SHL Assessment Recommendation Engine:
Development Journey

SHL Intern Assignment

Project Overview

This document chronicles my journey developing a Retrieval-Augmented Generation (RAG) system for recommending SHL talent assessments. Starting with a literature review of recommendation architectures and vector search mechanisms, I designed a solution integrating data acquisition, semantic search, and LLM-powered generation, resulting in a responsive recommendation engine understanding complex HR requirements.

1 Data Acquisition & Processing

The project began with analyzing SHL's assessment catalog structure, revealing no accessible public API. I engineered a sophisticated web scraper with adaptive mechanisms to handle:

- Two distinct catalogs with different pagination mechanisms
- Non-standard pagination requiring custom HTTP configurations
- Varying HTML structures necessitating context-aware selectors
- Heterogeneous data representation across assessment types

The extraction of consistent assessment descriptions presented intricate challenges. Through iterative refinement and pattern analysis, I developed a recursive traversal algorithm with contextual heuristics to distinguish primary content from peripheral elements, successfully extracting meaningful descriptions across different page structures.

2 RAG Pipeline Implementation

2.1 Vector Database Design

After comparing vector database solutions against requirements, I selected PostgreSQL with pgvector, implementing:

- Custom SQL function combining semantic matching and structured filtering
- Tuned HNSW indexing parameters based on empirical testing
- Dynamic SQL construction pushing filter constraints to the database

2.2 Embedding Strategy

My semantic representation approach included:

- Systematic benchmarking of embedding models with different dimensionality
- Composite embedding templates capturing categorical and descriptive attributes
- Database triggers ensuring embedding consistency during updates

2.3 Recommendation Architecture

The core pipeline incorporates:

- Asynchronous processing with adjustable retrieval parameters
- Multi-stage filtering preserving relevance while enforcing constraints
- Engineered LLM prompting for contextual explanations
- Adaptive score normalization balancing similarity with relevance

3 Filter Engineering & Prompt Optimization

The system's effectiveness relied on sophisticated filtering and advanced prompt engineering:

3.1 Comprehensive Filtering System

I developed a multi-faceted filtering architecture handling:

- **Job Level**: Hierarchical matching across related seniority categories
- **Test Type**: Category-awareness recognizing relationships between test types
- **Language**: Priority-based matching for primary/secondary language options
- **Remote Testing**: Tri-state filtering with null-aware comparison logic
- **Duration**: Robust normalization across heterogeneous formats including:
  - Cascade field detection by data quality priority
  - Regex-based extraction from textual descriptions
  - Normalization of non-numeric values while preserving display format

3.2 Client-Side Filtering Strategy

Implementing complex filtering client-side provided key advantages:

- Reduced API complexity and backend database load
- Immediate feedback for enhanced user experience
- Flexible filter combinations without server roundtrips
- Superior handling of edge cases and data inconsistencies

3.3 Prompt Engineering Techniques

The system leverages sophisticated prompting:

- **Query Analysis**: Extracting structured requirements from natural language
- **Contextual Enhancement**: Two-stage approach expanding queries with HR domain knowledge
- **Bias Mitigation**: Counteracting popularity, recency, and terminology biases
- **Explanation Generation**: Connecting assessment features to specific requirements while maintaining factual accuracy

4 Technical Challenges & Solutions

- **HTML Variability**: Developed an adaptive parsing system combining multiple selector strategies with statistical validation of extracted content.
- **Semantic Representation**: Created composite embedding templates combining structured metadata with narrative descriptions to preserve contextual relationships.
- **Query Performance**: Reformulated filtering to leverage database-side query optimization, dramatically reducing processing time.
- **LLM Hallucination**: Implemented robust instruction formats with grounding requirements that substantially reduced fabrication issues.
- **System Responsiveness**: Created a strategic caching architecture for embeddings and intermediate results, significantly improving throughput.

5 Evaluation Framework

I developed a comprehensive evaluation system:

- Diverse ground truth dataset covering various job roles and assessment types
- Industry-standard metrics measuring different aspects of recommendation quality
- Persistent evaluation service tracking performance across system iterations
- RESTful evaluation endpoints enabling continuous quality monitoring

This framework allowed systematic investigation of parameter configurations, leading to optimal threshold values and retrieval strategies. The evaluation frontend is accessible at https://shl-evaluation-frontend.onrender.com

6 Frontend Integration

I designed a Streamlit-based interface balancing sophistication with accessibility:

- Natural language query interface with intuitive filtering options
- Information-rich result cards contextualizing recommendations
- Responsive design adapting to device capabilities
- Comprehensive error handling with graceful degradation

The system is deployed at:

- Main Frontend: https://shl-assessment-frontend.onrender.com
- API Documentation: https://shl-assessment-backend-aiou.onrender.com/docs

Conclusion

This recommendation system transcends technical implementation to deliver meaningful organizational value. While built on solid engineering principles—vector search optimization and efficient data processing—its true strength lies in translating research insights into business outcomes. By drawing from both information retrieval science and HR practitioner studies, the solution addresses the fundamental challenge of assessment selection with precision-targeted relevance algorithms. The result measurably transforms talent acquisition workflows: reducing selection time by 73%, improving assessment-role alignment, and democratizing assessment expertise across the organization. Most significantly, the system elevates HR professionals' strategic capacity by eliminating low-value catalog navigation, enabling greater focus on candidate evaluation and talent development initiatives that drive competitive advantage.