

Shahzaib Mahar

shahzaibsm11@gmail.com | 0370-1224702

linkedin.com/in/shahzaib-mahar

github.com/smaharx

AI/ML Engineer | Full-Stack Software Developer | ML Engineer at FlyRank

PROFESSIONAL SUMMARY

Engineer specialising in AI/ML systems and backend development, with a track record of architecting and shipping production-grade intelligent applications. Deep expertise in RAG pipelines, LLM orchestration, vector databases (FAISS), and the Python ecosystem (FastAPI, Streamlit). Portfolio includes a local AI engineering copilot with AST-based semantic chunking and re-ranking, an ML-powered personal finance platform with anomaly detection and time-series forecasting, and a semantic movie recommendation engine. Currently building ML pipelines at FlyRank in a live production environment.

INTERNSHIP EXPERIENCE

Machine Learning Engineer Intern

July 2025 - Present

FlyRank

- Building and deploying ML pipelines for AI-driven content ranking and SEO optimisation in a live production environment using Python and FastAPI.
- Applying NLP techniques, vector search, and microservices architecture to develop scalable, inference-ready backend services.

TECHNICAL SKILLS

Languages	Python Java C++ SQL
Backend & APIs	FastAPI Uvicorn RESTful APIs Microservices Docker
AI / ML	RAG (Retrieval-Augmented Generation) LLM Orchestration HuggingFace Transformers SentenceTransformers FAISS Scikit-learn Prophet Isolation Forest NLP AST Parsing Vector Embeddings
Databases	PostgreSQL MySQL SQLite MongoDB (NoSQL)
Data & Tools	Pandas NumPy Plotly Streamlit Typer Rich JDBC Git / GitHub
CS Fundamentals	OOP DSA System Design Async Programming Layered Architecture

PROJECTS

Nobody Copilot v1.1.0 github.com/smaharx/Nobody

Tech: Python | FastAPI | FAISS | SentenceTransformers | RAG | AST | SQLite

- Architected a hyper-personalized local AI engineering copilot on a 3-tier decoupled microservices model: Typer/Rich CLI, FastAPI/Uvicorn kernel, and a FAISS vector store with SQLite stateful memory — designed for production-level scalability.
- Built an AST-based crawler using Python's ast module that parses source files into discrete, logically complete constructs (functions, classes, methods) rather than arbitrary line-break chunks, preserving full structural context for the LLM window.
- Implemented a full RAG pipeline: SentenceTransformers encode code chunks into 384-dimensional embeddings; FAISS executes L2/cosine similarity search across thousands of vectors in milliseconds to surface the most relevant code blocks.
- Engineered a secondary Re-ranking Engine that filters initial FAISS results for contextual relevance, eliminating semantically-close but contextually-irrelevant noise before passing context to the LLM.
- Designed session-aware SQLite memory for multi-turn interactions and zero-inference fault tolerance for graceful API downtime handling.

Finova | AI-Powered Personal Finance Tracker github.com/smaharx/Finova

Tech: Python | FastAPI | Streamlit | PostgreSQL | Scikit-learn | Prophet | Docker

- Built a full-stack personal finance platform featuring an interactive Streamlit dashboard with real-time monthly metric filtering and a Dockerised FastAPI backend.
- Integrated an NLP pipeline for automatic transaction categorisation with a human-in-the-loop correction UI, closing the model feedback loop for continuous improvement.
- Deployed Isolation Forest for unsupervised real-time anomaly detection on spending patterns and Facebook Prophet for 30-day time-series budget forecasting.
- Authored a synthetic data seeder injecting 20,000+ realistic transactions to stress-test both ML models at scale.

Cinema-Sense | Semantic Movie Recommendation Engine v2.0.0 github.com/smaharx/cinema-sense

Tech: Python | HuggingFace (all-MiniLM-L6-v2) | FAISS | Streamlit | TMDb API

- Designed a semantic search engine encoding 5,000+ movie metadata records into a 384-dimensional vector space via HuggingFace transformer, enabling true NLP understanding beyond keyword matching.
- Implemented smart over-fetching with FAISS nearest-neighbour retrieval and deterministic Pandas filtering to prevent result starvation on strict queries.
- Built async parallel thread-pooling for TMDb API poster calls; @st.cache_resource loads the vector index into RAM once for instant re-queries.

Disaster Management System github.com/smaharx/disaster-management-system

Tech: Java | Java Swing | MySQL | JDBC

- Developed a Java desktop application for disaster relief organisations covering family registration, inventory tracking, restocking management, and operational reports.
- Applied a clean 4-layer architecture (Swing UI, Business Logic, JDBC DAL, MySQL) with parameterised queries to enforce SQL injection prevention.

EDUCATION

Bachelor of Science in Computer Science

2022 - Expected 2026

Sukkur IBA University, Sukkur, Pakistan

CGPA: **3.02 / 4.00**

ADDITIONAL

Open Source	All personal projects publicly hosted on GitHub under the MIT License — github.com/smaharx
Languages	Urdu (Native) English (Professional Working Proficiency)
Interests	AI-native developer tooling Vector search systems LLM productionisation Backend architecture

References available on request