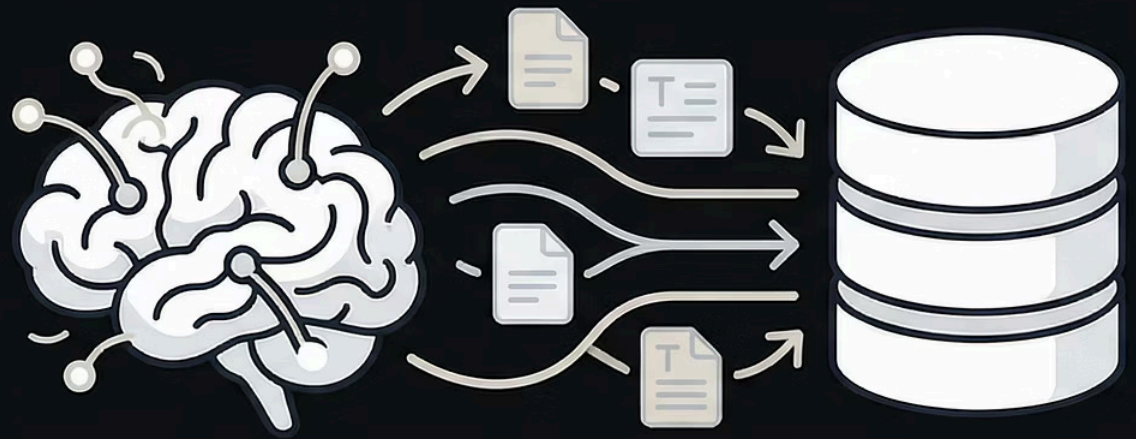
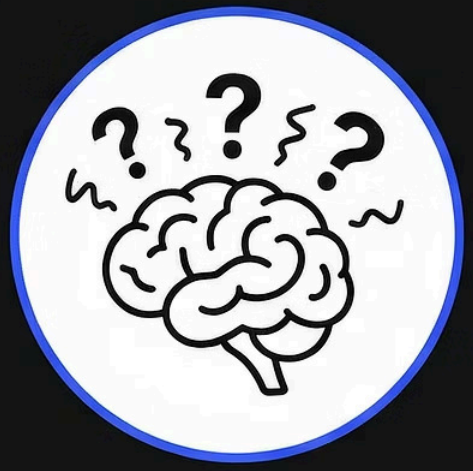


Retrieval-Augmented Generation

Revolutionizing AI with Knowledge Retrieval

Why Are We Learning RAG?



