

EDUCATION

Vellore Institute of Technology

Chennai, India

Computer Science and Engineering, GPA: 8.14/10

August 2020 – August 2024

- **Concentrations:** Artificial Intelligence & Machine Learning
- **Coursework:** Data Structures & Algorithms, Object-Oriented Programming, Deep Learning & Neural Networks, Business Intelligence, Computer Networks, Design of Algorithms, Database Management System, Statistics, Operating Systems, Software Engineering

WORK EXPERIENCE

Cloud Box Technology

Hyderabad, India

AI Systems Engineer

September 2025 – Present

- Architected scalable AI inference backend integrating **LLM APIs**, **Model Context Protocol (MCP)** for tool-use and reasoning, and **Python microservices** with async patterns to enable real-time model serving with sub-second latency across multiple concurrent requests.
- Designed end-to-end data pipeline combining **MS Graph data extraction**, **semantic processing**, and **contextual summarization** using **RAG** to deliver actionable insights; optimized query performance by 60% and reduced infrastructure costs through intelligent caching and batching.
- Built production-grade monitoring and observability using **logging**, **error tracking**, and **performance metrics** to ensure system reliability; implemented graceful degradation and fallback mechanisms for LLM failures, maintaining 99.2% uptime in production.

Sphere

Hyderabad, India

AI Engineer

August 2024 – September 2025

- Built an **AI-powered damage inspection system** by fine-tuning **Qwen2.5-3B** on 10K+ vehicle reports, enabling natural language queries for damage analysis with 94.2% accuracy. Deployed on production infrastructure with monitoring and fallback mechanisms.
- Developed a **computer vision pipeline** for VIN detection using **YOLOv8 with Oriented Bounding Box**, achieving 92.3% mAP50-75 and reducing inference time from 3s to 1.3s, enabling real-time processing at scale.
- Engineered an **automated email classification system** using **Phi-3mini** and **spaCy NER** for customer detail extraction, eliminating manual processing bottleneck and reducing turnaround time by 85%.

Forbes Advisor

Chennai, India

Data Analyst Intern

January 2024 – June 2024

- Built data infrastructure to stream **GA4** events to **BigQuery** for advanced segmentation and cross-platform analysis, significantly improving data accessibility and enabling faster insights.
- Implemented **GTM tracking** on the credit card segment and automated **KPI dashboards**, providing real-time visibility into user engagement across all business levels.

PROJECTS

- **Meat Quality Assessment and Distribution:** Developed a real-time meat quality assessment **framework** with 98.45% accuracy using **MobileNetV1** on 1,896 images. Integrated **Explainable AI** for segregation, reduced waste with a mobile app, and enhanced sustainability by **distributing** moderately fresh meat to NGOs.
- **ML Model Serving & Deployment Platform (Pulse):** Built a production recommendation serving system that handles multiple ML models simultaneously. Implemented **canary deployment** and **A/B testing** to safely roll out model changes, with automated rollback using **Celery-Redis**. Added real-time monitoring via **Prometheus-Grafana** to catch issues early. The system demonstrates how to bridge the gap between model development and reliable production ML.

SKILLS

Languages: Python, Go, SQL, JavaScript, Bash

AI/ML & LLM: PyTorch, Transformers, LangChain, LangGraph, RAG, YOLO, OpenCV, spaCy, BERT

Backend & APIs: FastAPI, Flask, gRPC, REST APIs, Microservices, Docker, PostgreSQL, MongoDB

ML Systems & Deployment: MLflow, Streamlit, AWS, React.js, Hugging Face, Label Studio, Git, Jupyter