**RAG Mini-Platform — Technical Assignment (1–2 Weeks)**

**Objective**

Design and implement a lightweight Retrieval-Augmented Generation (RAG) system with the following capabilities:

1. **Document Storage & Retrieval**
   * Use an open-source vector database to store documents.
   * Support KNN-based retrieval for queries.
   * Allow uploading of documents (TXT, PDF, or similar, up to a few MB).
2. **User Interface (React)**
   * Provide a web-based interface where users can:
     + Upload documents
     + Search/query the knowledge base
     + Interact through a simple chat-like experience
3. **Context Engineering / Memory Sharing**
   * Implement a mechanism to share memory between prompts to maintain continuity and improve responses across queries.
4. **Multi-Context Protocol (MCP)**
   * Define and implement a basic protocol where:
     + **LLM-A** performs first-pass retrieval and generates an initial response.
     + **LLM-B** refines the response, leveraging the context passed from LLM-A.
5. **Deployment**
   * Containerize both the React UI and Python API using Docker.
   * Provide a simple way to run the entire system locally (e.g., via docker-compose).

**Additional Notes**

* You may use **any tools or frameworks** (e.g., ChatGPT, v0.dev, or other coding assistants) to accelerate development.
* The main focus is **not just on code delivery** but also on:
  + How you research and evaluate different technical options.
  + How you translate requirements into a working solution.
  + How clearly you can explain your design choices.