Retrieval-Augmented Generation

Technical Interview

Eng. Miguel Salazar

July 2025

Agenda

1. Problem & Goal Statement
2. Solution Overview
3. High-Level Architecture Diagram
4. Component Breakdown
5. Implementation Strategy
6. Technical Challenges & Trade-offs
7. Demo Instructions & GitHub
8. Next Steps & Production Considerations

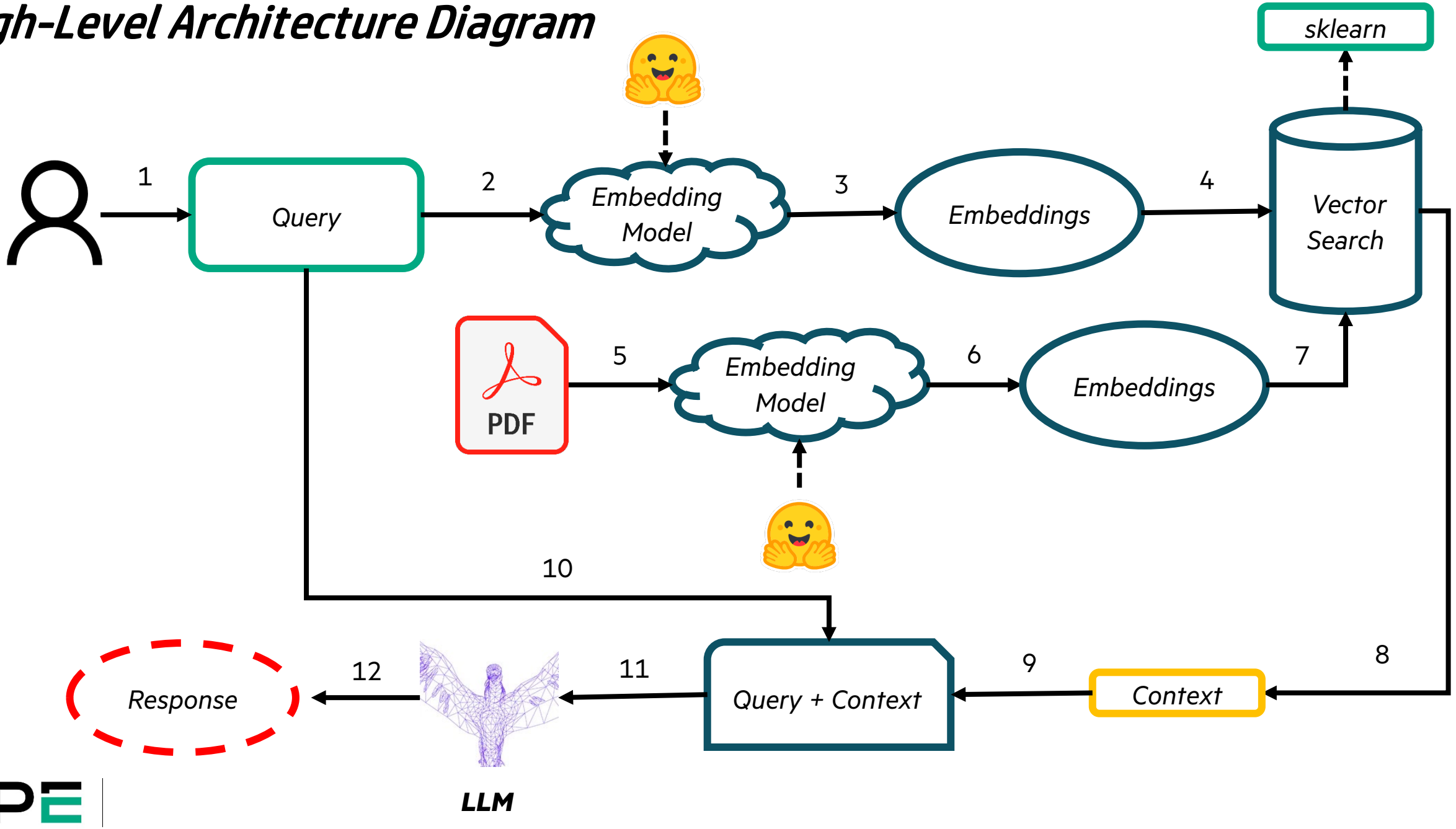
Problem & Goal Statement

- **Objective:** Build a RAG system using a local LLM and custom embedding-based retrieval.
- **Goal:** Answer user queries grounded on knowledge extracted from provided PDFs.
- **Constraints:**
  - Must run locally (no cloud APIs).
  - Efficient on CPU (low-resource hardware).
  - Deliver notebook + architecture PDF + GitHub link.

Solution Overview

- Local semantic search using SentenceTransformers (MiniLM).
- Local response generation using Falcon-RW-1B.
- Query flow:
    - Ingest PDF files.
    - Chunk and embed text.
    - Retrieve relevant chunks using cosine similarity.
    - Generate response via local LLM.

