Two innovative techniques to optimize RAG

1. Implement Hybrid Search with Keywords and Dense Vectors:

- Currently, the search is based solely on dense vector similarity. There is a possibility to enhance this by implementing a hybrid search that combines dense vector similarity with keyword-based search. This approach can capture both semantic similarity and exact keyword matches, potentially improving retrieval accuracy.

This hybrid approach could lead to more robust and accurate document retrieval, especially for queries that contain specific terms or phrases.

Reference - https://medium.com/@csakash03/hybrid-search-is-a-method-to-optimize-rag-implementation-98d9d0911341

2. Adding memory to maintain context across conversations:

- To add memory to the RAG chatbot and maintain context across conversations, we can use **Langchain** framework to implement a conversation history mechanism. The chain will be such that all the previous conversation with the LLM will be summarised and will be sent to the LLM in the next iteration of QnA. This method is better than the traditionally storing previous messages using any data structure like list or dictionary where if the conversation goes long , then the size of previous message increses incresing the tokens per minutes to LLMs.

I have tried to explain the mechanism in the following diagram.