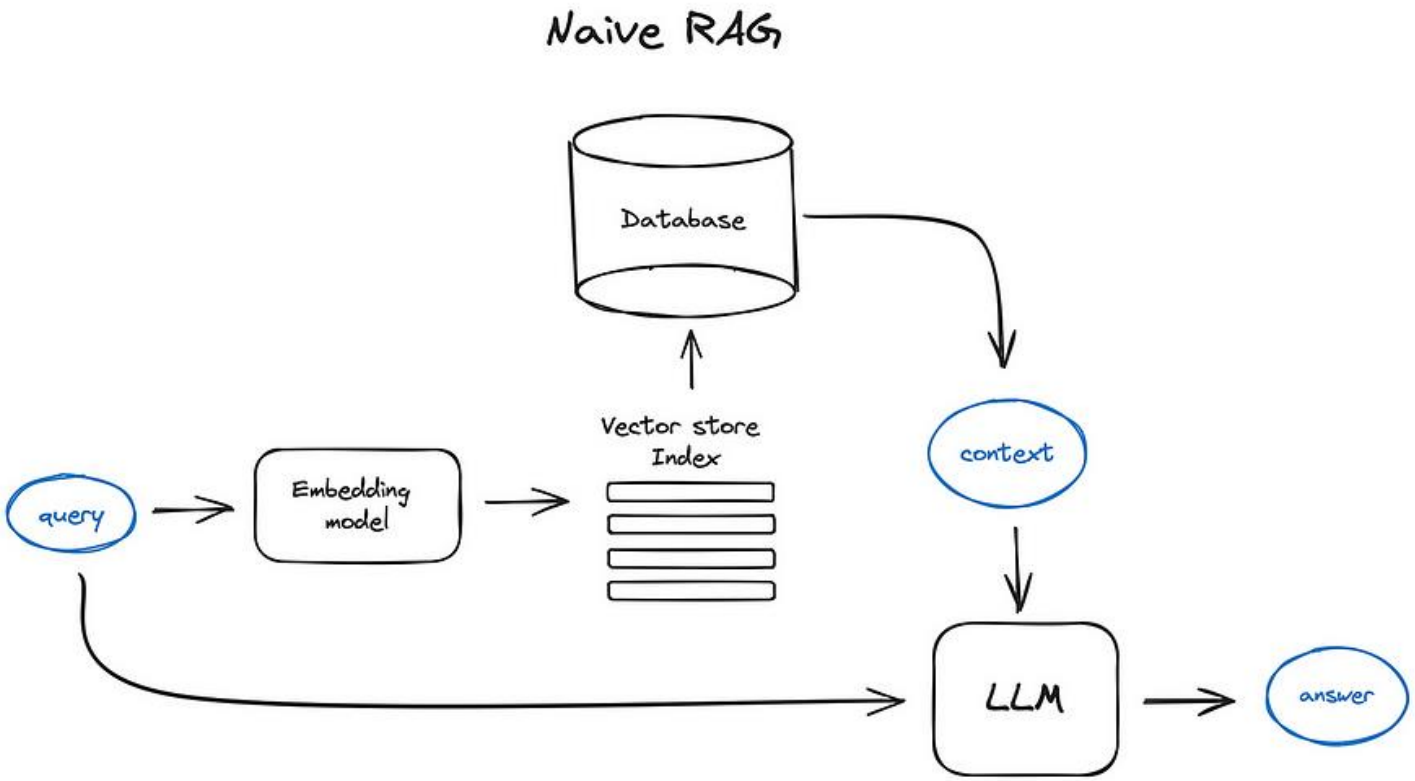
Context-Based Retrieval for Enterprise RAG Applications