Hallucinations

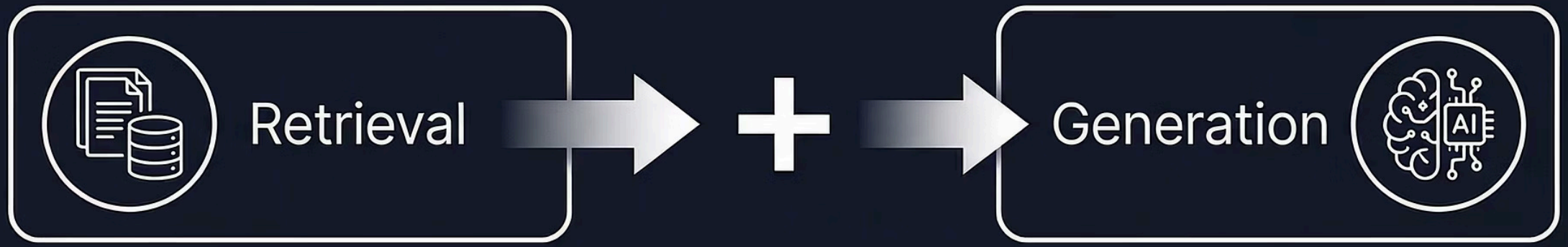
Lost Context

No Real-Time Data

Private Data Gap

RAG fixes this

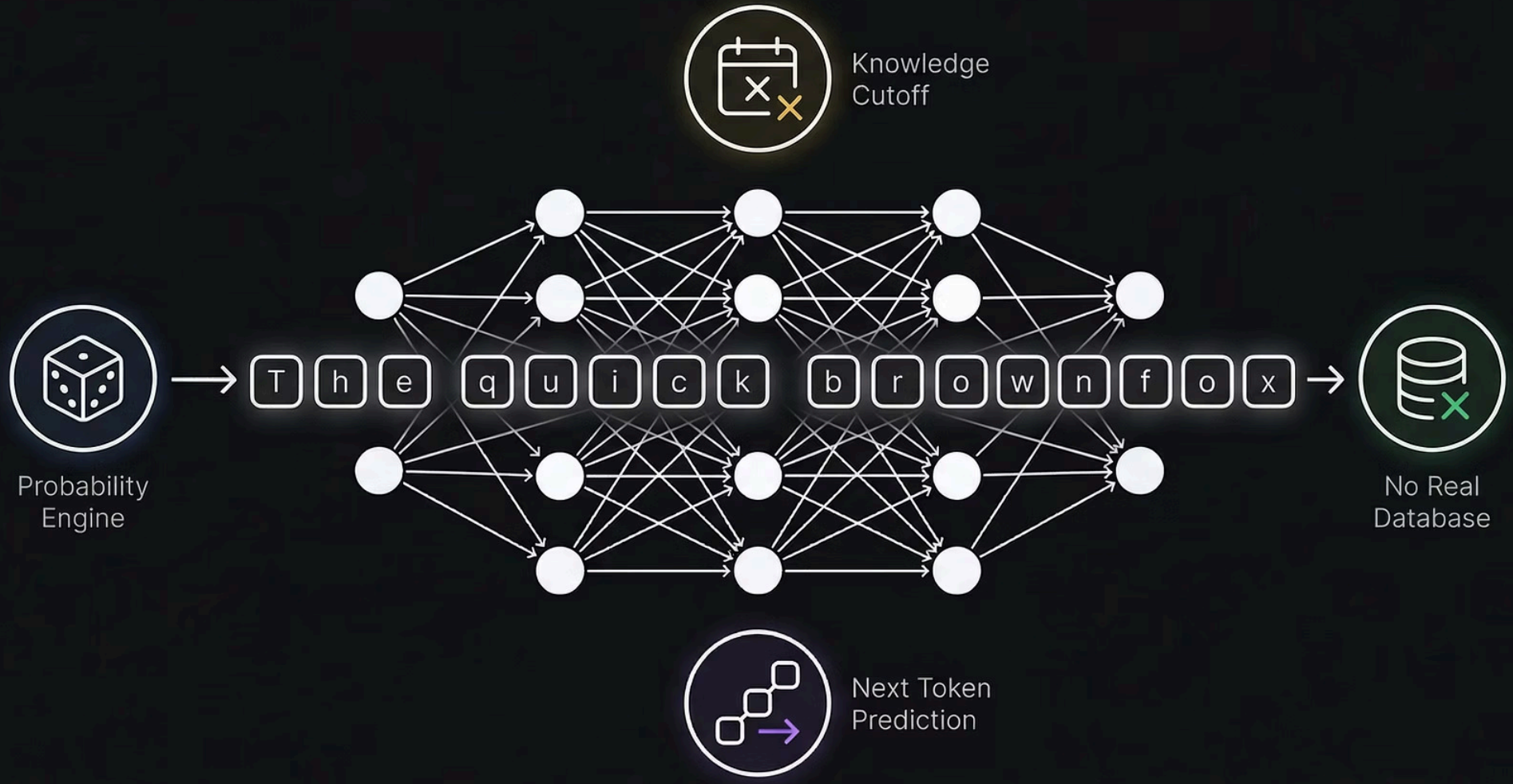
What is RAG?



RAG = Retrieval + Generation

"Retrieves relevant information and feeds it to the LLM"

