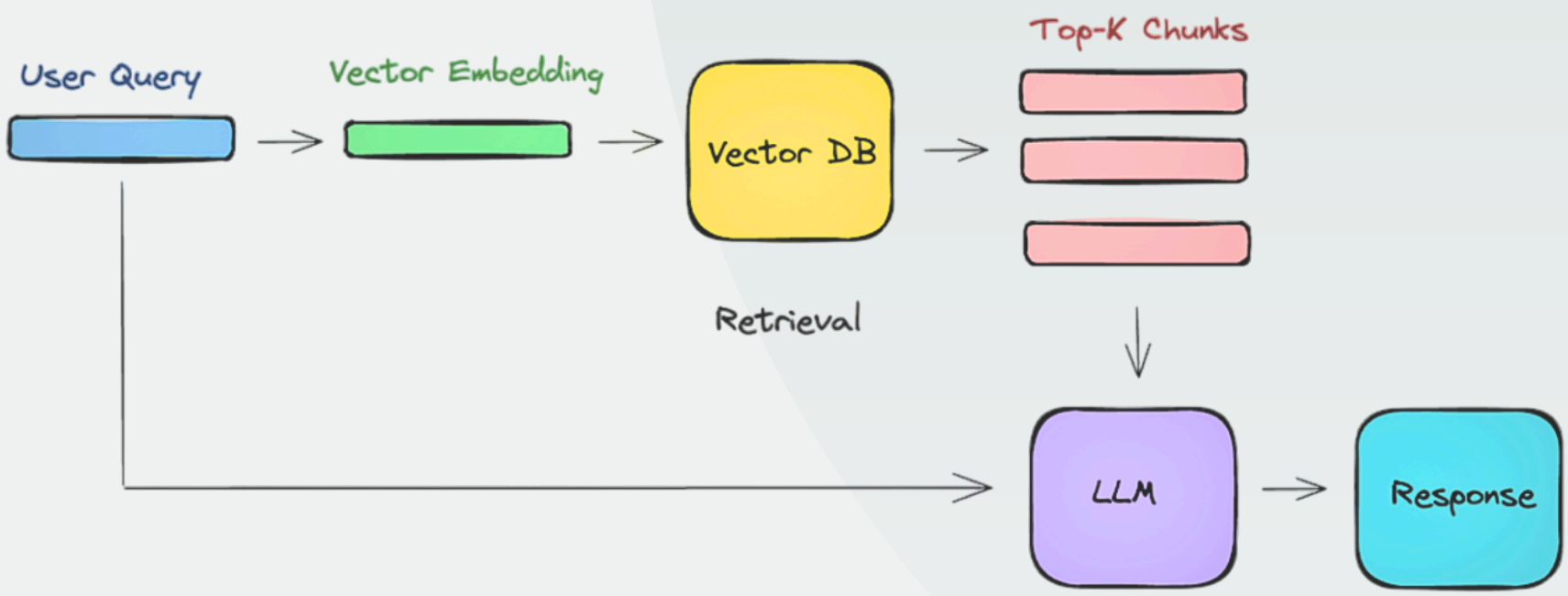
# PROMPT ENGINEERING

## PATTERNS FOR RAG IMPLEMENTATIONS




linkedin.com/in/that-aum

# INTRODUCTION

Generative AI models are powerful but often generate inaccurate or irrelevant responses.

- **Retrieval-Augmented Generation (RAG)** solves this by using external knowledge for better accuracy.
- **Prompt Engineering** is key to improving RAG performance.

Let’s explore how to craft better prompts for successful RAG implementation!

# RETRIEVAL PROMPT

In RAG, retrieval prompts enhance query quality before retrieving documents.

Three key techniques:

✅ Query Expansion

✅ Contextual Continuity

✅ Hypothetical Document Embeddings (HyDE)

# QUERY EXPANSION

- **What?** Improves query wording for better document retrieval.
- **How?** Add synonyms, related terms, and domain-specific keywords.

## Example

**"Expand the query {query} into 3 search-friendly versions using synonyms and related terms. Prioritize technical terms from {domain}."**

# CONTEXTUAL CONTINUITY

- **What?** Uses previous conversation history to refine the query.
- **Why?** Ensures continuity and relevance in retrieval.

## Example

**"Based on the user's previous query about {history}, rewrite their new query: {new query} into a standalone search query."**

# HyDE

## HYPOTHETICAL DOCUMENT EMBEDDINGS

- **What?** Generates a hypothetical answer to guide retrieval.
- **Why?** Helps find documents closer to the expected response.

## Example

**"Write a hypothetical paragraph answering {user query}. Use this text to find relevant documents."**

# GENERATION PROMPT

Once documents are retrieved, **generation prompts** guide the LLM to produce accurate responses.
Key techniques:

- ✅ Explicit Retrieval Constraints
- ✅ Chain of Thought (CoT) Reasoning
- ✅ Extractive Answering
- ✅ Contrastive Answering

# EXPLICIT RETRIEVAL CONSTRAINTS

- **What?** Forces LLM to generate answers only from retrieved documents.
- **Why?** Prevents hallucinations and ensures reliability.

## Example

**"Answer using ONLY the provided document sources: {documents}. If the answer isn't there, say 'I don't know.' Do not use prior knowledge."**

# COT

## CHAIN OF THOUGHT REASONING

- **What?** Breaks down complex reasoning step by step.
- **Why?** Helps generate structured and logical responses.

## Example

**"Based on the retrieved context: {retrieved documents}, answer {query} step by step, first identifying key facts, then reasoning through the answer."**

# EXTRACTIVE ANSWERING

- **What?** Extracts relevant text directly from retrieved documents.
- **Why?** Ensures precise and unchanged responses (useful for legal, medical use cases).

## Example

**"Extract the most relevant passage from {retrieved documents} that answers {query}. Return only the exact text without modification."**

# CONTRASTIVE ANSWERING

- **What?** Provides multiple perspectives on the same query.
- **Why?** Useful for debates, legal cases, and critical analysis.

## Example

**"Based on {retrieved documents}, provide a balanced analysis of {query} by listing pros and cons, with supporting evidence from the retrieved context."**

# CONCLUSION

🚀 **RAG enhances LLM accuracy by retrieving relevant knowledge.**

📌 **Well-crafted** prompts are key to improving both retrieval and generation.

Don’t forget to like,
comment, and save...

**linkedin.com/in/that-aum**