MASTERING LLM PRESENTS

COFFEE BREAK CONCEPTS

Improving RAG with Contextual Retrieval

Follow us on

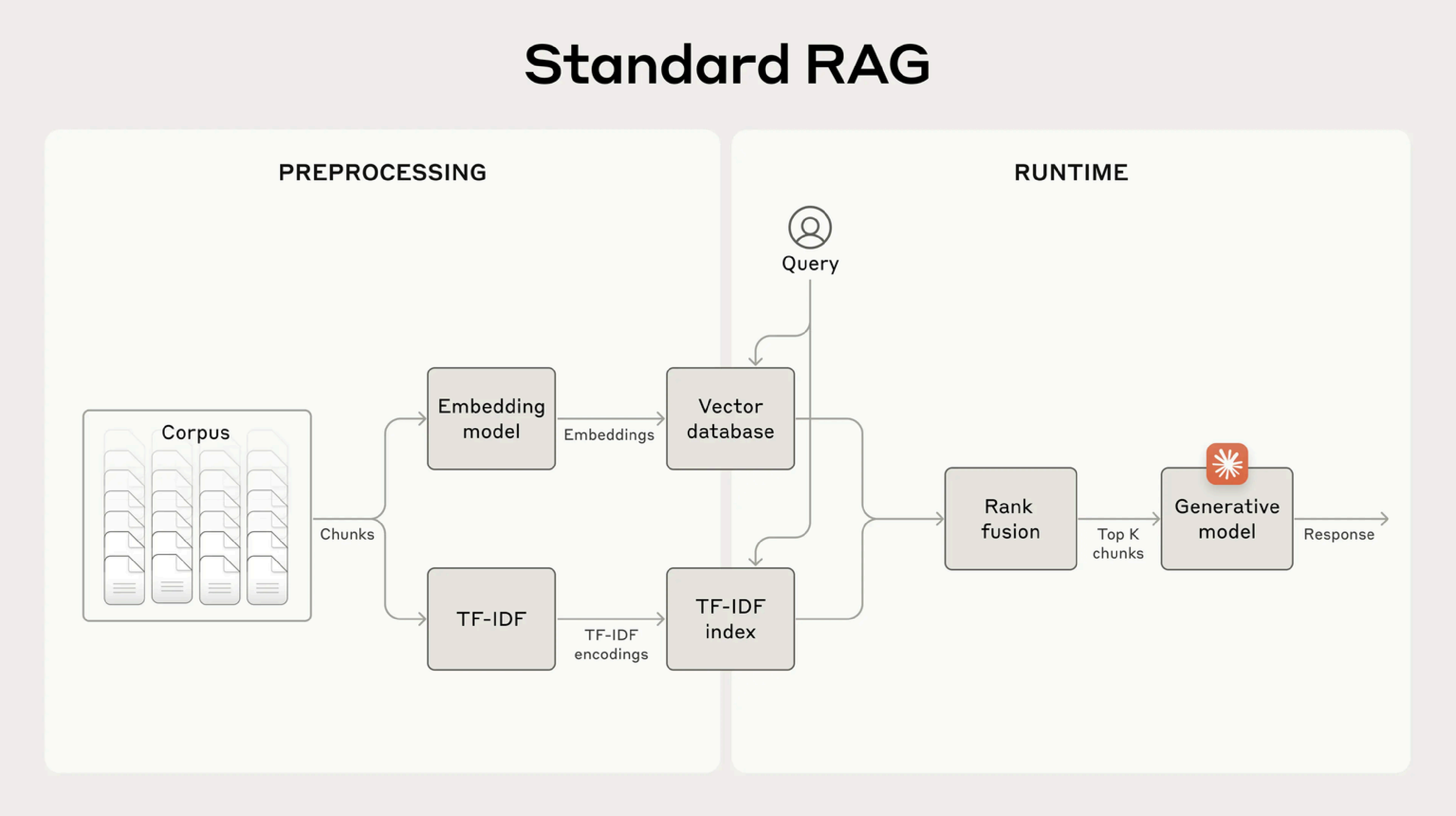
Challenges with RAG

1 Loss of contextual integrity due to chunking.

A financial document might contain the statement, "**The company's revenue grew by 3% over the previous quarter**," but without knowing which **company or quarter**, the context is lost.