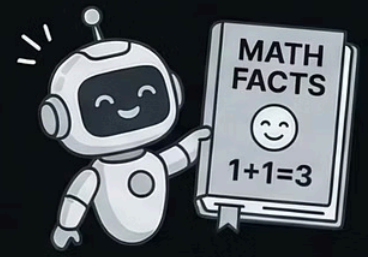
How LLMs Work



The Problem: Why LLM Answers Can Be Wrong

Disconnected from Real-Time Sources

Guesswork Transforms into Hallucinations



Confident Wrong Answer

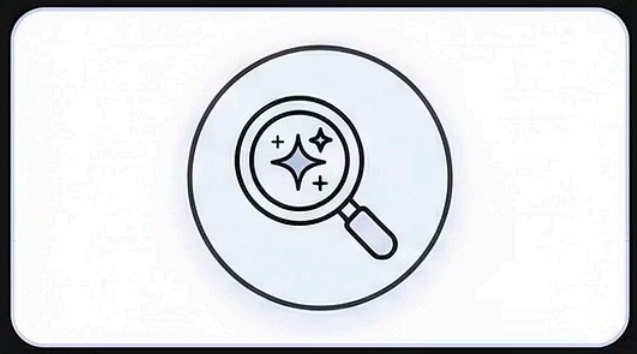
False Information

No external memory → hallucinations

RAG to the Rescue

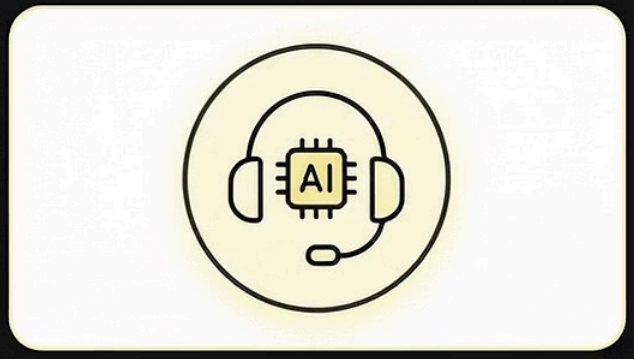
