HARSHA VARDHAN YELLELA

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

SUMMARY

Machine Learning Engineer with hands-on experience in **LLM fine-tuning (QLoRA, PEFT)**, **deep learning (LSTM, CNN, GNN)**, and **MLOps pipelines**. Built production ML systems using **PyTorch, TensorFlow, HuggingFace Transformers**, and **AWS SageMaker**. Skilled in **NLP**, **computer vision**, **time-series forecasting**, and deploying **scalable inference services** with **auto-retraining feedback loops**.

EXPERIENCE

Graduate Research Assistant – Agentic AI **Jan 2025 – Dec 2025**
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

- **ML Frameworks:** PyTorch, TensorFlow/Keras, PyTorch Geometric, HuggingFace Transformers, scikit-learn
- **Deep Learning:** CNN, LSTM, GRU, VGG19, Graph Neural Networks (GCN), Transfer Learning, Attention Mechanisms
- **LLM & Fine-tuning:** QLoRA, PEFT, TRL, 4-bit Quantization (NF4), Prompt Engineering, Instruction Tuning
- **NLP:** Sentiment Analysis (VADER, BERT), NER, Text Preprocessing, Embeddings, RAG Systems
- **Computer Vision:** Image Classification, Medical Imaging, OpenCV, Data Augmentation, YOLO, CLIP
- **MLOps & Cloud:** AWS SageMaker, Bedrock, ECS Fargate, Lambda, S3, Docker, Kubernetes, Terraform
- **Languages:** Python, SQL, Go, TypeScript | **Data:** Pandas, NumPy, DynamoDB, PostgreSQL, Qdrant

PROJECTS

Blindspot – AI-Powered Cognitive Bias Detector **Jan 2026**
Chrome Extension (Manifest V3), Claude API, JavaScript | Aurora Hackathon 2026 | GitHub

- Built **real-time cognitive bias detector** Chrome extension that identifies decision-making flaws using **Claude API** for AI-powered analysis.
- Implemented **proactive intervention system** monitoring high-risk sites (Amazon, eBay, DoorDash) detecting **6+ cognitive biases** including Sunk Cost Fallacy and Confirmation Bias.
- Designed **context menu integration** and keyboard shortcuts for on-demand bias analysis with actionable reframe suggestions.

Resume Optimizer – QLoRA Fine-tuned LLM for ATS Optimization **Dec 2025**
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 **Dec 2025**
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.

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.

FieldFuze Backend – Enterprise Go REST API **Dec 2024**
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.

Lambda Microservices Platform – Enterprise Serverless Backend **Dec 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.

EDUCATION

Lawrence Technological University **Jan 2024 – 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 **Aug 2018 – Aug 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)