**Project Scaffold: llm-fund-matcher (Updated Hybrid BM25 + ANN Version)**

**✅ Updated Project Goal**

Build a **production-ready, real-time, accurate fuzzy search system** for Indian mutual funds using a **hybrid search pipeline**:

* **BM25** lexical filtering
* **ANN-based dense semantic retrieval** (HNSW)
* **Metadata-aware reranking** All done using **small, efficient models**, modular Python tools, and no external infrastructure.

**🚀 Hybrid Search Architecture Flow**

flowchart TD

A[Query Input: "SBI infra"] --> B[Preprocessing: Lowercase, Tokenize, Typo Correction]

B --> C[BM25 Search (Top-K Filter)]

C --> D[Sentence Encoding (MiniLM/BGE)]

D --> E[ANN Search via HNSW over Top-K]

E --> F[Score Fusion: BM25 + Cosine Similarity + Metadata Boost]

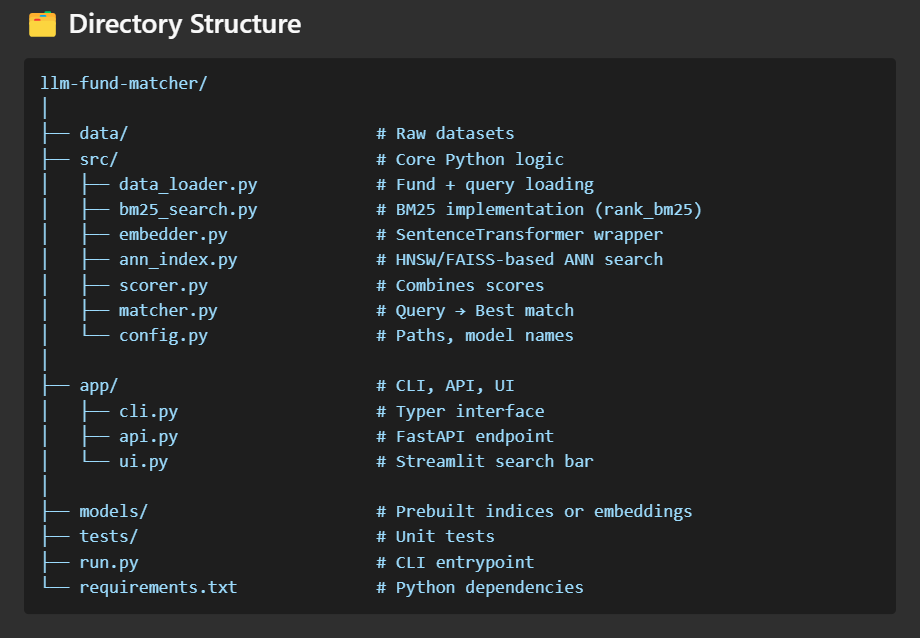
F --> G[Select Top-1 Best Match]

G --> H[Return Fund Name, Match Score, Explanation (Optional)]

**🧰 Refined Tech Stack**

| **Layer** | **Tool/Library** |
| --- | --- |
| **Embedding** | sentence-transformers (MiniLM, BGE) |
| **Lexical Search** | rank\_bm25 (pure-Python BM25) |
| **Vector Search** | hnswlib or faiss-cpu |
| **Preprocessing** | fuzzywuzzy, python-Levenshtein, re |
| **API/UI** | FastAPI, Streamlit, Typer |
| **Metadata Scoring** | Custom scoring in Python |

**🗂 Directory Structure**



**📦 requirements.txt (Updated)**

* pandas
* numpy
* sentence-transformers
* hnswlib
* rank\_bm25
* typer
* streamlit
* fastapi
* uvicorn
* loguru
* python-dotenv
* fuzzywuzzy
* python-Levenshtein

**✅ Summary of Changes from Original Plan**

* ❌ Removed sole reliance on FAISS dense search
* ✅ Added BM25 filtering step using rank\_bm25
* ✅ Integrated HNSW or FAISS ANN search **after filtering**
* ✅ Added a score fusion step: BM25 + Cosine + Metadata
* ✅ Kept setup fully offline, zero cost, no Elasticsearch/Postgres

**💡 Benefits of This Hybrid Approach**

* **Robust to typos, short/incomplete queries**
* **Handles semantic intent** (e.g., "infra" → "infrastructure")
* **Blazing fast** even on 10K+ funds
* **Simple and open-source**: install via pip, no billing, no containers
* **Modular**: swap components later (e.g., FAISS ↔ HNSW, BGE ↔ TinyLlama)