Devesh Chauhan

Greater Noida, India — +91-9794143181 — deveshchauhandk_cse23@its.edu.in
GitHub — LinkedIn — LeetCode — Portfolio

Professional Summary

AI Backend Engineer specializing in **production-grade Generative AI systems**, **Retrieval-Augmented Generation (RAG)**, and **scalable asynchronous backend architectures**. Hands-on experience building **end-to-end LLM-powered SaaS platforms**, vector retrieval infrastructure, and secure multi-tenant backends with emphasis on **latency, reliability, and clean system design**. Seeking engineering roles in high-impact AI product teams.

Core Technical Skills

Languages: Python, JavaScript, SQL
Backend: FastAPI, REST APIs, Async I/O, JWT Authentication, Redis, Background Workers
Generative AI: RAG Pipelines, LangChain, OpenAI-Compatible APIs, HuggingFace, Ollama, Embeddings
Data Systems: PostgreSQL, MongoDB, Vector Databases (ChromaDB, FAISS)
Cloud/DevOps: Docker, AWS (EC2, S3), Linux, Git

Selected Engineering Projects

EnterpriseRAG AI — Secure Enterprise Knowledge Assistant 2024 – Present

- Architected and built a **multi-tenant GenAI SaaS platform** enabling enterprises to query private knowledge using natural language.
- Developed **document ingestion and embedding pipelines** with semantic chunking and vector indexing for high-precision retrieval.
- Implemented **asynchronous FastAPI services** with JWT-based tenant isolation and real-time LLM answer streaming.
- Integrated **vector retrieval + LLM inference pipelines** for grounded answer generation with low-latency design.
- Designed **production-ready backend infrastructure** supporting concurrency, security, and cost-efficient inference.
- **Product Demo:** Enterprise Knowledge Assistant

OmniChat AI — LLM Orchestration Backend 2024

- Built a **vendor-agnostic LLM orchestration layer** enabling runtime switching across model providers.
- Designed **non-blocking async API services** for concurrent multi-session AI workflows.
- Implemented structured prompting and caching layers to optimize inference cost and response latency.
- Live: omnichat-ai.vercel.app

IntelliDocs AI — Document Intelligence Engine 2023

- Developed a **Retrieval-Augmented Generation (RAG) system** for semantic enterprise document search.
- Implemented embedding-based Top-K vector retrieval to improve answer relevance.
- Deployed Dockerized backend for scalable ingestion and querying workflows.
- Live: intellidocs-ai.vercel.app

Education

B.Tech — Computer Science & Engineering 2023 – 2027
ITS Engineering College (AKTU), Greater Noida