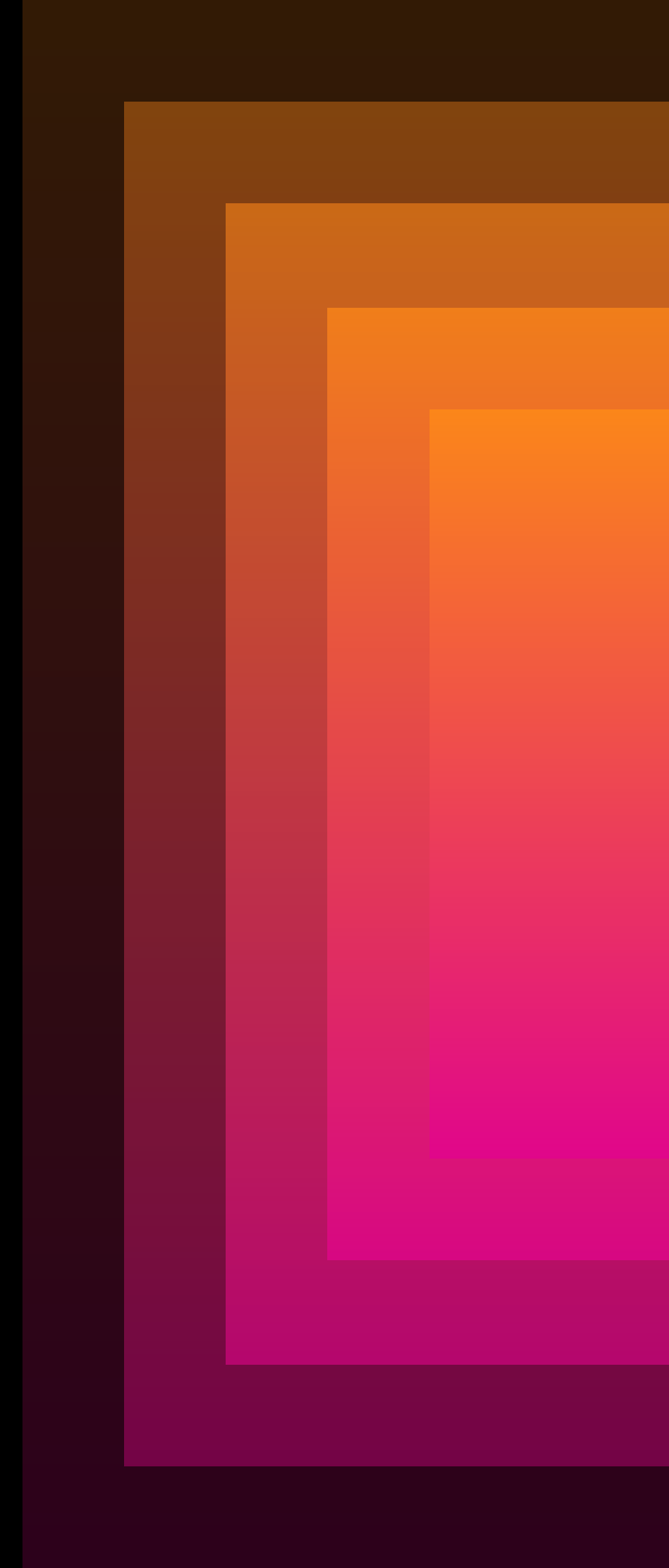