Real-World RAG Examples

ChatPDF

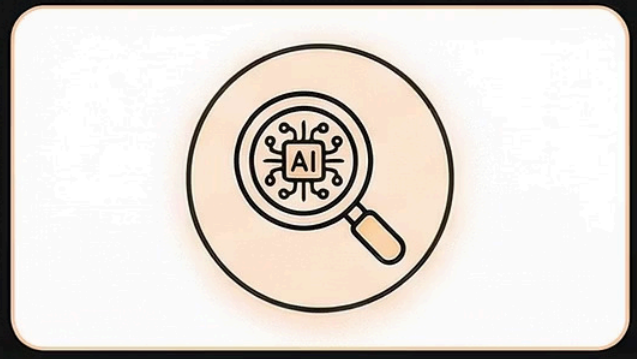


Perplexity

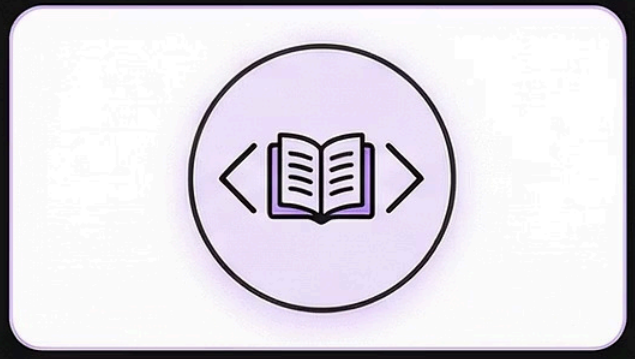
Enterprise Knowledge Bots



Customer Support AI

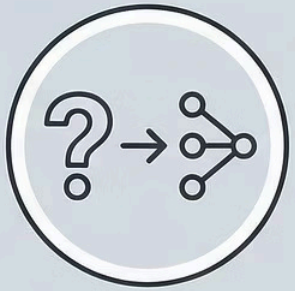


AI Search



Code Docs Bots

RAG Architecture Overview

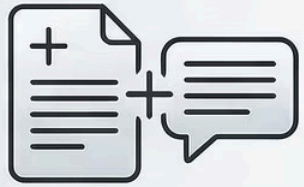


1 Query → Embedding

