

Vivek Sharma

+91 8828482583 | vivektusharma@gmail.com | linkedin.com/in/vivek-sharma-64951b24b | github.com/Vivek02Sharma

EDUCATION

MCC, University of Mumbai

BSc in Information Technology — CGPA: 7.37

Mumbai, India

2022 – 2025

GVM College

High School (HSC Science) — 65%

Mumbai, India

2020 – 2022

EXPERIENCE

Anvex AI Technologies Private Limited

Jr. Software Engineer Machine Learning

September 2025 – May 2026

Mumbai, India

- Built RAG system for AnvexSpeak, an AI Voice Agent platform with multi-source data extraction (client databases, Google Drive, web scraping), BAAI/bge-large-en-v1.5 embeddings, and advanced filtering with client data separation.
- Developed a unified LLM inference layer using vLLM to support open-source models (Gemma 3, Qwen, Mistral, Llama) alongside OpenAI models, with Qdrant (production) and ChromaDB (development) for real-time document-grounded responses.

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 and confidence scoring.

Techathon

Mulund College of Commerce

February 2025

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

Contributor (Data Science Template)

2024

GitHub Repository

- 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.

PROJECTS

MedVec-Scratch | Python, PyTorch, BPE Tokenizer, HuggingFace, Transformers

[Project Link]

- Built custom medical sentence embedding model from scratch using Siamese Transformer (4 layers, 256 dim) with custom BPE tokenizer (30k vocab), trained on 233k medical triplets using Triplet Margin Loss for healthcare semantic search.

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.

MLOps Pipeline with MLflow | Python, Scikit-learn, XGBoost, MLflow, FastAPI, Streamlit

[Project Link]

- Built an end-to-end MLOps pipeline for medical cost prediction, comparing 4 regression models with automated best-model selection via R2 score, MLflow experiment tracking, model registry, FastAPI serving, and Streamlit frontend.

TECHNICAL SKILLS

Languages: Python

Frameworks: TensorFlow, PyTorch, FastAPI, Flask, LangChain, LangGraph

Tools: Docker, Linux, Git, GitHub, MLflow, VS Code, Jupyter Lab, PyCharm

Libraries: Transformers, vLLM, Scikit-learn, pandas, NumPy, Matplotlib, Seaborn, Plotly

Vector Databases: Qdrant, ChromaDB