2 Precision issues, especially with exact matches

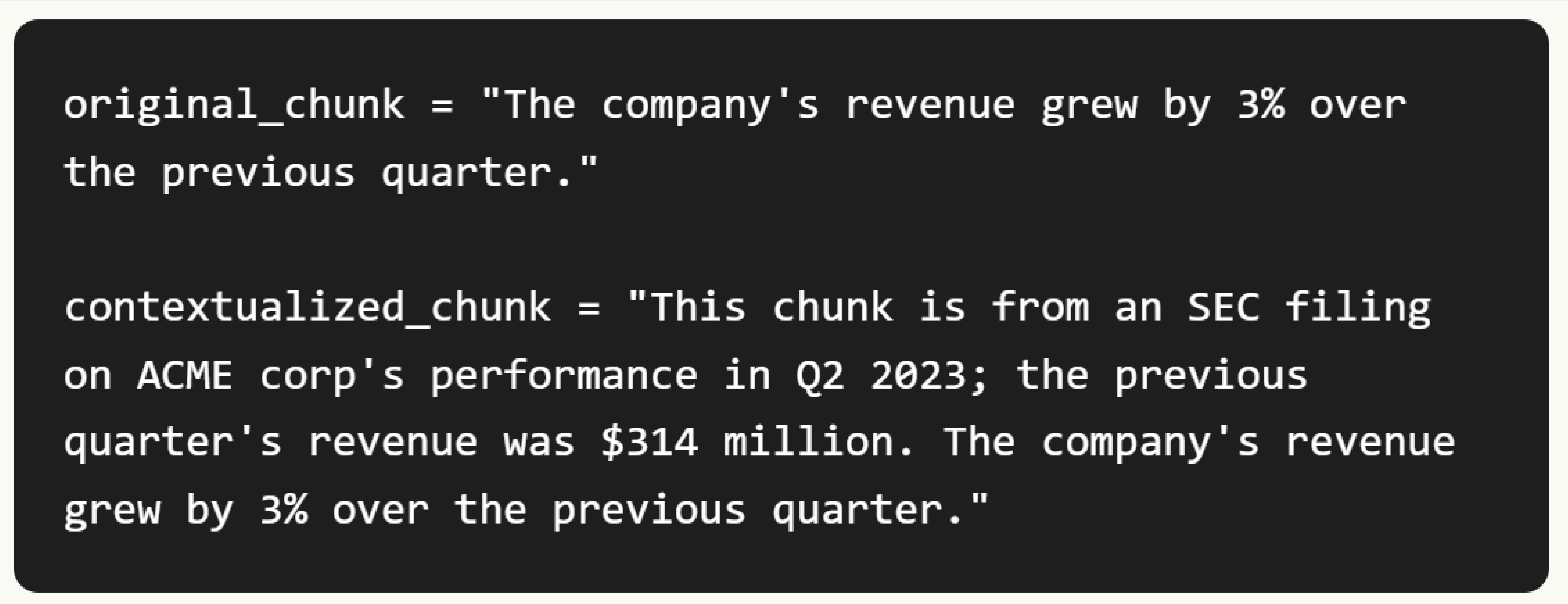
A user querying "**Error code TS-999**" may not retrieve the specific information related to that error code if conventional embeddings are used, as they might **return generalized content about error codes.**



Improved RAG with Contextual Retrieval

1. **Contextual Embeddings** - Prepending chunk-specific explanatory context to each chunk before embedding

