GAURAV GULATI

Delhi, IN | +91 9899559207 | gauravgulati619@gmail.com | LinkedIn | Github

TECHNICAL SKILLS

Programming Languages : Python, Java, C/C++, SQL, HTML/CSS, R
Web Development : HTML, CSS, Flask, FastAPI
Data Science/Machine Learning : Langchain, Pandas, NumPy, Matplotlib, Seaborn, Pytorch
Database and ORM : MySQL, MongoDB, PostgreSQL, SQLAlchemy
Misc : Linux, Git, Docker, AWS, Streamlit, BeautifulSoup, Postman, Gradio

EXPERIENCE

AI Engineer Intern Aug 2025 – Oct 2025
Sopra Steria *Noida, India*

- Developed advanced **Retrieval-Augmented Generation (RAG)** architectures for domain-specific chatbots, enhancing contextual accuracy and response relevance.
- Built and optimized scalable RAG pipelines using **LlamaIndex**, **LangChain**, **ChromaDB**, and **FAISS** for high-performance document retrieval and semantic search.
- Implemented modular APIs with **FastAPI** and integrated dynamic LLM workflows for prompt orchestration, advanced data retrieval, and response generation.

Generative AI Intern Jun 2024 – Aug 2024
Sansoftech Services Private Limited *New Delhi, India*

- Completed a 2-month **Generative AI** internship at **IGDTUW** in partnership with **Sansoftech Services Pvt. Ltd.**
- Engineered a full-stack **Video Transcript Summariser and Audio Summariser** pipeline leveraging the **Gemini API**, including data extraction, chunking, summarization, and UI integration.
- Built modular experimentation pipelines with **OpenAI**, **Anthropic Claude**, **Gemini**, and **Hugging Face Transformers** to compare LLM behavior and output quality.

PROJECTS

MediVox: Medical AI Assistant [**Live**][**Github**]

- Developed a project integrating **patient speech**, **images**, and **RAG** with **PDFs** for context-aware responses.
- Used **FastAPI** and **Gradio** for UI and backend integration, **LangChain** and **Groq** for RAG, **SpeechRecognition**, **torchaudio**, **pydub** for speech.
- Applied **transformers**, **torch**, **sentence-transformers** for inference, and **PyPDF2** for PDF extraction.
- Integrated **elevenlabs**, **gTTS** for text-to-speech, and **faiss-cpu** for similarity search in RAG.

Social Media API [**Live**][**GitHub**]

- Built a scalable **REST API** with **FastAPI**, **PostgreSQL**, **SQLAlchemy**, and **Alembic** supporting full **CRUD**, **JWT auth**, and a **voting/like system**.
- Developed modular endpoints for **posts**, **users**, **auth**, and **votes** with secure token-based access.
- Implemented **CI/CD** via **GitHub Actions**, cutting deployment time by **80%**.

Vehicle Insurance MLOps Pipeline [**GitHub**]

- Designed an end-to-end **production-grade MLOps pipeline** for vehicle insurance data including **data ingestion, validation, transformation, model training, evaluation, and deployment**.
- Integrated **MongoDB Atlas** for scalable data storage and automated model storage/retrieval using **AWS S3**, with deployment via **FastAPI**.
- Implemented **CI/CD automation** using **Docker**, **GitHub Actions**, and deployed on **AWS EC2** through **ECR** with self-hosted runner support.

EDUCATION

Degree/Certificate	Institute/Board	Grade	Graduation
B.Tech - CSE	Dronacharya College Of Engineering	76%	Ongoing
Senior Secondary	Lawrence Public School	82.8%	May 2022
Secondary	Lawrence Public School	90.8%	May 2020

CERTIFICATIONS & ACHIEVEMENTS

- Solved 200+ coding problems across LeetCode & GeeksforGeeks [**Profile**]
- Earned NPTEL Elite + Silver in Introduction to Large Language Models [**Badge**]
- IR4.0 Technologies - Programme of Microsoft, SAP, and Edunet Foundation [**Certificate**]
- Open-Source Models with Hugging Face - DeepLearning.ai [**Certificate**]