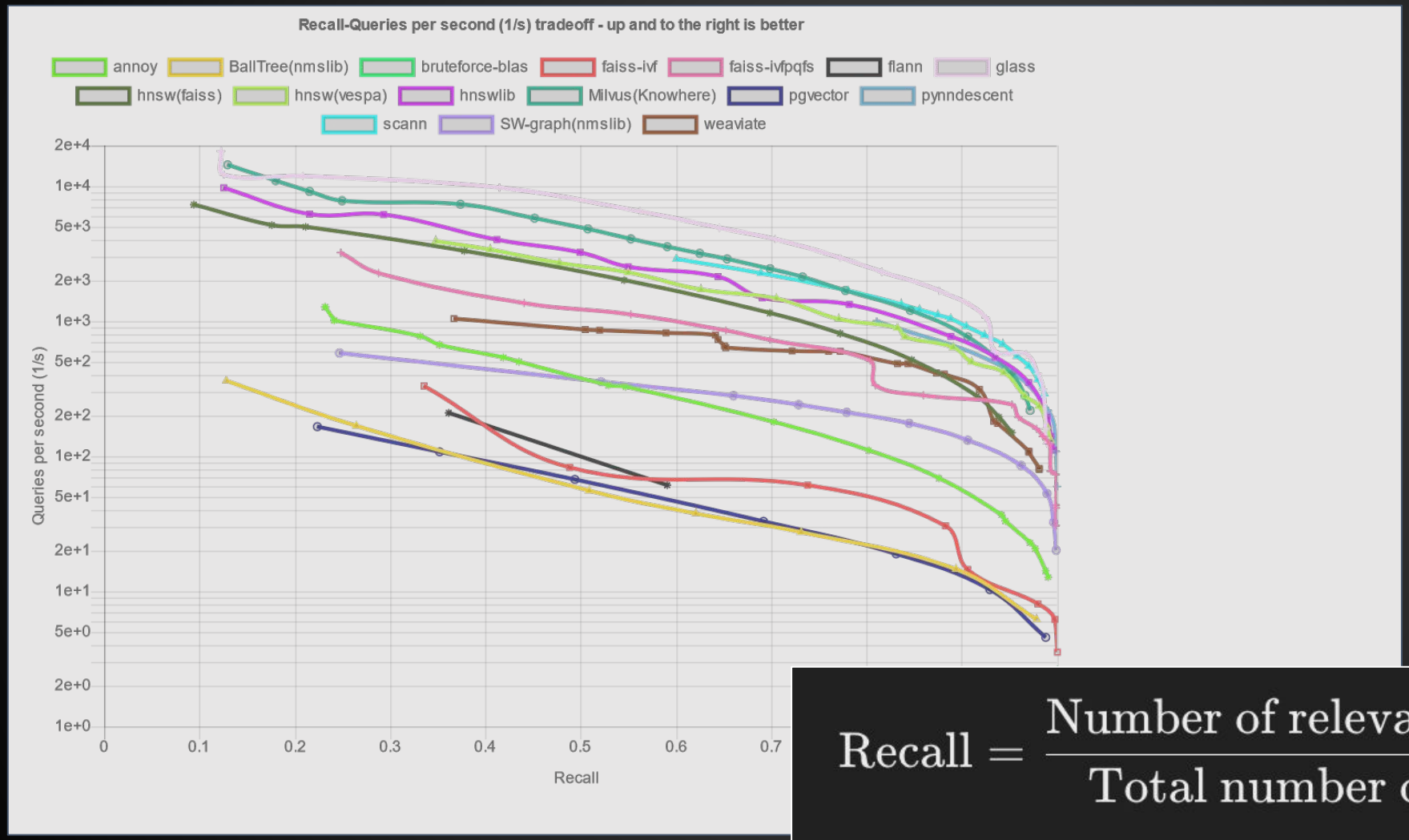


Voronoi Diagram (FAISS)

Vector Search: ANN

Approximate nearest neighbours Benchmarks

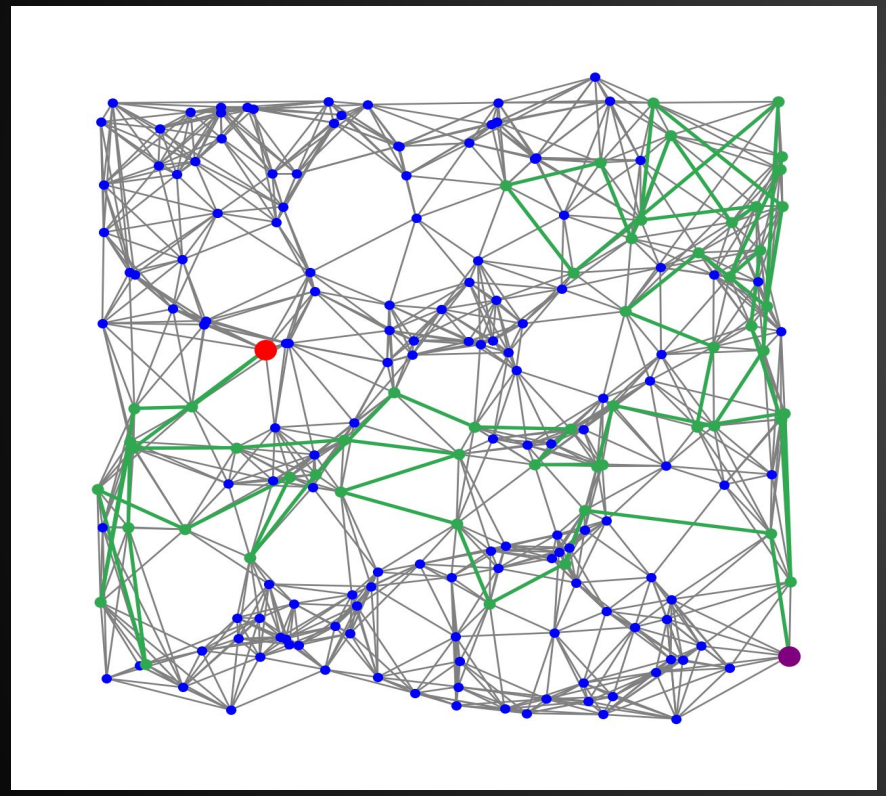
benchmarks



$$\text{Recall} = \frac{\text{Number of relevant items retrieved}}{\text{Total number of relevant items}}$$

Vector Search: ANN

Vector Indices Families



- Hash-based indexing
  - Locality-sensitive hashing
- Tree-based indexing
  - ANNOY
- Cluster-based or cluster indexing
  - Product quantization
- Graph-based indexing
  - Hierarchical navigable small world

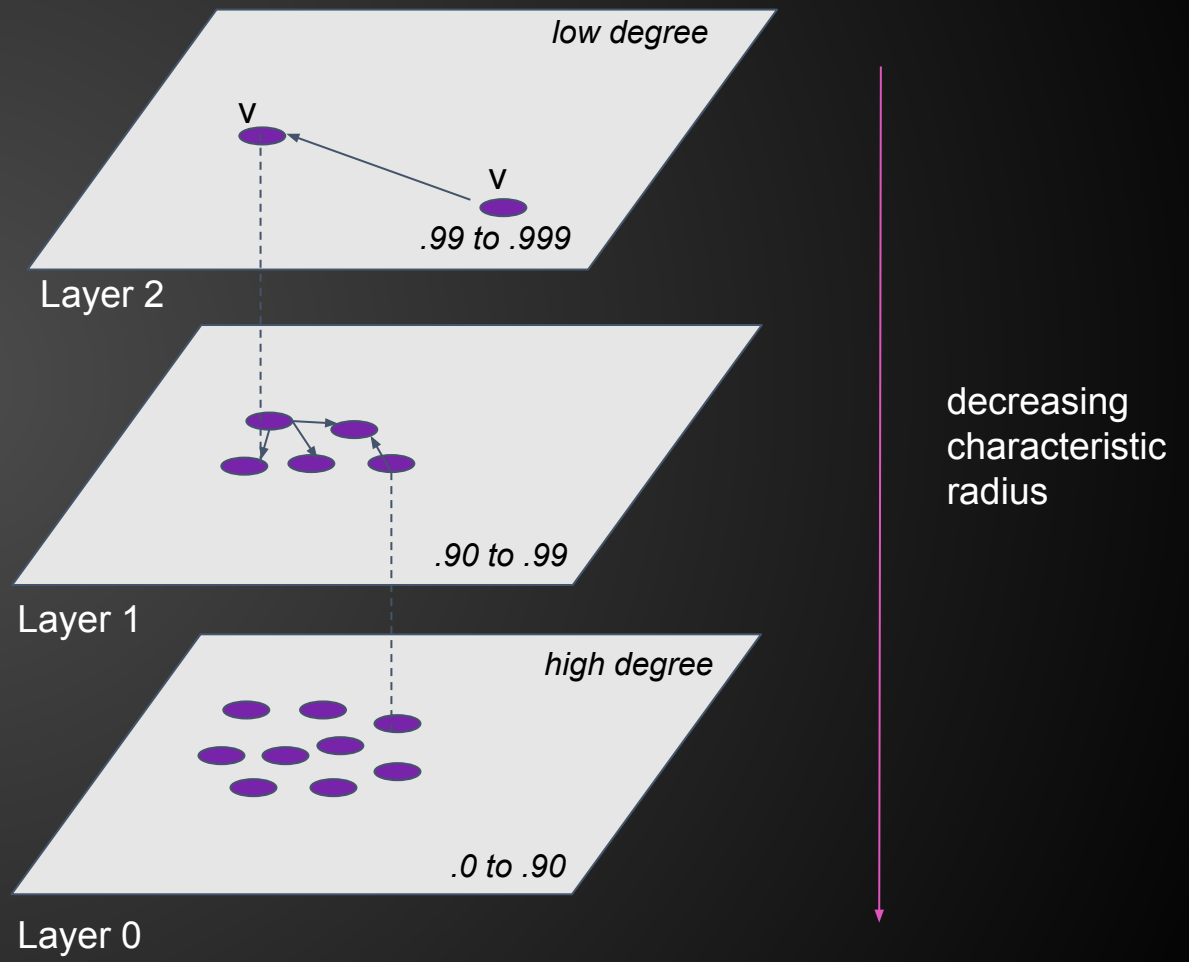
Vector Search ANN

Hierarchical Navigable Small World (HSNW)

