Ganesh Sharma

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

Summary

AI and Python Backend Developer experienced in building backend systems, RAG pipelines, and LLM-powered applications using modern AI and data infrastructure. Skilled in API development, vector search, model inference, and translating AI workflows into practical software systems.

Technical Skills

Languages: Python, C++

AI Engineering: LLM, RAG, Prompt Engineering, Vector Search, Embeddings, LangChain, LangGraph, OpenAI/Gemini/Anthropic/Groq/Mistral APIs

ML/Data: Pandas, NumPy, Matplotlib, Scikit-learn, ML Fundamentals

Databases: MySQL, PostgreSQL, Pinecone (Vector DB), ChromaDB (Vector DB)

Backend Engineering: FastAPI, REST APIs, SQLAlchemy, Authentication & Authorization, CRUD Architecture, Redis

Tools: Docker, Git/GitHub, Ollama, Linux, Jupyter Notebook, VS Code, Agentic IDEs

Projects

Full-Stack Blog Platform FastAPI, SQLAlchemy, PostgreSQL, Jinja2 Templates, Pydantic Validation

- Built a full-stack blog application with a modular backend architecture using FastAPI, SQLAlchemy, and PostgreSQL.
- Designed RESTful APIs for content creation, management, and retrieval while maintaining clean separation between application layers.
- Implemented database models, relationships, validation, and persistence workflows using SQLAlchemy ORM.
- Focused on maintainability, scalability, and production-oriented backend development practices.

RAG System (Smart India Hackathon Project) LangChain, FastAPI, ChromaDB, Whisper, CLIP, Docling

- Developed a retrieval-augmented generation pipeline capable of processing and retrieving information from multimodal data sources.
- Integrated document parsing, embedding generation, vector indexing, and semantic retrieval workflows.
- Utilized Whisper for audio processing and CLIP for multimodal understanding to support broader information retrieval capabilities.
- Exposed the system through FastAPI services to enable efficient interaction between retrieval and generation components.

MITRA AI Chatbot Assistant OpenAI/Gemini/Groq/Mistral APIs, LangChain, FastAPI, Ollama

- Built an AI assistant integrating both locally hosted and proprietary large language models for conversational workflows.
- Explored deployment and inference strategies for running quantized language models on consumer-grade hardware.
- Designed backend services to manage model interaction, prompt handling, and response generation.
- Applied practical LLM engineering concepts including inference optimization, model selection, and AI application integration.

Research Papers

MITRA AI Chatbot Assistant: Using 8-Bit Quantized Large Language Models on Consumer-Grade GPUs
Published in *International Journal for Research in Applied Science and Engineering Technology (IJRASET)*, Volume 14, Issue III, pp. 5059–5065, ISSN: 2321-9653.

Courses & Certifications

- Harvard CS50's Introduction to Programming with Python (edX)
- Applied Data Science Specialization – University of Michigan (Coursera)
- Machine Learning Crash Course – Google Developers
- Machine Learning Bootcamp Certification – IIT Bombay
- LLM and Generative AI Course – FreeCodeCamp
- AI Frameworks LangChain, LangGraph – Sheryians AI School

Education

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