Ailaysa

Technical Task:

AI/ML Intern - RAG-based Document Chatbot

Objective:

You are required to develop a **Retrieval-Augmented Generation (RAG) based Document Chatbot** that enables:

1. **Single-document chat**: Interaction with an individual document.

The chatbot should provide **efficient semantic retrieval** and generate **structured outputs** based on user queries.

Task submission deadline: April 07, 2025, 10:00 am (Monday)

Requirements:

Core Features:

- Implement a **RAG-based pipeline** that effectively retrieves relevant document sections and generates accurate responses.
- Ensure the chatbot **answers only within the context** of the provided document, avoiding hallucinations.
- The chatbot can be deployable on **Streamlit or any other suitable UI framework**.

Technical Considerations:

- The **retrieval mechanism** should efficiently extract the most relevant content.
- The **generation model** should produce **structured responses**, such as tables, bullet points, or well-formatted paragraphs when appropriate.
- The choice of embedding models, vector databases, or LLMs is **left to your discretion**.
- The chatbot should be capable of handling a reasonable amount of **document data** and processing user queries with low latency.

Deliverables:

- Source code with proper documentation.
- Instructions for running the chatbot.
- Brief technical write-up about:
    - Tech stack choices
    - Response structuring approach
    - Challenges faced and solutions implemented

Notes & Suggestions (Optional):

- Use Collab or Kaggle for GPU utilities.
- Use Hugging Face, Ollama libraries for open source LLMs and embedding models.

—