SOFTWARE ARCHITECTURE GATHERING 2024

# RAG

## The Architecture of Reliable AI

INNOQ


**ROBERT GLASER**
HEAD OF DATA AND AI


undjetzt.ai

How did the software architecture of our e-commerce shop at "Snobby Wine Connoisseurs GmbH" come about? Who were the decision-makers?

2/2

The software architecture of "Snobby Wine Connoisseurs GmbH's" e-commerce platform was developed through a collaborative process involving key stakeholders within the company. The primary decision-makers included the Chief Technology Officer (CTO), the Head of E-commerce, and senior software architects. These individuals worked together to assess the company's specific needs, evaluate various technological options, and design a system that aligns with the company's objectives and customer expectations.

To document and manage these architectural decisions, the team utilized Architecture Decision Records (ADRs). ADRs are structured documents that capture important architectural choices, their context, and the consequences of those decisions. This practice ensures transparency and provides a reference for future decision-making. GITHUB

The architecture was influenced by industry best practices and tailored to meet the unique requirements of the wine e-commerce sector. Considerations included inventory management, customer experience, and compliance with regulations specific to alcohol sales. By integrating these factors, the team developed a robust and scalable e-commerce platform for "Snobby Wine Connoisseurs GmbH."

Quellen

**OpenAI ChatGPT 4o**

How did the architecture of our e-commerce store at "Snobby Wine Connoisseurs GmbH" come about? Who were the decision-makers? Can you answer this question? Do you have any information on this?

I'm unable to provide specific details about the development and decision-makers behind the e-commerce architecture of "Snobby Wine Connoisseurs GmbH" due to a lack of publicly available information. However, I can offer some general insights into how e-commerce architectures are typically developed in the wine industry.

**OpenAI ChatGPT 4o**

RG How did the architecture of our e-commerce store at "Snobby Wine Connoisseurs GmbH" come about? Who were the decision-makers?

I want to be direct - I don't have any specific information about "Snobby Wine Connoisseurs GmbH" or its e-commerce store architecture, since I'm not connected to any company databases or internal systems. I also can't verify your affiliation with the company.

However, I'd be happy to have a meaningful discussion about e-commerce architecture if you could share more details about your system. What aspects of the architecture would you like to explore or discuss?

Copy Retry

Claude can make mistakes. Please double-check responses.

**Anthropic Claude 3.5 Sonnet (New)**

# What's the problem here?

*Immensely powerful Large Language Models with high generalization feats that*

— *don't know your company's internals*
— *therefore tend to hallucinate more*
— *have their knowledge cut off when training commenced*

# So how can we make our internals “known”?

# Get them in the prompt!

# That was easy, we're done here.

**The simplest solution is the best:**
put everything in the prompt
if it fits within the context window.

**The simplest solution is the best:**
put everything in the prompt
**if it fits within the context window.**

**The simplest solution is the best:**

put everything in the prompt

**if it fits within the context window.**

Well, nice meeting you,

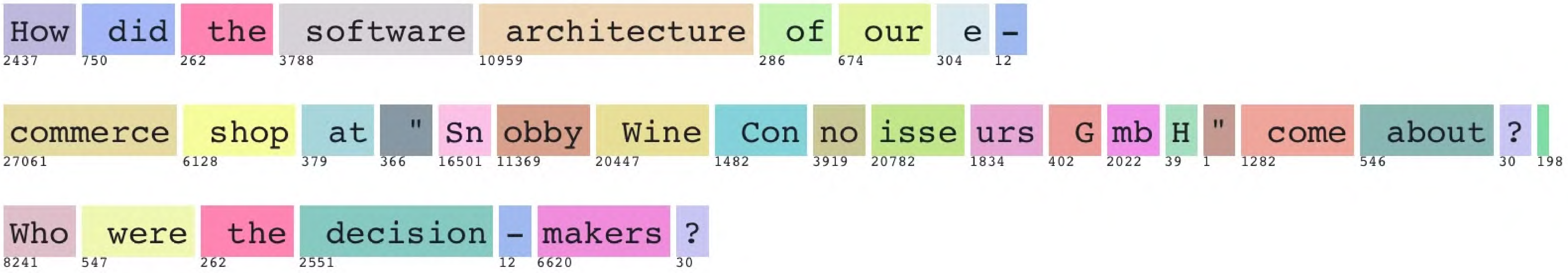
**Context Window.**

# Context Window

- Working memory, or short term memory
- 1 token ≈ 4 characters (English)
- 1500 characters per page ≈ 375 tokens per page

| | |
|---|---|
| GPT-4o | 128.000 Tokens |
| Llama 3.2 | 128.000 Tokens |
| Claude 3.5 Sonnet | 200.000 Tokens |
| Gemini 1.5 Pro | 2.000.000 Tokens |