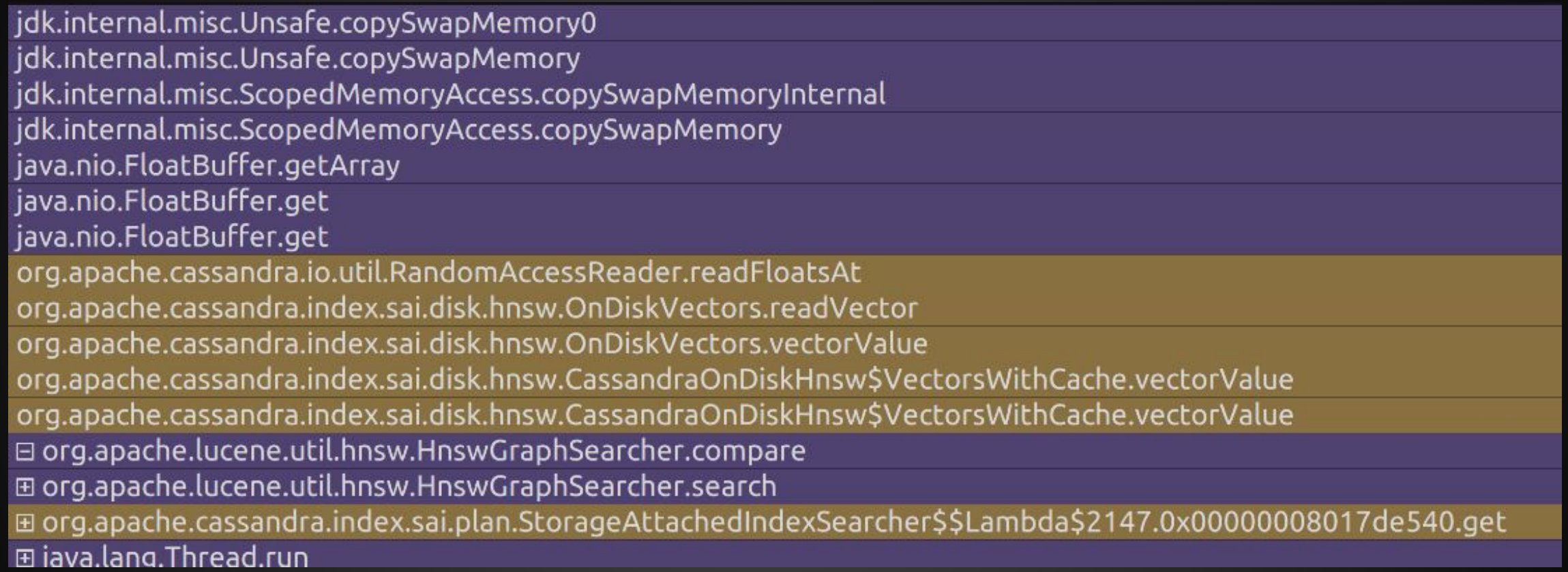
“Vector Index”

As seen in:

- Lucene (*Elastic, Solr, OpenSearch, MongoDB*)
- Weaviate
- Qdrant
- PGVector (August 2023)



HSNW on Cassandra (Wikipedia DataSet)



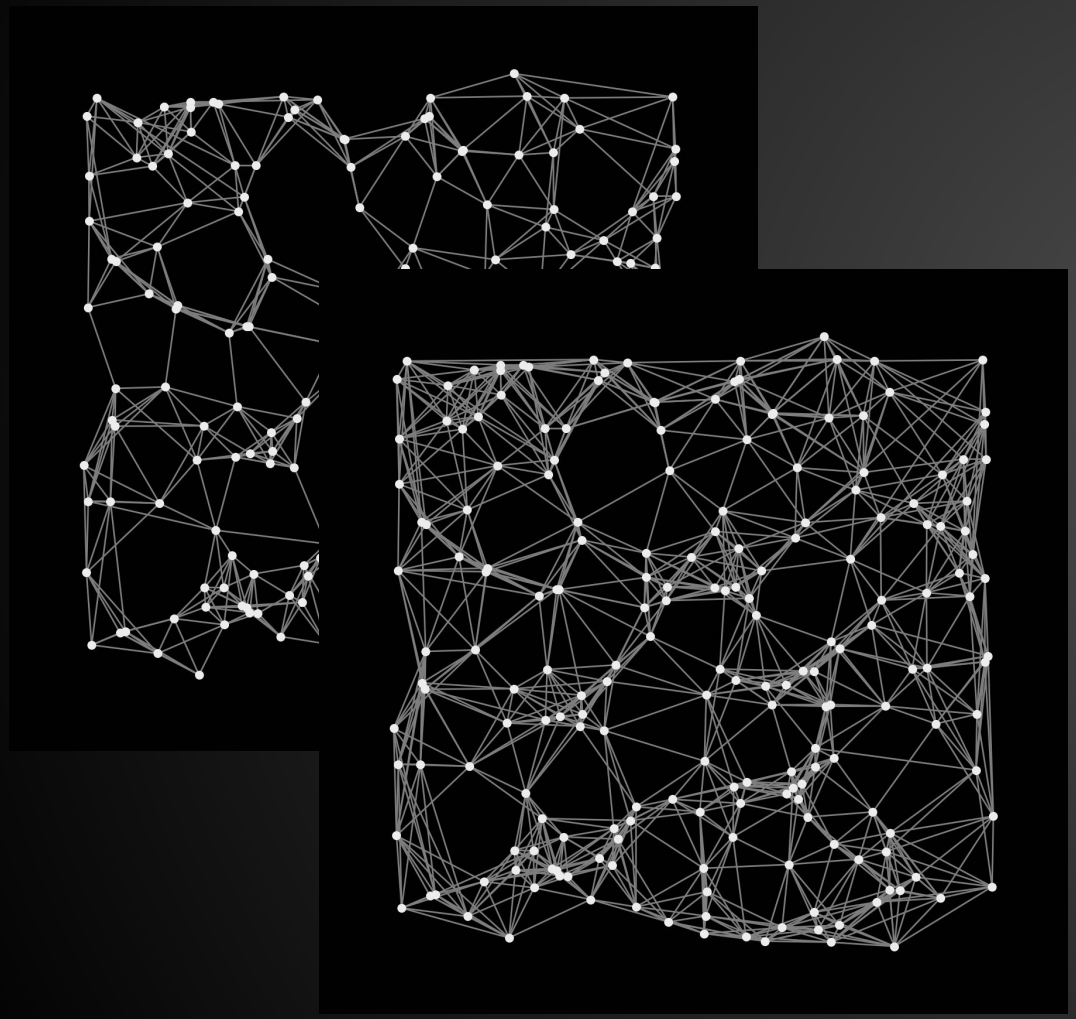
JVector
Under the hood

DiskANN =

Vamana +
Product Quantization +
Oversampling

JVector

Building the Index with Vamana

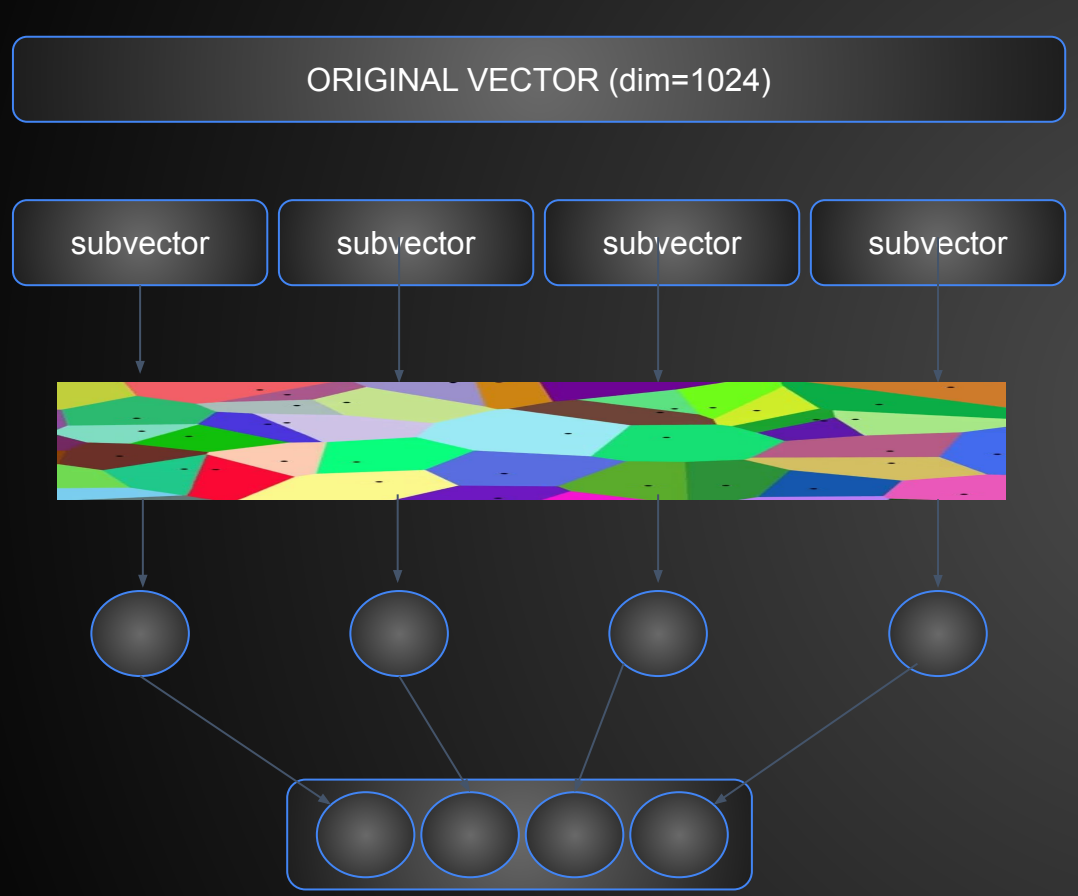


- **Data Compressed in memory**
- **Full vector on disk, less lookup**
- **Way better in high recall**

