RAG Pipeline Prototype

7th January 2026

Ruba Saad

1. **Objective:**
   To build a RAG system that is capable of retrieving factual data from historical documents.

2. **Overview:**
   Source Data: Nuclear History of 5 Countries (USA, UK, Russia, China and Pakistan).
   Framework: LangChain
   Vector Database: ChromaDB
   Embedding Model: Hugging Face - 'sentence-transformers/all-MiniLM-L6-v2' (Local & Open Source)
   LLM Model: Ollama - Llama3
   Execution Environment: Fully local deployment to ensure 100% data privacy and zero API costs.

3. **Design & Optimization:**

Feature	Selection	Reason
Splitting Strategy	Recursive Character Text Splitter	Avoids cutting sentences in the middle. Allows multiple separators. Splits in paragraphs/ lines to maintain the similarity coherence
Chunk Size	800 characters	Large enough to hold the facts and small enough to fit the LLM's context window
Search Metric	Cosine Similarity	Matches by meaning (semantic) rather than exact keywords
Retrieval (value of k)	k=3	Maximizes precision over noise. The small value of k keeps the system fast and reduces hallucination risks. During testing k values of 4 and 5 created noise which led to incorrect answers.

4. **Testing & Evaluation Results:**
   The retrieval system was tested with 8 synthetic questions to make sure the ChromaDB vector store correctly identifies and retrieves the relevant text chunks.
   - ➔ Test Query: "Besides the US and Russia, which country was the first to test a nuclear bomb?"
   - ➔ Result: Success. The retriever correctly identified the relevant passage in Document 2: "Britain became the third country to develop and test nuclear weapons."

   Overall System Accuracy: The LLM model achieved a success rate of 87.5% (7 out of 8 questions correctly answered).
   Successes: The LLM model provided precise and factual responses based on the timelines mentioned within the documents for all five countries.
   Failures: One query regarding the location of the research and production facilities of the nuclear weapons in the UK resulted in failure.
   - ➔ Observation: The LLM responded with "I do not have enough information based on the documents provided."
   - ➔ Reason of Failure: Retrieval Gap. The information existed in the source document, but the value of k set at 3 was too low to capture the specific chunk containing the location details.