Text transforming to vector dots

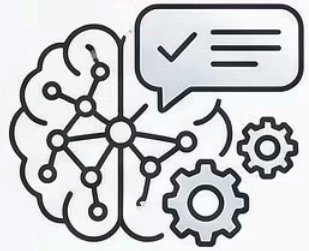
2 Retrieval

Database with search



3 Prompt Augmentation

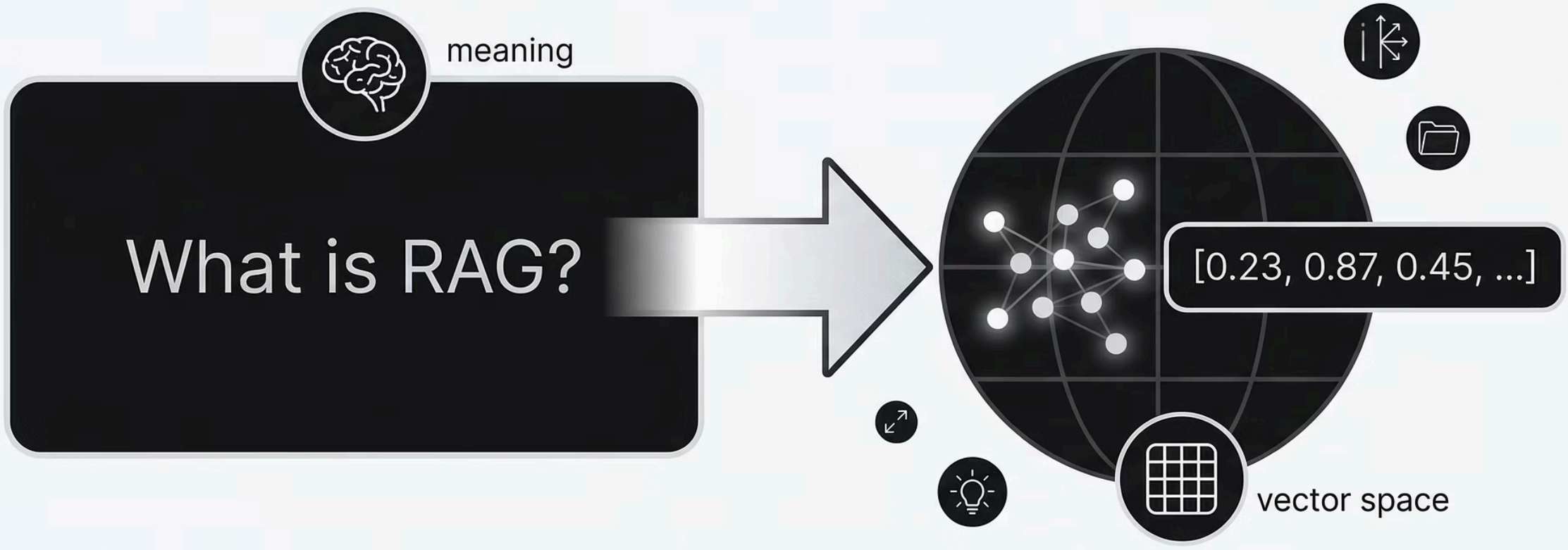
Context being added to prompt



4 Generation

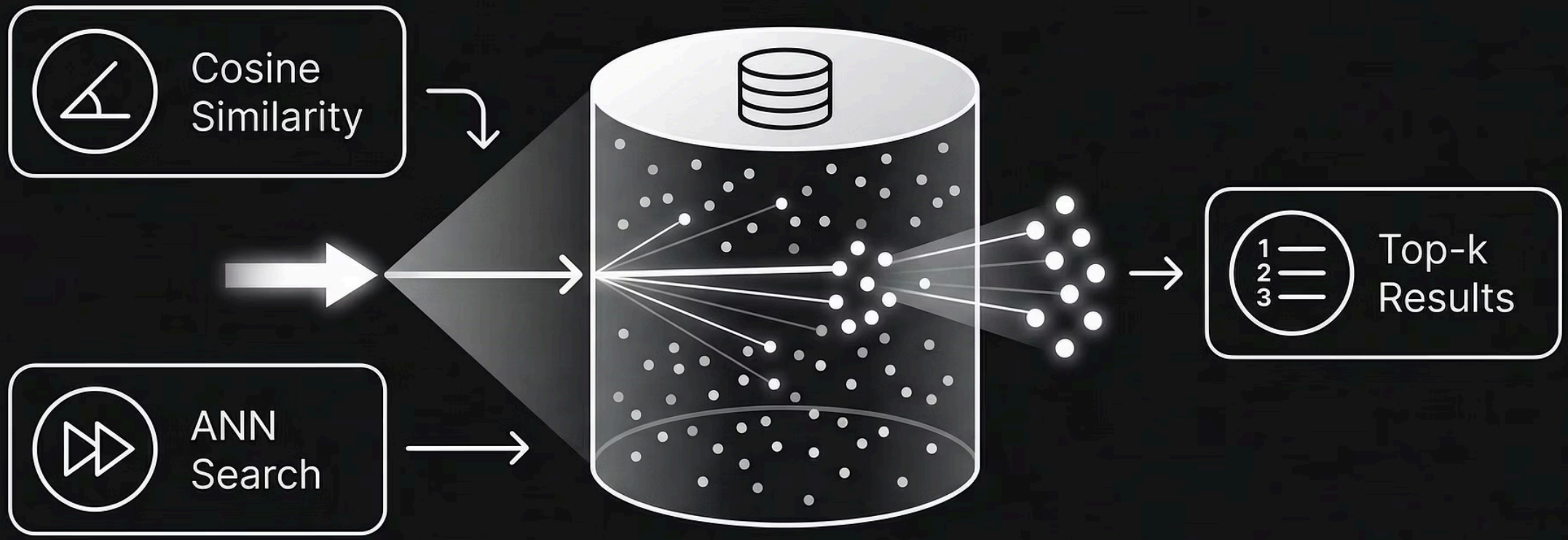
LLM producing answer

Stage 1: Query Embedding



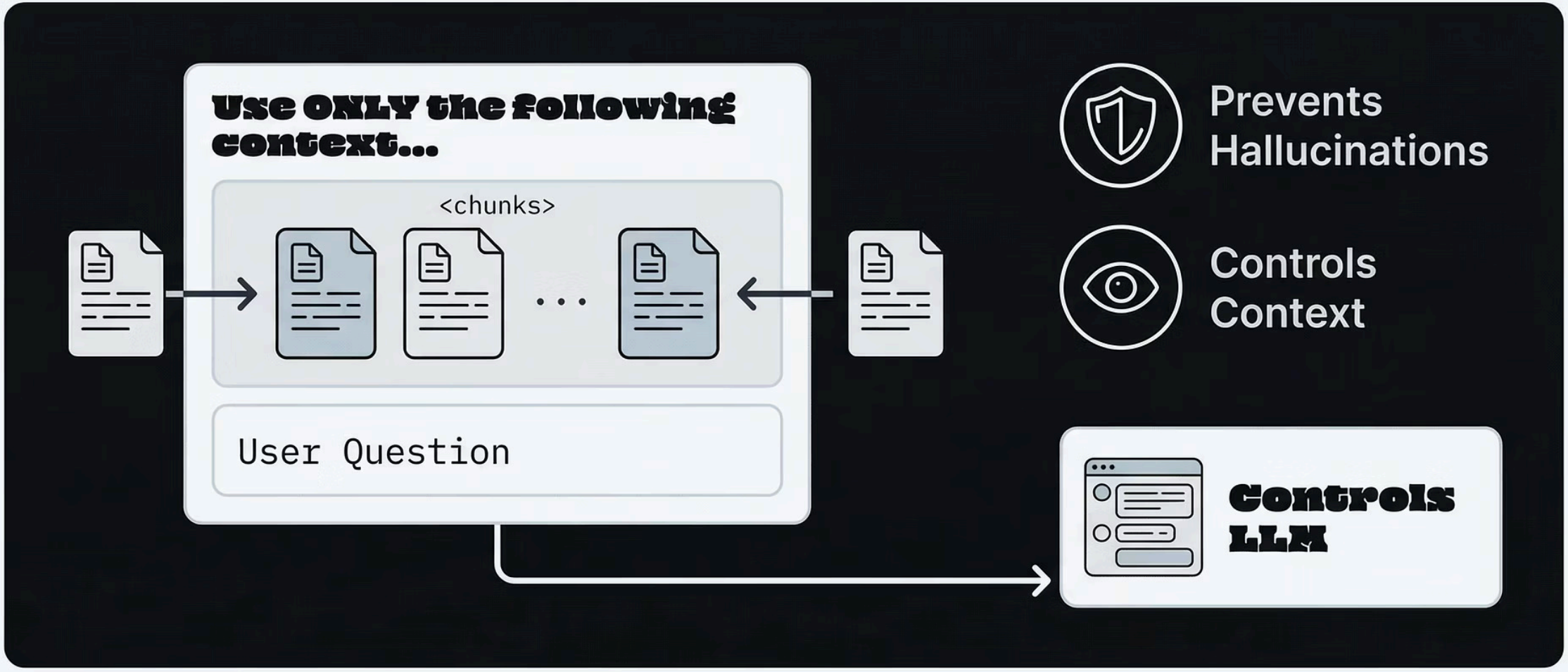
