HARSHITH DESHALLI RAVI

Santa Clara, CA | (224)-566 5998 | harshithdr10@gmail.com | linkedin.com/in/harshith-deshalli-ravi | github.com/HarshithDR

EDUCATION

Illinois Institute of Technology — Aug 2023 – May 2025
Master of Science in Data Science

Visvesvaraya Technological University — Aug 2019 – Jul 2023
Bachelor of Electrical and Electronics Engineering

SKILLS

Programming languages: Python | C | C++ | JS | shell
Frameworks and Libraries: TensorFlow | Pytorch | NLTK | Langchain | Langgraph | MCP | Transformers | Huggingface | OpenCV
AI/ML: NLP | CNN | RL | Tokenization embeddings | LLM Fine tuning | RAG | GraphRAG| Cross-Modal-learning | PEFT | vLLM
Cloud and llmOps: git | MLflow | Docker | WandB | AWS | Sagemaker | GCP | IBM | Jenkins | Kubeflow | Airflow
DB and Big Data: Apache Spark | Hadoop | SQL | MongoDB
Math: Calculus | Time Series | Advanced Statistics

WORK EXPERIENCE

AI/ML Intern – Alpha Ventures, Bengaluru — May 2022 - Apr 2023

- Pioneered a **Random Forest Regressor**, achieving 86% test accuracy in predictive performance for automated irrigation, working with team of 5 as project **lead**
- Optimized a smart sprinkler system reducing water consumption by 65% compared to traditional methods, outperforming competing "smart" solutions by 15%

AI RESERCH INTERN - Indian Institute of Science (IISC), Bengaluru — Feb 2023 - Apr 2023

- Engineered **CNN** models integrated with IoT robotics for real-time body-balancing task guidance, boosting system accuracy and improving the learning rate by 50%
- Optimized **TensorFlow Lite** deployment on Raspberry Pi with **AWS S3** data storage to shrink model size and enable low-latency inference, reducing cloud round-trip by 40%
- **Fine-tuned** GPT-3 to serve as a virtual educator and applied customized **reinforcement learning** algorithms, enhancing skill acquisition by 60% through adaptive, real-time feedback

Machine Learning Intern - DHI Flagship and Innovation Centre, Bengaluru — Aug 2020 - Dec 2021

- Developed an AI product to assist visually impaired individuals by improving accessibility to public transportation, **currency recognition**, and road-crossing safety through real-time **object-detection** and interaction
- Created **YOLOv5** and advanced Computer-Vision models (**Mobilenet**) deployed on edge device, achieving 95% accuracy in identifying public buses and currency denominations in diverse conditions
- Designed and integrated an **audio-based interaction** system on the **Raspberry-Pi**, enhancing user experience by reducing response time by 30% and ensuring real-time communication

PROJECT

Auto Finetune, Rag Pipeline Deployment on Aws with Jenkins

- Streamlined the training and deployment cycle through a "one-click" platform that automatically converts unstructured data into trainable formats (via Guardian models) and applies advanced tuning (**LoRA**, **QLoRA**, **knowledge distillation**)
- Implemented dual **Retrieval-Augmented Generation (**RAG**) pipelines**, supporting up to 1M+ data points in both standard **vector DBs** and **graph DBs** (**Neo4j**), boosting knowledge retrieval efficiency by 35%
- Forged **Jenkins CI/CD** on **AWS** and created an **Auto-LLM deployment** pipeline with **LORAX**, cutting post-training API rollout time by 50% and increasing inference efficiency by 70% for enterprise AI solutions

Customer Support Chatbot with RAG

- Built a GPT-4 powered chatbot leveraging **Zendesk APIs** for real time Auto-support ticket handling and **Pinecone's vector search** for efficient knowledge retrieval, reducing manual ticket resolution time by 30%
- Established a confidence-based **fallback mechanism** to trigger human escalation for predictions below a 0.7 threshold, enhancing **customer satisfaction**
- Designed a real-time monitoring pipeline with **Prometheus** and **Grafana** to track latency, accuracy, and escalation rates, improving model performance by 15%

Social media Driven Podcast generation platform using OpenAI and ElevenLabs

- Developed a comprehensive podcast creation solution that automatically fetches data from Twitter and Reddit, reducing **content-scripting** time by 70% and storing user profiles in **MongoDB** for personalized episodes
- Streamlined **Eleven Labs** for speech synthesis with **GPT-4** via OpenAI's API using **Replit** for **cloud deployment**, enabling on-demand podcasts generated in under 2 minutes

ACHIEVEMENTS AND CERTIFICATIONS

- Constructed a **multimodal AI video generation** system integrating news scraping and rapid video production, finishing the project in 48 hours; recognized as the best out of 70 teams, **winning** the **OraHacks** hackathon
- Secured Top 10 **Finalists** position in **TikTok** TechJam Hackathon for **Gen AI** project, ranking among the **top 0.3%** of global participants
- Certifications in Fine-tuning, RAG, Docker and AI from Udemy and DeepLearning.AI