

Project: Multi-user Conversational RAG System

<u>Instructor</u>

Dipanjan Sarkar

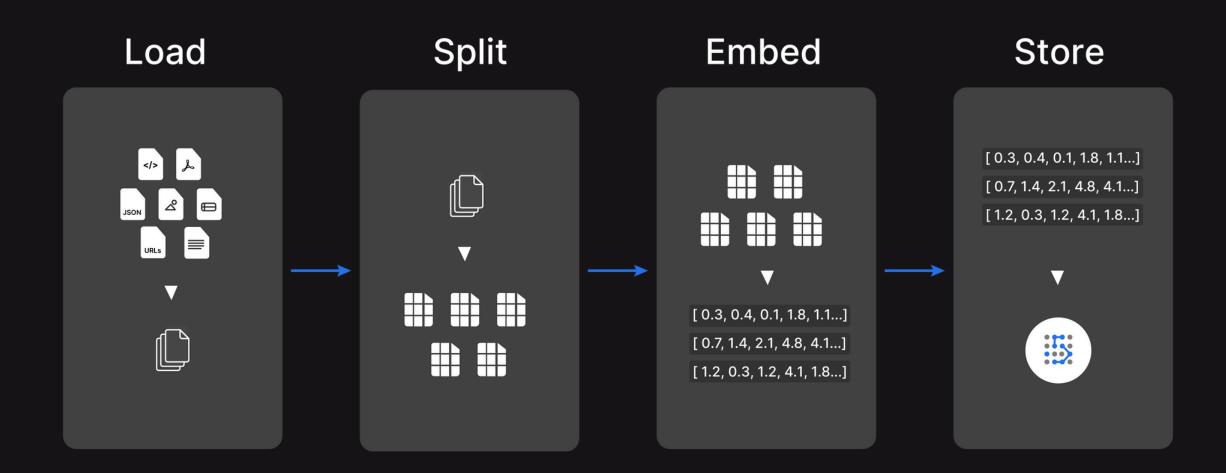
Head of Community & Principal Al Scientist at Analytics Vidhya

Google Developer Expert - ML & Cloud Champion Innovator

Published Author

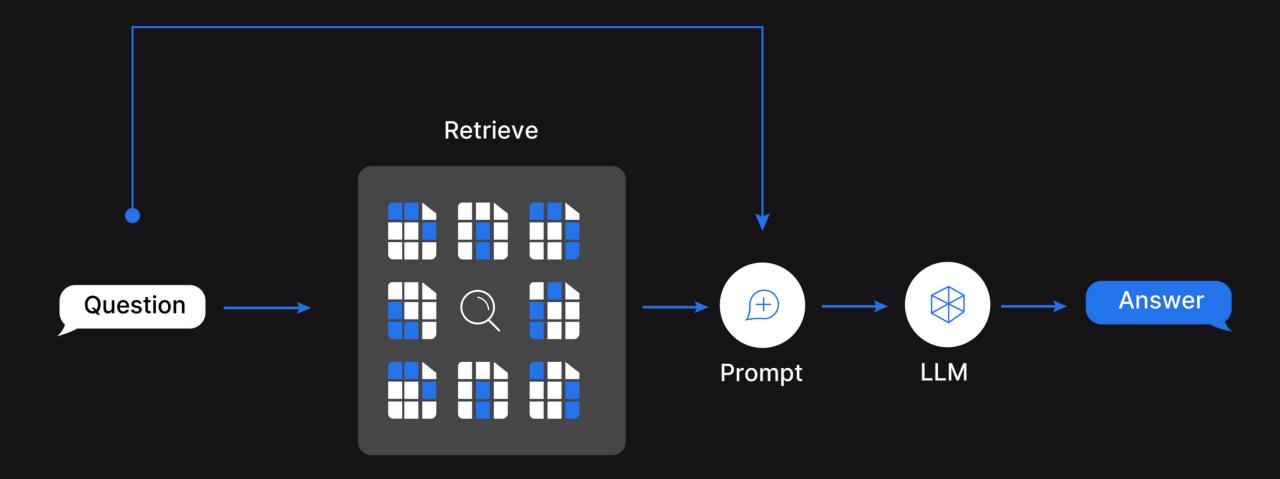


RAG System Architecture





RAG System Architecture





What is a Conversation?

HumanMessage Explain to me Generative AI in 3 bullet points Creation of New Content: Generative AI uses machine learning models to create new data similar to existing data. It can generate text, images, music, and other types of content by learning patterns from large datasets. • Advanced Neural Networks: It leverages advanced neural networks, such as Generative Adversarial Networks (GANs) and Transformer models (like GPT-4), to produce realistic and **AIMessage** high-quality outputs that mimic human creativity and intelligence. • Wide Applications: Generative AI has diverse applications, including content creation, design, entertainment, drug discovery, and more, providing innovative solutions across various **Conversation History** industries by automating and enhancing creative processes. or Memory (小) □ ① ◇ ◇ HumanMessage Now do the same for Deep Learning · Hierarchical Learning: Deep learning uses neural networks with multiple layers (deep neural networks) to automatically learn hierarchical representations of data. Each layer extracts increasingly complex features from the raw input data. • State-of-the-Art Performance: It achieves state-of-the-art performance in various tasks such as image and speech recognition, natural language processing, and autonomous driving by **AIMessage** leveraging large amounts of data and computational power. • End-to-End Training: Deep learning models are trained end-to-end, meaning they can learn directly from raw data to final output without the need for manual feature extraction, allowing for more accurate and efficient processing and prediction. 小日日日日 New Prompt or Query HumanMessage

Need for Conversational Memory

With Conversational Memory

I am interested in integrating LLMs with external knowledge

LLMs are great at generating human-like text. Yet, integrating external knowledge can enhance their capabilities even more.

