Harsha Vardhan Yellela

United States | +1-248-497-9965 | harsha.yellela@gmail.com | har5ha.in | LinkedIn | GitHub

SUMMARY

Software Engineer with expertise spanning **Machine Learning**, **Cloud Infrastructure**, and **Full-Stack Development**. Specialized in **LLM fine-tuning (QLoRA)**, **MLOps pipelines**, and **production Kubernetes deployments**. Built scalable systems including **94 AWS Lambda functions**, **8-microservice architectures**, and **enterprise Go backends**. Experienced in **ROS robotics**, **Graph Neural Networks**, and **Computer Vision**. Passionate about building innovative solutions from research to production.

EXPERIENCE

Graduate Research Assistant – Agentic AI — Jan 2025 – Present

Lawrence Technological University — *Southfield, MI*

- Built and compared **no-code (n8n)** vs. **coded multi-agent systems (CrewAI + LangChain MCP)** for workflow automation and intelligent decision-making.
- Deployed persistent **MCP agent services** on **AWS Fargate** and **Amazon EKS**, integrating **OpenSearch Serverless** for semantic search and RAG.
- Designed hybrid pipelines combining **Bedrock-hosted models** with custom tools, achieving up to **70% reduction in manual process time**.

Infor India Pvt. Ltd. — Apr 2022 – Dec 2023

LN Technical Consultant — *Hyderabad, India*

- Developed modular, production-ready tools for global clients (**Ferrari, Boeing, Triumph**) by extending **Infor LN ERP** workflows.
- Integrated **Infor ION** process flows with **AWS S3, Lambda, and API Gateway** for asynchronous file transfer and event-driven automation.
- **Containerized** business logic services using **Docker** and simulated Kubernetes-like orchestration with enterprise systems.

TECHNICAL SKILLS

- **Languages:** Python, Go, TypeScript, JavaScript, SQL, Bash, Dart, C++
- **ML/AI:** PyTorch, TensorFlow/Keras, PyTorch Geometric | LLM Fine-tuning (QLoRA, PEFT, TRL) | Transformers, HuggingFace
- **LLM & Agents:** GPT-4, Gemini, Claude, Llama, Qwen | LangChain, LangGraph, CrewAI | RAG, Vector DBs (Qdrant, pgvector)
- **Backend:** FastAPI, Flask, Gin (Go), Express.js | REST APIs, OpenAPI/Swagger, HATEOAS | JWT, OAuth
- **Cloud (AWS):** Lambda, ECS Fargate, EKS, SageMaker, Bedrock | S3, DynamoDB, ECR | API Gateway, Kinesis, SNS
- **DevOps:** Docker, Kubernetes | Terraform, CloudFormation, CDK | Jenkins, GitHub Actions | Prometheus, Grafana
- **Robotics & Vision:** ROS (Robot Operating System), OpenCV, YOLO, CLIP | Point Cloud Processing, Sensor Fusion
- **Frontend/Mobile:** React Native/Expo, Next.js, React | Flutter/Dart | TypeScript, CSS Modules
- **Databases:** PostgreSQL, DynamoDB, MongoDB | Qdrant, pgvector (Vector) | Redis

PROJECTS

Resume Optimizer – QLoRA Fine-tuned LLM for ATS Optimization — Oct 2024 – Present

PyTorch, QLoRA, PEFT, TRL, Transformers, FastAPI, Ollama | GitHub

- Fine-tuned **Qwen3-4B** using **QLoRA (4-bit NF4 quantization)** with LoRA rank 16, alpha 32, reducing GPU memory to **18-22GB peak VRAM**.
- Processed **1,800+ resumes** to create **1,304 training examples** with structured JSON output, achieving **9.5/10 quality score** (GPT-5.1 evaluation).
- Built **FastAPI REST API** for resume generation with **3-5 second inference time** on RTX 3090.