Mayur Hooli

Basic RAG Architecture



Why Context Matters

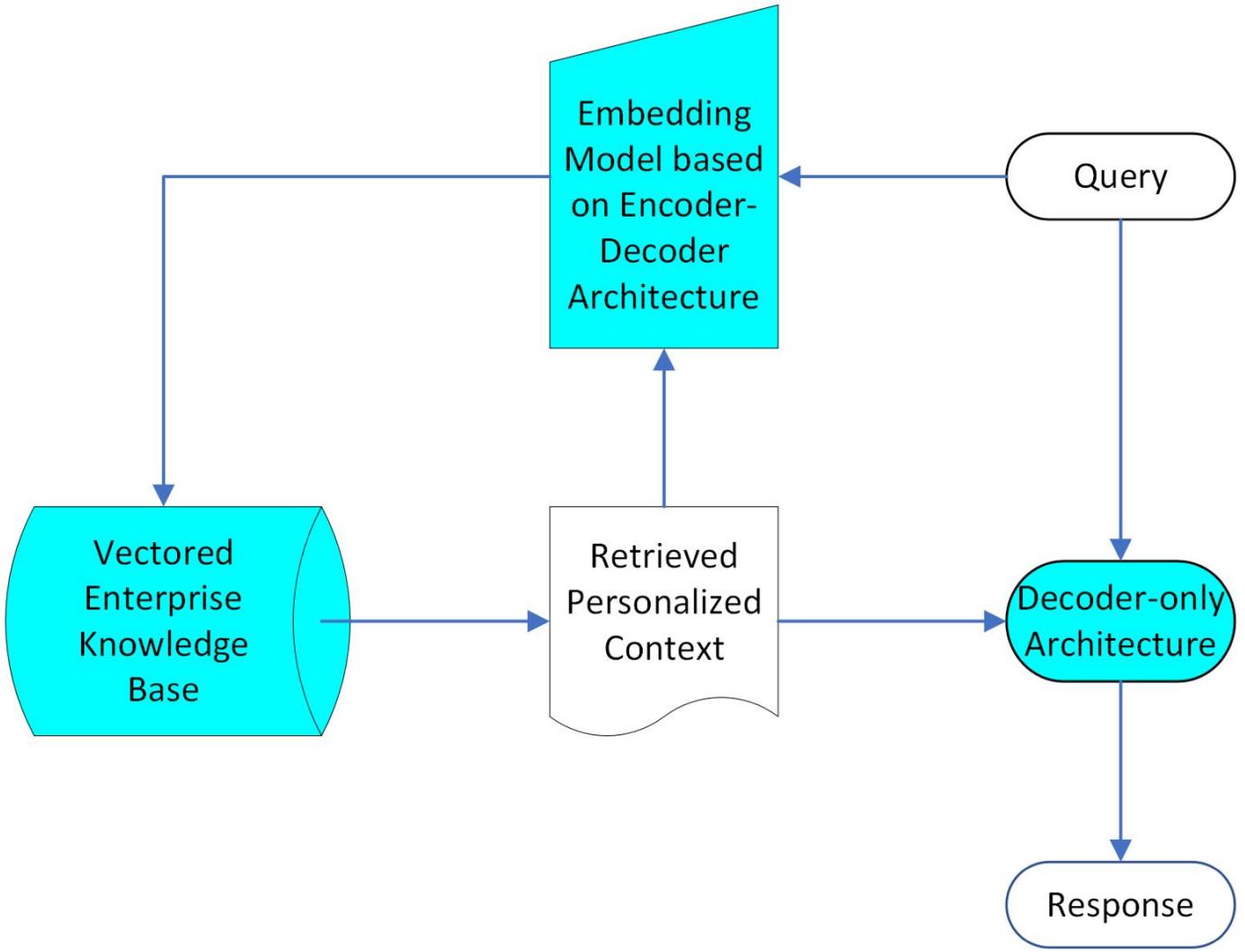
Telecom: Engineers need detailed technical specifications, such as 3GPP standards.

Finance: Financial analysts require highly relevant, recent market data.

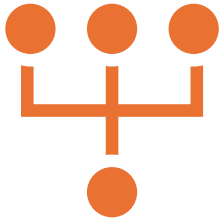
Retail: Retail managers need customer trend analyses specific to their region.