High-Level Architecture Diagram



Component Breakdown

1. PDF Loader:
Uses PyMuPDF to extract text.

2. Text Chunker:
Splits documents into ~200 token chunks with sentence-awareness.

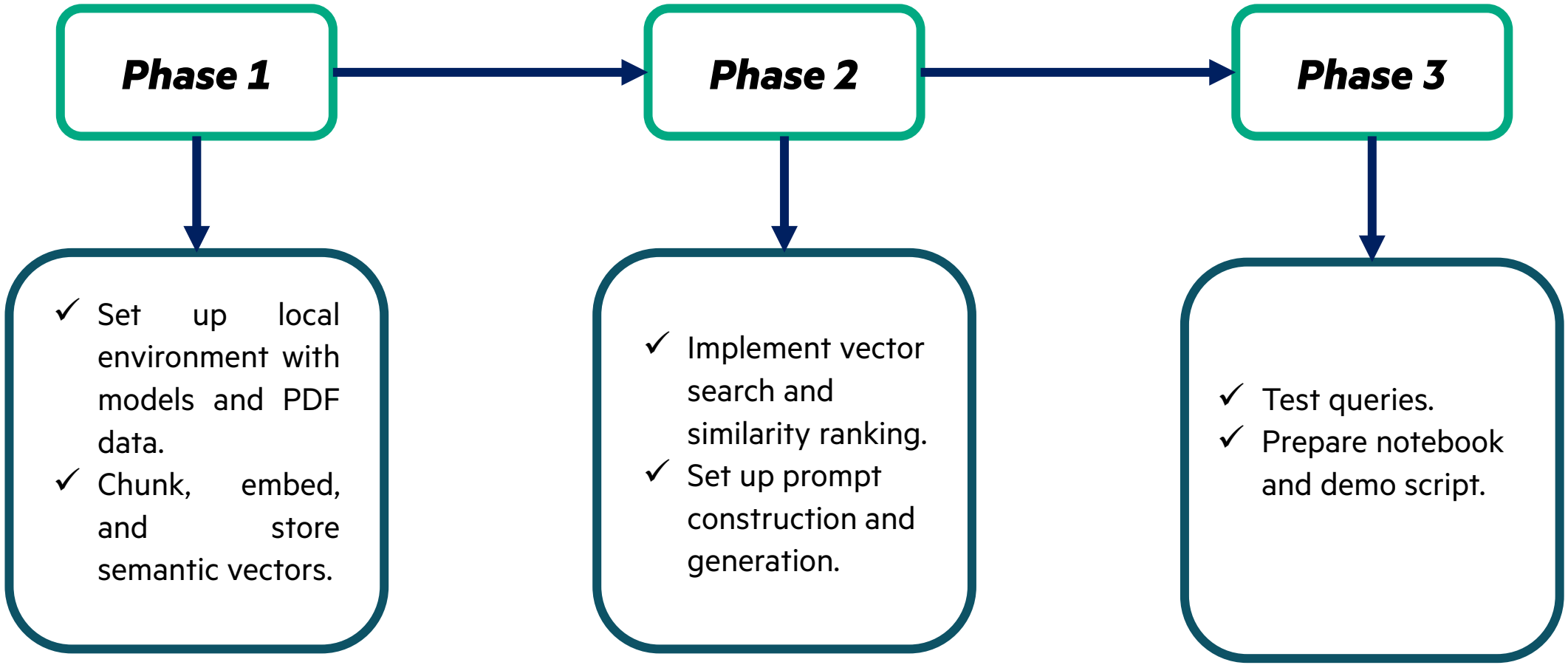
3. Embedding Generator:
SentenceTransformer (MiniLM-L6-v2) locally hosted.

4. Vector Search:
Cosine similarity using sklearn between question and stored chunks.

5. Prompt Constructor:
Concatenates top-k retrieved chunks with the user query.

6. Generator:
transformers.pipeline() with Falcon-RW-1B.

Implementation Plan



Technical Challenges & Trade-Offs

Area	Challenge	Resolution
Model Size	Limited RAM/CPU	Used MiniLM + Falcon-RW-1B
Chunk Quality	Overlapping/redundant chunks	Refined token-based chunking
Output Length	Truncated/incomplete answers	Tweaked max_new_tokens, prompt
No GPU	Inference speed	Used efficient models + batching

Demo & Repo Instructions

- ✓ **Jupyter Notebook:** notebooks/rag_demo.ipynb
- ✓ Models are downloaded and used via local paths.
- ✓ **Repo:** https://github.com/miguelsa12/rag-project.git
- ✓ **To test:**

pip install -r requirements.txt
python rag_pipeline.py
jupyter notebook notebooks/rag_demo.ipynb

- ✓ **Dataset:** Inside /data folder, 3 HPE-related PDFs

HPE

Round Table

- Anything you would like to add?
- Any feedback?

Thank you!
Let's Build What's Next — Together

miguel-andres.salazar@hpe.com
+1 833 698 1869 Ext. 2031163