2. **Contextual BM25** - Uses lexical matching to identify precise terms



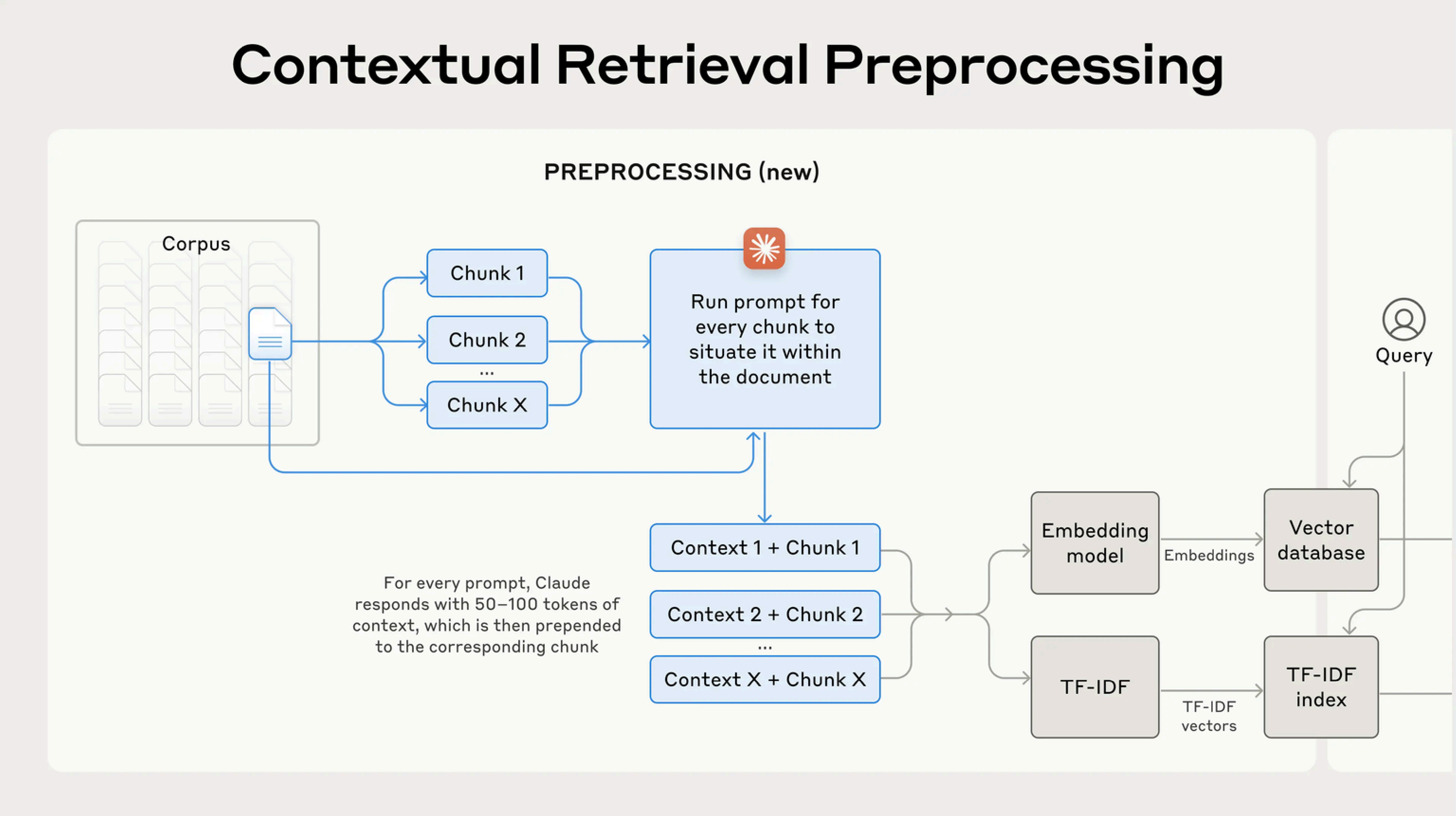
```
original_chunk = "The company's revenue grew by 3% over
the previous quarter."

contextualized_chunk = "This chunk is from an SEC filing
on ACME corp's performance in Q2 2023; the previous
quarter's revenue was $314 million. The company's revenue
grew by 3% over the previous quarter."
```

Improved contextual integrity

Improved Precision

Implementing Contextual Retrieval

1. Decompose documents into smaller, manageable chunks.
2. Prepend relevant contextual data to each chunk.
3. Generate both embeddings and BM25 indexes for advanced retrieval.

