Overview[¶](https://www.kaggle.com/code/hinepo/q-a-chatbot-with-llms-harry-potter#Overview)

* Use [Langchain](https://python.langchain.com/en/latest/index.html) to **build a chatbot that can answer questions about** **Medications**
* Flexible and customizable RAG pipeline (Retrieval Augmented Generation)
* In this project I will only use Langchain as LLM
* Use [FAISS vector store](https://python.langchain.com/docs/integrations/vectorstores/faiss) to store text embeddings created with [Sentence Transformers](https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2)
* Use [Retrieval chain](https://python.langchain.com/docs/modules/data_connection/retrievers/) to retrieve relevant passages from embedded text
* Summarize retrieved passages
* Chat UI with [Gradio](https://www.gradio.app/guides/quickstart)

**Approach:**

1. **Document and Vector Management:**
   * I will set up processes to load, split, and vectorize documents from various formats. These documents will be stored in a Chroma vector database to support semantic searches and efficient document retrieval.
2. **Development of Custom Prompt Templates:**
   * I will develop multiple prompt templates to guide the chatbot’s responses under various scenarios such as handling general inquiries, executing document retrievals, performing named entity recognition (NER), responding to graph queries, and summarizing chat histories.
3. **Agent and Service Integration:**
   * I will implement an **Agent** class that utilizes these templates within different operational contexts. This agent will be integral in determining the appropriate type of response based on the user's query.
   * I will also develop a **Service** class to manage the overall interaction flow, utilizing the Agent to handle direct inquiries and an LLMChain to generate summary responses based on chat history.
4. **Operational Flow:**
   * I will design the system to enrich the user's current message by summarizing recent chat history, providing a more informed and contextually relevant response. This will be especially beneficial in multi-turn dialogues.
5. **User Interface:**
   * I will use Gradio to create and share a web-based interface, ensuring that the chatbot is accessible and easy to use for individuals seeking medical consultation.

**Timeline:**

The project will be completed over a period of one month, with milestones set for the completion of document management systems, development of chat functionalities, integration of the agent and service classes, and final deployment of the chatbot interface.