Tarun Sai Goddu

+91 123456789 | tarunsaiaa@gmail.com | linkedin.com/in/tarunsaigoddu | github.com/tarun7r

Education

Indian Institute of Technology, Bombay Mumbai, India
Bachelor of Technology, Electrical Engineering *July 2019 – May 2023*

Experience

Data Scientist August 2023 – Present
Jio Platforms Ltd *Hyderabad, India*

HR Assistant Platform | *Built enterprise-scale AI chatbot serving 10k+ employees, streamlining employee support*

- Engineered a multi-format document extraction pipeline, processing 1,200+ files, with OCR and semantic analysis
- Architected an efficient Graph RAG pipeline with robust chunking strategy, improving retrieval accuracy by 30%
- Designed modular retrieval pipelines leveraging Qdrant, Graph RAG connectivity and Cross-Encoder re-ranking
- Leveraged LangChain for multi-turn chats, improving coherence, clarity of responses, and personalization by 20%

Real-Time TTS Engine | *Delivered robust and scalable cross-lingual TTS APIs powering HelloJio and Jio Translate*

- Engineered end-to-end data preprocessing with enhanced normalization to enable multilingual TTS model fine-tuning
- Developed a Python API wrapper to orchestrate LLM-based semantic tokenization and a VQGAN-based vocoder
- Enabled 100+ concurrent TTS streams on NVIDIA MIG instances, sustaining 0.2 RTF at 90% GPU utilization
- Deployed **Gunicorn/Nginx** with round-robin distribution, reducing API errors to 0.5% during peak traffic spikes

Multilingual OCR System | *Developed a regional multi-language OCR with a high-accuracy language classifier model*

- Fine-tuned the PaddleClas PULC-based language classifier model for 10 languages, achieving 93.54% F1 score
- Designed a multi-stage OCR pipeline with a confidence-based adaptive model routing, improving accuracy by 25%
- Engineered asynchronous OCR task handling using Celery with Redis broker, webhook callbacks, and retry logic
- Validated the OCR pipeline with MLOps practices, delivering 85% accuracy for 10 languages in production workflows

Object & Landmark Detection | *Enhanced Multi-Object and Landmark Recognition with Scalable Deployment*

- Fine-tuned RT-DETR (AP = 0.884 for object detection) and DINOv2 (93.51% F1 score for landmark classification)
- Designed a high-performance pipeline using Flask API, JWT, batch processing, and optimized for real-time response
- Devised scalable Docker/Kubernetes deployments with HPA auto-scaling and ELK stack, handling 4x peak loads

Projects

Finance Domain Language Model | *Domain-Specific LLM Fine-tuning* April 2025 – May 2025

- Fine-tuned LLaMA 8B on 500k+ financial instructions to create a specialized financial LLM with domain expertise
- Leveraged Unsloth, 4-bit quantization, and PEFT/LoRA techniques for memory-efficient LLM fine-tuning processes
- Implemented efficient model serving using FP16/INT4 GGUF quantization and Ollama with domain-specific prompts

Telephonic Voice Assistant | *Speech-Driven Generative AI* January 2025 – March 2025

- Developed a voice assistant using Twilio's API, integrating ASR, LLM inference, and TTS pipelines for sub-2s latency
- Implemented LLM-driven user intent recognition and robust response validation to enable context-aware interactions
- Automated export of user interaction data as JSON logs for integration with analytics pipelines and CRM systems

COVID-19 Detection System | *Deep Learning-Based Image Classification* August 2021 – October 2021

- Developed a modular PyTorch Dataset & DataLoader pipeline to efficiently preprocess and batch 10k+ CT scans
- Architected a hierarchical EfficientNet-V2 CNN with a custom image classifier, achieving 94% validation accuracy
- Implemented ROC-AUC analysis, confusion matrix, & F1-score evaluation for robust diagnostic model assessment

Technical Skills & Certifications

Languages & Tools: Python, C, C++, SQL, Git, Kubernetes, Docker, Flask, FastAPI, Redis, ELK Stack, GCP, Azure
ML/DL Frameworks: PyTorch, TensorFlow, Scikit-learn, LangChain, NumPy, Pandas, OpenCV, Transformers, vLLM
Online Courses: Data Structures and Algorithms, Machine Learning Specialization and Deep Learning Specialization
Domain Knowledge: Generative AI, Computer Vision, Natural Language Processing, Neural Networks, Transformers