Brief Report: Offline AI-Powered RAG Knowledge Portal

1. Introduction

Organizations maintain a large amount of internal knowledge in the form of documents such as manuals, reports, research papers, and compliance files. Accessing useful information from these documents is often slow and inefficient, especially when using traditional keyword-based search. Many modern AI solutions depend on cloud-based services, which creates privacy and security concerns when handling sensitive organizational data.

This project proposes an **Offline AI-Powered RAG Knowledge Portal** that enables intelligent question answering over internal documents while ensuring complete offline operation, data privacy, and trustworthy responses.

2. Problem Description

The main challenges addressed in this project are:

- Difficulty in searching and understanding large document collections
- Inability of keyword search to handle semantic queries
- Risk of data exposure when using cloud-based AI systems
- Hallucinated or unverifiable answers generated by generic language models

The solution must operate without internet access, generate answers strictly from documents, and provide citations for transparency.

3. Proposed Solution

The proposed system follows a **Retrieval-Augmented Generation (RAG)** approach implemented fully offline.

Users upload documents in formats such as PDF, DOCX, or TXT. These documents are stored locally and processed page-wise to preserve original page numbers. The extracted text is divided into overlapping chunks to maintain semantic meaning across sections.

Each chunk is converted into a vector embedding using an offline sentence embedding model. Along with embeddings, metadata such as document ID, page number, and chunk ID is stored. These embeddings and metadata are indexed using a **local FAISS vector index**.

When a user submits a query, the query is embedded and compared with stored document embeddings. The most relevant chunks are retrieved and combined with the query. This augmented input is passed to a **single offline Large Language Model**, which generates the final response using only the retrieved document content.

Citations are generated using the stored metadata, allowing users to verify answers at the document and page level.

4. System Architecture Overview

The system consists of two main pipelines:

- **Offline Indexing Pipeline:**
  Handles document upload, text extraction, chunking, embedding generation, and indexing. This process can be performed once or incrementally when new documents are added.
- **Offline Query and Answering Pipeline:**
  Processes user queries, retrieves relevant document chunks using FAISS, and generates answers through the offline LLM using an augmented prompt.

All components function locally, ensuring zero dependency on external APIs or internet connectivity

5. Implementation Plan

The solution is implemented using a modular design:

- Frontend: Next.js with Tailwind CSS for the chatbot interface
- Backend: FastAPI for handling requests and processing
- Embeddings: SentenceTransformers (offline)
- Vector Search: FAISS (local)
- LLM: Quantized open-source offline model such as Mistral or Phi

This modular structure simplifies development, testing, and future enhancements.

6. Expected Outcomes

The system is expected to:

- Provide accurate, document-grounded answers
- Ensure complete offline operation
- Protect sensitive organizational data
- Display responses with clear file-wise and page-wise citations

System performance can be evaluated using answer accuracy, retrieval relevance, citation correctness, and response time on local hardware.

7. Conclusion

This project presents a practical and secure approach to knowledge management using offline AI. By combining semantic retrieval with controlled language generation, the proposed system ensures accurate, explainable, and privacy-preserving access to internal documents, making it suitable for enterprise and regulated environments.