ML Sentiment Feedback Loop – Production MLOps Microservices — Nov 2024 – Present

AWS (ECS Fargate, SageMaker, S3), Terraform, GitHub Actions, Docker | GitHub

- Built **8-microservice architecture** (API Gateway, Inference, Feedback, Model Registry, Evaluation, Retraining, Notification, Model Init) with **independent scaling**.
- Implemented **complete ML feedback loop** with **auto-retraining**, model versioning, and **SageMaker integration** for training jobs.
- Configured **GitHub Actions CI/CD** with **Terraform IaC** for automated infrastructure provisioning and container deployments.

Lambda Microservices Platform – Enterprise Serverless Backend — 2024

AWS Lambda, DynamoDB, API Gateway, Terraform, CloudFormation | Stripe, Twilio, DocuSign, QuickBooks

- Developed **94 AWS Lambda functions** for complete SaaS platform with **Terraform** and **CloudFormation** infrastructure.
- Integrated **10+ third-party services**: Stripe Connect (payments), DocuSign (e-signatures), Twilio (communications), QuickBooks (accounting), EagleView (aerial imagery).
- Built **automated deployment pipeline** with Git change detection, deploying only modified functions to reduce deployment time.

FieldFuze Backend – Enterprise Go REST API — Aug 2024 – Present

Go (Gin), AWS DynamoDB, JWT, Docker, GitHub Actions

- Built **production-ready REST API** in **Go 1.23** with **multi-tenant RBAC architecture** and **permission inheritance**.
- Designed **modular architecture** (controller → service → repository pattern) with **comprehensive middleware** (JWT auth, CORS, logging).
- Implemented **infrastructure automation worker** for DynamoDB table/index management with **extensive unit tests** for all layers.

Stretch2 Robot – Autonomous Navigation & Grasping — Dec 2024

ROS, Python, OpenCV, YOLO, CLIP, Point Cloud Processing | GitHub

- Built **autonomous object cluster detection** using **ROS services** with **multi-strategy arm manipulation** (mean/max/random fallback).
- Integrated **YOLO object detection** with **CLIP segmentation** for multi-modal understanding on **Stretch2 mobile manipulator**.
- Developed **point cloud processing** with farthest point sampling and **location change detection** to avoid redundant scans.

Traffic Flow GNN – Graph Neural Network Anomaly Detection — Nov 2024

PyTorch Geometric, SUMO Traffic Simulator, NetworkX, Pandas | GitHub

- Built end-to-end traffic analysis pipeline from **SUMO simulation** to **Graph Convolutional Network (GCN)** anomaly detection.
- Designed **graph-based representation** of traffic networks (nodes=intersections, edges=roads) using **PyTorch Geometric**.
- Integrated **SUMO TraCI** for real-time data collection with modular architecture for extension to real-world traffic data.

EDUCATION

Lawrence Technological University — Expected Dec 2025

Master of Science in Computer Science · GPA: ***3.6/4.0*** — *Southfield, MI*

- **Relevant Coursework: Machine Learning**, **Artificial Intelligence**, **Natural Language Processing**, **Intelligent Robotics (ROS)**, **Agentic AI Research**

Geethanjali College of Engineering & Technology — Graduated: August 2022

Bachelor of Technology in Computer Science & Engineering · GPA: ***7.5/10 (~3.0/4.0)*** — *Hyderabad, Telangana*

- **Relevant Coursework: Deep Learning & Python**, **Machine Learning Foundations**, **Software Engineering**, **Internet of Things**

ACHIEVEMENTS

- **Selected for Amazon Nova AI Challenge: Trusted AI Track** (2025)
- **Participated in RSNA Pneumonia Detection Challenge**; ranked in upper quartile with VGG19 transfer learning model (2024)
- **Built production SaaS platform** with 94 Lambda functions serving real customers for field service management (2024)
- **Gold Medalist in Indian National Mathematical Olympiad (INMO)** (2012)