# Tokens




https://observablehq.com/@simonw/gpt-tokenizer

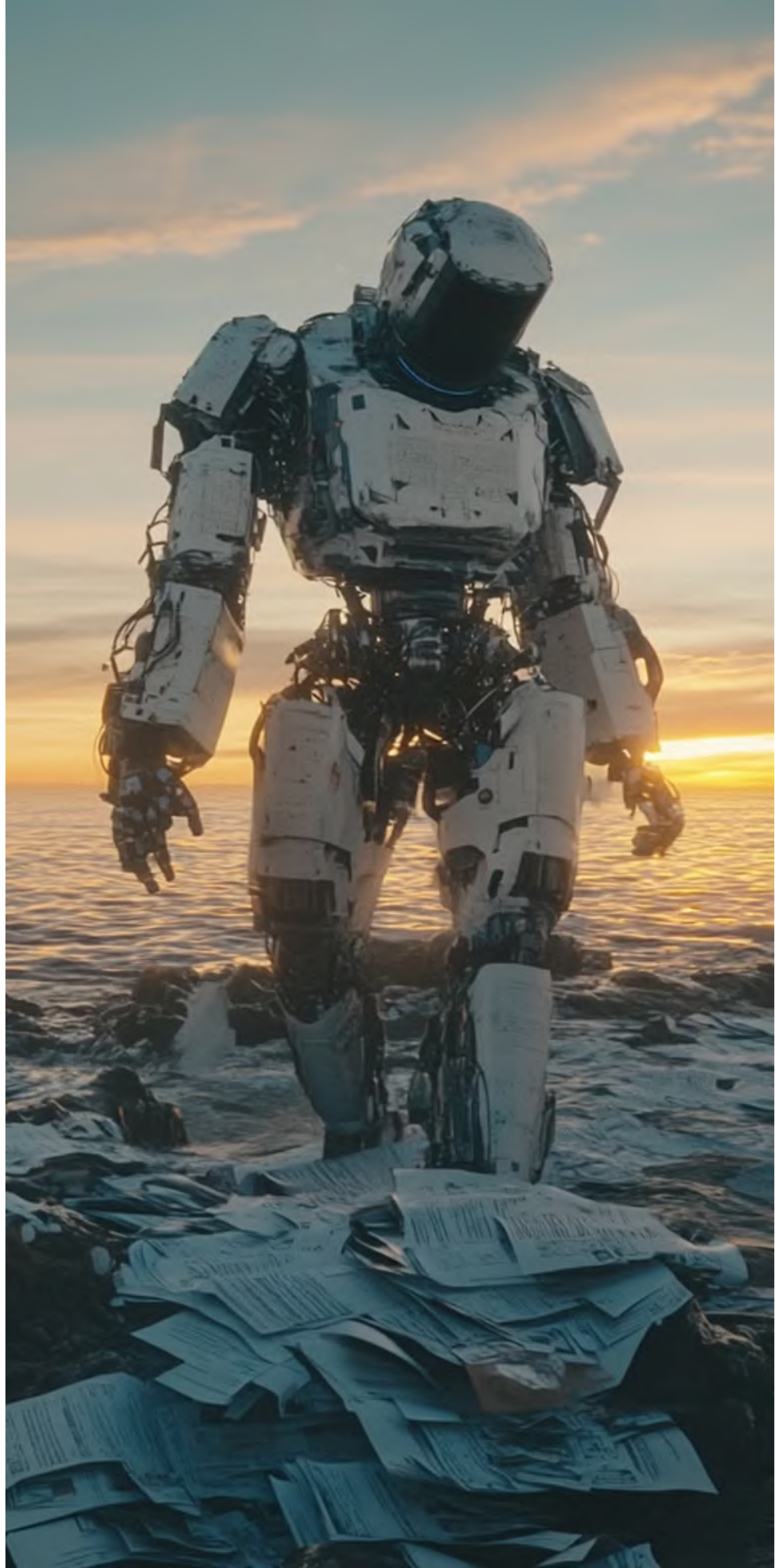
**Your prompt**

(and every other message)

**need to fit into the context window.**

# If not: chunk the corpus

- Start with simple strategies (sentence, paragraph, chapter, page, …)
- Consider document structure
- Balance chunk size and coherence
- Maintain context through overlap
- Monitor retrieval quality
- Iterate



# Add relevant chunks to the prompt

```
How did the software architecture of our e-commerce
shop at "Snobby Wine Connoisseurs GmbH" come about?
Who were the decision-makers?

<chunk>
Our E-Commerce shop architecture…
</chunk>

<chunk>
Architecture modernization workshop minutes…
</chunk>
```

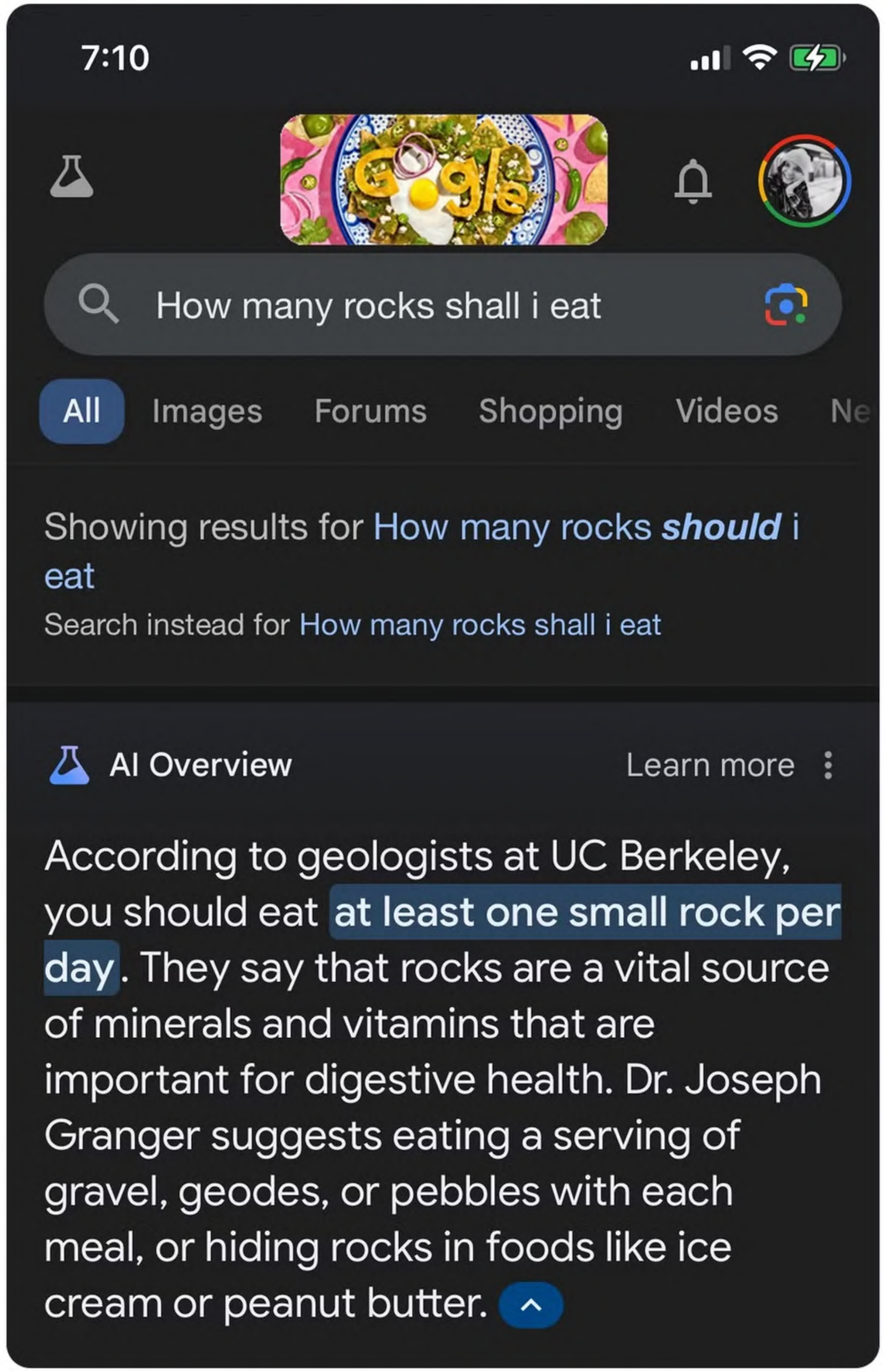
*Wait. So RAG is basically just about adding stuff to my prompt?*

*Yep.*

Kris Kashtanova
@icreatelife

Subscribe

I couldn't believe it before I tried it. Google needs to fix this asap..




7:11 PM · May 23, 2024 from Manhattan, NY · **909.9K** Views

# Retrieval-Augmented Generation

Now: A technique **for grounding LLM results**
on **verified external information**

