Analysis on Multi-Turn Memory Rag Model

- Multi Turn Memory Rag is a conversional memory rag model with conversation storing capability, that help keep flow and interconnected results based on conversation and context.

- Name of the Model File:
  - Memeory_multiturn_RAG.ipynb

- Document Used:
  - IB ACIO Job Notification pdf

- Chunking Strategy:
  - Recursive Character Chunking with
    - Chunk size = 500
    - Overlapping = 100

- Vector DB used:
  - FAISS

- Embedding Model used:
  - HuggingFace
    - all-MiniLM-L6-v2

- Model used:
  - Mistral AI's
    - mistral-large-latest

- Memory used for Storing Conversations:
  - ConversationBufferMemory from langchain

- Analysis of Working Flow:
  - Memory is used for storing conversation between User and RAG.

  - Each and every conversations are stored on memory until exit of runtime of RAG.

  - Rag look on memory and context from vector db and give the result based on user query like - i forget about last query, could you generate my last asked queries ?

    - Rag will look at memory and give the user their past conversations
    - Conversations on Memory persist until exit of runtime, means memory will be wiped after exit of RAG
    - User can look and see what conversations are stored at Memory
        - Using memory.buffer or memory.buffer_as_message
    - Noticed fast performance and accurate answers from RAG.
    - Calculated response time by using time module from python.

- Look at Rag_with_memory_conversation excel document for before continues conversations and after some time conversations for data persist on memory when data at buffer or cache

- RAG and relevant reports at : Gen-AI/model_project/Progression_report_Memory_analysing at main · sayed2174/Gen-AI

Reported by:
SAYED MOHAMMED