$$\text{recall} = \frac{\text{Number of docs matchings retrieved}}{\text{Number of docs matchings available}}$$

SINGER LAYER, DENSER CONNECTIONS

JVector
Product Quantization (PQ)



(m=4)

split into “subspaces”

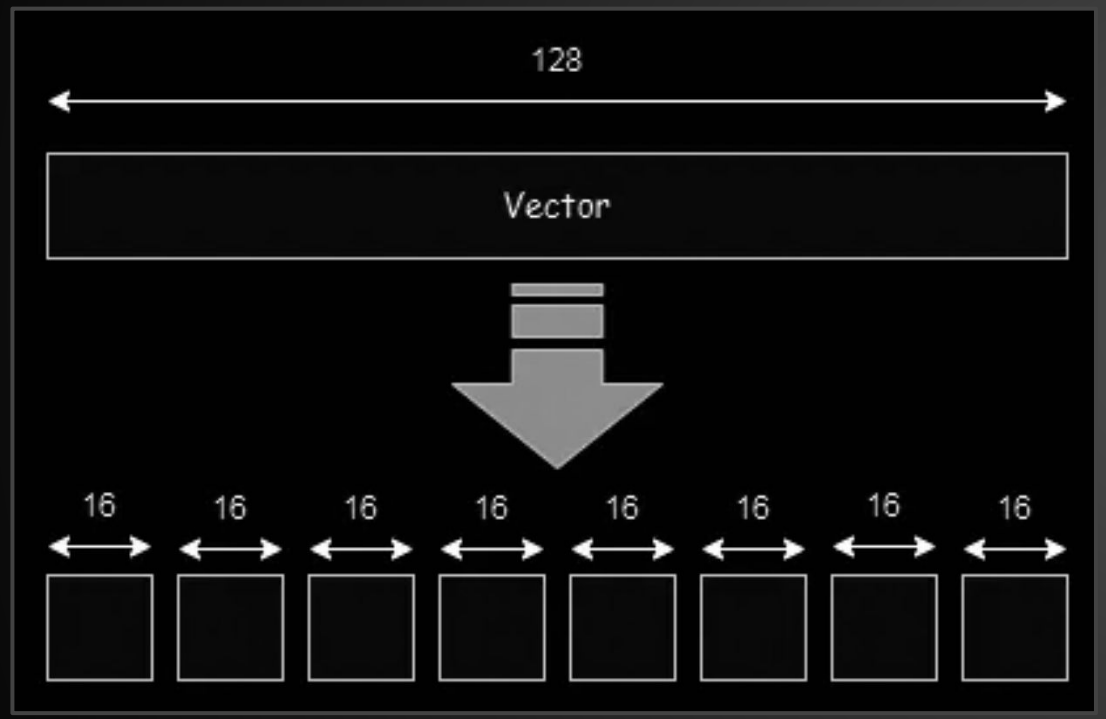
K-mean for all subspaces

replace with closest centroid (1 double)

output vector

LOSSY COMPRESSION FOR VECTORS

JVector
Product Quantization (PQ)



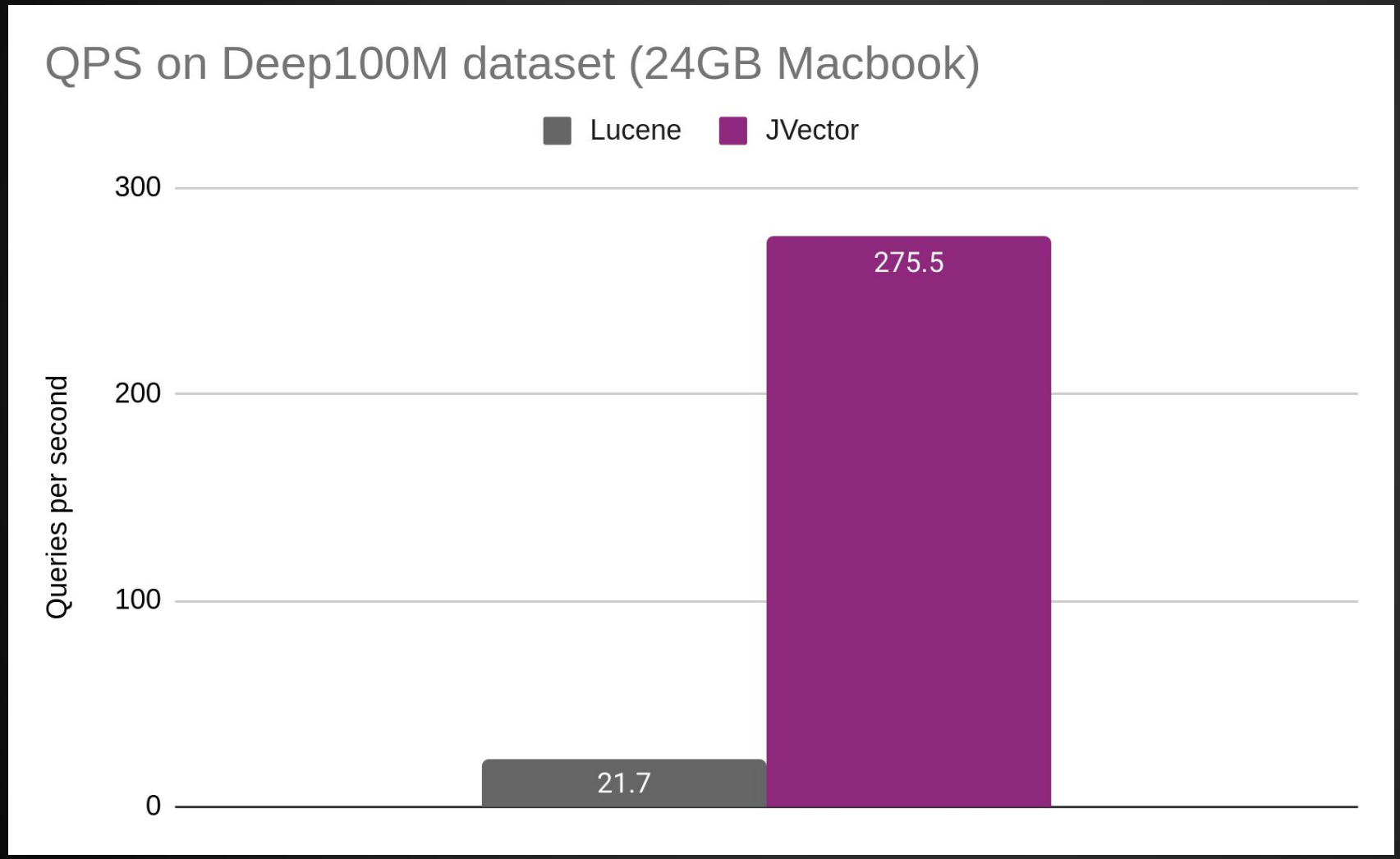
- **Efficiency**: Reduces storage requirements significantly as compared to storing full vectors.
- **Speed**: Allows for fast approximate nearest neighbors
- **Scalability**: Facilitates the handling of very large datasets by reducing the dimensionality and data redundancy.

