Ganesh Sharma

Mumbai, India | ganesh27sharma09@gmail.com | +91 90760 99303
linkedin.com/in/ganeshsharmaa | github.com/GaneshSharmaa | ganeshsharma.tech

Summary

Python backend and AI engineer with hands-on experience building LLM deployments, RAG pipelines, and production APIs. Expertise in FastAPI, PostgreSQL, Docker, and LLM inference optimization. Built locally-hosted systems that combine AI capabilities with robust backend infrastructure.

Technical Skills

Languages: Python, C++

AI Engineering: LLM, Prompt Engineering, Retrieval Strategies, RAG, Vector Embeddings, LangChain, LangGraph, LlamaIndex, OpenAI API, Anthropic API, Hugging Face Inference API

Data & ML: Pandas, NumPy, Matplotlib, Seaborn, Scikit-learn, Feature Engineering, ML Algorithms, ML Pipelines, Model Evaluation

Databases: MySQL, PostgreSQL, Pinecone, ChromaDB, Redis (caching)

Backend & DevOps: FastAPI, SQL, Docker, Git/GitHub

Tools: Ollama, Jupyter Notebook, VS Code

Projects

Unified Multi-Modal RAG System Python, LangChain, CLIP, Whisper, FAISS, FastAPI

– Built a fully offline RAG pipeline capable of ingesting PDFs, DOCX, images (via OCR), and audio recordings for semantic search and Q&A.

– Aligned text and visual embeddings into a shared vector space using OpenAI CLIP, enabling cross-modal retrieval such as text-to-image search.

– Integrated OpenAI Whisper for automated speech-to-text transcription with timestamp-based source citations in generated responses.

– Optimized local LLM inference using GGUF/AWQ quantization to run Llama 3 8B on consumer hardware, ensuring full data privacy.

MITRA – Locally Hosted AI Chatbot Python, Ollama, Quantized LLMs, RTX 4060

– Deployed a conversational AI chatbot powered by a locally hosted quantized open-source LLM (4-bit GGUF) on an NVIDIA RTX 4060.

– Engineered the full prompt handling and response pipeline; benchmarked inference throughput and memory usage on consumer-grade hardware.

Plagiarism Detector Python, FastAPI, Scikit-Learn

– Developed a REST API using FastAPI to accept text input and return document similarity scores via an ML-based comparison model.

– Evaluated model performance using a confusion matrix, precision-recall metrics, and accuracy benchmarks.

Certifications

- **Harvard CS50** (edX) – Introduction to Programming with Python
- **University of Michigan** (Coursera) – Introduction to Data Science in Python
- **Google Developers** – Machine Learning Crash Course
- **FreeCodeCamp** – LLM & AI Fundamentals

Education

Bharat College of Engineering Mumbai, India

B.E. in Computer Science & Engineering (AI & ML) | University of Mumbai *2022 – 2026*