Task 1:

Your objective is to develop a chatbot system powered by large language model (LLM)

inference on your local machine taking advantage of GPU acceleration to

optimize performance and including these machine details in your submission

documentation. If you don't have access to a machine with GPU, CPU inference is also sufficient.

The interface of the chatbot should be designed for ease of use by human

users. Establish a local git repository on your machine for version control, and employ

professional best practices by committing updates regularly as you progress through the

task. Select a suitable LLM for the chatbot and consider integrating established

packages from trusted sources such as public GitHub repositories or PyPI. In your

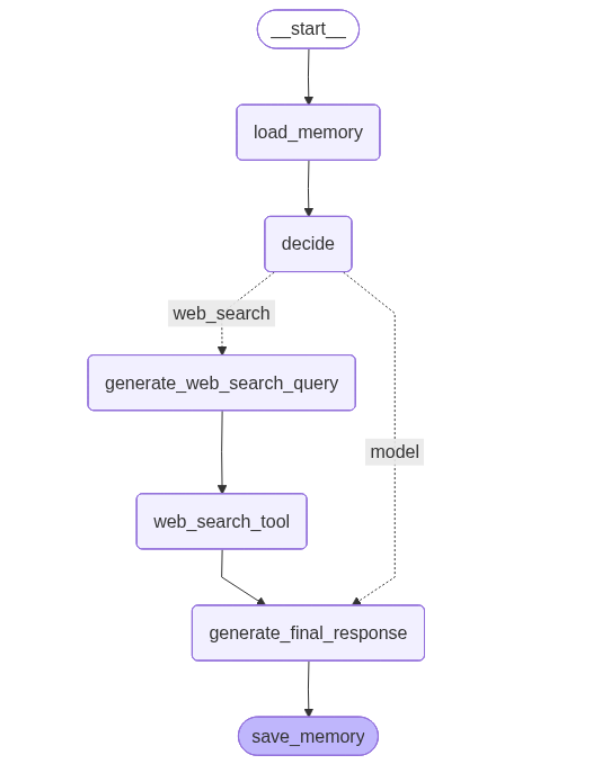
documentation, justify your selection of the model and packages. Ensure the repository

remains local and not uploaded to any public platforms like GitHub.

Approach:

* The chatbot inference is performed in **CPU** (12th Gen Intel(R) Core(TM) i5-1235U (1.30 GHz)) with 16 GB RAM.
* The chatbot is built using **Langgraph** and served using **Gradio** for user friendly interface.
* **Model — llama3.2 3B (instruct-tuned):** A compact, instruction-tuned 3B model chosen to balance responsiveness on CPU with reasonable instruction-following quality.
* **Web search tool:** A real-time search integration lets the bot fetch and incorporate up-to-date information during conversations.
* **Short-term session memory:** Session-scoped memory stores recent messages and small user preferences to keep the conversation coherent
* Used **Sentence Transformer(sentence-transformers/all-mpnet-base-v)** as embedding model to store and retrieve conversational memory

Below is the workflow:



Chatbot Interface and Example Usage

