SHL Assessment Recommender – System Overview

Data Collection & Representation

- **Crawling**:
  I scraped SHL's product catalog using **BeautifulSoup** to extract assessment metadata such as title, description, tags, job level, remote/adaptive support, and duration.
    - I also crawled individual assessment detail pages to capture richer content like **job levels** and **test traits**.

- **Representation**
    - After crawling, I enriched each assessment with **semantic tags** using the **Gemini LLM**. Tags include hard skills (e.g., Python), soft skills (e.g., communication), and test traits (e.g., scenario-based).
- **Storage**:
    - All enriched assessments were stored in a `CSV`, and vector embeddings (using `all-MiniLM-L6-v2`) were stored in **Pinecone**, with full metadata attached for reranking.

Retrieval & Reranking Pipeline

- **Query Input**:
    - Users can enter a **natural language query**, paste a **job description**, or provide a **JD URL**.
    - A preprocessing module intelligently classifies input and uses **Gemini** to extract the core search intent from JDs.
- **Vector Retrieval**:
    - Top 50 matches are fetched using **semantic vector similarity** from Pinecone.
- **LLM-based Reranking**:
    - A custom prompt is sent to **Gemini** to rerank the top candidates based on semantic alignment with:
        - Technical + soft skills
        - Duration constraints
        - Job levels and **test types (mapped)**

Stack: Emerging LLM Tools Used

Tool	Role
Gemini (Google Generative AI)	Query understanding, tag generation, reranking
Pinecone	Vector storage and similarity search
LangChain (partial)	Vector doc wrappers (not used fully)
SentenceTransformers	Embedding generation
FastAPI	REST API (health & `/recommend`)
Gradio	Frontend UI (text input + CSV download)

Evaluation & Tracing

- Evaluated on **3 real hiring queries** using both:
  - LLM-based reranking
  - Hybrid scoring (vector + metadata)
- Manually validated top matches using:
  - Duration alignment
  - Skill/tag overlap
  - Real-world test naming relevance
- Prompt engineering, Gemini response cleaning, and ID mapping were logged to trace each stage.

FLOW DIAGRAM: