optimization needed: use azure functions for scaling implement caching to reduce API calls optimize data

deploying the RAG web Application on azure and OpenAI API costs around $332.50—

latency measurement

For latency mesurement I have created a function which will measure retrieval latency and generation late
path: `backend/logs/latency_data.csv` -> backend > logs > latency_data.csv

Technologies Used

- **FastAPI** → Backend framework
- **Streamlit** → UI for search queries
- **FAISS** → Vector search for job retrieval
- **Gemini** → LLM for career advice
- **Huggingface** → Embeddings Model - "sentence-transformers/all-MiniLM-L6-v2"
- **LangChain** → To access embedding model
- **OpenAI GPT** → Evaluate
- **RAGAS** → Evaluate the performance of the RAG system

[ ]:

[ ]: