Steps to RUN back-end API

venv\Scripts\activate

cd .\backend\

uvicorn main:app --reload --host 0.0.0.0 --port 8000

Steps to RUN UI (front end app)

myenv\Scripts\activate

cd .\frontend\

streamlit run .\app.py

RAG pipeline will work like this:

1. Parse JSON → Extract each job role as a single chunk.
2. Embed the full job profile using a vector model.
3. Store in Vector DB for retrieval.
4. Query retrieval using similarity search.

* Each job role is already a self-contained chunk.
* The text fits within embedding limits (most models handle 512–1024 tokens well).

Since I am using one job role as a single chunk, I do not need a text splitter in this case.

By treating one job role as a single chunk, you ensure:  
- Better context preservation – The entire job description remains intact.  
- Efficient retrieval – Queries will match complete job roles rather than fragmented text.  
- Faster processing – No need for extra chunking logic.

|  |  |
| --- | --- |
| **FAISS** | Fast in-memory search, best for local retrieval. |

|  |  |
| --- | --- |
| **FAISS** | Fast in-memory search, best for local retrieval. |