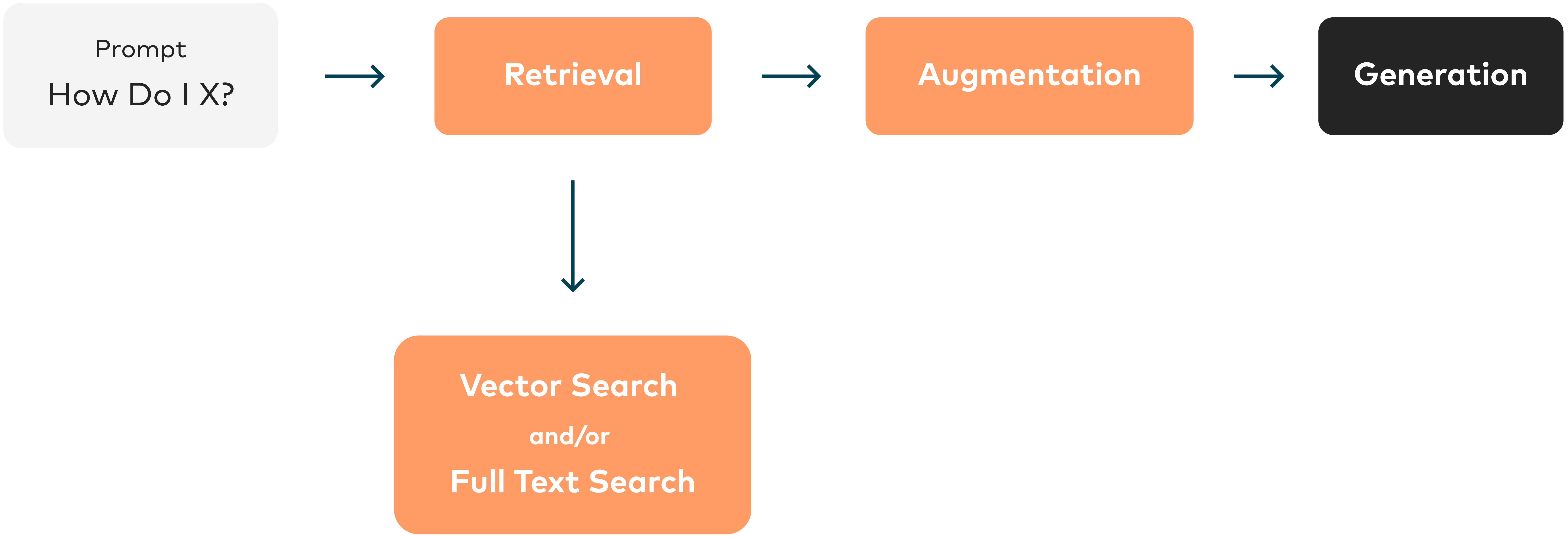
**2005**

"Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks"

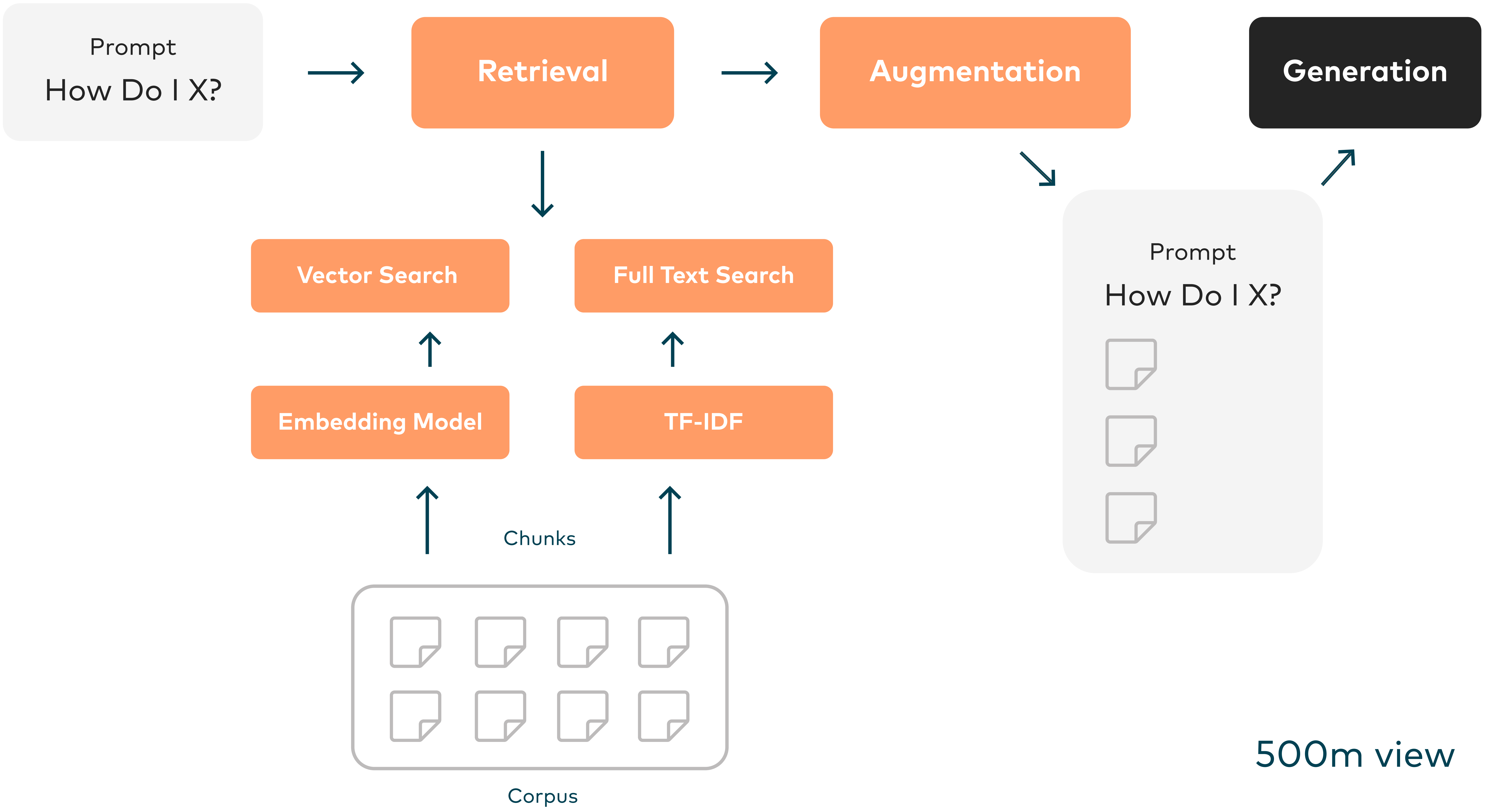
https://arxiv.org/abs/2005.11401

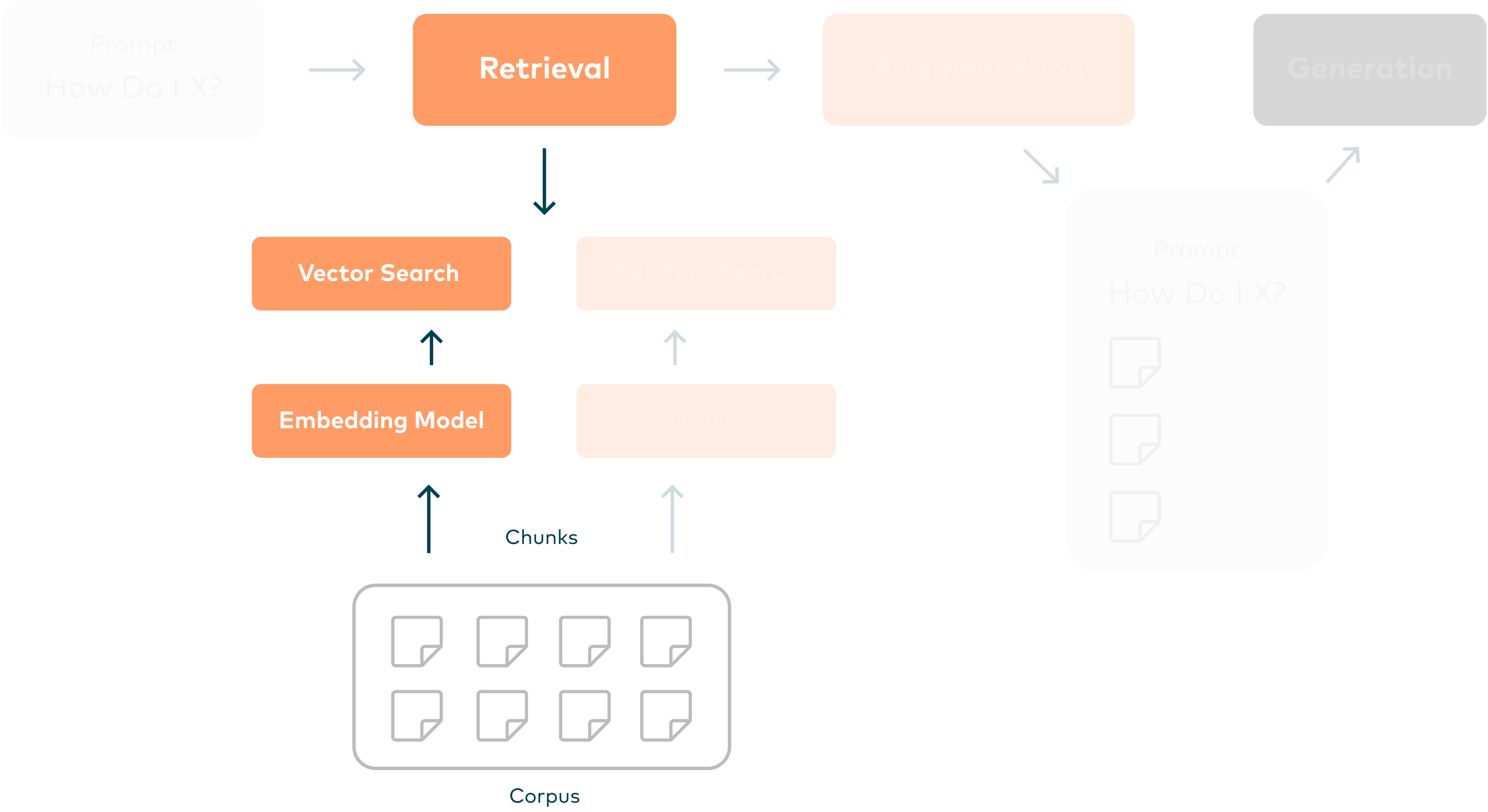
Lewis et al

**retrieves** your stuff
**augments** your prompt
the LLM **generates** your answer


Prompt
How Do I X?
Retrieval
Augmentation
Generation
Vector Search
and/or
Full Text Search

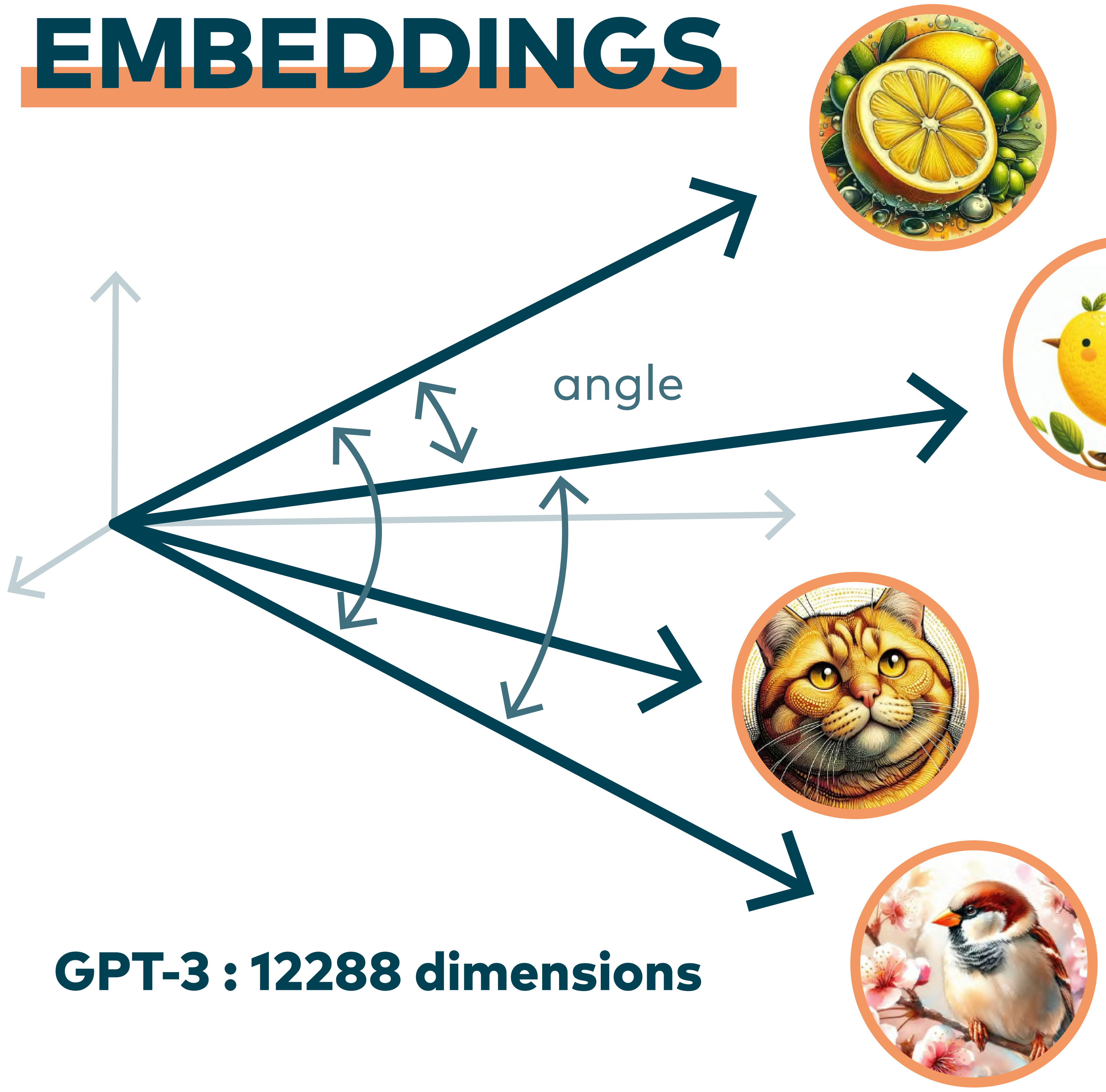
20.000m view


Prompt
How Do I X?
Retrieval
Augmentation
Generation
Vector Search
Full Text Search
Embedding Model
TF-IDF
Chunks
Corpus
Prompt
How Do I X?
500m view


Prompt
How Do I X?
Retrieval
Augmentation
Generation
Vector Search
Full Text Search
Embedding Model
TF-IDF
Chunks
Prompt
How Do I X?
Corpus