LOSSY COMPRESSION FOR VECTORS

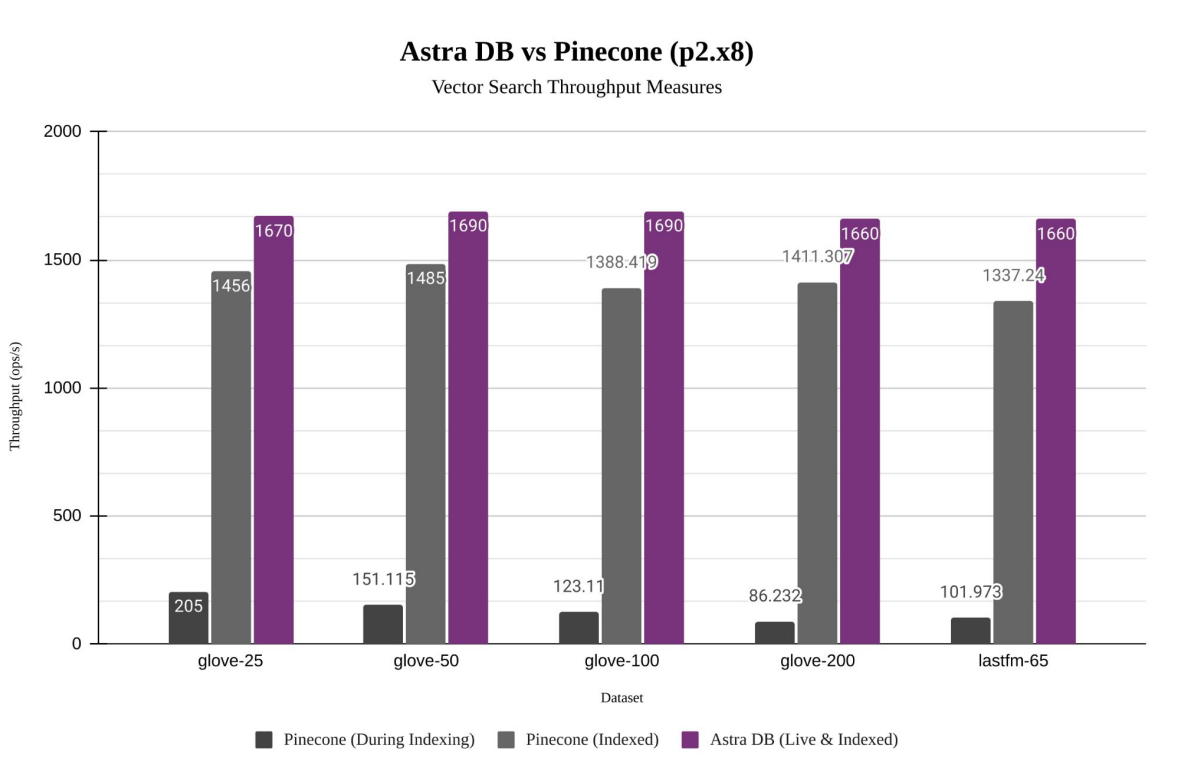
JVector
OverSample

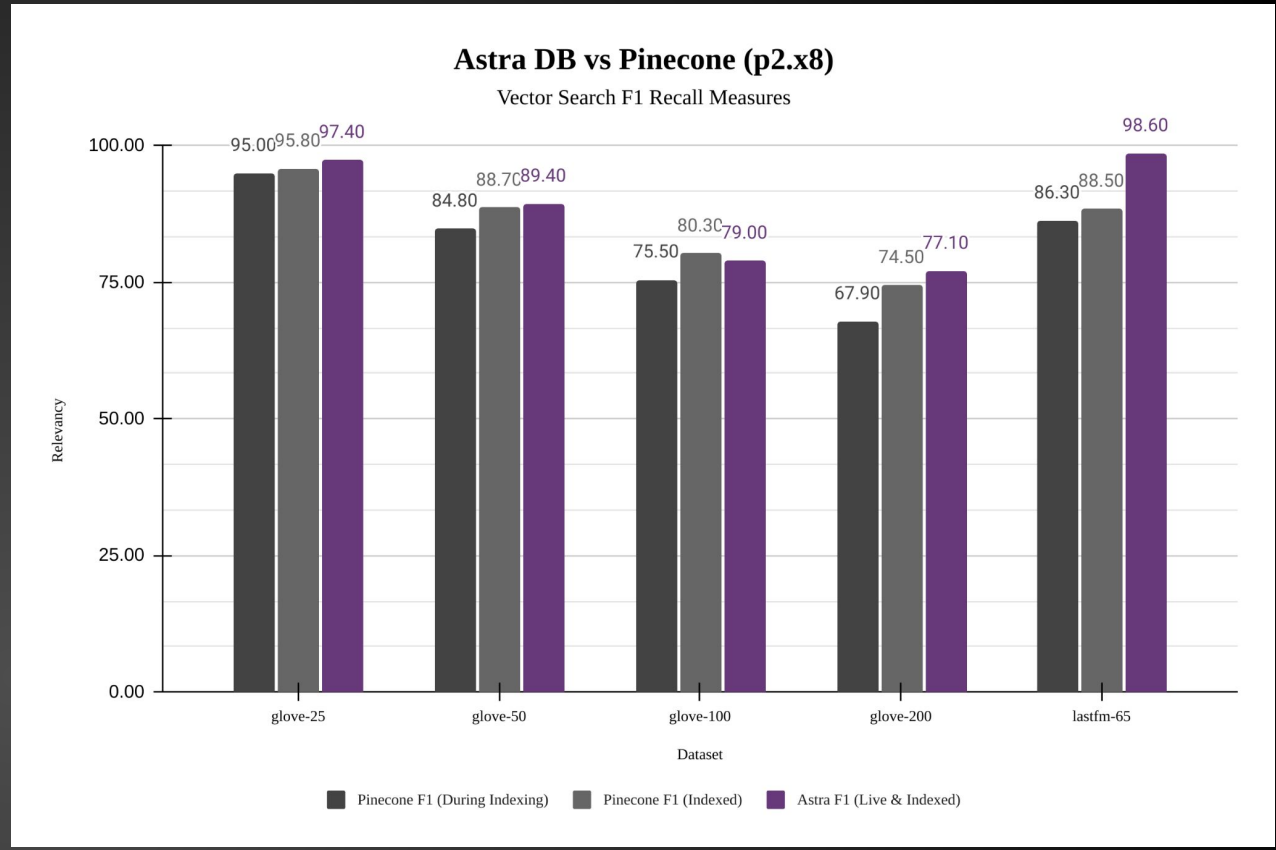
- Instead of searching for closest K, search for closest 2K *(using compressed comparisons)*
- Read uncompressed vectors from disk during search whenever a candidate is added to the resultset
- Reorder the resultset (of 2K) using uncompressed vectors, and return the top K

JVector
Some results



JVector
Some results





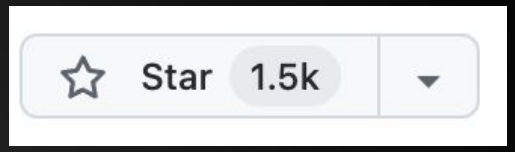
$$\text{Precision} = \frac{\text{Number of relevant items retrieved}}{\text{Total number of items retrieved}}$$

$$\text{Recall} = \frac{\text{Number of relevant items retrieved}}{\text{Total number of relevant items}}$$

$$F1 = 2 \times \left(\frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \right)$$

JVector

5 hard problems addressed



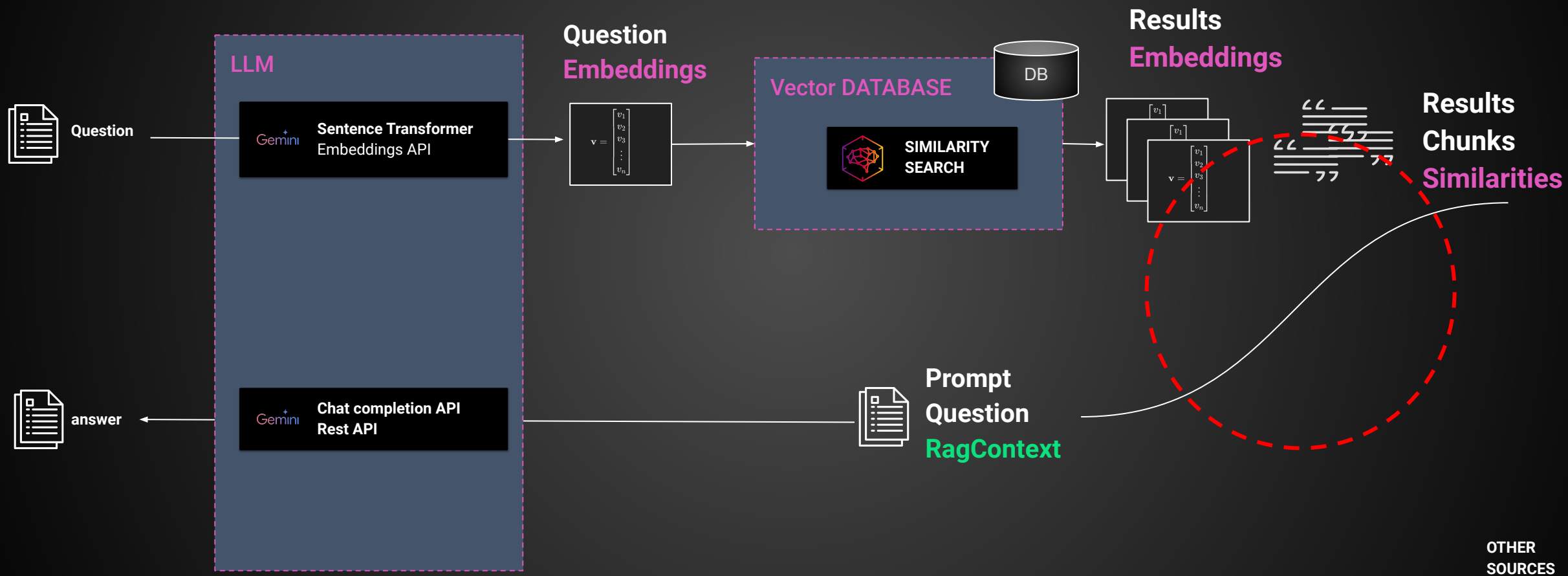
- **Scale-Out Capabilities:** No upper limits
- **Garbage Collection:** Pruning obsolete index information
- **Effective Use of Disk:** Enabling high throughput
- **Composability:** Predicates, term-based searches. Aka Hybrid Search
- **Concurrency:** Non-blocking, multi-threaded index construction

https://thenewstack.io/5-hard-problems-in-vector-search-and-how-cassandra-solves-them/

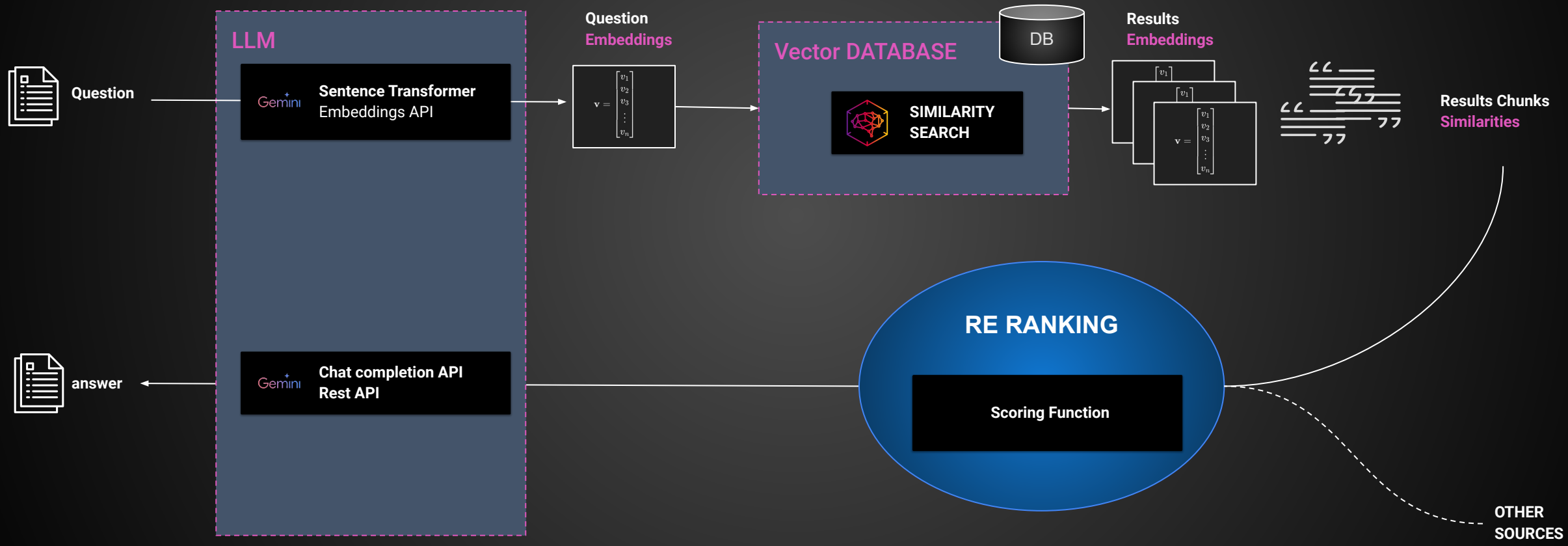
https://github.com/jbellis/jvector

Vector Databases
Forrester Wave

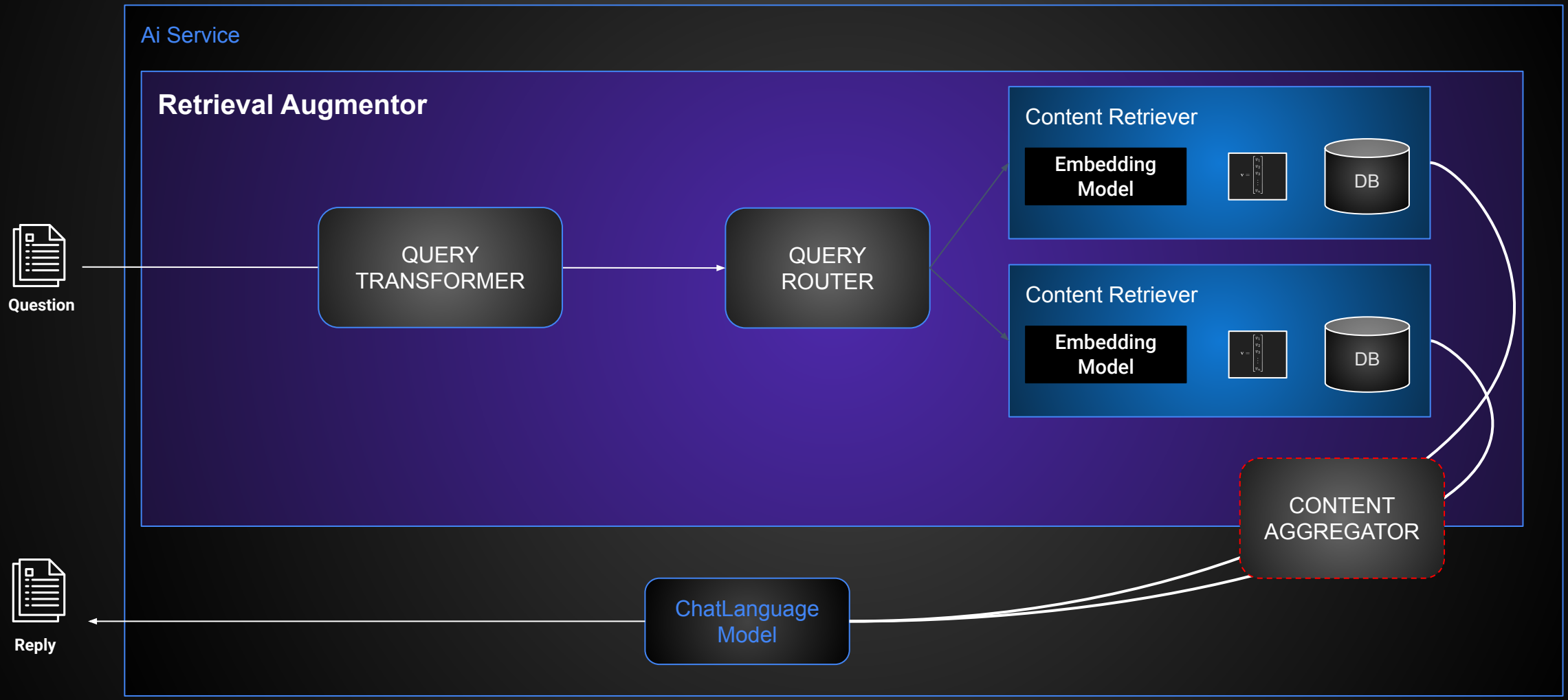
Query Post-Processing
Overview



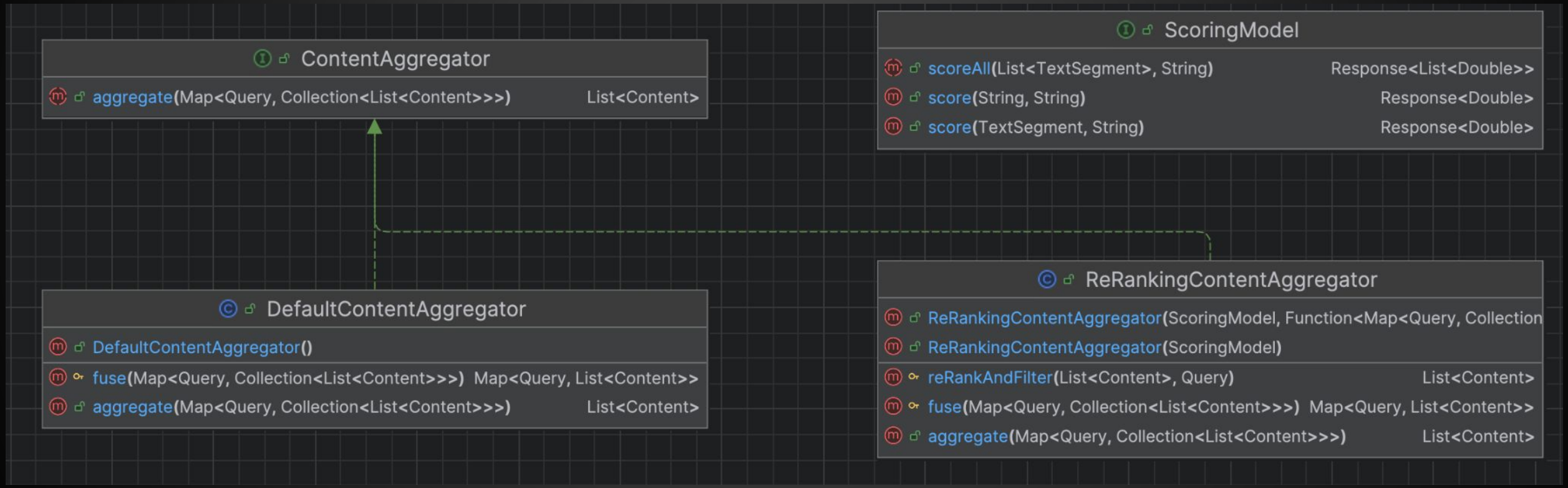
ReRanking
General Principle



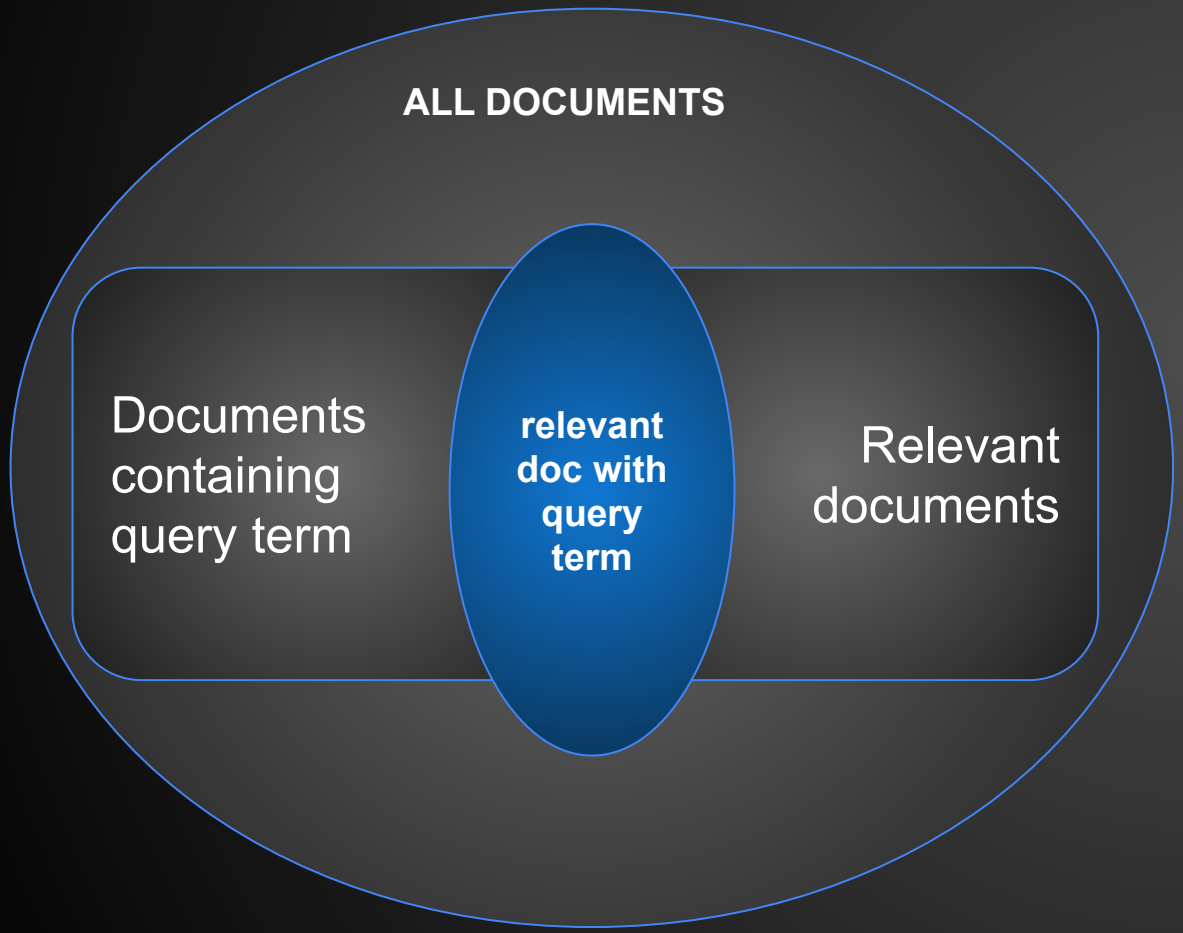
ReRanking
LangChain4j “Content Aggregator”



ReRanking
LangChain4j “Content Aggregator”



ReRanking
BM25 (Best Matching 25)



Parameters:

- How often do the query terms appear in the document (Term Frequency, TF)
- Inverse document Frequency (IDF)