Traditional RAG models don’t differentiate between these needs, causing generic results that may not meet the exact requirements of diverse professionals.

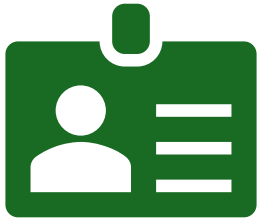
Proposed Solution



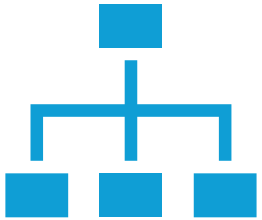
Why RAG Needs Enhancement



Lack of Flexibility: Encoders retrieve information in a general manner, without nuanced adaptation to context.



Limited Personalization: The retrieved data doesn't account for individual user profiles or industry-specific needs.



Enterprise Scalability Issues: Large organizations with varied departments face inconsistent retrieval results.

Core Innovation

Instead of a traditional encoder-decoder model, the model I am using is an encoder-decoder-decoder model

The encoder model is good for retrieval, but my experiments have yielded better results by replacing an encoder-only retrieval model with an encoder-decoder model

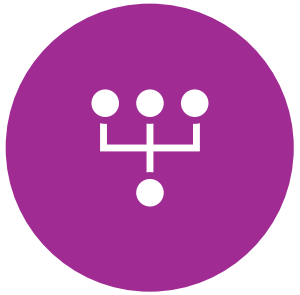
Personalize responses: By utilizing user feedback, the decoder dynamically modifies the retrieved information, leading to a more tailored and accurate output for the user.

Key Benefits

Enterprise-Specific Context: Retrieves industry-relevant data, such as legal precedents for lawyers or financial trends for analysts.

Personalization: Adaptable to each user's historical queries and preferences.

Increased Accuracy: The system learns from interactions and refines its responses to become more accurate and contextually appropriate over time.



Scalability: Capable of handling the retrieval needs of large organizations with diverse departments.

Use Case: Telecom Industry

Problem: A telecom engineer seeks 5G specifications.

Traditional RAG: Retrieves general telecom data, potentially irrelevant.

Proposed Model: Retrieves specific 5G standards from **3GPP documents** and relevant articles, adapting to the engineer's context and previous queries, ensuring the result is both timely and specific.

Continuous Learning

User Feedback: Each user interaction refines the system's understanding of what constitutes relevant and useful information.

Personalization Over Time: As the system learns, it adjusts responses to be more accurate and personalized, improving the user's experience over time.

Enterprise Learning: The system can learn patterns for entire departments or industries, providing better results across the organization.

Scalability

Sector-Specific Adaptability: The system can retrieve relevant information across multiple departments, such as finance, legal, engineering, or sales, all within the same organization.

Efficient Performance: Handles massive data sources and returns personalized results in real-time, optimizing performance for large organizations.

Competitive Advantage

Better Relevance: Provides more personalized and contextually relevant responses, customized for enterprise use cases.

Adaptability: Works across diverse industries, unlike traditional RAG models which may not specialize in any specific context.

Continuous Improvement: Learns from user interactions, offering increasingly better results over time.

Future Applications

Retrieve	Healthcare: Doctors and medical professionals can retrieve patient-specific data, treatment histories, and recent research based on the context of their practice.
Retail	Retail: Retailers can query sales trends, customer behavior data, and product-specific insights, receiving personalized recommendations for marketing strategies.
Retrieve	Legal: Lawyers can retrieve case law and legal precedents, refined by region, jurisdiction, and case type.

Summary

The **Context-Based Retrieval System** enhances **traditional RAG models** by improving retrieval flexibility, **contextual relevance**, and **personalization**.

The solution is scalable for enterprise use, adaptable across industries, and continuously improving based on user feedback.

This system addresses the critical gaps left by traditional **encoder-only** retrieval processes.