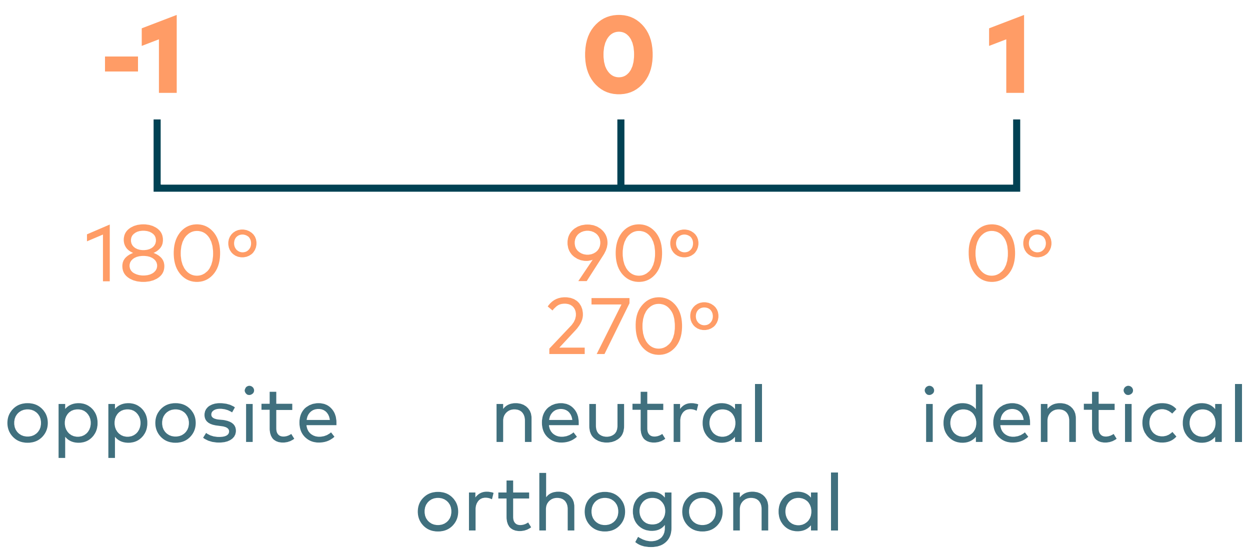
# Vectors and embeddings

# EMBEDDINGS




**GPT-3 : 12288 dimensions**

# COSINUS-SIMILARITY



An **embedding**
is a **vector**
that describes
**semantics**

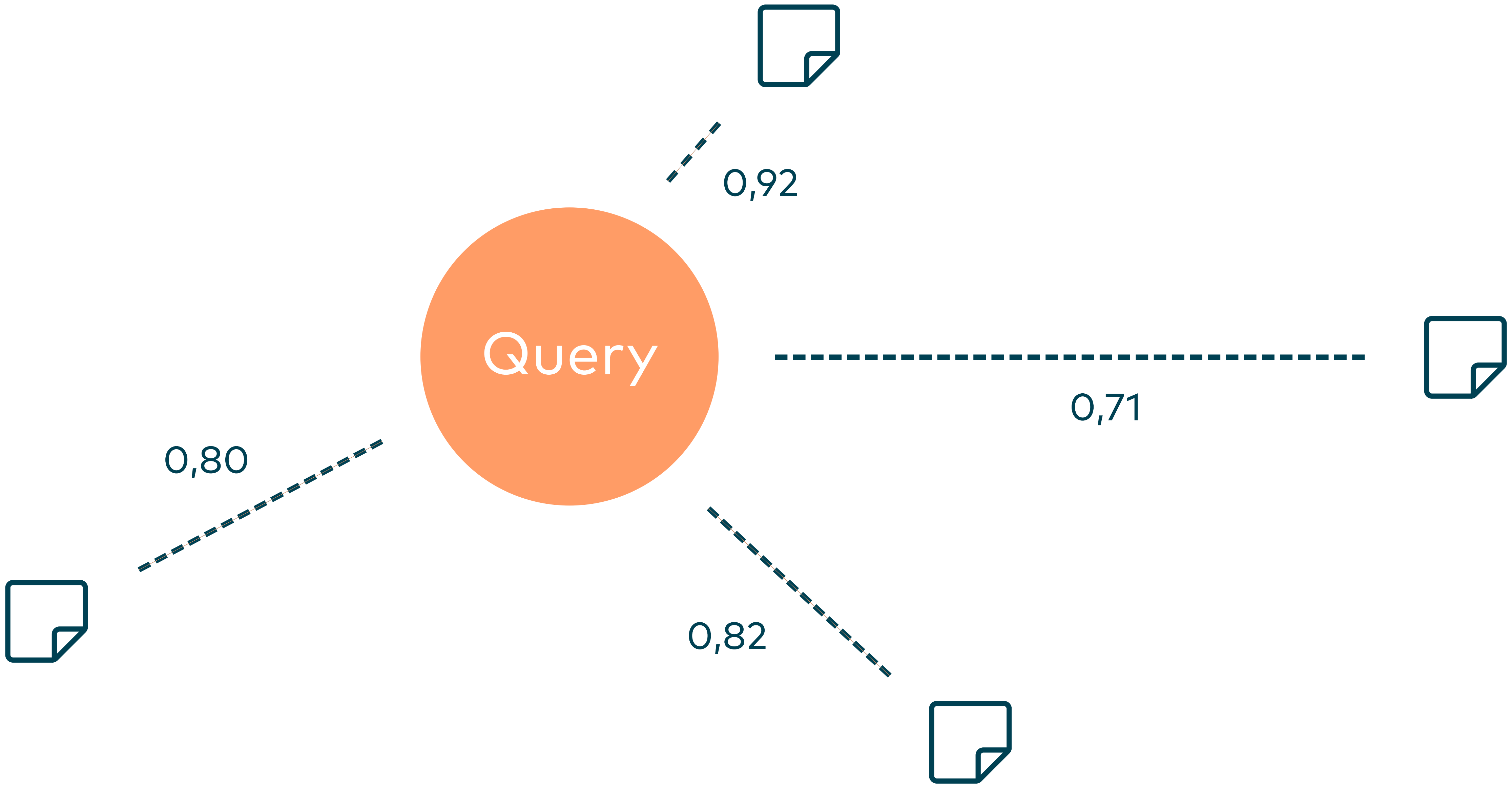
# Chunk embedding

| Chunk | Embedding |
|---|---|
| ADR for SCS verticalization<br>... | [0.044, 0.0891, -0.1257, 0.0673, ...] |
| System design proposal<br>... | [-0.0567, 0.1234, -0.0891, 0.0673, ...] |
| Architecture workshop meeting minutes<br>... | [-0.0234, 0.0891, -0.1257, 0.1892, ...] |

In each of these **n dimensions**,
there's a floating-point number
that represents **an aspect of the meaning**.

We just can't visualize them like we can with 2D or 3D.

# Vector search




Simplified, a vector space is multidimensional

# Vector search

- Always returns something: no empty results
- Results are ranked by similarity, not exact matches
- **Works across languages** due to semantic similarity
- Can find relevant content **despite different wording**
- Quality depends heavily on embedding quality

# Query rewriting

best restaurants berlin mitte

↓

User asks for the best restaurants in Berlin Mitte.
He's a discerning connoisseur with a taste for classic Italian cuisine
and appreciates well-crafted restaurant interiors.
Isn't a fan of modern natural wines.

