Financial Chatbot — Project Documentation

1. Introduction

The Financial Chatbot is an AI-powered assistant designed to answer questions based on a company's quarterly financial presentations. These reports include structured financial metrics and unstructured data such as CEO commentary. This chatbot integrates Retrieval-Augmented Generation (RAG) to support intelligent querying across both data types.

2. Objective

To create a robust AI system that enables users to interact with financial reports in natural language and extract useful insights using both structured SQL data and unstructured PDF content.

3. Key Features

- Dual-retrieval logic using FAISS (unstructured) and SQLite (structured)

- Auto-detection of trend-related metric questions and rendering of line plots

- Editable metric table with support for adding, renaming, and deleting columns

- Accurate LLM-based answers using GPT-3.5 Turbo via OpenAI API

- Interactive Gradio frontend with Submit & Clear functionality

- Deployed in Hugging Face spaces environment for a seamless deployment.

4. Technologies Used

- Python

- Gradio (for web UI)

- FAISS (vector database for unstructured content)

- SQLite (structured financial metric storage)

- OpenAI GPT-3.5 Turbo (for answer generation)

- Hugging Face Spaces (for deployment)

5. Challenges Faced and Solutions

• Challenge 1: FAISS loading errors in Hugging Face due to folder structure.

→ Solution: Moved all FAISS files to root directory and updated app paths.

• Challenge 2: SQLite threading error during Gradio runtime.

→ Solution: Enabled `check_same_thread=False` and safely managed DB access per request.

• Challenge 3: API key exposure risk in open-source deployment.

→ Solution: Used `os.environ['OPENAI_API_KEY']` sourced from Hugging Face Secrets.

6. Deployment

Users can simply visit the chatbot's deployed interface on Hugging Face and interact with it by typing natural language financial questions related to Tracxn's quarterly performance. The chatbot intelligently retrieves both structured metrics and unstructured commentary to respond.

1. The URL for the deployed chatbot is available below:

   https://huggingface.co/spaces/TGChandu/RAG_LLM_Financial_QA_Chatbot

   Copy and paste the above URL into your browser. You will be directed to the deployed chatbot interface.

2. Once on the page, you can enter your question in the input box.
   - You can ask about revenue, EBITDA, net profit, margins, or even CEO commentary for any specific quarter.
   - The chatbot can answer metric-specific queries, retrieve commentary, and generate insights across 10 quarterly reports.
3. After entering your question, click the **Submit** button.
   - For trend-related queries (e.g., "trend in EBITDA over the last 4 quarters"), the chatbot will also return a line chart visualizing the trend.
   - For other types of questions, it will return a text-based answer.
4. If the user can't see the app or it appears inactive, they should **restart the space** on Hugging Face by clicking the "⋮" menu and selecting **"Restart Space"**.

7. Conclusion

This project successfully demonstrates the integration of a full-fledged Retrieval-Augmented Generation (RAG) pipeline capable of answering intelligent queries using both structured and unstructured data from quarterly financial presentations. By combining visual data extraction, a customizable metric editor, a vector-based semantic search system, and a conversational interface powered by GPT-3.5 Turbo, the chatbot delivers reliable, context-aware responses to a wide variety of business questions. The inclusion of trend visualization for metric-related queries adds further value by helping users interpret patterns over time. The system is designed with flexibility, scalability, and user experience in mind — from local testing in Colab to final deployment on Hugging Face Spaces. Overcoming critical challenges like image-to-text extraction, SQLite integration, and dual data retrieval showcases the robustness of the solution. This project not only fulfills all the assignment objectives but also reflects a production-ready design suitable for real-world financial analytics applications.