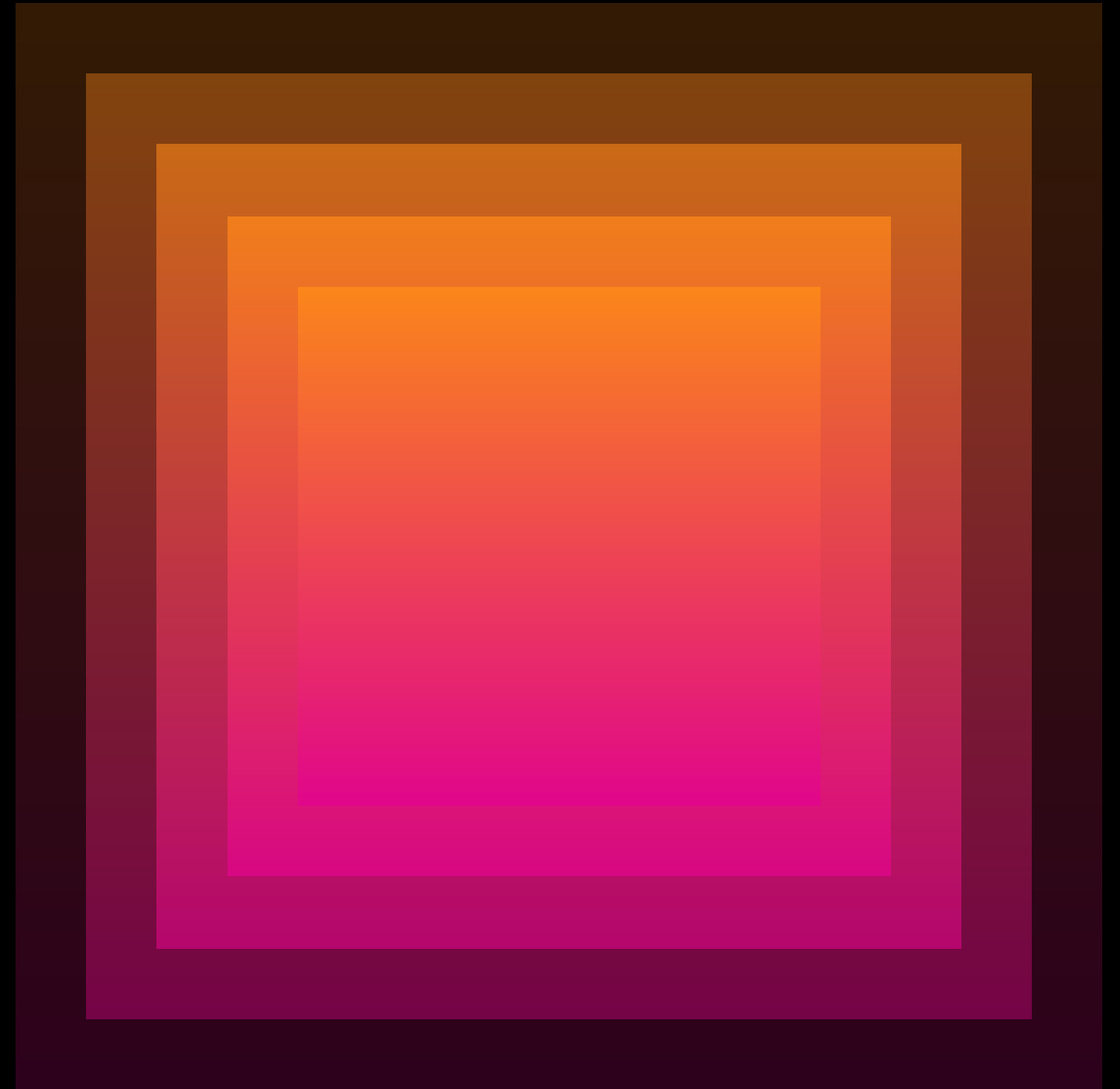

