Task: Build a Chatbot Using a Locally Deployed Open-Source LLM and a PDF Knowledge Base

**Objective:**

Develop a chatbot that responds to user queries based on the content of a given PDF document. The chatbot should use an open-source LLM deployed locally and retrieve relevant information from the PDF to generate accurate responses.

**Requirements:**

1. LLM Deployment:

* Use an open-source model such as Lama-3, Mistral, Llama-2, Falcon.
* Deploy the model locally using tools like Ollama, vLLM, LM Studio.
* Ensure the model can process user queries and return responses via an API or function call.

2. Knowledge Base Integration (PDF Processing):

* Load and process the given PDF document to extract its text content.
* Store the extracted text in a retrieval-friendly format, such as a vector database or an indexed text search.
* Implement a retrieval mechanism to find relevant information from the PDF before passing it to the LLM.

3. Chatbot Functionality:

* Create a simple UI using Streamlit, Flask, or FastAPI where users can input queries.
* When a user asks a question, retrieve the most relevant information from the PDF knowledge base and pass it along with the query to the LLM for response generation.
* Display the LLM’s response in the UI.

4. Backend Processing:

* Implement a pipeline to:
  1. Extract text from the PDF.
  2. Chunk the extracted text and store it in a vector database for fast retrieval.
  3. Retrieve relevant sections based on the user query.
  4. Feed the retrieved context into the LLM to generate an accurate response.

5. Optimization & Enhancements (Bonus Points):

* Optimize inference speed using quantization techniques.
* Enable streaming responses for a better user experience.
* Improve search accuracy using ranking or embedding’s.

**Expected Deliverables:**

* A GitHub repository with the Chatbot code.
* Instructions on how to set up and run the Chatbot locally.
* A short README explaining the approach, technologies used, and any limitations.
* A public link to access the Chatbot. (Plus Point)
* A recorded video link of demo.