Vivek Sharma

8828482583 | vivektusharma@gmail.com | linkedin.com/in/vivek-sharma-64951b24b | github.com/Vivek02Sharma

Education

MCC, University of Mumbai — Mumbai, India
BSc in Information Technology; CGPA: 7.37 — *2022 – 2025*

GVM College — Mumbai, India
High School (HSC Science); 65% — *2020 – 2022*

Projects

MedVec-Scratch | Python, PyTorch, BPE Tokenizer, HuggingFace, Transformers [Project Link]

- Engineered medical sentence embedding model from scratch with Siamese Transformer architecture (4 encoder blocks, 256-dim embeddings, 4-head multi-head attention, 1024-dim FFN, mean pooling) using custom BPE tokenizer (30k vocab, 128 max seq length), trained on 233k medical triplets (anchor-positive-negative) with Triplet Margin Loss for semantic similarity learning, enabling clinical document retrieval, medical literature search, and symptom-diagnosis matching in healthcare applications.

Federal Registry RAG Agentic System | Python, Groq LLM, MySQL, Streamlit, Asyncio [Project Link]

- Developed full-stack RAG system for US Federal Registry queries with Groq LLM, async pipelines, MySQL backend, and agentic tool calling architecture for real-time document retrieval.

Student Performance Analysis and Prediction | Python, Scikit-learn, XGBoost, Plotly [Project Link]

- Built Streamlit dashboard for academic performance prediction using XGBoost with role-based access, trend analysis, and MongoDB backend for SGPA forecasting.

Experience

Anvex AI Technologies Private Limited — September 2025 – Present
Jr. Software Engineer Machine Learning — *Mumbai, India*

- Built RAG system for AnvexSpeak, a telephony AI platform using Twilio/Plivo with multi-source data extraction (client databases, Google Drive, web scraping), BAAI/bge-large-en-v1.5 embeddings, and advanced filtering.
- Created custom LLM wrapper supporting open-source (Gemma3, Qwen, Mistral, Llama) and closed-source models (OpenAI) using LangChain/LangGraph; deployed with Qdrant (production) and ChromaDB (development) for real-time document-grounded responses.

Anvex AI Technologies Private Limited — June 2025 – August 2025
Machine Learning Intern — *Mumbai, India*

- Developed AVA, a RAG-powered policy chatbot using LangChain/LangGraph for query routing, Qdrant vector store, and SambaNova Llama-4 across 26+ policies with FastAPI endpoints achieving 800-1500ms response time and confidence scoring.

Techathon — February 2025
Mulund College of Commerce — *Mumbai, India*

- Built and trained machine learning models to predict equipment failures using sensor data, historical maintenance records, feature engineering, hyperparameter tuning, and integrated into backend using FastAPI for real-time predictive maintenance.

Open Source Contribution

Universal-Box — 2024
Contributor (Data Science Template)

- Developed and added Data Science template with starter configurations for Jupyter, pandas, scikit-learn, structured workflows, and example notebooks, collaborating with maintainers to align with best practices.

Technical Skills

Languages: Python, SQL
Frameworks: PyTorch, TensorFlow, Flask, FastAPI, LangChain, LangGraph
Developer Tools: Docker, Linux, Git, GitHub, VS Code, Jupyter Lab, PyCharm
Libraries: Scikit-learn, pandas, NumPy, Matplotlib, Seaborn, Plotly
Vector Databases: Qdrant, ChromaDB