Retrieval Augmented Generation (RAG)

Prepared by: Abdallah Hammad



Overview

- Why do we need RAG?
 - RAG
 - The Retriever
 - Chunking
 - Vector DB
 - RAG relevance
 - Frameworks
-



Why do we need RAG?

- Knowledge Cutoff
 - Proprietary Data
 - Hallucination
 - Source Attribution
 - Generic Responses
-

Why do we need RAG?

- Customer Support
 - Research
 - Legal and Healthcare
 - Education
-

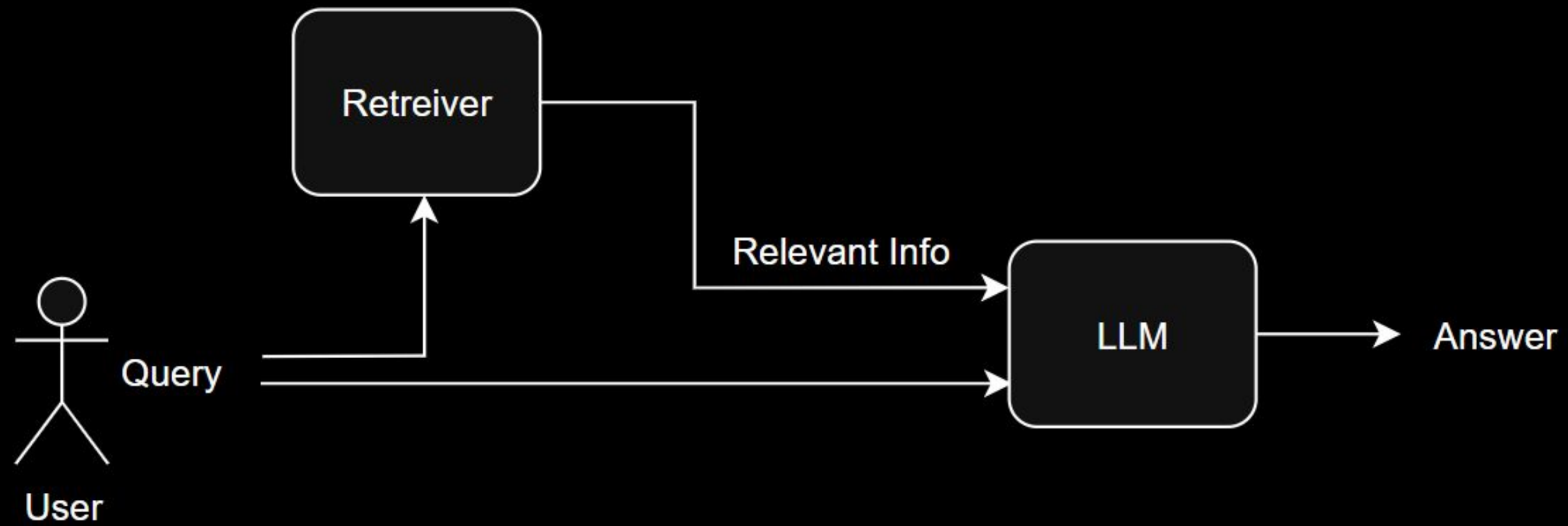
Without RAG



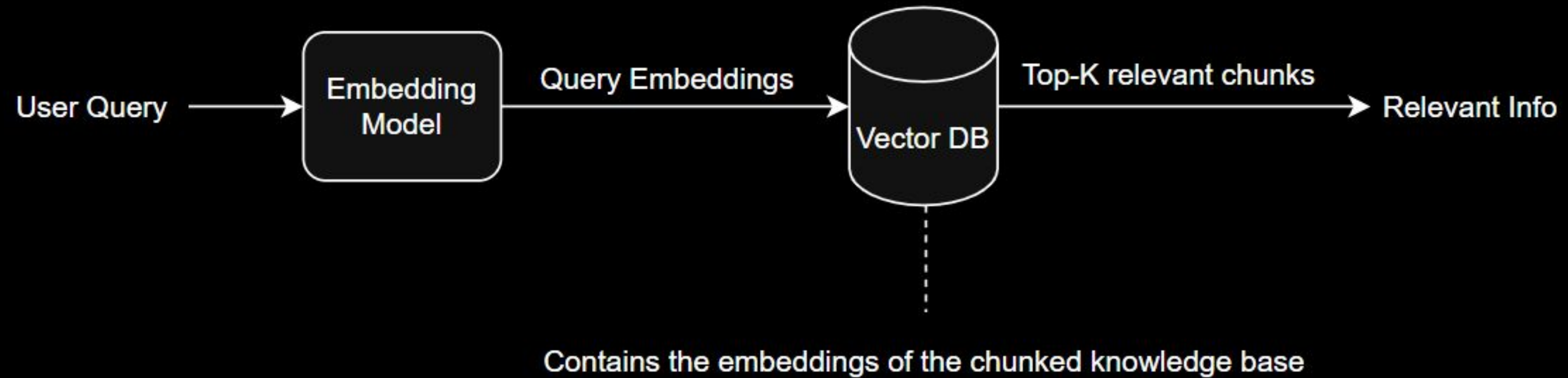
RAG

What if we use the user query to retrieve relevant information from a knowledge base that the LLM can then use as context?

RAG



The Retriever



The Retriever

1. The user query is embedded
 2. The embedded user query is compared to the embedded chunks
 3. Top-K relevant chunks are returned
-

Chunking

- Cost saving
 - Hallucinations
 - Meaningful Embeddings
-

Chunking methods

- Sliding-Window