- The length of the document (DL)
- The average length of all documents in the collection (AVDL)

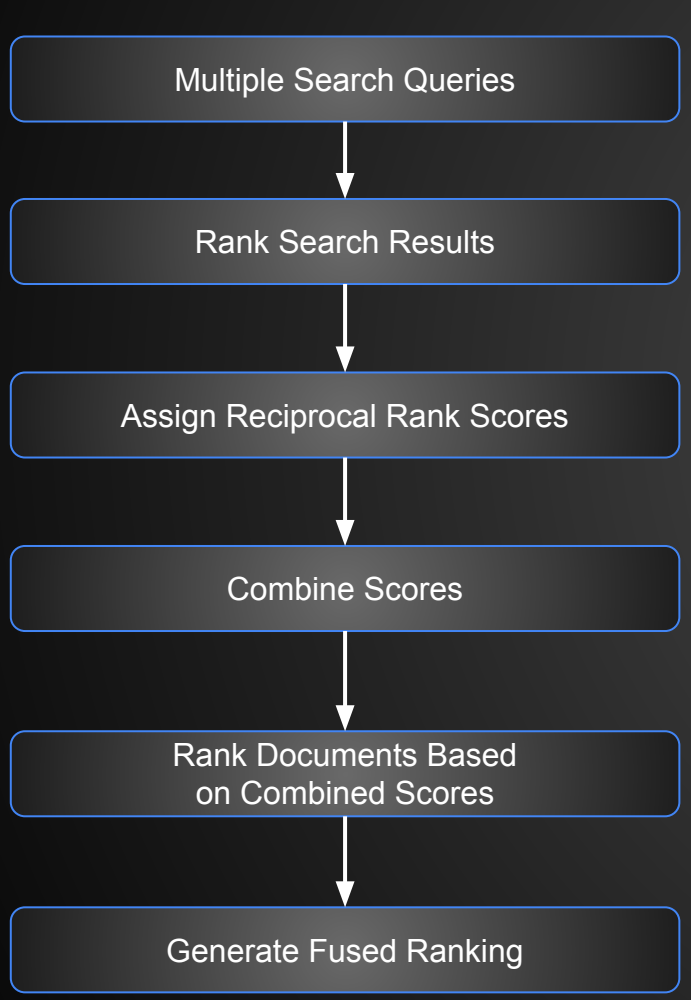
Pro:

- Dynamic rankings
- Good for long queries

Cons:

- No Semantic
- No personalization

ReRanking
Reciprocal Rank Fusion (RRF)

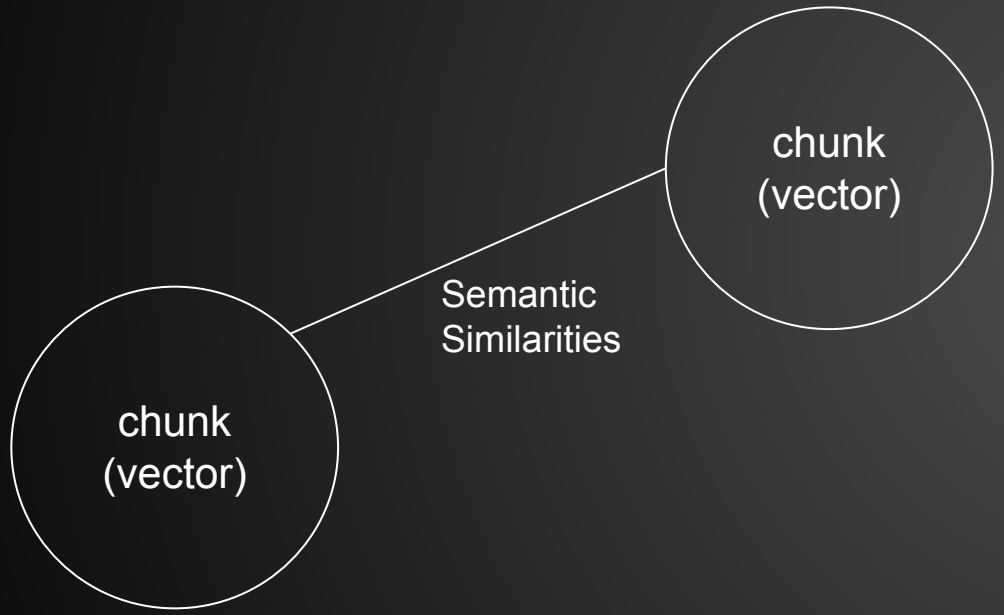


- Different Similarity Searches with different filters
- Similarity Search + BM25

$$\text{RRF Score} = \sum_{k=1}^{N} \frac{1}{k + rank(d)}$$

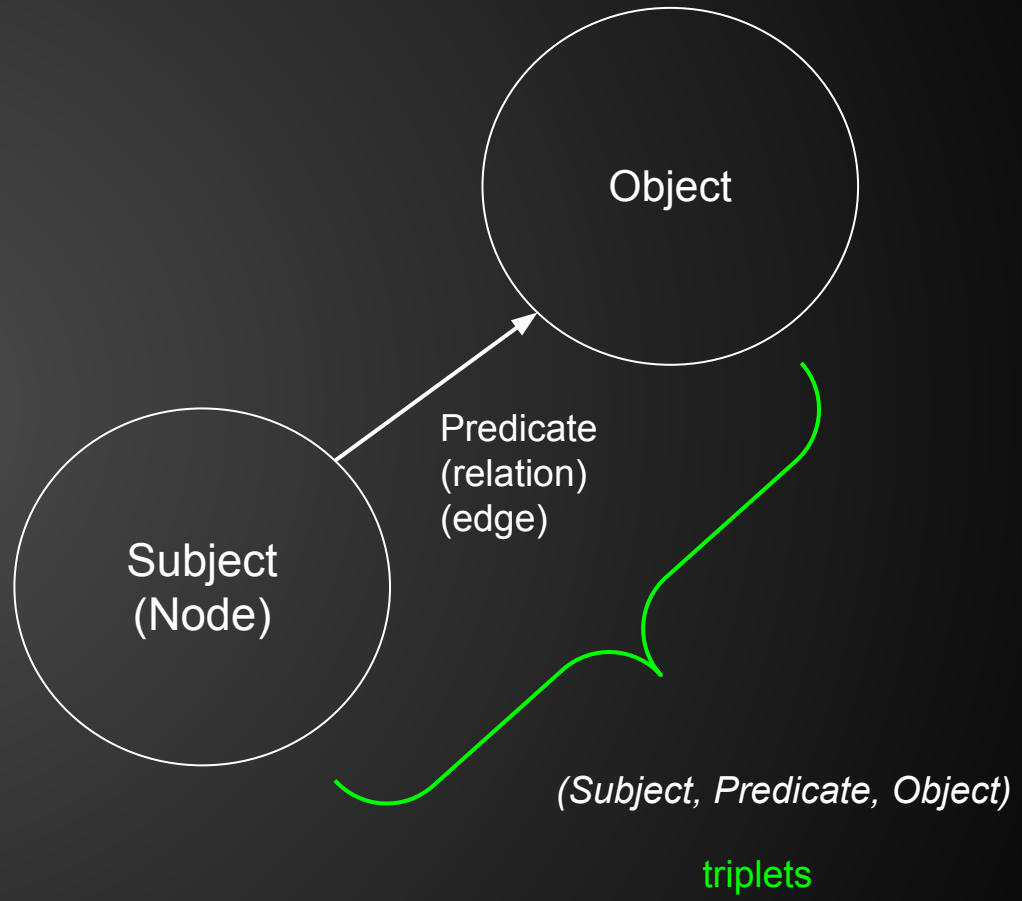
- k constant
- *rank(d)* rank position of doc d in each series (N)