Text → Vector → Semantic Search

Stage 2: Vector Search

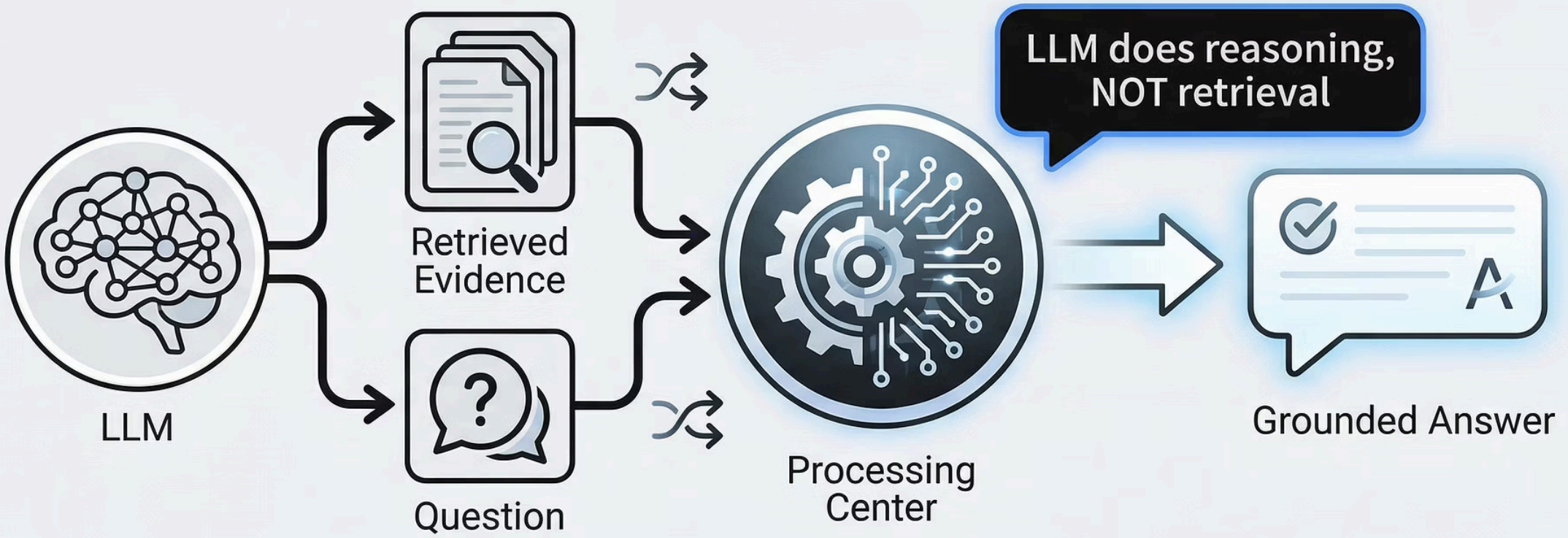


FAISS Chroma Pinecone Qdrant

Stage 3: Prompt Augmentation

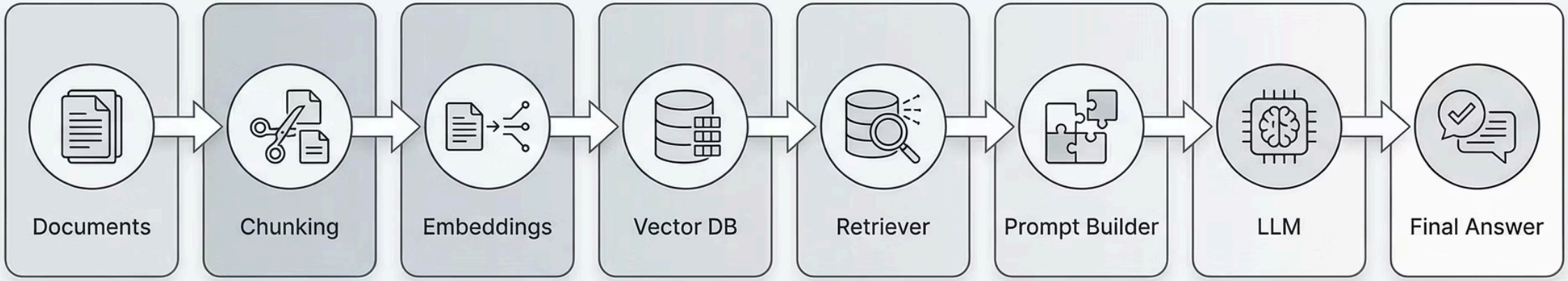


Stage 4: Generation



Retrieved Evidence Question → Grounded Answer

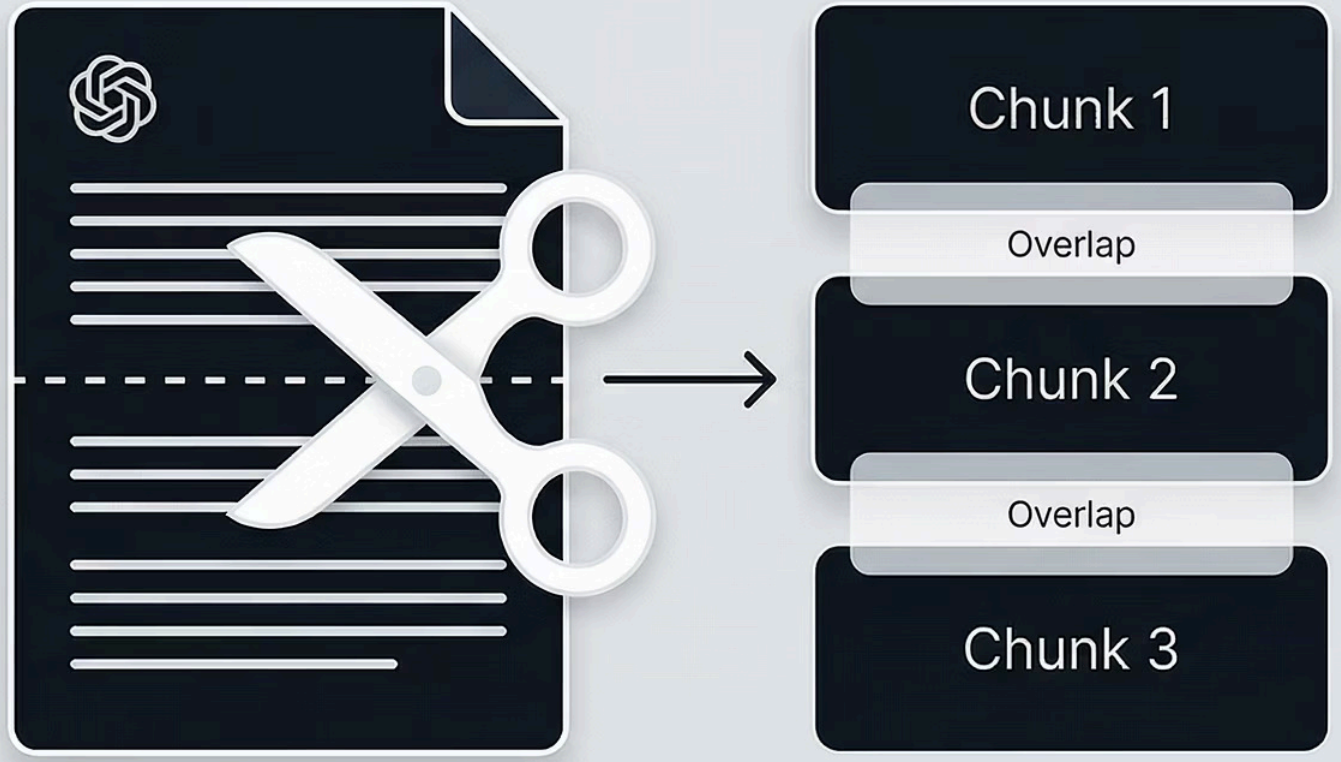
Full Architecture Diagram



Chunking (Important Detail)