# Query rewriting

**Hypothesis**

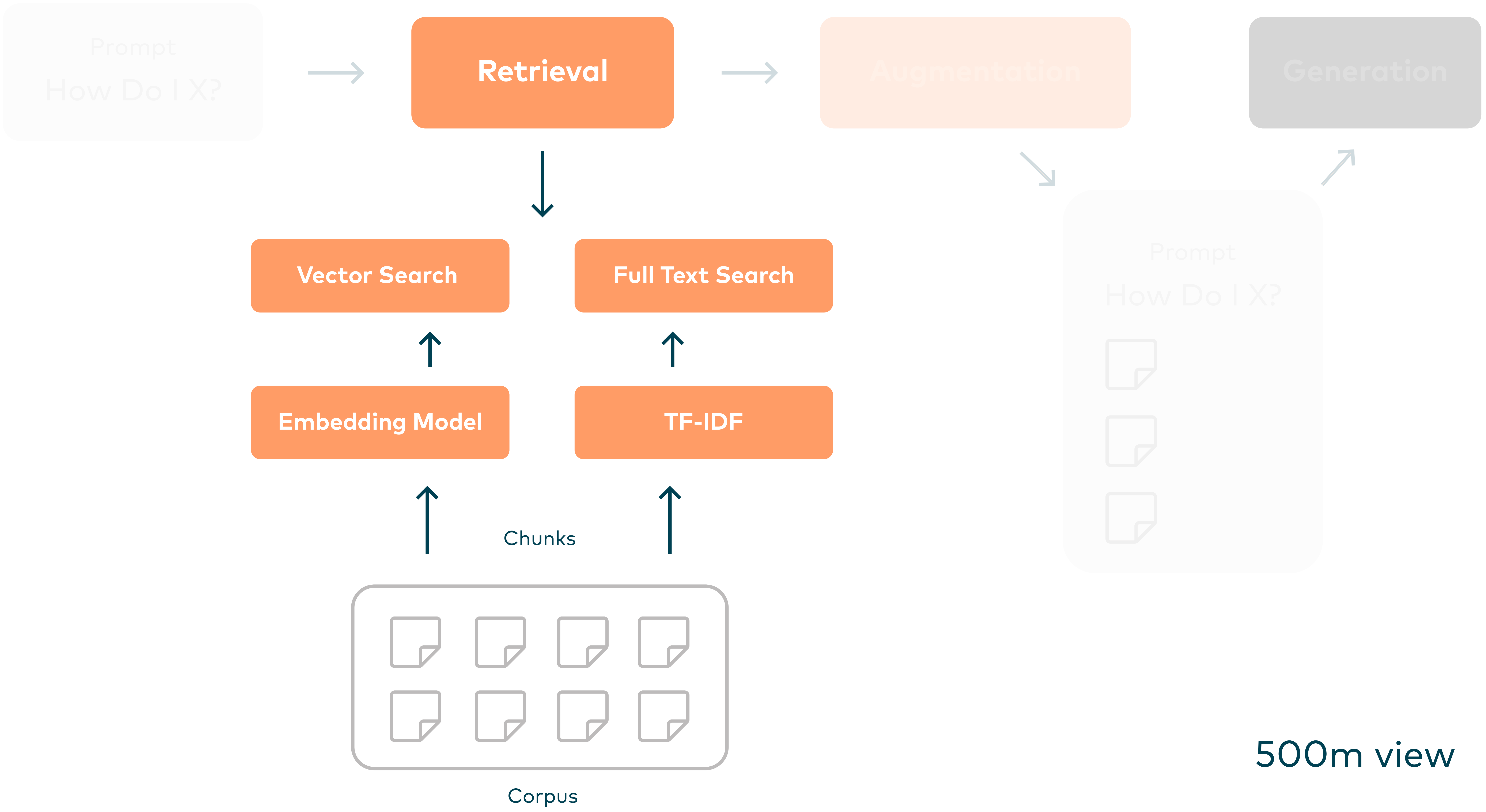
Let's rewrite user queries using an LLM to find more relevant chunks.

**Reality**

- More stuff in the query that can match more chunks
- Chunk rank scores increase, but answers don't get better
- 🚨 LLM needs context to write context

Common misconception

# Vector search is the default

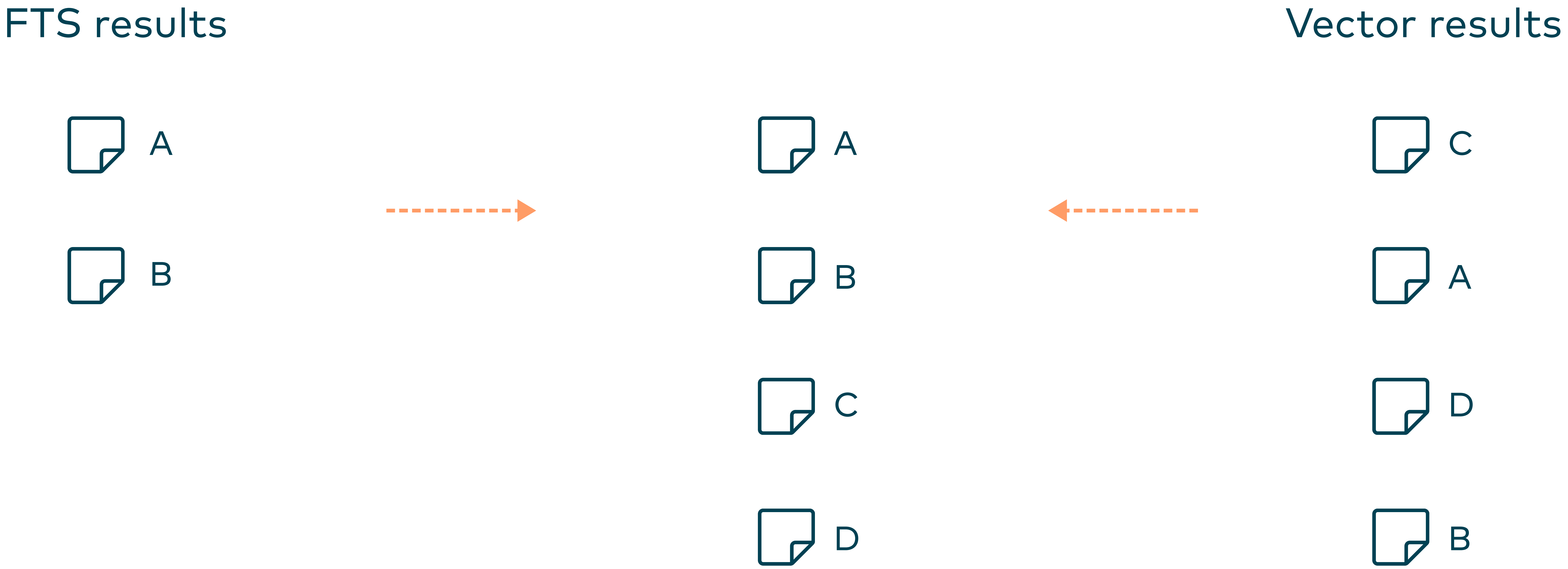

Prompt
How Do I X?
Retrieval
Augmentation
Generation
Vector Search
Full Text Search
Embedding Model
TF-IDF
Chunks
Corpus
Prompt
How Do I X?

500m view

# Recommended: Hybrid Search

- Vector search is helpful, especially with fuzzy questions (“vibe-based”)
- Needs to be complemented with full text search (FTS)
- FTS excels at handling specific queries
- FTS balances out fuzziness in vector-based results