- Semantic
- Agentic
- Custom

```
Adding: 'The month is October.'  
No chunks, creating a new one  
Created new chunk (51322): Date & Times
```

```
Adding: 'The year is 2023.'  
Chunk Found (51322), adding to: Date & Times
```

```
Adding: 'I was a child at some past time.'  
No chunks found  
Created new chunk (a6f65): Personal History
```

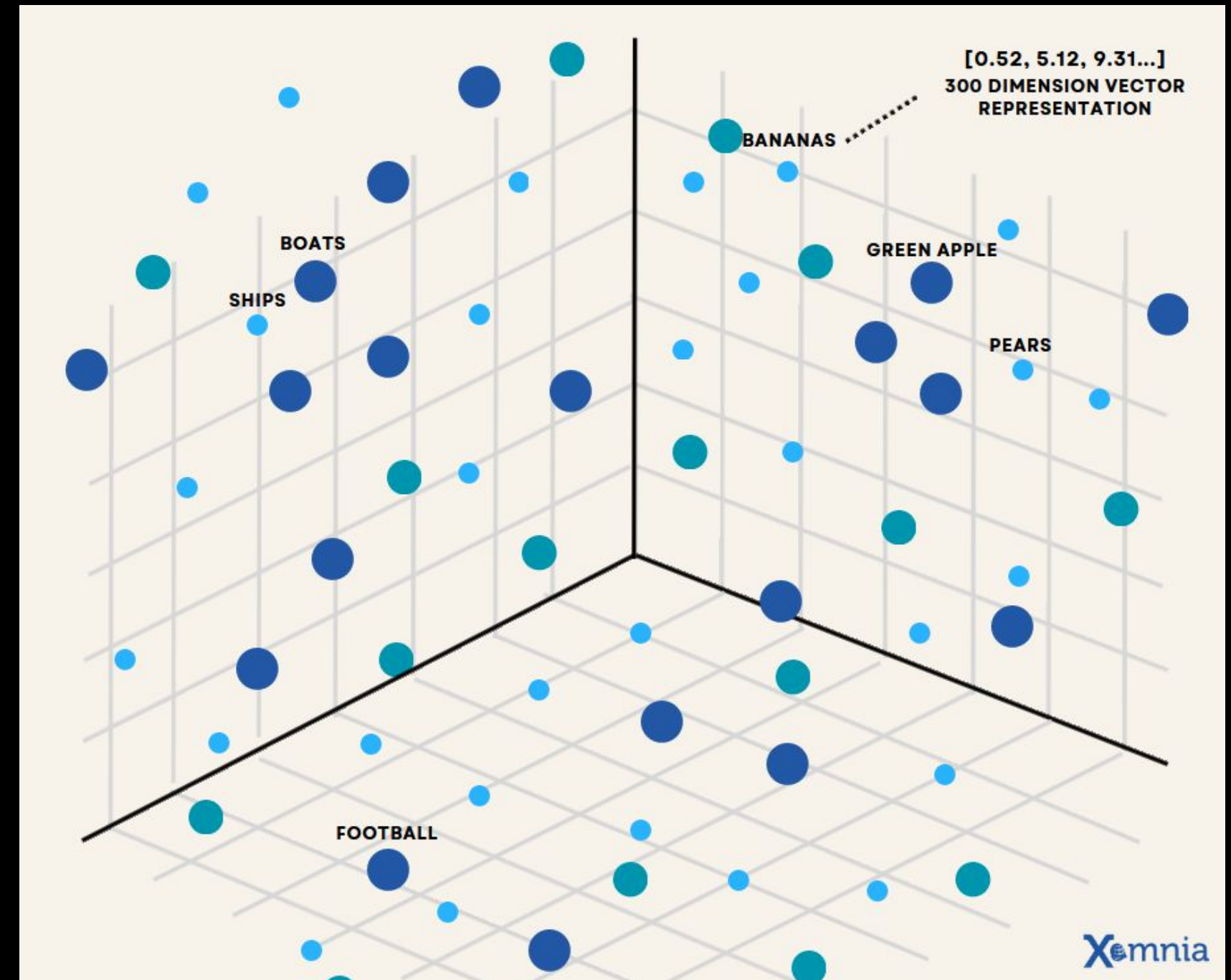
```
Adding: 'At that past time, I did not understand something important about the world.'  
Chunk Found (a6f65), adding to: Personal History
```

```
Adding: 'The important thing I did not understand is the degree to which the returns for performance are superlinear.'  
No chunks found  
Created new chunk (4a183): Performance & Returns Relationship
```

















```
Adding: 'Teachers and coaches implicitly told us the returns were linear.'  
Chunk Found (4a183), adding to: Performance & Returns Relationship
```

