**Sahil Wani**

SUID: 466848553

Oct 1st report

**Chatbot app with short term memory:**

A screenshot of a computer screen

Description automatically generated

A screenshot of a computer

Description automatically generated

**Chatbot code:**

A screenshot of a computer program

Description automatically generated

**Best Practices**

**Manage Conversation History:** To maintain efficient performance, it's important to limit the amount of conversation history passed to the model. Restricting the memory to the last 5-10 exchanges ensures the chatbot retains relevant context without exceeding model limits.

**Clear Termination Commands:** Provide users with straightforward options to exit the conversation, such as typing `exit` or `quit`. This enhances the user experience by making the interface intuitive.

**Token Usage Awareness:** Be mindful of token limitations for the LLM model in use. Exceeding these limits may result in incomplete responses or errors. Ensuring prompts stay within these bounds helps avoid disruptions.

**Optimize API Requests:** To reduce latency and ensure smoother responses, structure API calls efficiently. Focus on balancing token usage with model accuracy to ensure prompt replies while managing resource consumption.

**Challenges**

**Managing the Context Window:** One of the primary challenges is handling large conversation histories, as they can exceed the model's token limit. Properly managing and truncating older parts of the conversation is essential to maintaining relevant context.

**Token Limit Issues Despite Balance:** A recurring issue is encountering token limit errors or requests to upgrade the plan, even when there are available tokens in the account. This can disrupt functionality and often requires a review of account settings or contacting API support to resolve the discrepancy.