📄 **VISIONQ CHATBOT PROJECT DOCUMENTATION**

🔍 **About the Project**

The VisionQ Chatbot is an intelligent, context-aware question-answering system designed for VisionQ Technology’s website. It uses modern Retrieval-Augmented Generation (RAG) techniques to provide precise answers based on real company data. The chatbot processes and understands content using advanced AI models. It is accessible via an API built with FastAPI and tested through Swagger UI. This solution allows for easy integration into any frontend or web application.

✨ **Key Features**

 ✅ Accurate answers based on company-specific scraped content

 ⚡ FastAPI-based RESTful chatbot service

 🧠 Uses the **Mistral language model** via Ollama for efficient local inference

 📂 Data preprocessed, cleaned, and embedded with FAISS for semantic retrieval

 🌐 Easily testable through Swagger UI or integrate with any frontend

**🧱 Technologies Used**

* **FastAPI** – for building the REST API
* **LangChain** – for chaining LLM, retrieval, and prompt templates
* **FAISS** – vector store used to search relevant document chunks
* **Ollama** – for local LLM execution
* **Mistral (4.1GB)** – the primary quantized LLM used to generate answers
* **Python 3.10+**

**🧠 How It Works**

1. **Data Source**:  
   The chatbot relies on data scraped from the official VisionQ Technology website.
2. **Cleaning & Preparation**:  
   The raw HTML/text is cleaned using clean.py to remove noise and ensure context clarity.
3. **Chunking & Embedding**:  
   Cleaned text is split into small chunks and embedded using the **OllamaEmbeddings** method with the Mistral model. FAISS is used to store and retrieve similar chunks during queries.
4. **Retrieval & Generation**:  
   When a user asks a question, the system retrieves the most relevant content chunks using FAISS. These are passed along with the query into the **Mistral** LLM using **LangChain's RAG pipeline**.
5. **Answering**:  
   Mistral generates a response grounded in the retrieved context. The response is cleaned to ensure professionalism and relevance before returning to the user.

# **Installation and Setup**

Follow these steps to run the chatbot:

1. Install dependencies:

pip install -r requirements.txt

2. Run the scraper:

python scrape.py

3. Clean the text:

python clean.py

4. Download and install Ollama from : <https://ollama.com/download>

5. Pull the Mistral model:

ollama pull mistral

6. Start the FastAPI server:

uvicorn main:app --reload

7. Visit Swagger UI at: <http://127.0.0.1:8000/docs>

**🧪 Usage Example**

Ask:

*"Where is VisionQ Technology located?"*

Answer:

*"Our company VisionQ Technology is located at Sidharth Apartment, Chinchapada, Kalyan East, Mumbai, Maharashtra - 421306."*