Tanish Vardhineni

+1 (201) 448-6556 | tv2291@nyu.edu | LinkedIn | GitHub | Portfolio | New York City, NY

Summary

AI Software Engineer and MS Candidate specializing in **ML Systems Infrastructure and SW/HW Co-design**. Experienced in building high-performance inference pipelines, optimizing distributed LLM architectures (RAG/SBERT), and engineering patent-pending sensor fusion systems. Proven expertise in PyTorch, C/C++, and performance optimization of large-scale ML models. Seeking to apply system-level analysis and hardware acceleration techniques to scale Meta's R&D infrastructure.

Education

New York University — New York, NY
Master of Science in Computer Engineering — *May 2027*
- Coursework: **Distributed Systems**, Computer Architecture, Machine Learning, Network Security, Big Data.

Anurag University — Hyderabad, India
Bachelor of Technology in Computer Science; GPA: 3.8/4.0 — *May 2025*

Experience

HSBC Software Development India — Pune, India
Software Developer Intern (AI Systems) — *Jan 2025 – Mar 2025*
- Engineered high-throughput **Adversarial Safety Classifiers** (BERT/XGBoost) for real-time monitoring of internal communication, achieving a 45% improvement in risk detection accuracy via performance tuning.
- Developed scalable **Out-of-Distribution (OOD)** detection frameworks for financial news feeds, optimizing model robustness against domain-specific data drift and systemic noise.
- Architected a multi-modal **AI Infrastructure** pipeline integrating OCR and STT modules; benchmarked and optimized inference latency to ensure real-time compliance across high-volume video streams.

Thunder Client — Dublin, Ireland
Software Developer Intern (Systems & Infrastructure) — *Apr 2023 – Aug 2025*
- Architected a **gRPC-based** testing interface and high-performance Swagger-to-OpenAPI converter in TypeScript, automating core infrastructure workflows and reducing manual documentation overhead by 90%.
- Implemented an LLM-based documentation synthesis engine using **Retrieval-Augmented Generation (RAG)**, integrating **FAISS/SBERT** for low-latency semantic search across 10,000+ API technical specifications.
- Designed a **distributed Node.js mock server** to simulate complex financial transactions, optimizing developer environment efficiency and reducing testing setup latency by 40%.
- Engineered secure SSO systems using OAuth2 and optimized data-driven analysis pipelines to monitor community platform vulnerabilities via BERTopic modeling.

Selected Projects & Research

Patent: AI-based Modular Positioning System | *SW/HW Co-design, Sensor Fusion*
- Invented a multi-sensor fusion framework (Indian Patent 202541025471) integrating visual inputs with proximity sensors for **real-time autonomous navigation** and collision-free positioning.
- Developed hardware-aware control algorithms in C++ to map physical layouts, optimizing for **low-power consumption and embedded compute constraints** in autonomous reconfiguration tasks.

BuildBot AI: Local LLM Infrastructure | *Llama 3.1, Ollama, PyTorch*
- Architected a local LLM code generation system utilizing **Llama 3.1**, implementing syntax-constrained validation loops to ensure executable React/Node.js outputs.
- Optimized **vector store memory retention** (Faiss), reducing context-drift in long-sequence generation tasks and improving the efficiency of the inference orchestration engine.

Embedded AI Traffic Safety Pipeline | *YOLOv8, OpenCV, Fail-Safe Logic*
- Designed a **computer vision pipeline** for on-device congestion control, implementing fail-safe logic to maintain system reliability during sensor occlusion or detection hardware failure.
- Optimized model inference for **edge deployment**, reducing idle wait times by 60% through data-driven analysis of traffic flow patterns and real-time detection feedback loops.

Technical Leadership & Recognition

1st Place, Iterate NYC: Real-time Audio Alignment System | *Python, WebSockets* — Dec 2025
- Engineered a **low-latency feedback loop (¡200ms)** using LLM-based sentiment scoring to steer speaker tone in real-time, demonstrating expertise in high-performance streaming AI systems.

Runner Up, FitchGroup Codeathon: Physics-Constrained ML Models | *Ensemble Models* — Nov 2025
- Resolved massive Pareto-distributed data skew in ESG datasets by fusing "Economic Physics" constraints with gradient boosting, ensuring predictive stability across imbalanced distributed systems.

Technical Skills

ML Systems & Infra: **PyTorch**, TensorFlow, Distributed Systems, **Hardware Acceleration**, RAG, FAISS, SBERT, Model Interpretability, Performance Optimization, LLM Inference.
Languages & Backend: **C/C++**, Python, Java, SQL, **gRPC**, REST, GraphQL, Node.js, TypeScript.
Systems & Tools: Linux, **Docker**, AWS, Kafka, Redis, CI/CD, Git, OpenCV, YOLOv8, Embedded Systems, TCP/IP.