Vector DB

- Stored embedded chunks
- Groups similar chunks
- Enables searching for similar chunks



Vector DB examples

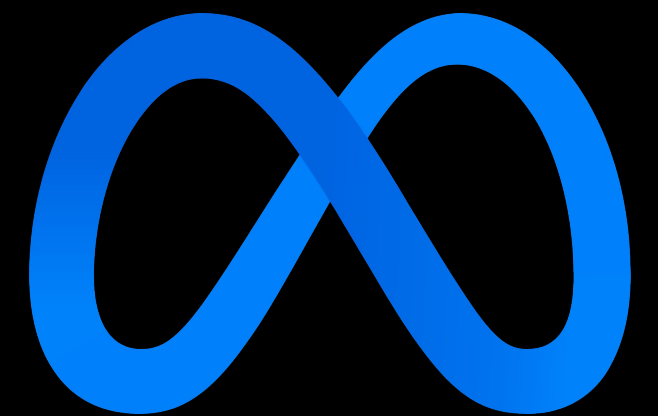
	Dedicated vector databases	Databases that support vector search
Open source (Apache 2.0 or MIT license)	 chroma  vespa  LanceDB  marqo  drant  Milvus	 OpenSearch  ClickHouse  PostgreSQL  cassandra
Source available or commercial	 Weaviate  Pinecone	 elasticsearch  redis  [ROCKSET]  SingleStore

Is RAG still relevant?