200-500 tokens

Overlap for context



Recursive

Why needed: Context windows have limits

Embedding Models

all-MiniLM-L6-v2

Fast

text-embedding-3-large

Quality

bge-small / bge-large

Scale or Size Options

nomic-embed

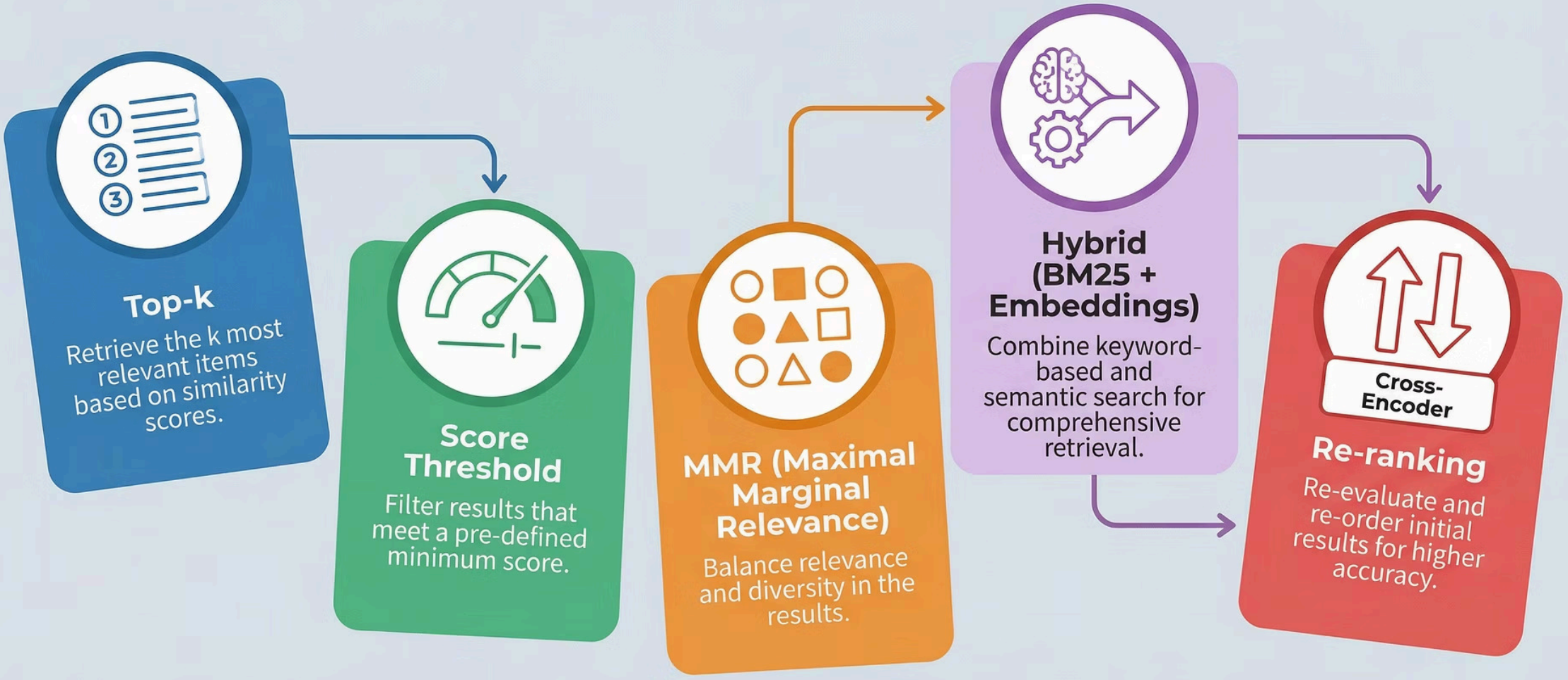
Efficient

Dimensionality

Quality Differences

Vector Databases

Retrieval Strategies

