Project Report: Fashion Search AI - AI-driven Generative Search System

Project Overview

The Fashion Search AI project implements a Retrieval-Augmented Generation (RAG)-based intelligent search system for
fashion products using the Myntra Fashion Dataset (https://www.kaggle.com/datasets/djagatiya/myntra-fashion-product-dataset).
It allows natural-language queries like "Find me a women's cotton midi dress under Rs 2000" and returns semantically relevant
matches with a concise generative summary.

Objectives

- Build a 3-layer AI search system (Embedding -> Search -> Generation)
- Implement cache, re-ranking, and LLM-based generation
- Evaluate using 3 self-designed queries and 6 screenshots
- Document design decisions, challenges, and insights

System Design

Architecture: User Query -> Embedding Layer -> ChromaDB Vector Store -> Search Layer -> Generation Layer -> Final Answer

Technologies: Python, Sentence-Transformers, Cross-Encoder, ChromaDB, OpenAI GPT-4o-mini, SQLite Cache, FastAPI (optional UI)

Implementation

1. Embedding Layer: Preprocessed dataset, created text chunks, embedded using 'all-MiniLM-L6-v2'.
2. Search Layer: Retrieved top-20 documents, re-ranked top-3 with 'ms-marco-MiniLM-L-6-v2', implemented SQLite cache.
3. Generation Layer: Prompt-based LLM summarization or fallback extractive method.

Queries Tested

1. women summer cotton midi dress under 2000 rupees
2. men running shoes with breathable mesh black color

3. kids winter hoodie warm fleece for boys

Challenges

- Dataset inconsistencies
- Balancing speed and accuracy
- Re-ranking latency
- Handling API dependency
- Memory optimization for large datasets

Lessons Learned

- Embedding choice impacts accuracy.
- Cross-Encoders enhance retrieval quality.
- Caching improves runtime performance.
- Prompt design is key for coherent generation.
- Modular pipeline enables experimentation.

Future Enhancements

- Add feedback-driven re-ranking.
- Integrate price filtering and multimodal embeddings.
- Deploy FastAPI-based UI for live interaction.

Conclusion

The Fashion Search AI project demonstrates how LLMs and embeddings can power intelligent, context-aware fashion search.

It bridges retrieval and reasoning to deliver relevant, human-like recommendations.