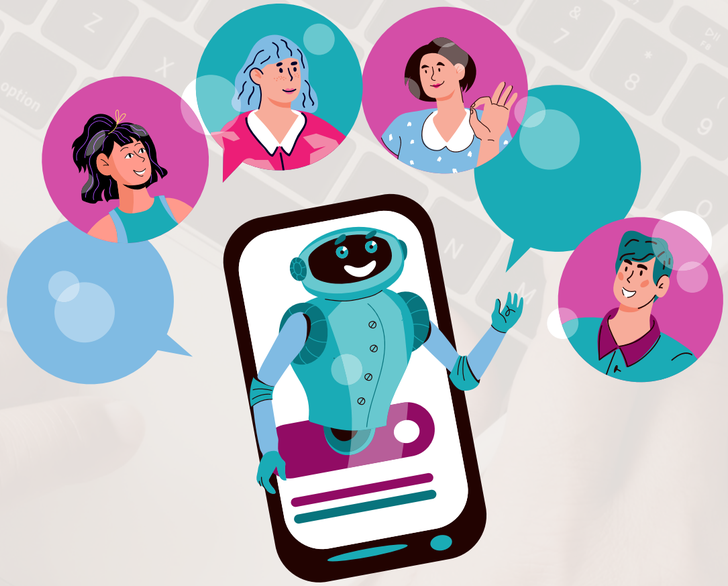
RAG

METHODOLOGY



Introduction

RAG METHODOLOGY

Imagine you’re having a conversation with an AI assistant, and it confidently answers your question—but you realise the response is outdated or completely made up. Frustrating, right? That’s where Retrieval-Augmented Generation (RAG) steps in to save the day! Instead of relying solely on pre-trained knowledge, RAG pulls in real-time, relevant information to provide more accurate and reliable responses.

WHY RAG ?

Reduces Hallucinations

Explainability & Trust

Domain Specific

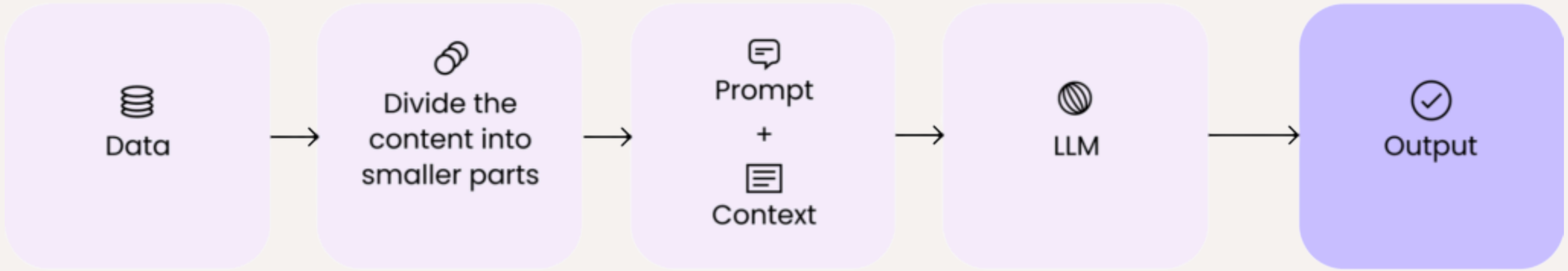
Keeps info UP - TO - DATE

Function Calling

Different Versions of RAG

- **Standard RAG –** Simple retrieval + generation pipeline.
- **Multiquery RAG –** Generates multiple queries from a single input to improve retrieval quality (used in our case).
- **Hierarchical RAG –** Uses structured knowledge retrieval, like summaries from different document levels.
- **Conversational RAG –** Optimized for multi-turn dialogue, ensuring context is preserved across interactions.

The RAG process



RAG is Like Mobile Phones

Just as different companies release different models of smartphones, different frameworks implement varied RAG architectures to optimize performance based on specific needs.

Applications of RAG

- **Standard RAG –** Simple retrieval + generation pipeline.
- **Multiquery RAG –** Generates multiple queries from a single input to improve retrieval quality (used in our case).
- **Hierarchical RAG –** Uses structured knowledge retrieval, like summaries from different document levels.
- **Conversational RAG –** Optimized for multi-turn dialogue, ensuring context is preserved across interactions.

Big challenge we face in LLM’s

Hallucinations

What's the capital of Mars?

The capital of Mars is Muskland.

RAG solves hallucinations

- **Uses Real-Time Data:** Fetches actual information from external sources to ensure accuracy.
- **Reduces Memory Reliance:** Checks up-to-date info, minimizing errors from outdated knowledge.
- **Provides Relevant Answers:** Generates responses based on documents it retrieves, making them more accurate.
- **Clarifies Unclear Questions:** Pulls in useful documents to give better responses and avoid mistakes.

Conclusion

In a world full of information overload, RAG acts as a smart guide, always checking the latest facts before giving you a response. It’s like having a well-informed assistant who never relies solely on memory but instead taps into the best sources for clear, relevant, and accurate answers every time. With RAG, there's no room for ambiguity—only reliable, up-to-date insights!

Thank you