SHL GenAI Assessment Recommendation System
Technical Approach & Evaluation

Author: Shashwat Pandey
Role Targeted: AI / GenAI Research Engineer Intern

1. Problem Understanding

Hiring managers frequently struggle to identify appropriate assessments using keyword-based filters. Such approaches fail to capture nuanced role requirements such as the balance between technical proficiency and behavioral competencies. The objective of this project is to design a GenAI-powered recommendation system that maps natural language hiring queries or job descriptions to the most relevant SHL Individual Test Solutions in an accurate, explainable, and scalable manner.

2. Data Ingestion & Catalog Preparation

The SHL product catalog was programmatically scraped from SHL's public product catalog. Only Individual Test Solutions were retained, while pre-packaged job solutions were explicitly excluded. Each assessment was parsed into a structured schema containing assessment name, URL, description, test type (Knowledge/Skills or Personality/Behavior), duration, and delivery attributes. This structured representation enables reliable retrieval and downstream reasoning.

3. Retrieval-Augmented Recommendation Architecture

The recommendation engine follows a Retrieval-Augmented Generation (RAG) paradigm. Assessment descriptions are embedded using the sentence-transformers MiniLM model (all-MiniLM-L6-v2), selected for its strong semantic performance and efficiency in low-latency settings. User queries are embedded using the same model to ensure alignment in the embedding space. Cosine similarity is used to retrieve the top-N candidate assessments.
A lightweight re-ranking layer is applied to enforce recommendation balance when queries span multiple competency dimensions. For example, queries referencing both technical skills and collaboration requirements result in a balanced mix of Knowledge/Skills (K) and Personality/Behavior (P) assessments.

4. Evaluation Methodology

System performance was evaluated using Mean Recall@10, the primary metric specified by SHL. A labeled training dataset consisting of hiring queries and human-annotated relevant assessments was used to measure retrieval effectiveness. Recall@10 was computed per query and averaged across all queries to obtain Mean Recall@10.

Evaluation was applied at both the retrieval stage and the final recommendation stage. Iterative experimentation was performed by varying retrieval depth, adjusting re-ranking logic, and refining query interpretation. These iterations led to measurable improvements in recall while maintaining recommendation diversity and relevance.

5. API & Deployment

The system is exposed via a RESTful API implemented using FastAPI. A health check endpoint verifies service availability, while a recommendation endpoint accepts hiring queries and returns structured assessment recommendations in JSON format. A Streamlit-based web interface was built on top of the API to allow recruiters to interactively test the system.

6. Design Choices & Trade-offs

A compact embedding model was chosen to balance semantic accuracy with inference efficiency, making the system suitable for real-time usage. Explicit keyword matching was avoided in favor of semantic retrieval to improve generalization across diverse query formulations. The modular architecture allows future extension to more advanced LLM-based re-ranking or hybrid retrieval approaches.

7. Conclusion & Future Work

This project demonstrates a complete GenAI-driven recommendation pipeline grounded in sound engineering principles and measurable evaluation. Future enhancements could include hybrid lexical-semantic retrieval, richer LLM-based reasoning for complex queries, and continuous learning from recruiter feedback.