

Overview

Your task is to build a **question-answering AI assistant** that can answer queries based on a provided set of documents. The system should use **Retrieval-Augmented Generation (RAG)** so that responses are grounded in the documents instead of relying only on the LLM's knowledge.

What You Need to Do

1. Data Preparation

- Take the provided set of documents (markdown/PDF files with product guides & FAQs).
- Preprocess and chunk them so they're suitable for vector search.

2. Retrieval System

- Store embeddings in a vector database (FAISS / Qdrant / Weaviate — your choice).
- Implement similarity search to retrieve the top-k relevant chunks.

3. Answer Generation

- Use an open-source LLM (e.g., Llama-2, Mistral) or an API-based model (OpenAI/Groq/etc. if you prefer).
- Create a pipeline where:
Query → Retrieve Docs → Construct Prompt → Generate Answer.

4. Response Features

- Each answer must **cite the source** (e.g., doc name or snippet reference).
- For queries not in the docs, return a safe fallback response like:
"I couldn't find this information in the provided documents."

5. Testing

- Create 5 example queries and show the system's responses.
- Document any limitations and ideas for improvement.

Bonus (Optional if you have time)

- Add streaming responses (token-by-token output).
 - Build a simple UI (Streamlit / React).
 - Add basic guardrails (e.g., profanity filter, hallucination check).
-

Deliverables






1. Working Project

- Codebase with clear setup instructions (README).
- The assistant should be runnable locally or via a hosted endpoint.

2. Loom Video

- Record a **5–10 min Loom video** explaining:
 - Your overall approach.
 - How the code works (key components).
 - Demo of the system answering queries.
-

Evaluation Criteria

-  Correctness: Does the assistant retrieve and answer accurately?
-  Code Quality: Is the code clean, modular, and documented?
-  Functionality: Does it handle in-scope & out-of-scope queries properly?
-  Communication: Is the Loom video clear in explaining the thought process and demo?
-  Bonus: Extra features like UI, streaming, or guardrails.

