

Kushal Varma Gangaraju

(585) 202-9561 | gangarajukushalvarma@gmail.com | [linkedin.com/in/kushalgangaraju](https://www.linkedin.com/in/kushalgangaraju) | github.com/kushalX13

Machine Learning Engineer with 2 years building ML systems across LLMs, diffusion models, and medical imaging. Experienced in model fine-tuning, RAG pipelines, and MLOps using PyTorch, Hugging Face, Docker, and AWS.

Skills & Certifications

- **Programming Languages:** Python, C++, SQL, TypeScript, MATLAB, Bash
- **ML & AI:** PyTorch, TensorFlow, Hugging Face Transformers, LoRA/PEFT, LLM fine-tuning, RLHF/DPO, RAG pipelines, LangChain, FAISS, diffusion models, computer vision, model evaluation, MLflow, Scikit-Learn, Pandas, NumPy
- **Cloud & MLOps:** AWS (EC2, Lambda, Batch, S3, ECR, IAM), Docker, CI/CD, GitHub Actions, Databricks, Delta Lake
- **Data & Infra:** PostgreSQL, Apache Spark, ETL pipelines, REST APIs, Server-less architecture, Distributed Systems
- **Certifications:** [Machine Learning](#) - Stanford | [Advanced Learning Algorithms](#) - Stanford | [SQL for Data Science](#) - UC Davis

Experience

ROC HCI Lab Rochester, NY

ML Engineering (LLM) - Research Assistant, March 2026 – Present

- Fine-tuned Qwen2.5-14B-Instruct using LoRA (PEFT) for multi-class discourse classification on imbalanced datasets, improving accuracy from 76% to 89% and macro F1 from 0.63 to 0.81; applicable to production LLM customization pipelines
- Engineered rationale supervision and threshold tuning, boosting minority-class F1 by 18% and recall by 14% on noisy data

Earth Imaging Lab Rochester, NY

MLOps & Cloud Engineering - Research Assistant September 2025 – December 2025

- Deployed cloud-native ML pipeline on AWS (Lambda, Batch, EC2) with Docker, processing 2,000+ waveform samples
- Engineered S3 + DocumentDB layer indexing 10,000+ metadata entries, enabling production-grade MLOps infrastructure

Global Health & Medical Device Laboratory Rochester, NY

Graduate Machine Learning Research Assistant March 2025 – July 2025

- Shipped end-to-end video segmentation pipeline (MedSAM2 + CycleGAN) processing 400+ frames with Gradio inference UI
- Automated annotation by integrating MATLAB ROI labeling with segmentation backend, eliminating ~70% of manual labeling
- Built structured Dice and IoU evaluation harness achieving 0.85 Dice score with strong temporal consistency across frames

Indian Space Research Organization (ISRO) Sriharikota, India

Machine Learning Research Intern February 2024 – March 2024

- Developed real-time weather forecasting models using SARIMAX, Random Forest, and Gradient Boosted Trees ensembles
- Enhanced forecast accuracy by 21% and reduced launch delays by approximately 15% via systematic feature engineering

Camplain Innovations Bengaluru, India

Quantitative Data Science Intern April 2023 – June 2023

- Built and deployed a real-time stock prediction dashboard using LightGBM and REST APIs, reducing manual analysis time by ~30%; implemented lag/rolling features and TA-Lib indicators, improving forecast accuracy by 20% through tuning

Education

University of Rochester Rochester, NY

Master of Science (MS) – Computer Science August 2024 – May 2026

Honors: Awarded Merit Scholarship for Academic Excellence

BMS Institute of Technology and Management Bengaluru, India

Bachelor of Engineering (BE) – Electronics and Communication Engineering December 2020 – May 2024

Projects

RefineVL | PyTorch, Transformers, LLaVA, HA-DPO, LoRA (PEFT), Multimodal LLMs

- Built end-to-end multimodal alignment pipeline comparing LLaVA-1.5-7B vs HA-DPO fine-tuned LoRA adapters using PEFT
- Achieved +0.157 mean score gain with 11–8 aligned wins via rubric scoring across hallucination, reasoning, and correctness

CarDiag-RAG | PyTorch, FAISS, SentenceTransformers, BM25

- Built production-style RAG pipeline with sparse retrieval, candidate re-ranking, and evaluation over NHTSA records
- Achieved Recall@1 0.94, Recall@10 1.00, and MRR 0.967 on 81 paraphrased real-world automotive diagnostic queries

3D OCT Latent Diffusion Model | PyTorch, CUDA, MONAI, 3D Generative Modeling

- Architected 530M-parameter 3D latent diffusion system (66M VAE + 464M UNet) for volumetric OCT using Rectified Flow
- Engineered mixed-precision GPU training and memory-efficient patch scheduling for reproducible generative modeling

Medical Image Classification with KAN-Integrated Transformers | PyTorch, Vision Transformers, CUDA, Grad-CAM

- Architected MedViTV2 integrating FasterKAN with Vision Transformers, training 4 variants from 5M to 50M parameters
- Achieved 98% on BloodMNIST and 97% on Czech WBC via hybrid cross-domain training with Grad-CAM validation