Harsh Jain

+91 7247429069 | jainh2113@gmail.com | LinkedIn | GitHub | CodeChef | Leetcode | Portfolio

EDUCATION

Indian Institute of Information Technology, Bhopal *November 2022-May 2026*

Bachelors of Technology – Electronics and Communication Engineering(ECE) ***CGPA - 8.42***

PROFESSIONAL EXPERIENCE

AI Engineer Intern, SarvM.AI *June 2024– September 2024*

- Conducted comparative evaluation of multiple open-source LLMs (LiquidAI/LFM2-1.2B, LFM-700M, Phi-3.5, LLaMA-3, Gemma-3B) on downstream NLP tasks within the agent pipeline
- Integrated SPRINGLab/k2 streaming ASR model using Icefall to transcribe audio files in real-time via WebSocket communication
- Preprocessed and converted `.mp3` files to 16kHz `.wav` format for accurate speech-to-text output using the Zipformer-based decoding pipeline
- Engineered a scalable RAG pipeline over 15,000+ documents, integrating hybrid search and model-specific performance tuning, reducing query-error rate by ~30%
- Streamlined LLM inference by capping context length and using pre-computed FAISS embeddings for faster response times

SKILLS

- **Languages and Databases:** C++, Python, SQL
- **Frameworks:** Hugging Face, LangChain, Flask, FastAPI, Streamlit, PyTorch, TensorFlow, OpenAI
- **Tools:** Debugging, Version Control, Git, Docker, AWS EC2, MLflow, DVC, VS Code
- **Concepts:** Machine Learning, Deep Learning, LLMs(large language models), RAG, NLP, Computer Vision, Transfer Learning, MLOps, ML Lifecycle
- **Relevant Coursework:** DSA , OS , OOP , DBMS, AI & Statistics
- **Soft Skills:** Leadership, Team Collaboration, Priority Management, Optimisation Mindset, Communication

PROJECTS

Multi-Disease Medical Image Classification using CNNs | Python, TensorFlow, MobileNetV2, Xception, Flask **GITHUB**

- Developed and trained CNN models on 10,000+ labeled scans across four disease categories, delivering test accuracies of 97% (Pneumonia), 94% (Breast), 85% (Lung), and 73% (Skin)
- Employed data augmentation techniques and 5-fold cross-validation to enhance model generalization
- Deployed CNN models as Flask APIs with serialized `.h5` weights for real-time inference and integrated endpoints for category-wise predictions

Student Performance Indicator(Regression ML Pipeline) | Python, scikit-learn, Flask, DVC **GITHUB**

- Designed and benchmarked linear, ridge, and lasso regression pipelines on a 1,000-record student dataset, achieving R^2 =~ 0.87 and MAE < 3.0
- Automated data versioning with DVC and built a Streamlit web app for on-demand performance predictions

LangSQL-Assistant | Python, LangChain, GROQ LLaMA-3, Streamlit, SQLAlchemy **GITHUB**

- Architected an LLM-powered chatbot that translates natural language queries into SQL for MySQL/SQLite3, enabling real-time database interaction and automated result retrieval
- Enabled automatic schema extraction from MySQL/SQLite3 databases to guide LLM-based SQL generation
- Created a Streamlit UI interface for users to input natural language queries and view real-time SQL responses and outputs

ACHIEVEMENTS

- Solved 400+ algorithmic challenges on LeetCode and GFG,attaining a Highest LeetCode rating of 1,705
- Achieved a 4★ rating on CodeChef with a Highest rating above 1,800
- Secured 3rd place in IIIT-Bhopal's internal Hack-O-Sprint 2.0 (SIH 2024), demonstrating strong competitive innovation skills **(Certification)**
- Developed "NyaaySaarthi," an AI-powered legal assistant for India's Ministry of Law & Justice using LLaMA 2, RASA NLU, Whisper API, Tesseract OCR, and PyPDF2;deployed as a full-stack Flask+React.js application Github