Graph RAG
Overview



vector indices (vector databases)

Structured Knowledge Graph



Graph RAG
Overview

Building the Graph (Knowledge Extraction)

- Link content based on hyperlinks in HTML
- How to: Link content based on common keywords (using Keybert)
- How to: Link content based on named entities (using GLiNER)
- How to: Link content based on document hierarchy

Contextual Embeddings

- Embedded the triplet
- Context-aware embeddings

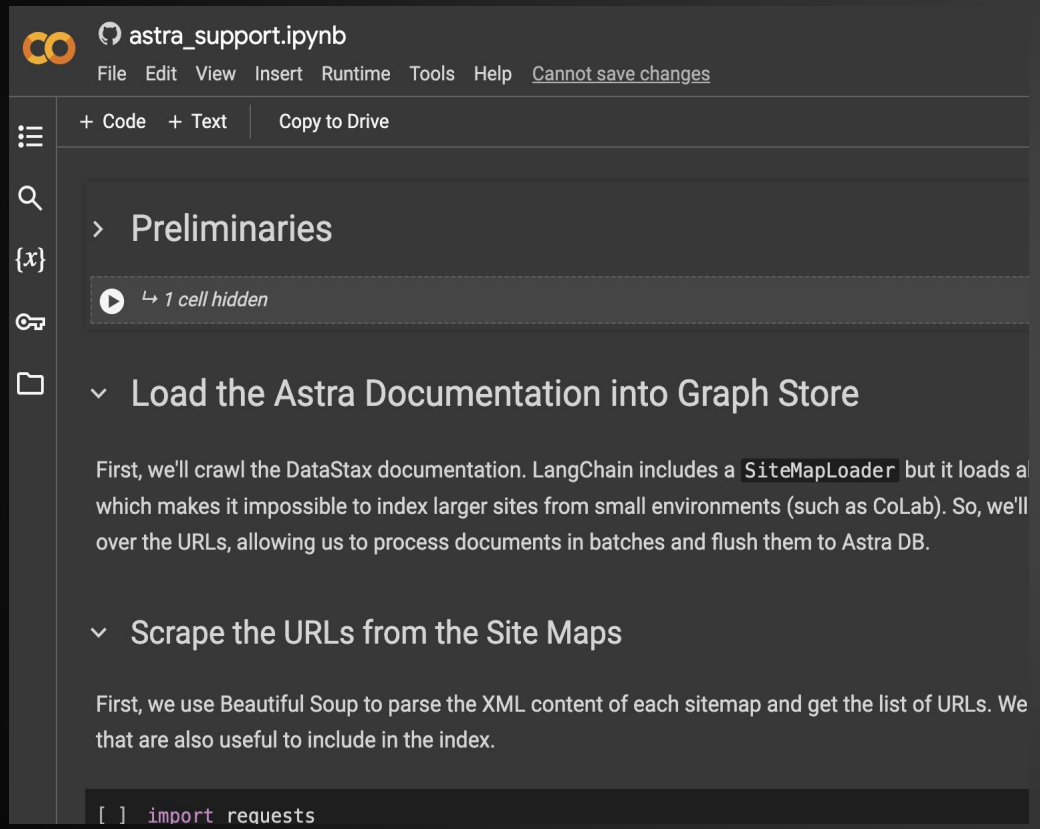
BUILD GRAPH QUERY WITH LLM

GREMLIN / CYPHER

GRAPH TRAVERSAL

Triplets
Embeddings

Graph RAG
CassandraGraphStore



[link]

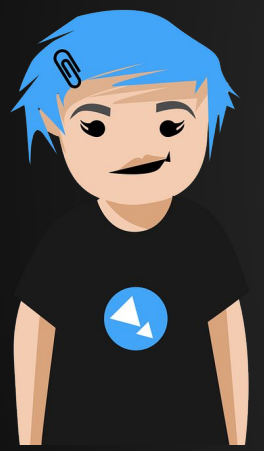
```
CREATE TABLE astra_docs_nodes (
    content_id text PRIMARY KEY,
    kind text,
    link_to_tags set<tuple<text, text>>,
    links_blob text,
    metadata_blob text,
    text_content text,
    text_embedding vector<float, 1536>
)
```

```
CREATE TABLE graph_targets (
    kind text,
    tag text,
    target_content_id text,
    target_text_embedding vector<float, 1536>,
    PRIMARY KEY ((kind, tag), target_content_id)
);
```

5. Advanced Concepts

- Functions Calling
- Agents, Agentic RAG

Function calling



What's the weather like in Paris?

Chatbot app

Gemini

user prompt + getWeather(String) function contract

call getWeather("Paris") for me please 🙏

getWeather("Paris")

External API or service

{"forecast":"sunny"}

function response is {"forecast":"sunny"}

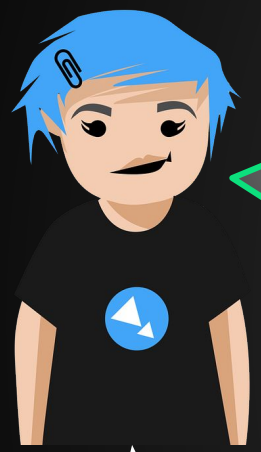
Answer: "It's sunny in Paris!"

It's sunny in Paris!

AI Agents

- Different types, and capabilities
  - **Reflection**
    - Chain-of-Thought, self reflection & correction, self grading
  - **Planning**
    - Create a multi-step plan of action
  - **Tool use**
    - Multiple function calling
  - **Multi-agent collaboration**
    - Chain several LLMs and/or RAG searches

Agentic RAG



Berlin's origins, population, geographic situation

🛠️ History/Geography Tool 🛠️

1) Execute RAG search
2) Call topic assistant to summarize topic

Vector database

TOPICAL REPORT

TOPICAL REPORTS

🧠 Agentic Assistant 🧠

1) Identify topics
2) Create questions
3) RAG search
4) Collect answers & generate final report

FINAL REPORT

🧠 Topic Assistant 🧠

1) Study topic answers
2) Create a report summary on the topic

_49_agentic_RAG

6. Quality and Data Governance

- RAG Evaluation
- Security
- Data Lifecycle

RAGAS evaluation metrics

GENERATION

faithfulness

how factually acurate is
the generated answer

answer relevancy

how relevant is the generated
answer to the question

RETRIEVAL

context precision

the signal to noise ratio of retrieved
context

context recall

can it retrieve all the relevant information
required to answer the question

DeepEval evaluation metrics

- Introduction
- G-Eval
- Summarization
- Answer Relevancy
- Faithfulness
- Contextual Precision
- Contextual Recall
- Contextual Relevancy
- Tool Correctness
- Hallucination
- Bias
- Toxicity
- RAGAS
- Custom Metrics
- Multimodal Metrics
  - VIEScore
- Conversational Metrics
  - Conversation Completeness
  - Conversation Relevancy
  - Knowledge Retention

Evaluation
MRR and NDCG

Essential to compare each of the final generated answers in response to user queries.

- MRR (Mean Reciprocal Rank):
  - Measures how quickly the first relevant result appears by averaging the reciprocal of its rank across queries.
- NDCG (Normalized Discounted Cumulative Gain):
  - Assesses the overall ranking quality by assigning higher importance to relevant items at top positions, considering both relevance and position.
- Other Techniques
  - Bilingual Evaluation Understudy (BLEU) : Text translation
  - Recall-Oriented Understudy for Gisting Evaluation (ROUGE) : Text Summarization
  - BERTscore

LLM as Judge

- Prepare a dataset of questions and **golden responses**
- Use your RAG pipeline to answer those questions
- Use an **LLM as a judge** to gauge the quality of your RAG results, against a set of metrics

OWASP Top 10 for LLM Applications

LLM01

Prompt Injection

This manipulates a large language model (LLM) through crafty inputs, causing unintended actions by the LLM. Direct injections overwrite system prompts, while indirect ones manipulate inputs from external sources.

LLM02

Insecure Output Handling

This vulnerability occurs when an LLM output is accepted without scrutiny, exposing backend systems. Misuse may lead to severe consequences like XSS, CSRF, SSRF, privilege escalation, or remote code execution.

