I experimented with three different methods for implementing RAG on financial data:

GPT-4 with Pinecone: The accuracy in this approach was not satisfactory, possibly due to issues with chunking.

Ollama: I tested Ollama for RAG, and it provided better results than GPT-4 with Pinecone. However, it required significant computational resources, which my current setup couldn't handle efficiently. Initially, I used pdfplumber to extract tables from PDFs, but the results were not as expected. Instead, I converted the PDFs into DOCX files and then extracted the tables, which yielded better results.

GPT-4 with Chroma: This approach gave the best results among all three while also being more resource-efficient. However, some further fine-tuning is still needed.

While writing the Streamlit script, I encountered an issue related to ChromaDB, where the error indicated that the "collections" table was missing in SQLite. This caused failures when querying the database. After researching the issue, I found that it might be related to ChromaDB’s internal system database and was able to resolve it through a Google search.

while running docker use http://localhost:8501/