Hitesh Shanmukha

+91-7794808900 | hiteshshanmukha@gmail.com | linkedin.com/in/hiteshshanmukha | github.com/hiteshshanmukha

Education

Indian Institute of Technology Jodhpur — Jodhpur, India
Bachelor of Technology in Artificial Intelligence and Data Science; CGPA: 7.65 — *Aug. 2022 – May 2026*

Narayana e-Techno School — CBSE Board
Senior Secondary; Percentage: 90.0% — *2022*

Experience

AI Intern — May 2025 – July 2025
Stealth Startup — *Remote*

- Built a Retrieval-Augmented Generation (RAG) system with Azure OpenAI and Ollama, integrating HyDE, step-back prompting, and query analysis for 10–15% performance gain (RAGAs)
- Developed Hybrid Search and persistent Chat Memory using Qdrant, Weaviate, and Mem0 with metadata-enriched retrieval pipelines
- Designed Router Agent for multi-task LLM routing via LiteLLM and containerized deployment with Docker for production readiness

Projects

Neural Machine Translation (Seq2Seq with Attention) | Github — Mar. 2025 – May 2025

- **Tools & Technologies:** PyTorch, TorchText, SpaCy, LSTMs, Attention Mechanism
- Developed Seq2Seq model with attention mechanism for German→English and Hindi→English translation
- Trained on 29K German–English and 10K Hindi–English pairs, achieving a BLEU score of 36.05 on German→English
- Applied scheduled teacher forcing, gradient clipping, and checkpointing to improve stability over 100 epochs
- Validated on 500+ test sentences, showing robust handling of grammar, word order, and context

Multi-Object Tracking on MOT17 Dataset | Github — Jan. 2025 – Mar. 2025

- **Tools & Technologies:** PyTorch, OpenCV, YOLOv5/v8, Faster R-CNN, DeepSORT, Hungarian Algorithm
- Analyzed three real-time object tracking pipelines using tracking-by-detection methodology
- Built Faster R-CNN + SORT pipeline with anchor tuning, focal loss, and adaptive Kalman filtering
- Integrated YOLOv8 with DeepSORT for strong ID preservation, achieving MOTA 0.529 and IDF1 0.578
- Benchmarked on MOT17 using MOTA, MOTP, IDF1, and FPS metrics, optimizing speed–accuracy trade-offs

COVID-19 Chest X-ray Analysis | Github — Feb. 2024 – Apr. 2024

- **Tools & Technologies:** TensorFlow, Keras, Scikit-learn, OpenCV, Pandas, Matplotlib, Seaborn
- Preprocessed and analyzed 9,535 chest X-ray images, utilizing PCA for dimensionality reduction and lung segmentation
- Engineered ML models (Random Forest, XGBoost, SVM) and DL models (VGG-19, ResNet-50, EfficientNet-B3), achieving 93% and 97.5% accuracy
- Optimized hyperparameters with GridSearchCV and fine-tuned deep learning models using transfer learning

Technical Skills

Programming: C/C++, Python, SQL
AI/ML Frameworks: PyTorch, TensorFlow, LangChain, LangGraph, HuggingFace Transformers, Keras-OCR, PEFT Techniques (LoRA, QLoRA, Adapters), Weights & Biases
Developer Tools: Git, Jupyter Notebook, Google Colab, Linux, Windows, Docker, Spark, VS Code
Libraries: Pandas, NumPy, Scikit-learn, Matplotlib, Seaborn, OpenCV
Web Skills: HTML, CSS, JavaScript

Achievements

Certified as Oracle Cloud Infrastructure 2025 Generative AI Professional — *2025*
Certified as Oracle Cloud Infrastructure 2025 Data Science Professional — *2025*
Selected for Amazon ML Summer School held in July — *2024*
Ranked in top 0.7% out of 1 million+ students in JEE Mains — *2022*
Achieved All India Rank of 6616 in JEE Advanced, placing in top 4% — *2022*