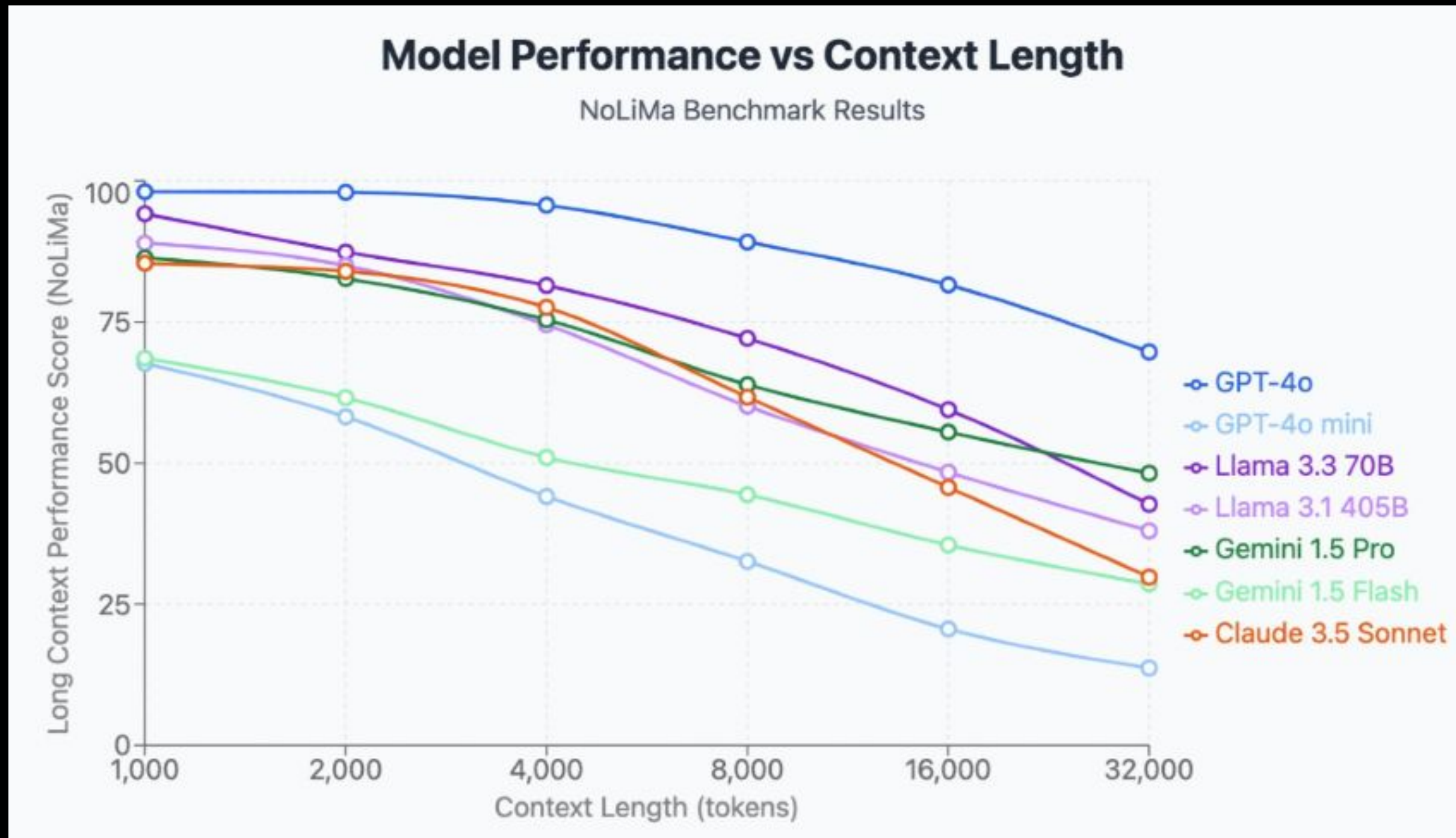
- Large context windows (1m Tokens)
- Lower costs
- Higher intelligence

Gemini

Closed
~~OpenAI~~



Yes, yes it is



RAG frameworks

- LangChain
 - Defy
 - RAGFlow
 - LlamaIndex
 - smolagents
-

Thank you!