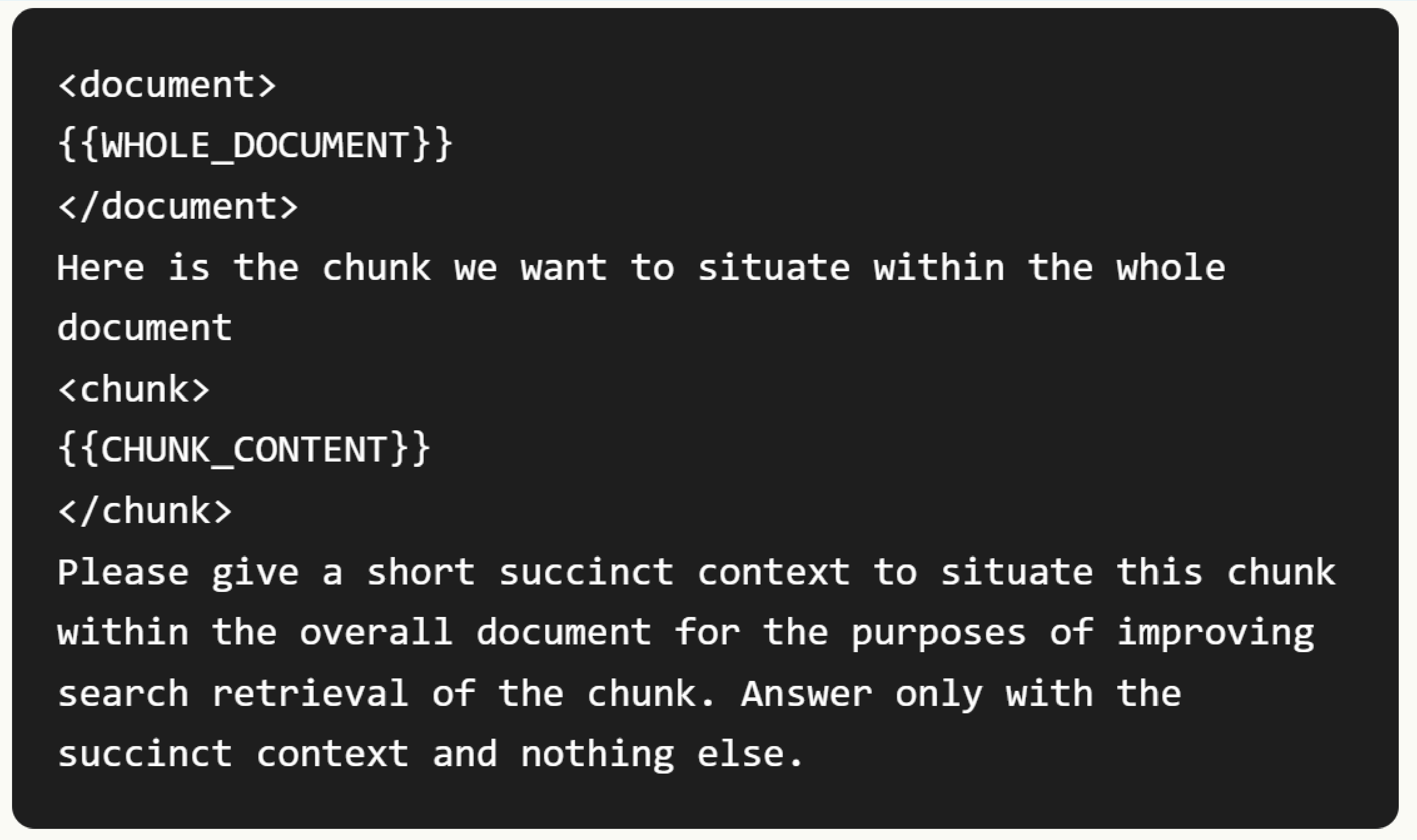


Contextual Retrieval Preprocessing

For every prompt, Claude responds with 50–100 tokens of context, which is then prepended to the corresponding chunk

But how do i create contextual relevant chunks?