LLM03

Training Data Poisoning

This occurs when LLM training data is tampered, introducing vulnerabilities or biases that compromise security, effectiveness, or ethical behavior. Sources include Common Crawl, WebText, OpenWebText, & books.

LLM04

Model Denial of Service

Attackers cause resource-heavy operations on LLMs, leading to service degradation or high costs. The vulnerability is magnified due to the resource-intensive nature of LLMs and unpredictability of user inputs.

LLM05

Supply Chain Vulnerabilities

LLM application lifecycle can be compromised by vulnerable components or services, leading to security attacks. Using third-party datasets, pre- trained models, and plugins can add vulnerabilities.

LLM06

Sensitive Information Disclosure

LLMs may inadvertently reveal confidential data in its responses, leading to unauthorized data access, privacy violations, and security breaches. It's crucial to implement data sanitization and strict user policies to mitigate this.

LLM07

Insecure Plugin Design

LLM plugins can have insecure inputs and insufficient access control. This lack of application control makes them easier to exploit and can result in consequences like remote code execution.

LLM08

Excessive Agency

LLM-based systems may undertake actions leading to unintended consequences. The issue arises from excessive functionality, permissions, or autonomy granted to the LLM-based systems.

LLM09

Overreliance

Systems or people overly depending on LLMs without oversight may face misinformation, miscommunication, legal issues, and security vulnerabilities due to incorrect or inappropriate content generated by LLMs.

LLM10

Model Theft

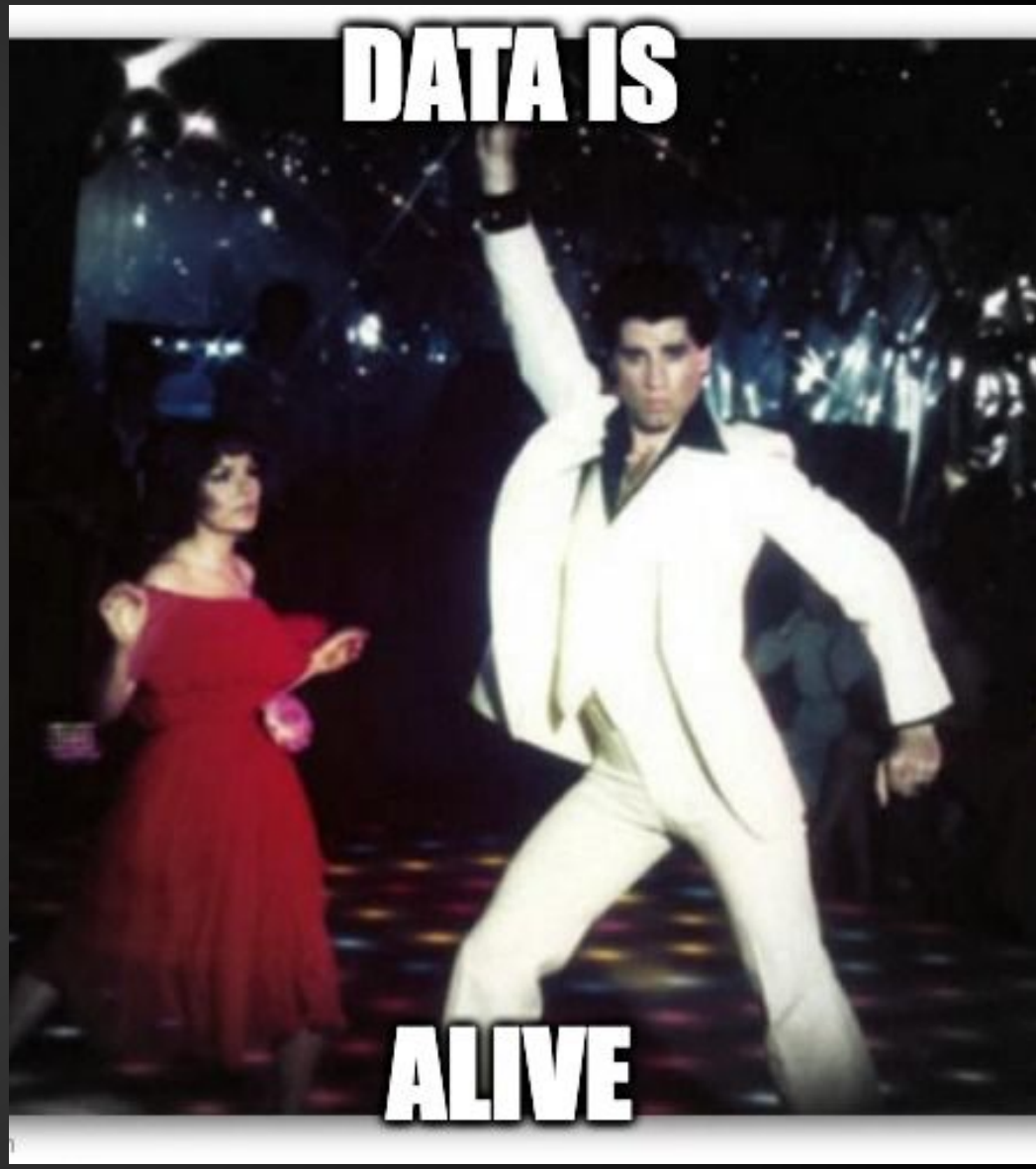
This involves unauthorized access, copying, or exfiltration of proprietary LLM models. The impact includes economic losses, compromised competitive advantage, and potential access to sensitive information.

Security & Data Privacy

- Anonymize data *(for ex. with Google Cloud Data Loss Prevention)*
- Don’t log PII details
- Use local models when possible
- Separate tenants for compliance with data protection laws

Data Lifecycle

- Your **data isn't stale**, it's alive
- When a document is updated,
  - chunking has changed
  - old chunks need to be retired
- Chunk metadata should track document origin, last update timestamps or document versions
- Prepare an update schedule



Conclusion

- It's hard… no "*one size fit all*" solution
- The different types of questions *(Multi-hop & reasoning tasks)*
- LLMs with large context windows are great at reasoning!

Lots of techniques, which one to pick?

There are easy questions... and hard ones!

Type	Description	Example
Yes/No	Answer is a Yes or No	Has Lady Gaga ever made a song with Ariana Grande?
Comparative	Compare 2 items by an attribute	Is Mont Blanc taller than Mount Rainier?
Generic	Simple questions	Where was Michael Phelps born?
Intersection	Requires multiple conditions	Which movie was directed by Denis Villeneuve and stars Timothee Chalamet?
Ordinal	Based on item's position in a list	Who was the last Ptolemaic ruler of Egypt?
Count	Answer requires counting	How many astronauts have been elected to Congress?
Difference	Contains a negation	Which Mario Kart game did Yoshi not appear in?
Superlative	Max or Min of given attribute	Who was the youngest tribute in the Hunger Games?
Multi-hop	Requires multiple steps to answer	Who was the quarterback of the team that won Super Bowl 50?

Mintaka: A complex, natural, and multilingual dataset for end-to-end question answering. arXiv preprint arXiv:2210.01613

LLM w/ large context + advanced database + agentic

Combine the best of both worlds!

- Implement **Retrieval Augmented Generation** with a capable **vector database**
- Use **multistep agentic reasoning** with **LLMs with large context windows**

Gemini

Astra DB

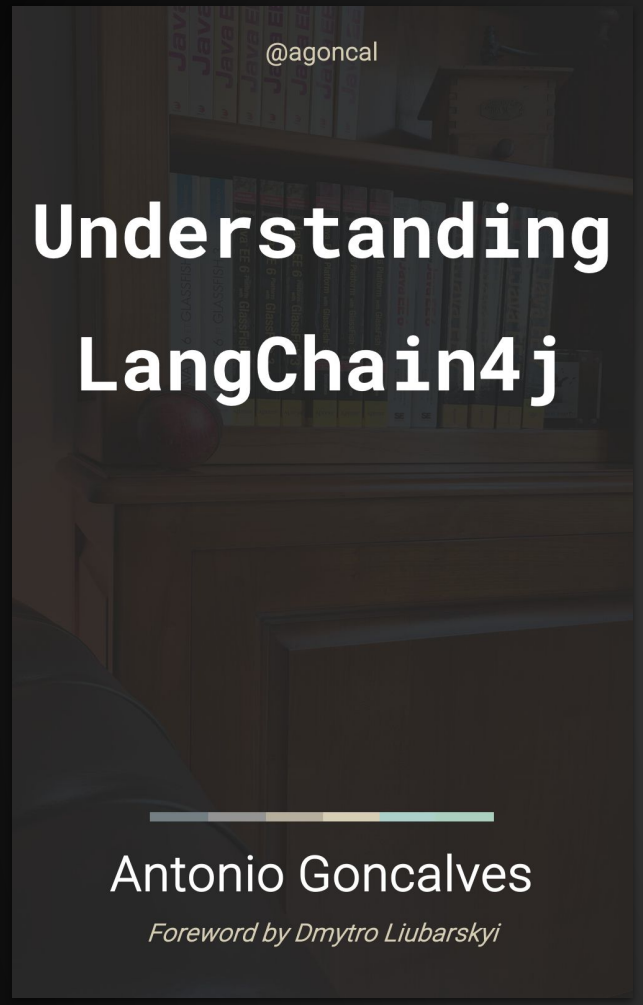


Table of Contents (220 pages):

https://amazon.com/author/agoncal

https://agoncal.teachable.com

Thanks for your attention!

(is all you need?)

github.com/
datastaxdevs/
conference-2024-devoxx