What are the different possible methods for doing this?

You could use pre-existing knowledge graphs, allow LLMs access to tools like APIs, or retrieval augmentation with vector DBs!

······ Conversation History ·····

Interesting! What was it I wanted to know about?

You were interested in integrating LLMs with external knowledge.

Without Conversational Memory

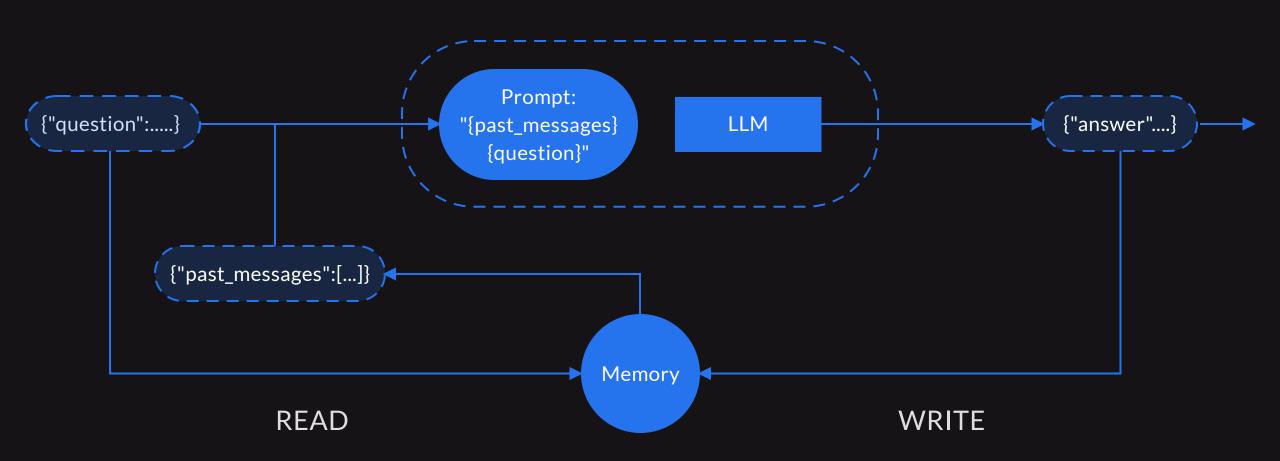
(No conversation history is stored)

Interesting! What was it I wanted to know about?

Sorry, I have no idea what you're talking about!

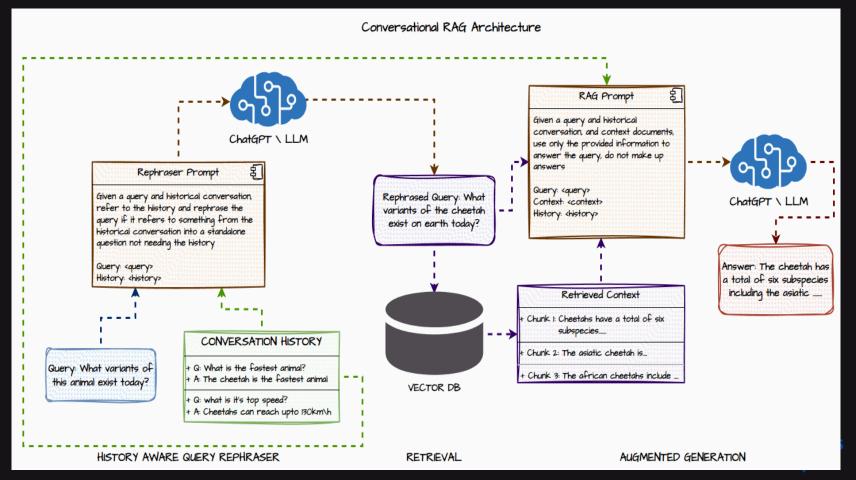


Conversation Chain with Memory in LCEL





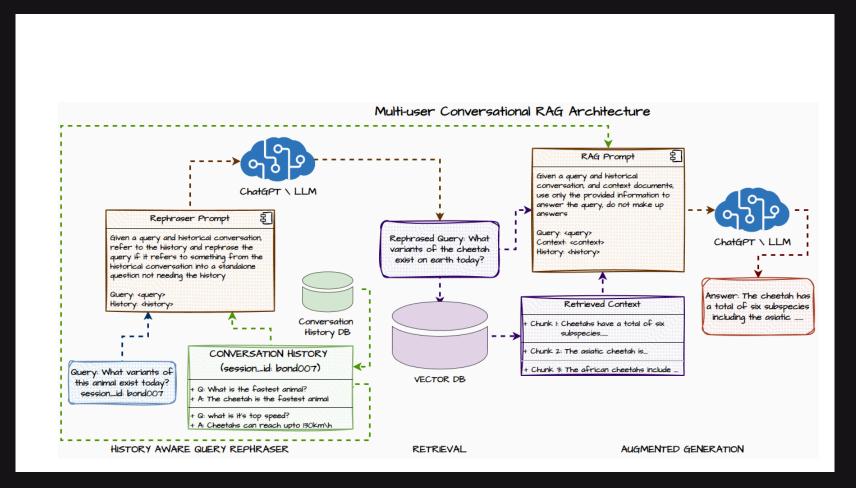
Conversational RAG System



Conversational RAG System rewrites the user query based on the historical conversation into a more contextual query and plugs it into the RAG System



Multi-User Conversational RAG System



Multi-user Conversational RAG System maintains a conversation history per user-session ID and hence can cater to multiple user queries at the same time



Thank You