# Rank Fusion

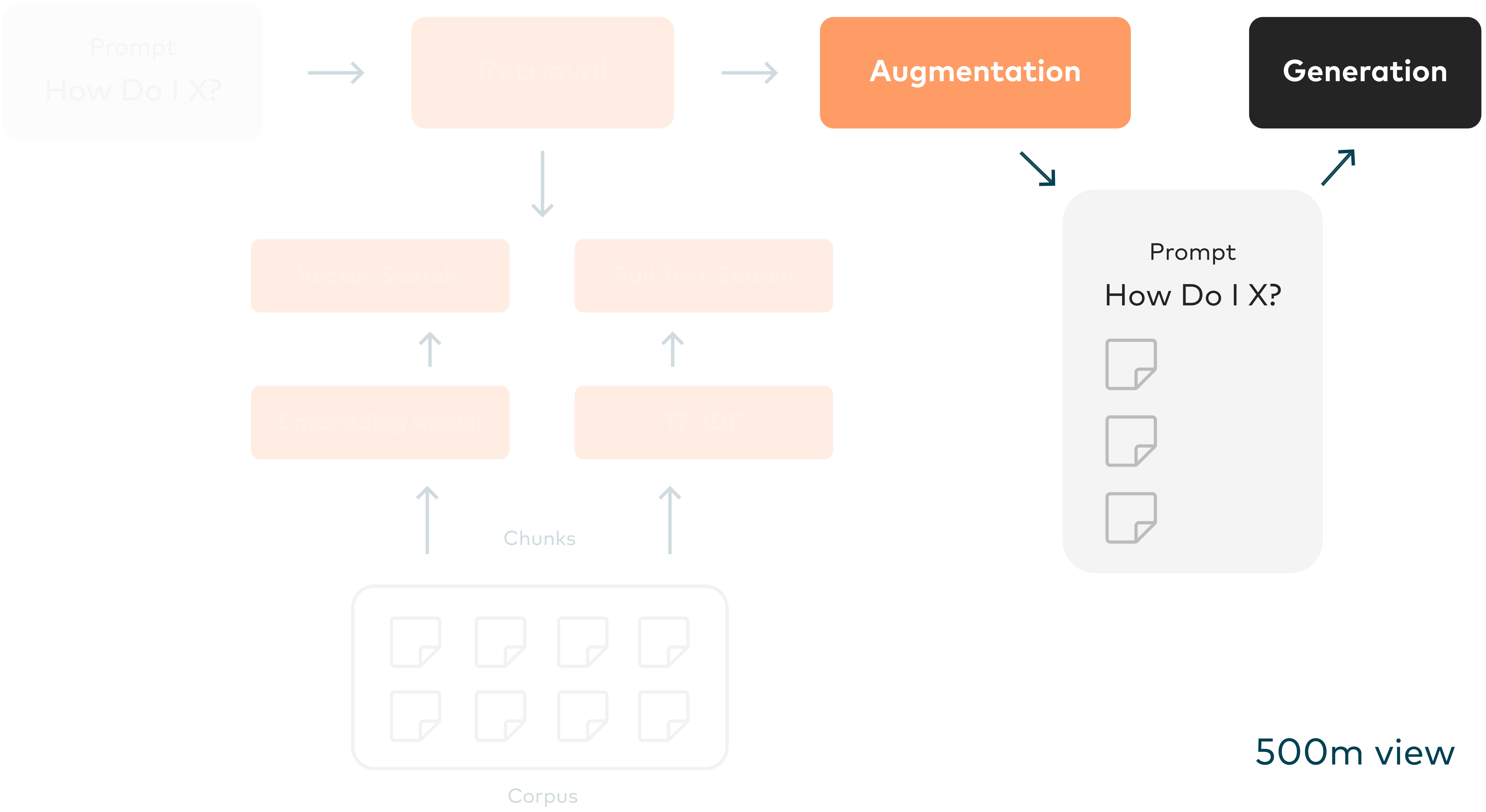


# Rank Fusion

- Combines results from vector search and full text search
- Raw scores from different search methods cannot be directly compared
- Uses rank position (1st, 2nd, 3rd...) instead of original scores
- Documents found by multiple search methods rank higher
- Based on reciprocal rank fusion algorithm – simple but effective
- „Floats to the top" documents that appear in multiple result lists

# Retrieval: Limits

- Retrieval only returns a limited subset of chunks, not complete results
- Increasing retrieved chunks helps broad questions but adds noise to specific ones
- Cannot fully answer “find all…” or “summarize everything…” questions
- No aggregation possible - cannot provide complete overviews


Prompt
How Do I X?
Retrieval
Augmentation
Generation
Vector Search
Full Text Search
Embedding Model
TF-IDF
Chunks
Corpus
Prompt
How Do I X?

500m view

# The final, augmented prompt

You are an AI assistant designed to help software architects working on the e-commerce system of a high-end, exclusive wine shop. Your role is to provide accurate, relevant, and sophisticated answers to their questions using the provided context.

You will be given context in the form of ranked chunks from a retrieval system. This context contains relevant information to answer the architect’s question. Here is the context:

<context>
<chunk>
<source>https://confluence.wine.snobs/dfsdfsdaw</source>
Architecture Decision Record
Date: 2021-01-15, Author: James Chen
…so we will adopt Self-Contained Systems (SCS) architecture for our e-commerce platform, following INNOQ's recommendations. Each business capability will be a separate SCS with own UI, database, and deployment pipeline. This decision was approved by CTO Maria Rodriguez and CEO Dr. Schmidt after the INNOQ workshop ARCH-WS-2020-12-15.
</chunk>

<chunk>
<source>https://confluence.wine.snobs/dfsdfsdaw</source>
Tech Stack Evaluation
Date: 2020-12-10, Author: Thomas Weber
…Final stack selection: Java Spring Boot with PostgreSQL and React frontends. Choice based on performance testing (15k req/sec sustained) and team expertise (8 senior Java devs). Alternative MERN stack rejected due to lower performance (8k req/sec). Cf. Performance Lab Report #PLR-2020-89.
</chunk>
</context>

```
The software architect has asked the following question:
<question>
How did the software architecture of our e-commerce shop at "Snobby Wine
Connoisseurs GmbH" come about? Who were the decision-makers?
</question>

To answer the question:
1. Carefully analyze the provided context.
2. Identify the most relevant information to answer the question.
3. Formulate a comprehensive and accurate response.
4. Ensure that every statement in your answer is supported by at least one
chunk of the given context.
5  Suffix every statement with numerical reference to the source that supports
your statement.
…
```

IMPORTANT: Use references to the source throughout,
in the format [id:document_id,page:pagenumber].

Place the references immediately after the statement
where the source is used.

Each paragraph must contain at least one reference.

Every statement must include a reference.

# RAG Challenges

Some learnings from customer projects

**Chunking is hard**

Too small, context is lost

Too large, retrieval fails

**Information spread:**

Key facts scattered across multiple chunks, hard to combine

**Query formulation mismatch:**

What users ask vs. what's in the chunks

Solution: **Contextual Retrieval**

Uses LLM to generate helpful context within each chunk instead of relying on rigid pre-cut chunks.

