Monish Raman Vishakraman

mv2734@nyu.edu +1 (718) 909-6728 linkedin.com/in/v-r-monish-raman github.com/saltnpepper12

monishraman.com

Summary

AI/ML Engineer who builds production LLM agent systems and the model-efficiency work that keeps them cheap at scale: multi-agent orchestration, RAG, and **INT4/INT8 quantization** of LLMs and vision models. NYU MS Computer Engineering (2026); 4 publications, 1 patent.

Technical Skills

- **ML & Deep Learning:** PyTorch, TensorFlow, Hugging Face Transformers, scikit-learn, CUDA, ONNX Runtime, Quantization, ViT, CNNs
- **LLMs & Agents:** LLM inference & deployment, vLLM, RAG, LangGraph, LangChain, CrewAI, Tool-Calling, LangSmith, Pinecone
- **MLOps & Infrastructure:** AWS, Azure, Docker, Kubernetes, Ray, MLflow, Modal, CI/CD, Prometheus, Grafana, HPC
- **Languages & Backend:** Python, SQL, TypeScript/JavaScript, C, Java, FastAPI, Flask, REST APIs, PostgreSQL, Git

Professional Experience

Blood & Treasure 01/2026 – Present | New York City, United States

AI Engineer

- Shipped production **LLM agents** for multiple concurrent enterprise clients, owning the full lifecycle from agent orchestration and retrieval design through evaluation, deployment, and post-launch iteration.
- Engineered stateful **multi-agent** systems in **LangGraph** using a parallel supervisor/researcher/coder/reviewer pattern, with tool-calling and human-in-the-loop control.
- Grounded agent and **voice-agent** outputs in **RAG** over proprietary data; added automated testing, CI/CD, and observability.

Incluya 05/2025 – 08/2025 | New York City, United States

Founding AI Engineer

- Launched and scaled a private-beta financial-wellness LLM agent to **100+ monthly active users**, owning orchestration, prompts, retrieval, and deployment end to end.
- Cut inference latency and cost with multi-layer caching, batched vector-DB queries, and serverless autoscaling; built an adaptive coaching layer with long-term memory for cross-session personalization.

INDIAN SPACE RESEARCH ORGANISATION (ISRO) 01/2024 – 06/2024 | Bangalore, India

Machine Learning Intern

- Designed a multi-model forecasting pipeline (**LSTM, Autoencoder, ARIMA**) over **150K+** satellite telemetry records for anomaly detection and predictive maintenance.
- Improved anomaly-forecasting accuracy by **35%** with a hybrid Autoencoder-LSTM; cut feature dimensionality by **40%**.

FORTINET 06/2023 – 08/2023 | Bangalore, India

SDE Intern

- Shipped the FortiDAST Chrome extension (Chrome Web Store), cutting manual security-testing time by **60%**.

Projects

Autonomous UAV for Low-Light Human Activity Recognition | PyTorch, OpenCV, YOLOv8, MAVLink

- Built an autonomous night-vision quadcopter with real-time **YOLOv8/CNN** pipelines, reaching **78.4%–94.5%** activity-recognition accuracy across conditions.

Fairness-Aware Compression for Vision Models | PyTorch, ONNX Runtime, Modal

- Fine-tuned **ViT, Swin, and ResNet** on FairFace and profiled per-demographic accuracy across FairFace and UTKFace (7 groups).
- Quantified the fairness cost of **INT8 post-training quantization** (ONNX Runtime): ViT's per-group accuracy gap widened 14 points (31 to 45), while Swin held a 1-point drop vs. 4 for ViT.

Hybrid Outlier Smoothing Technique for Efficient LLM Quantization | PyTorch, CUDA, Hugging Face
- Built a layer-wise framework assigning **SmoothQuant, QuaRot, or SpinQuant** and **INT4/INT8** per transformer submodule by reconstruction fidelity.
- Reached **3.4x weight compression** over FP16 (4.7 average bits) while preserving accuracy on the most sensitive 17% of 224 submodules.

Education

New York University (NYU) 09/2024 – 05/2026 | New York City, United States
Master of Science- MS, Computer Engineering | GPA:3.74/4.0
- Coursework: Deep Learning, MLOps, Efficient AI and Hardware Accelerator Design, Computing Systems Architecture.

M.S.Ramaiah Institute of Technology 12/2020 – 06/2024 | Bangalore, India
Bachelor of Engineering - BE, Computer Science and Eng.|GPA:3.73/4.0

Patent & Publications

- **Patent:** Autonomous UAV System for Enhanced Human Target Detection and Acquisition (App. No. 202541052751, published 2025).
- **Publications (4):** Wiley book chapter (2025); Springer LNNS (2025); Springer CCIP (2024); IJACSA (2023).