1. Ask LLM to provide concise, chunk-specific context.
2. The contextual text should be 50-100 tokens long.



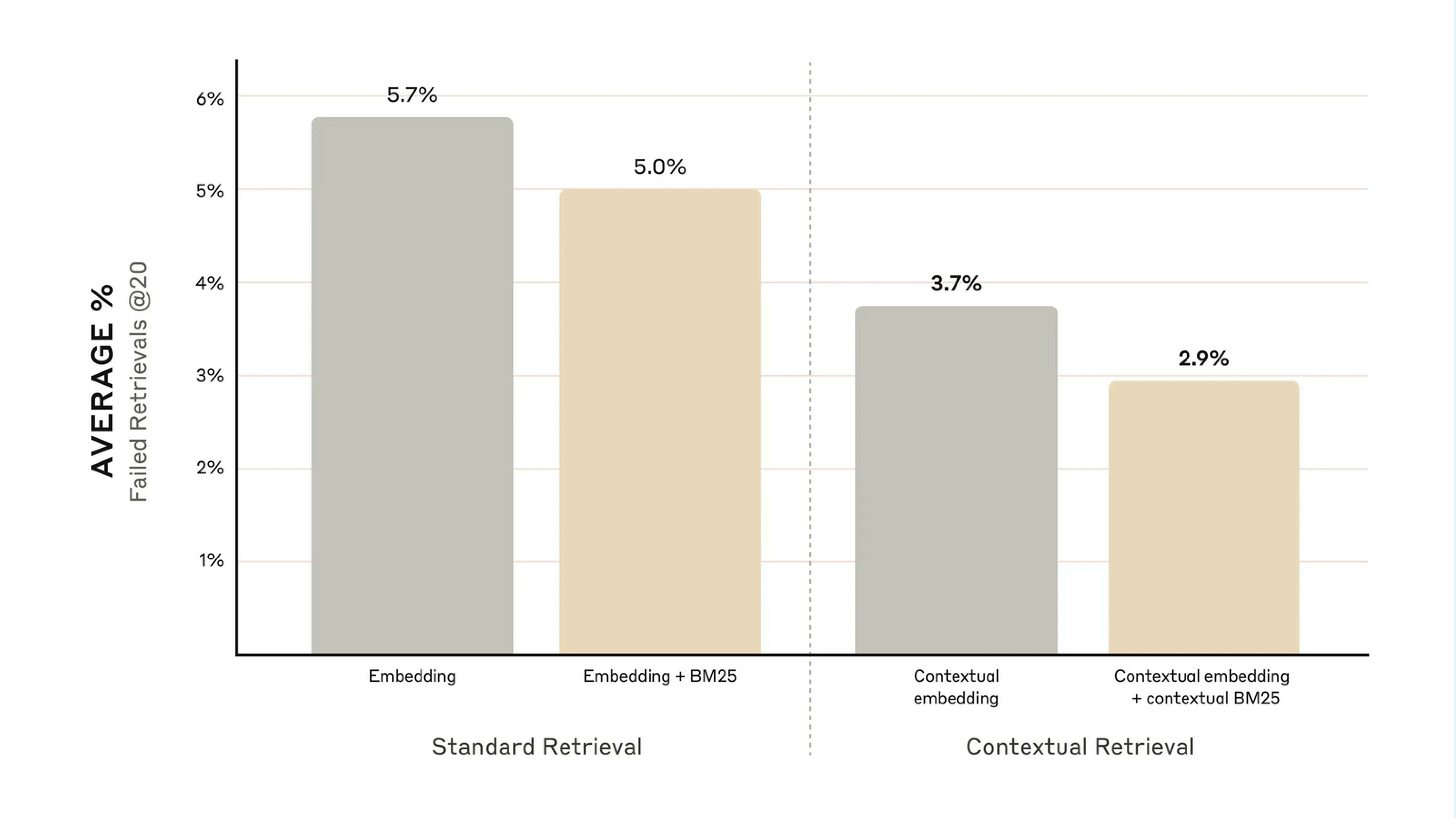
```
<document>
{{WHOLE_DOCUMENT}}
</document>
Here is the chunk we want to situate within the whole
document
<chunk>
{{CHUNK_CONTENT}}
</chunk>
Please give a short succinct context to situate this chunk
within the overall document for the purposes of improving
search retrieval of the chunk. Answer only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

Prompt to create contextual retrieval

Performance improvements

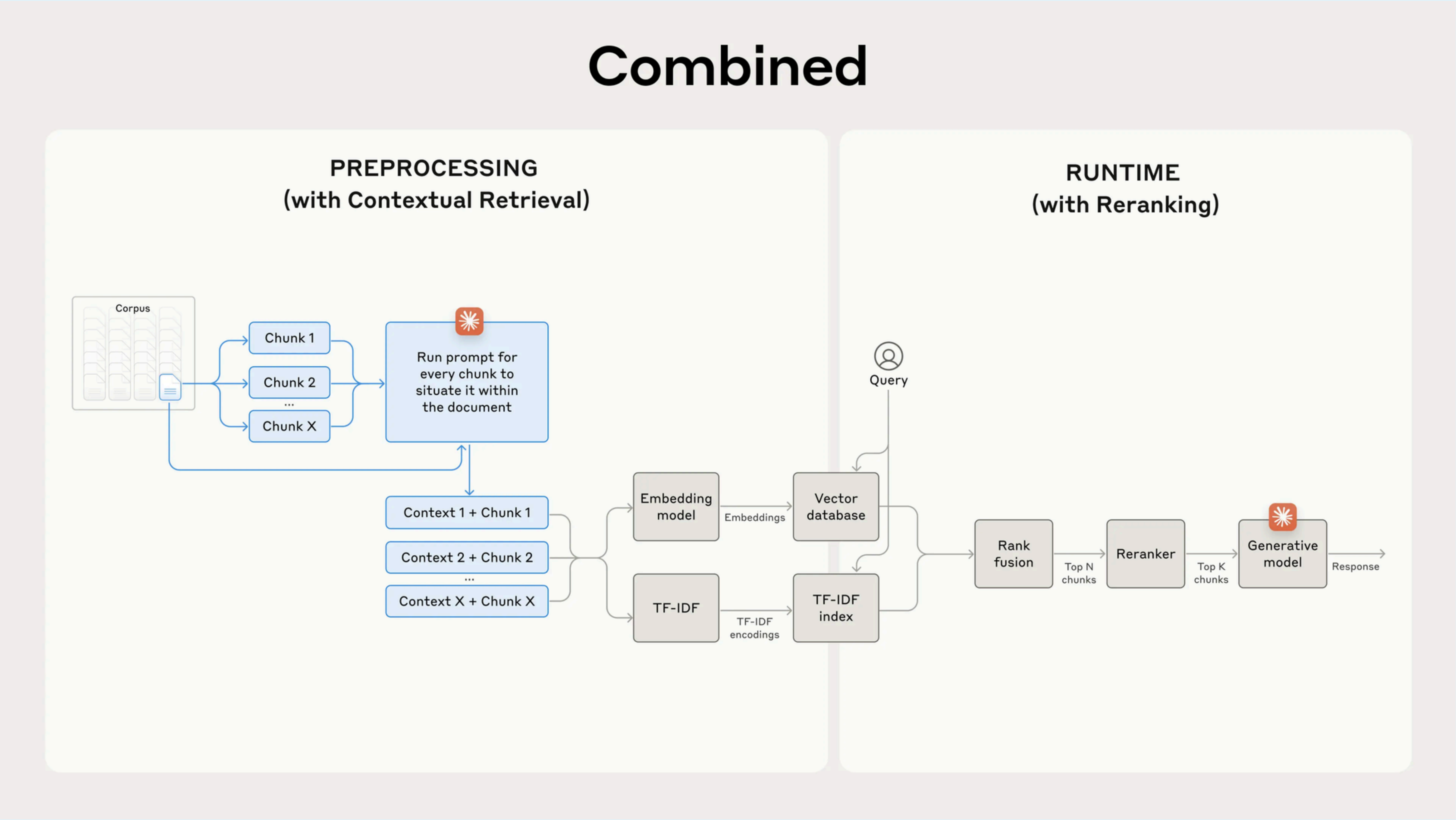
1. Contextual Embeddings: **35% reduction** in retrieval failures.

2. Contextual Embeddings + Contextual BM25: **49% reduction** in retrieval failures.



Further boosting performance with Reranking

1 Incorporating reranking: **67% reduction** in retrieval failures.



Implementation Considerations

1. **Chunk Boundaries:** Chunk size, boundary, and overlap affect retrieval performance.

2. **Embedding Model:** Some models perform better; Gemini and Voyage were effective.

3. **Custom Prompts:** Tailored prompts may yield better results, e.g., including glossary terms.

4. **Number of Chunks:** More chunks improve context but can distract models; 20 chunks was effective.

5. **Always Evaluate:** Improve response generation by distinguishing context from the chunk.

www.masteringllm.com

50% OFF

LLM Interview Course

Want to Prepare yourself for an LLM Interview?

✓ 100+ Questions spanning 14 categories with Real Case Studies

✓ Curated 100+ assessments for each category

✓ Well-researched real-world interview questions based on FAANG & Fortune 500 companies

✓ Focus on Visual learning

✓ Certification

HOT SALE!

Coupon Code - LLM50

Coupon is valid till 30th Oct 2024

www.masteringllm.com

AgenticRAG with LlamaIndex

Want to learn why AgenticRAG is future of RAG?

- ✓ Master **RAG fundamentals** through practical case studies
- ✓ Understand how to overcome **limitations of RA**G
- ✓ Introduction to **AgenticRAG** & techniques like **Routing Agents, Query planning agents, Structure planning agents, and React agents with human in loop**.
- ✓ **5** real-time **case studies with code walkthroughs**