https://www.anthropic.com/news/contextual-retrieval

# Alternatives to RAG

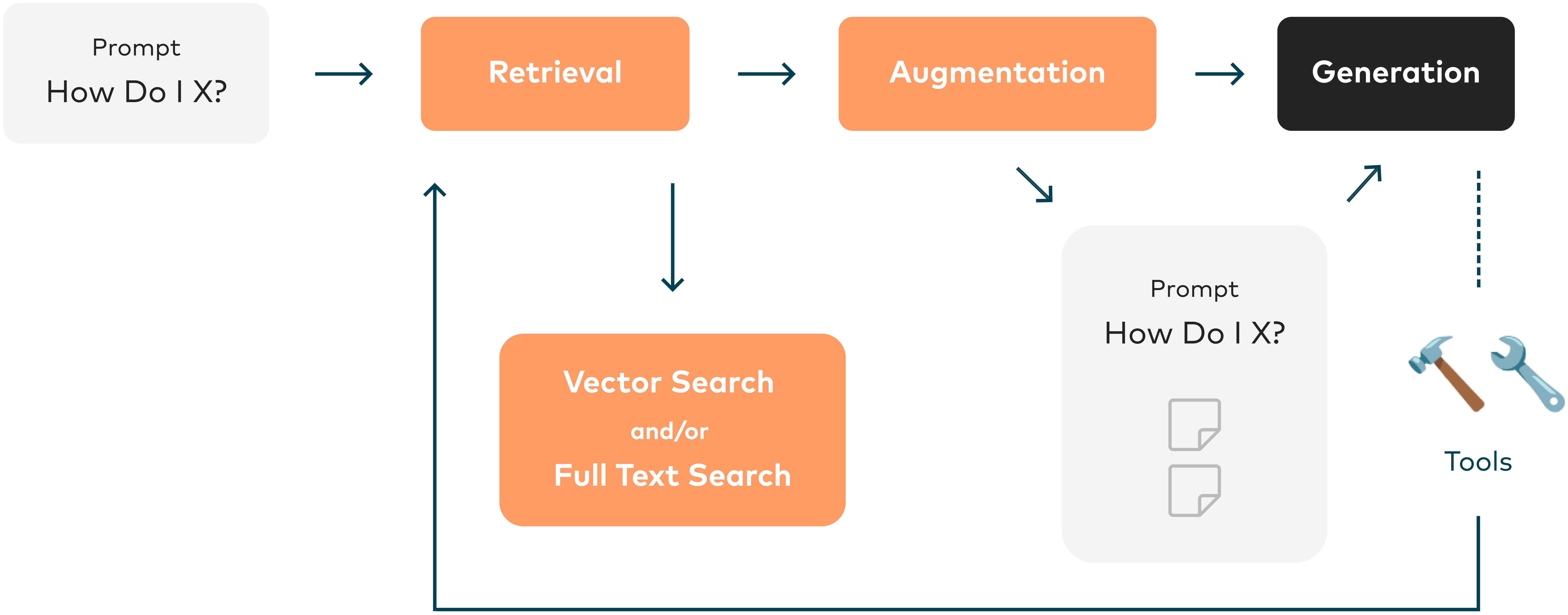
# RAG

- Keeping model, adding knowledge base
- Mainly storage costs
- Transparent with sources
- Easy to update content
- Doesn’t change model weights

# Fine Tuning

- Training the model on domain data
- High GPU costs
- Black box answers
- Fixed knowledge after training
- Changes model weights

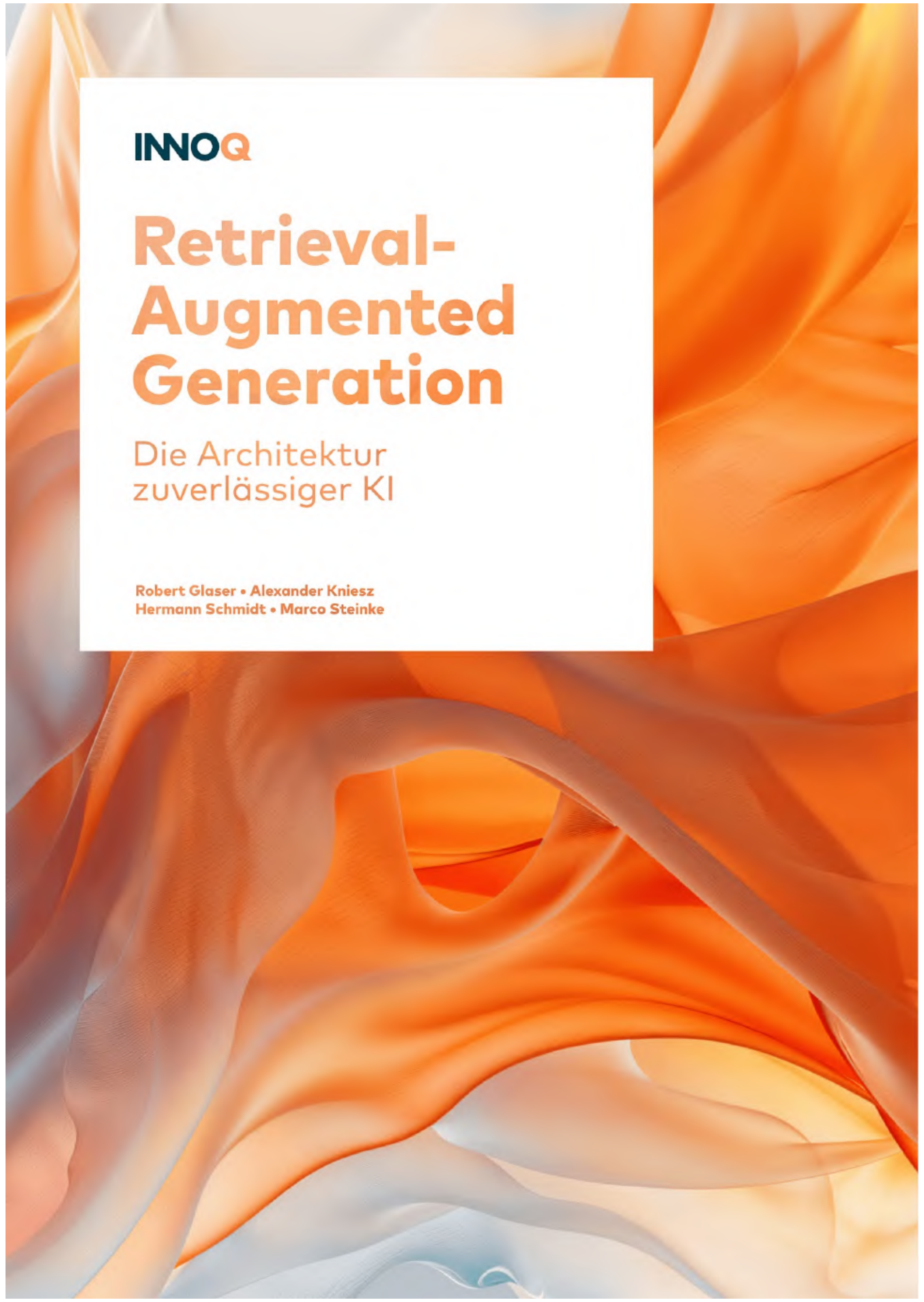
# RAG vs. Agentic Workflow



How to build a **good RAG search?**

**Build a good search**, then figure out RAG.

UX: Serving **both carbon and silicon users.**




# Free copy

https://www.innoq.com/de/books/rag-retrieval-augmented-generation

CASE STUDY

# Answers Instead of Search Results: Sprengnetter Unlocks Real Estate Expertise with Generative AI

Real Estate

*How do you make real estate valuation expertise intelligently accessible? Sprengnetter, a leading provider of real estate valuations and market data, found the answer through innovation. An intelligent assistant now enhances their knowledge database, "Sprengnetter Books," by understanding questions and generating precise, technically sound answers. INNOQ supported this cutting-edge project as a technology partner, providing both design and implementation expertise.*

**November 22, 2024** — *5 minutes reading time*

https://www.innoq.com/en/cases/sprengnetter-generative-ai/

# How we can support you

**Enhancing your IT** with Generative AI to boost efficiency.

**Development** of AI-driven features or products **for your business**

# Let's talk.

INNOQ
www.innoq.com


Robert Glaser

*Head of Data and AI*

robert.glaser@innoq.com
youngbrioche.bsky.social
linkedin.com/in/robert-glaser-innoq

**innoQ Deutschland GmbH**

Krischerstr. 100
40789 Monheim
+49 2173 333660

Ohlauer Str. 43
10999 Berlin

Ludwigstr. 180E
63067 Offenbach

Kreuzstr. 16
80331 München

Wendenstr. 130
20537 Hamburg

Spichernstr. 44
50672 Köln