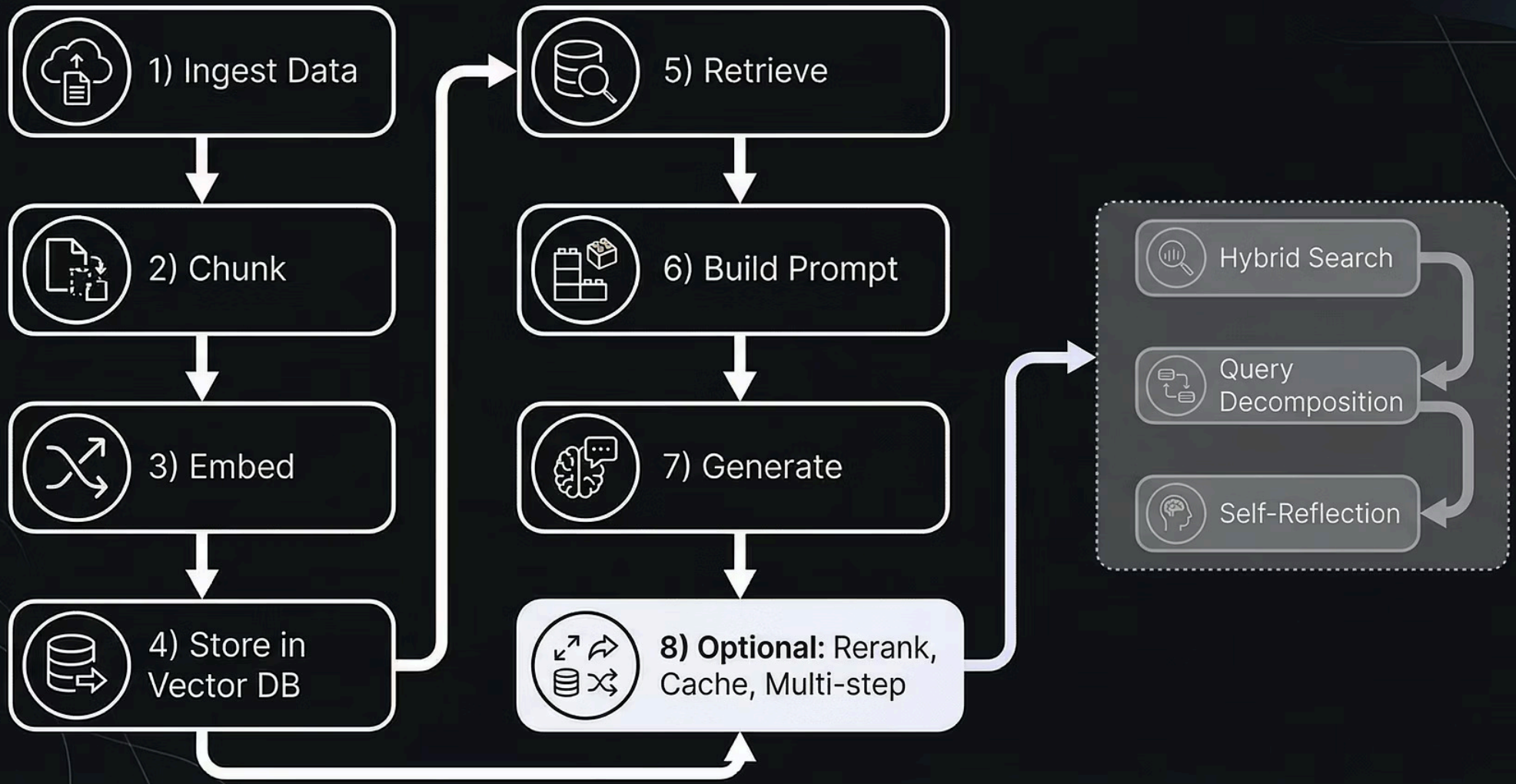


Prompt Engineering for RAG

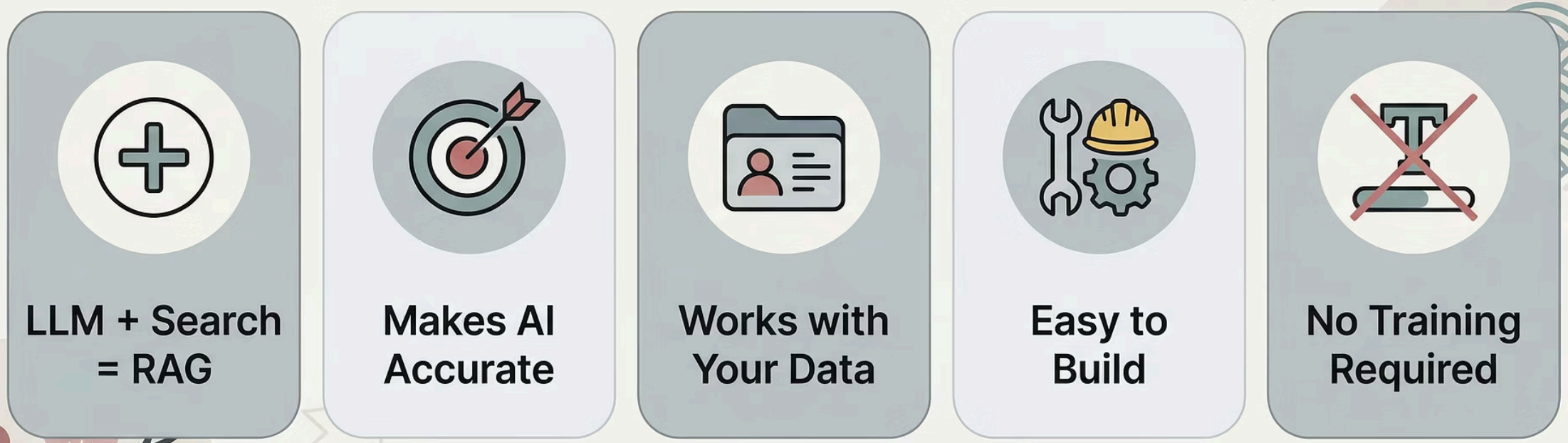
- **Clear Instructions**
- **Provide Context**
- **Use Separators**
- **Ask for Citations**

```
Instructions
Be clear and direct

Context
---
<retrieved_docs>
---
Question

Ask for citations
```

Putting It All Together



Summary + Next Steps



→ Start building