Aayush Kumar Singh

Ranchi, India | aayushkr646@gmail.com | LinkedIn | GitHub

Professional Summary

AI/ML Engineer skilled in deep learning, generative AI, and MLOps. Proficient in Python, C++, and end-to-end AI deployment. Experienced in computer vision, NLP, RAG systems, and building production-ready solutions with measurable improvements in accuracy, latency, and cost efficiency.

Skills

- **Languages & Core**: Python, C++ (OOP, DSA, STL), Problem-Solving, System Design, Debugging
- **AI/ML & Deep Learning**: TensorFlow, PyTorch, Scikit-learn, Keras, OpenCV, YOLOv8, DeepFace, CNN, RNN, LSTM, GAN, BERT, GPT, T5, Transfer Learning, Fine-tuning, Clustering (HDBSCAN), Classification, Regression
- **Generative AI & LLMs**: RAG, LangChain, Vector DBs (FAISS, Pinecone, Chroma), Prompt Engineering, Hugging Face, Mistral, STT (Faster Whisper), TTS (ElevenLabs), Intent Classification
- **MLOps & Deployment**: Docker, CI/CD, Model Optimization (ONNX, TensorRT), MLflow, DVC, FastAPI, Streamlit, Flask, REST APIs
- **Data & Tools**: MongoDB, PostgreSQL, Redis, Pandas, NumPy, Librosa, YAMNet, Git, Linux, AWS/GCP

Education

- **Birla Institute of Technology, Mesra** — B.Tech. in Civil Engineering (AI/ML-Focused) *CGPA: **7.61** | 2023–2027*
- XII (CBSE) ***76.2%** | 2022*
- X (CBSE) ***94%** | 2020*

Academic Projects

Agentic Insurance Voice Agent

Mistral 7B, RAG, FAISS

A fully functional AI voice agent with speech-to-text, text-to-speech, and intelligent query handling.

- Integrated **STT (Faster Whisper) and TTS (ElevenLabs)** for seamless voice interaction and natural conversation flow.
- Developed **31 intent classification models** with custom utterances, boosting accuracy by **18%**.
- Implemented **local inference with FAISS vector DB** for cost-free RAG-based testing and contextual responses.

AI-Based Music Recommendation System

FastAPI, Librosa, HDBSCAN, YAMNet

An end-to-end content-based music recommender leveraging audio embeddings and clustering techniques.

- Built recommendation engine using **YAMNet & Librosa embeddings**, improving similarity accuracy by **32%**.
- Clustered **10K+ songs with HDBSCAN** to auto-discover subgenres and deliver personalized suggestions.
- Integrated **FastAPI backend with JioSaavn API** achieving under **250ms latency** for real-time recommendations.

DeFi AI Assistant

LangChain, Pinecone, Redis

A modular AI assistant for decentralized finance queries with intelligent routing and caching.

- Built **modular FastAPI backend** for DeFi queries using RAG and vector DB thresholds for accurate responses.
- Added **Redis caching with 300s TTL** to persist partial action details and reduce latency by **40%**.
- Designed **cost-efficient query routing** with small/tiny models for clarification, reducing API costs by **40%**.

Achievements

- **Innovate-A-Thon 3.0 (National Level)** — Ranked among **Top 30 out of 1000+ participants** for developing the DeFi AI Assistant project with advanced RAG capabilities
- **Smart India Hackathon 2024** — Team Leader (College Level Qualifier). Led team of 6 to design AI-driven solution for Humsafar, surpassing 50+ competing teams

Certifications

- **Data Science & ML Bootcamp** — Krish Naik (Deep Learning, ML Algorithms, Model Deployment)
- **GenAI & LangChain Course** — Krish Naik (RAG, Vector DBs, LLM Integration, Prompt Engineering)
- **Docker & MLOps Specialization** (Containerization, CI/CD, Model Versioning, Production Deployment)