AI Intern Assignment: Document Q&A Assistant (RAG)

Goal

Build a web-based AI assistant that allows users to upload any document and ask questions that are answered strictly based on the document's content.

Overview

The intern will design and implement a Document Question & Answer (Q&A) system using Retrieval-Augmented Generation (RAG). The system should extract knowledge from documents, store it in a vector database, and answer user questions with clear citations.

Functional Requirements

1. Knowledge Setup (Upload & Index)

- Upload supported formats: PDF, DOCX, TXT, Markdown

- Display upload and indexing status

- Extract text from documents

- Chunk content into manageable sections

- Generate embeddings for each chunk

- Store embeddings with metadata (document name, page/section)

2. Chat Interface

- Simple chat UI

- User asks questions based on uploaded documents

- System retrieves relevant chunks

- LLM generates answers strictly from retrieved content

- Each answer must include citations

- If answer is not found, respond with:

  "I couldn't find this information in the uploaded document."

3. Core Features

- Chunk size: 500–1000 tokens with overlap

- Top-k retrieval (default = 5)

- Clear chat option

- Reset knowledge base option

Technical Expectations

- Use OpenAI or Azure OpenAI (or equivalent)

- Use a MongoDB

- Frontend can be Streamlit, React, or Next.js

- Backend can be Python (FastAPI) or Node.js

Deliverables

1. Working application

2. README with setup instructions

3. Architecture diagram

4. Test questions and expected answers

5. Short demo video (3–5 minutes)

Evaluation Criteria

- Accuracy & grounding (40%)

- Citation quality (20%)

- User experience (20%)

- Code quality & documentation (20%)